DUPLEX-MPE: BENCHMARKING MULTI-PARTY INTERACTION IN FULL-DUPLEX DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time full-duplex speech models can listen while speaking, enabling natural interaction without rigid turn boundaries. Existing benchmarks evaluate turn-taking, interruption handling and multi-round dialogue, but largely centre on a designated user rather than an assistant participating in a shared conversation among several people. We introduce **Duplex-MPE** to evaluate when such an assistant should answer, remain silent or stop speaking. The benchmark contains 2,000 scenarios with three or four human speakers and one assistant, each paired across explicit and implicit addressing of the same request. Models receive continuous conversation audio without transcripts or supplied turn boundaries. Four scores measure fresh response initiation, answer accuracy, silence preservation and stopping when a human resolves a request. We evaluate five open-weight speech systems: MiniCPM-o 4.5, Moshi, FLM-Audio, Voila and Freeze-Omni. MiniCPM-o 4.5 leads on three scored capabilities, while frequent speech from other systems can coexist with inaccurate answers or failures to remain silent. A transcript-based Gemini 3.1 Pro reference responds 64.3 percentage points more often to explicit than implicit requests; paired tests detect no significant response-rate difference for the speech systems. Audio examples are provided in the supplementary material.

1 INTRODUCTION

Spoken assistants should let people communicate without having to wait for a rigid sequence of listening and speaking turns. A user may need to correct a misunderstanding, add information or stop an answer while the assistant is still speaking. Supporting these interactions requires the assistant to keep listening during its own response and adapt to what it hears, rather than wait until that response finishes (Défossez et al., 2024; Wang et al., 2026). Full-duplex speech models provide this capability by processing incoming audio while generating speech (Nguyen et al., 2023; Défossez et al., 2024). Their value therefore depends on more than producing fluent answers: they must also decide when to speak, when to listen and when to stop.

These decisions become especially important when an assistant joins a conversation among several people, as in a meeting, living room or car. Participants may address one another, another device or the assistant, and speech not addressed to the assistant can still supply information needed for a later answer (Carletta et al., 2005; Jovanovic & op den Akker, 2004). Another participant may even answer a question while the assistant is responding, making further assistant speech unnecessary. An assistant in this setting must follow the shared conversation while deciding separately whether its participation is needed.

However, current benchmarks provide only part of the evidence needed to assess this behaviour. Most spoken-language benchmarks evaluate isolated inputs for which an answer is expected (Yang et al., 2021; 2024; Wang et al., 2025; Chen et al., 2026). Full-Duplex-Bench v1.5 tests interruptions, backchannels, side conversations and background speech by introducing overlap into an ongoing user–assistant exchange (Lin et al., 2026a). HumDial includes third-party speech and speech directed at others, but evaluates whether the model rejects these utterances while serving a designated user (Wang et al., 2026). These benchmarks test important duplex behaviours, yet leave open how well an assistant participates in a shared conversation where any speaker can request its help, pro-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

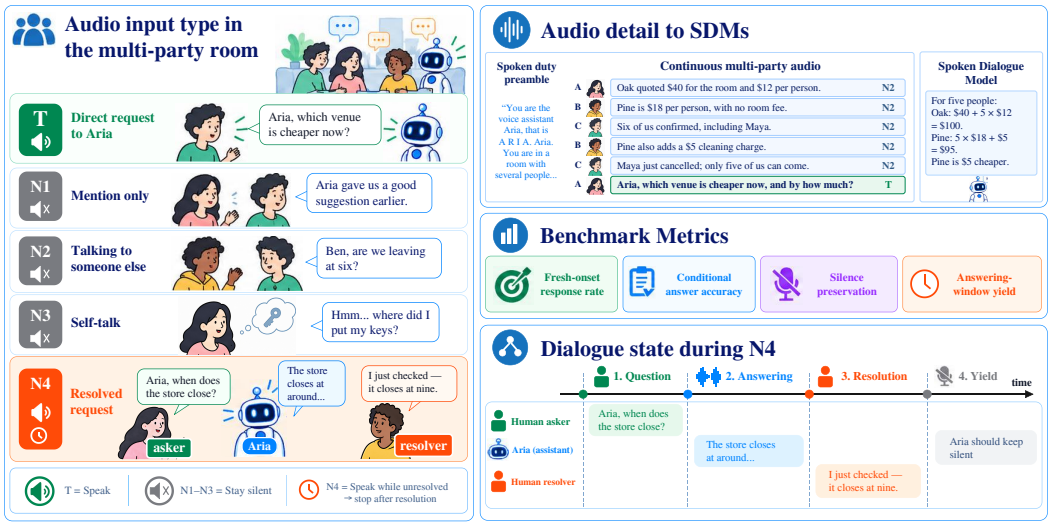


Figure 1: Duplex-MPE overview: turn types (left), audio input and metrics (top right), and N4 timing (bottom right).

vide relevant evidence or resolve a request. Evaluating that setting requires checking both whether the assistant understands the conversation and whether it speaks at the appropriate moments.

We introduce **Duplex-MPE** to evaluate this selective participation in continuous multi-party dialogue (Figure 1). It contains 2,000 continuous multi-party spoken scenarios, each with three or four humans and assistant Aria. Each scenario contains one unresolved request and labelled turns requiring silence. The benchmark also tests whether the assistant stops speaking when another participant resolves a request addressed to it. An addressing-inverted counterpart preserves the scenario-level task and intended answer while changing whether the request explicitly names Aria. Every speech system receives the same input: a spoken duty preamble followed by continuous room audio, with no transcript, speaker label, turn boundary, or candidate endpoint.

We evaluate MiniCPM-o 4.5, Moshi, FLM-Audio, Voila and Freeze-Omni and find that frequent speech does not imply appropriate participation. MiniCPM-o 4.5 leads on three scored capabilities, while other systems exhibit inaccurate answers, missed requests or speech during turns that require silence. To assess how addressing affects response decisions when the words and speakers are known, we also evaluate Gemini 3.1 Pro on the corresponding speaker-attributed transcripts. It responds 64.3 percentage points more often to explicit than implicit requests; the five speech systems show no statistically significant paired response-rate difference.

Our contributions are threefold. **First, a benchmark for shared multi-party dialogue:** Duplex-MPE provides paired explicit and implicit requests within conversations where the assistant must infer whom each turn addresses. **Second, separate measures of participation:** four scores assess fresh response initiation, answer correctness, silence preservation and stopping after a human resolves a request (Table 2). **Third, an automated construction and evaluation pipeline:** generated scripts supply labels and answers, speech synthesis supplies audio boundaries, and scoring combines timing checks with semantic judgments, with human validation of sampled data and outputs.

2 Related Work

Duplex-MPE draws on three lines of work: spoken-dialogue evaluation, addressee recognition and selective responding to non-addressed speech. Apps. A1.1 and A1.2 provide additional context and adjacent tasks. Table 1 compares the task interfaces of representative benchmarks from these three lines rather than their model scores.

Table 1: Protocol comparison. ✓: included; ×: absent; n/a: not applicable. Non-addressed speech covers input utterances directed elsewhere or to no particular addressee, including label-prediction tasks. Streaming denotes interactive speech evaluation; "no fixed user" applies to an assistant serving multiple participants.

Benchmark	Streaming	Non-addressed speech	No fixed user	Yield on interruption
Talking Turns	✓	×	×	✓
Full-Duplex-Bench v1.5	✓	✓	×	✓
Inoue et al. (2025)	×	✓	n/a	n/a
Fukuda et al. (2026)	×	✓	n/a	n/a
HumDial	✓	✓	×	✓
WearVox	×	✓	×	n/a
Duplex-MPE (ours)	✓	✓	✓	✓

Spoken dialogue and full-duplex evaluation. Talking Turns evaluates when a spoken system starts speaking, briefly acknowledges the user or stops after an interruption during a conversation with one human (Arora et al., 2025). Full-Duplex-Bench probes pause handling, backchannelling, turn-taking and interruption management (Lin et al., 2025). FD-Bench generates duplex test material procedurally (Peng et al., 2025), and MTR-DuplexBench extends evaluation to multiple rounds (He et al., 2026). Together, these protocols evaluate when a spoken system should take, hold or yield the floor.

Addressee recognition and device-directed speech. Addressee recognition asks whom an utterance is directed to. Classical work predicts an addressee for a supplied utterance in meeting or chat corpora (Jovanovic & op den Akker, 2004; Jovanovic et al., 2006; Ouchi & Tsuboi, 2016; Gu et al., 2021). Device-directed speech detection studies the corresponding deployed decision of whether an utterance is intended for an assistant or device (Wagner et al., 2024; Rudovic et al., 2024; Palaskar et al., 2024; Garg et al., 2022; Kim et al., 2026). Across these settings, the evaluated unit is normally a supplied utterance and the output is an explicit addressee or device-directed label. These works establish addressee recognition as an existing multi-party and deployed-device problem; Duplex-MPE treats that recognition as a latent decision controlling an end-to-end model's speech rather than as a new classification task.

Selective responding and non-addressed speech. A third line of work makes silence a correct system output. On the text side, several datasets pose the speak-or-stay-silent decision over multi-party transcripts (Bhagtani et al., 2026; Nama et al., 2026; Liu et al., 2025; Patel et al., 2025). Audio-native benchmarks introduce non-addressed speech into interactive protocols. Full-Duplex-Bench v1.5 (Lin et al., 2026a) devotes two of its four conditions to non-addressed speech: "talking to others" and background speech, alongside user interruption and user backchannel. In both non-addressed conditions, the desired behaviour is to filter the inserted overlap and resume the model's preceding response. HumDial (Wang et al., 2026) defines a complete *rejection* category with four sub-scenarios that cover third-party speech and speech directed at others in dual-channel human conversations. Its interruption tasks also include explicit requests for the model to stop speaking. WearVox (Lin et al., 2026b) records 3,842 multi-channel egocentric sessions on AI glasses with bystanders present, and one of its five tasks is side-talk rejection. These audio benchmarks therefore score whether a system rejects speech outside a designated user interaction. Duplex-MPE instead evaluates selective participation in a shared multi-party dialogue: speech from other participants can establish context for a later request or resolve a request already addressed to the assistant, so the model must follow every participant while deciding separately whether to speak.

3 The Duplex-MPE benchmark

Duplex-MPE evaluates whether a spoken model responds selectively while receiving continuous multi-party audio. It assigns an expected action to each turn, pairs each scene across two addressing

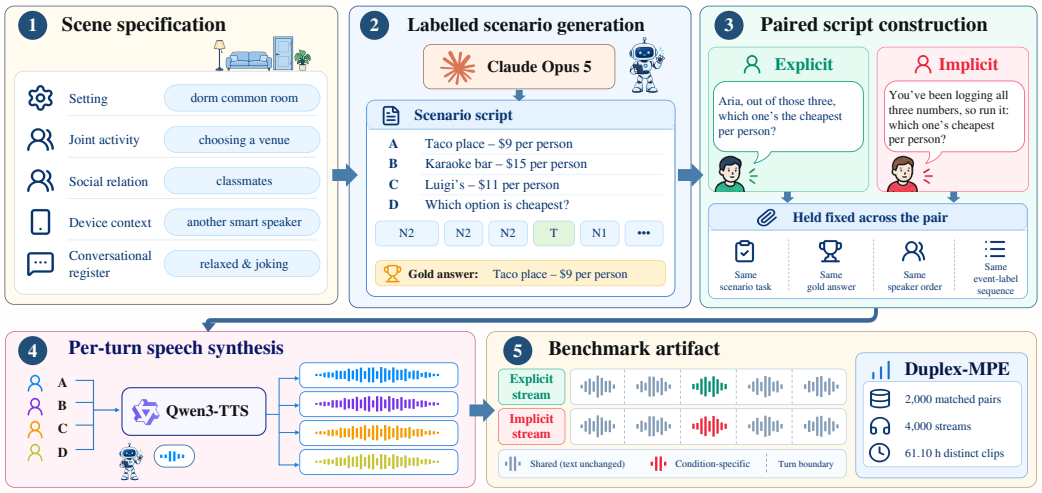


Figure 2: Duplex-MPE dataset construction.

forms, streams the resulting audio without side information, and scores four capabilities on separate denominators. The evaluation unit is one scenario under one addressing condition, giving 4,000 evaluated conversations from 2,000 scenario pairs.

3.1 Turn types and expected assistant behaviour

Each scenario is a sequence of turns spoken by three or four humans, and every turn receives exactly one label. **T**: a direct, still-unresolved request addressed to Aria. T is the only ordinary-turn label that requires a response. Each scenario contains exactly one T.

N1: Aria is mentioned but not asked to do anything. **N2**: the turn is addressed to someone or something other than Aria, including another human, or assistant. **N3**: speech without a designated addressee, including self-talk and thinking aloud. N1, N2 and N3 require silence.

N4 tests whether Aria stops when its answer is no longer needed. It begins when a human participant, the *asker*, poses a *question* to Aria ($N4_Q$). The subsequent human utterance, the *resolution* ($N4_R$), answers the question or explicitly tells Aria that it need not answer. The person who speaks this resolution, either the asker or another participant, is the *resolver*. The interval from the end of $N4_Q$ to the start of $N4_R$ gives the model 3 s to respond. Aria should answer during this interval and stop or remain silent when it hears $N4_R$.

3.2 Dataset construction

Figure 2 summarises the construction process, from scene attributes to paired audio streams. Claude Opus 5 generates multi-party conversation scripts from combinations of five attributes. These specify where the conversation takes place (setting), what the participants are doing (activity), their relationship, what devices are present (device context), and their manner of speaking (register). For example, the scenario in Table 4 places classmates in a university dorm common room, choosing a venue in a relaxed, joking conversation, with another smart speaker in the room alongside Aria. Each script specifies the human speakers, their utterances and turn labels, and the gold answer to T.

Each script is constrained to contain exactly one T at an assigned position bucket. T is never first and is always followed by further conversation, preventing end-of-scene timing from serving as a response cue. When present, an N4 question is addressed to Aria by design but remains a separate temporal event whose request is later resolved. These constraints hold in all 2,000 scenarios under both conditions; App. A2.2 reports the position distribution.

Qwen3-TTS synthesises every human turn separately under a deterministic speaker-to-voice map. Turn-wise synthesis supplies construction-level gold boundaries (App. A2.3).

We group the completed scenarios by T's addressing form: $2{,}000$ explicit versions and $2{,}000$ implicit versions, with one of each per scenario pair. These groups define the conditions in all result and duration tables. The dataset contains 61.10 hours of distinct human-speech clips.

3.3 Paired explicit and implicit conditions

For each generated scenario, we construct a counterpart by reversing whether T explicitly names Aria, giving $2{,}000$ matched pairs and $4{,}000$ audio streams. Each pair contains one explicit condition, in which T names Aria, and one implicit condition, in which the intended addressee must be inferred from conversational context. The scenario-level task is held fixed, and the gold answer is identical in every pair.

The rewrite targets T, adding the name for explicit addressing or replacing it with a contextual cue for implicit addressing. Simply deleting the name can leave a request plausibly directed to a human in the room. When T cannot naturally carry a cue that identifies the assistant, the immediately preceding turn is also revised to establish whom the speaker is addressing. In step 3 of Figure 2, "Same turn count" and "Same event-label sequence" mean that the addressing rewrite preserves the number of turns and each turn's label. Subsequent N4 edits and separate synthesis introduce differences in wording, event counts and audio between the final versions (App. A2.4). We therefore compare response presence between complete paired scenarios, without attributing the difference solely to the presence of the name. App. A2.1 gives a complete scenario and its paired request.

3.4 A single audio-only protocol

Each speech system begins a scenario from fresh streaming state and receives:

$$\underbrace{\text{spoken duty preamble}}_{\text{role and response policy}} + \underbrace{\text{continuous multi-party scenario audio}}_{\text{persistent model state}}.$$

The preamble identifies Aria and states when it should answer, remain silent, and stop. Scoring begins with the first scenario turn; the preamble and the silence immediately following it are outside the scoring windows. The model receives no transcript, speaker label, turn boundary, addressee label or gold decision.

The scheduler controls when each human audio clip is presented, playing ordinary turns in order with silence between them. App. A3.1 lists the timing settings. After T, a model not already speaking has $5\,\text{s}$ to begin a fresh voiced onset. Once a response is present, the scheduler waits for $2\,\text{s}$ of confirmed silence before advancing and caps that wait at $60\,\text{s}$. After the complete N4_{Q} question ends, the scheduler streams $3\,\text{s}$ of silence to give the model time to answer. It then plays the scripted human answer or statement that Aria need not answer (N4_{R}), whether or not the model has finished speaking. If the model is still speaking, the evaluation measures whether it stops after hearing N4_{R}.

We evaluate Gemini 3.1 Pro on text inputs as a *transcript-conditioned reference*. Given the duty instruction, the current utterance and preceding speaker-attributed text, it chooses whether Aria should respond or remain silent. For T requests on which it chooses to respond, it also generates a text answer. This reference evaluates the response decision with lexical content and speaker attribution supplied explicitly; it is neither an acoustic system nor an upper bound on speech-model performance. Its paired effect establishes that the intervention changes a transcript-conditioned decision, against which Sec. 4.2 compares the speech systems.

3.5 Metrics

Duplex-MPE scores four capabilities: fresh-onset response rate, conditional answer accuracy, silence preservation and answering-window yield. Each is a success rate on the turns where that capability is defined, and higher is better. Response presence and window response are reported alongside them to expose denominator coverage and timing composition, but neither is scored as an additional capability. Table 2 fixes the six numerators and denominators before any results are examined. Figure 3 illustrates the speech timings used by fresh-onset response rate, response presence, window response and answering-window yield. No aggregate is formed because success on one denominator cannot compensate for failure on another.

Table 2: Metric definitions.

Metric	Definition
Scored capabilities *higher is better*	
Fresh-onset response rate	Fraction of all 2,000 T requests with no speech at request end and a new speech onset within 5 s.
Conditional answer accuracy	Fraction correct among T requests with speech in the response window.
Silence preservation	Fraction of N1, N2 and N3 windows with no speech or only a brief acknowledgement that does not take the floor.
Answering-window yield	Among N4 events with no speech during the question, first speech within 3 s afterward, and speech still active at $N4_R$ onset: the fraction silent at $N4_R$ onset + 2 s and through the rest of its observation window.
Coverage diagnostics	*Speech coverage used to interpret the capability scores, not additional scores.*
Response presence	Fraction of all 2,000 T requests with any speech in the response window, including speech already underway. These requests form the conditional answer accuracy denominator.
Window response	Fraction of N4 events with any speech during the question or the following 3 s answering window. Answering-window yield scores only the subset that first speaks after the question and is still speaking at $N4_R$ onset.

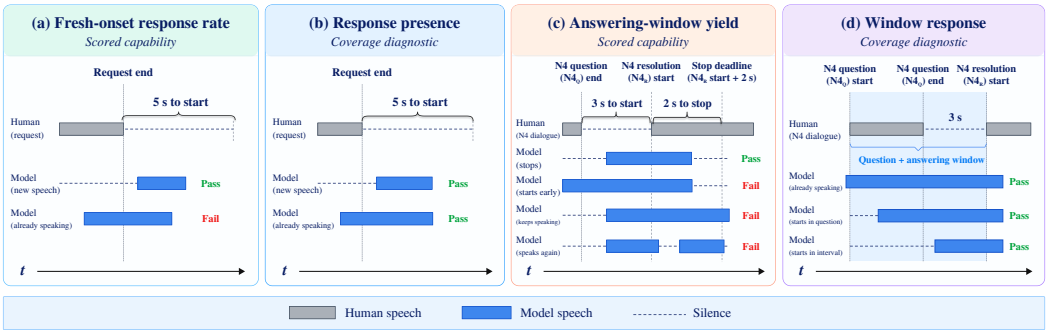


Figure 3: Schematic speech timelines for the response-timing metrics.

Fresh-onset response rate counts a T turn only when a fresh voiced onset begins within 5 s after T ends and no assistant speech is active at that boundary. **Response presence** uses the same 2,000 T turns but also counts speech already active when T ends. The two quantities therefore distinguish a new decision to speak from permissive speech coverage without assigning continuation a second capability score (App. A3.2).

Conditional answer accuracy asks whether the speech counted by response presence correctly answers the T request. Qwen3-ASR-1.7B transcribes the decoded response, and Claude Opus 5 compares its meaning with the gold answer while ignoring filler, politeness and transcription noise. All T turns remain in the fresh-onset response rate and response presence denominators; conditional answer accuracy is evaluated only where speech is available because a silent T turn supplies no answer content to judge. Unparseable answer-correctness verdicts count as incorrect; App. A3.3 describes the semantic scoring procedure.

Silence preservation is evaluated on every N1, N2 and N3 window. A window passes if Silero VAD detects no assistant speech. Otherwise, detected speech fragments are joined, transcribed with Qwen3-ASR-1.7B, and classified by Claude Opus 5; the window passes only if the output is classified as a brief acknowledgement that does not take the floor. An empty transcript or missing classification does not receive this exemption (App. A3.3). N4 is excluded because its expected action changes during the event.

Table 3: Results by T addressing condition. Bold marks the highest reportable value per condition and scored metric; answering-window yield uses model-specific eligible events.

		Scored capabilities				Coverage diagnostics	
Model	T addressing	**Fresh-onset response rate ↑**	**Conditional answer accuracy ↑**	**Silence preservation ↑**	**Answering-window yield ↑**	**Response presence**	**Window response**
MiniCPM-o 4.5	Explicit	**0.9500**	**0.4730**	**0.9242**	0.5964	0.9545	0.9513
	Implicit	**0.9435**	**0.4665**	**0.9289**	0.5925	0.9485	0.9461
Moshi	Explicit	0.6340	0.0123	0.2734	0.1918	0.8155	0.9226
	Implicit	0.6700	0.0170	0.2676	0.2107	0.8220	0.9309
FLM-Audio	Explicit	0.5380	0.0011	0.3957	0.0082	0.9505	0.9927
	Implicit	0.5425	0.0005	0.3991	0.0059	0.9540	0.9937
Voila	Explicit	0.6465	0.0031	0.8351	**0.8484**	0.6515	0.6217
	Implicit	0.6670	0.0075	0.8307	**0.8046**	0.6690	0.6196
Freeze-Omni	Explicit	0.2525	0.0035	0.0002	n/a $(n=1)$	0.9955	1.0000
	Implicit	0.2580	0.0015	0.0002	n/a $(n=1)$	0.9975	0.9995

Human validation. We validate both benchmark construction and model evaluation with six reviewers, with two assigned to each item resolving disagreements through discussion. For *construction*, data-validity checks on 100 sampled conversations achieve 98%–100% pass rates, and all 1,282 checked TTS turns are intelligible and preserve the script's meaning. For *evaluation*, all 200 checked ASR transcripts preserve the model output's meaning, and final reviewer labels agree with Opus 5 on $99/100$ T-answer correctness judgments and $98/100$ N1–N3 acknowledgement-versus-intrusion judgments. App. A3.4 details the sampling, review procedure and results.

N4: which events are scored and what counts as stopping. **Window response** records whether any assistant speech is present from the start of $N4_Q$ through the end of its 3 s answering window.

Answering-window yield scores only events in which the model remains silent during $N4_Q$, starts speaking within the next 3 s, and is still speaking when $N4_R$ begins. These events form its denominator. Events in which the model speaks before the question ends, never speaks in the answering window, or finishes before $N4_R$ begins are excluded from this rate rather than counted as stopping failures.

For each scored event, the stopping deadline is 2 s after $N4_R$ begins. The event passes if the model is silent at that deadline and, if the $N4_R$ clip plus its following 0.5 s silence extends beyond the deadline, remains silent for the rest of that interval. An earlier end to the human clip does not shorten the 2 s deadline (App. A4.1). Answering-window yield is the fraction of scored events that pass these checks, regardless of whether the model's answer is correct. Sec. 4.3 reports the counts, and App. A4.2 gives the full stopping analysis.

4 RESULTS

Table 3 reports all six quantities under explicit and implicit addressing for the systems listed in App. A1.3. We examine distinct response failures, addressing effects, stopping behaviour, request timing, and robustness to scoring choices and sampling uncertainty.

4.1 SEPARATE METRICS EXPOSE DISTINCT BEHAVIOURAL PROFILES

Presence does not identify fresh initiation. Under explicit addressing, Freeze-Omni has the highest response presence at 0.9955 but the lowest fresh-onset response rate at 0.2525. FLM-Audio has response presence of 0.9505 and fresh-onset response rate of 0.5380. By contrast, MiniCPM-o differs by only 0.0045 between response presence (0.9545) and fresh-onset response rate (0.9500). The gap is exactly the share of T turns on which speech was already active at request completion, so equal-looking response coverage can arise from different floor-taking behaviour (App. A3.2).

Presence does not identify answer accuracy. Under explicit addressing, MiniCPM-o and FLM-Audio have similar response presence (0.9545 and 0.9505), yet their conditional answer accuracy values are 0.4730 and 0.0011. FLM-Audio produces only 2 correct answers among $1{,}901$ response-present requests. Response presence therefore establishes that speech was available to judge, not that the request was answered.

Response presence cannot distinguish selective responding from indiscriminate speech. Under explicit addressing, MiniCPM-o and Freeze-Omni both have high response presence (0.9545 and 0.9955), yet their silence preservation values are 0.9242 and 0.0002. Freeze-Omni preserves silence in only 5 of $20{,}360$ silence-requiring windows, so its near-perfect response presence coexists with speech in almost every window where silence is required. The contrast is not confined to this extreme case: Moshi has higher response presence than Voila (0.8155 versus 0.6515) but much lower silence preservation (0.2734 versus 0.8351). Response presence alone therefore cannot distinguish a system that produces speech after T while remaining quiet elsewhere from one that has speech available almost everywhere. This comparison concerns observable speech allocation across scored windows rather than the models' internal mechanisms. App. A6.1 illustrates allowed acknowledgements and penalised speech.

4.2 Explicit versus implicit addressing

The transcript-conditioned reference first checks whether the paired manipulation changes a decision when lexical content and speaker identities are supplied. Gemini 3.1 Pro receives the speaker-attributed transcript and duty instruction but no audio (App. A5.2).

The response rate is 0.9470 when the assistant is named and 0.3040 when it is not, a paired difference of 0.6430. The exact McNemar test gives $p < 10^{-371}$ (App. A5.2); App. A5.3 reports the reference's silence preservation by turn type. This contrast compares the complete paired versions defined in Sec. 3.3, including the context edits and N4 realisations documented in App. A2.4.

For the five speech systems, subtracting the implicit response rate from the explicit rate gives differences from -1.75 to $+0.60$ percentage points. A positive difference means more frequent speech in the explicit versions; a negative difference means more frequent speech in the implicit versions. We call a difference statistically significant when $p < 0.05$. All five p values exceed this threshold (the smallest is 0.2365), so these tests do not establish a systematic response-rate difference between the addressing conditions (App. A5.1). In our experiments, the speech systems' response rates change little between requests that name Aria and those that do not. This pattern could reflect either successful inference of the addressee from context or a failure to recognise the name as an addressing cue, leaving the model similarly inclined to speak or remain silent in both conditions.

We also examine requests directed to other humans or devices, where the assistant should remain silent. In scenarios with explicit T requests, MiniCPM-o has speech present after 95.45% of T requests addressed to Aria, while it fails to preserve silence on 216 of $4{,}704$ requests directed to other humans or devices (4.6%; App. A5.3). It thus responds frequently when addressed and usually preserves silence when someone else is addressed.

4.3 Floor release is a coverage-conditioned outcome

In scenarios with explicit T requests, MiniCPM-o produces speech in $1{,}818$ of $1{,}911$ N4 windows. In $1{,}736$ of these events, it stays silent while the N4 question is being played and starts speaking within 3 s after the question ends. In $1{,}633$ of these events, the model is still speaking when $N4_R$ begins, so its ability to stop is scored by answering-window yield. Freeze-Omni has speech in all $1{,}911$ windows, but $1{,}784$ are continuations from before the question and only one event reaches the answering-window yield denominator. Thus similar window response can produce radically different eligible sets (App. A4.1).

For the explicit condition, the answering-window yield denominators are $1{,}633$ for MiniCPM-o, 219 for Moshi, 977 for FLM-Audio, 620 for Voila and 1 for Freeze-Omni. The corresponding answering-window yield estimates for the first four systems are 0.596, 0.192, 0.008 and 0.848. We report this rate only when at least 30 events meet its scoring conditions; Freeze-Omni has only one such event, on which it fails to stop, so its rate is not displayed (App. A4.2). For the implicit condition, the

four reportable estimates are 0.593, 0.211, 0.006 and 0.805, respectively (Table 15). Voila's high conditional rate therefore describes stopping on its eligible events, not overall N4 performance.

In the explicit condition, FLM-Audio is still speaking 2 s after $N4_R$ begins in 968 of its 977 scored events. In 97 of Moshi's 177 failed events, the model is silent at that deadline but speaks again before the $N4_R$ observation window ends (Table 15). These cases show why a brief pause is insufficient: the model must remain silent from the deadline through the end of that window (App. A4.2).

4.4 Response performance by request time

We compare model responses to requests occurring earlier, midway and later in the conversation.

Response behaviour. From the earliest to the latest group, Voila's response presence falls from 71.64% to 59.55% under explicit addressing and from 76.21% to 55.82% under implicit addressing. FLM-Audio's fresh-onset response rate falls from 67.61% to 47.67% under explicit addressing and from 66.97% to 48.96% under implicit addressing, while its response presence remains above 94% in every group. These patterns show that some models become less likely to respond or initiate a fresh response when requests occur later in the conversation.

Answer accuracy. From the earliest to the latest group, MiniCPM-o's conditional answer accuracy falls from 50.69% to 46.82% under explicit addressing and from 51.27% to 44.41% under implicit addressing. The other four models have answer accuracy below 2% in every group, leaving little room for a further measurable decrease. The results thus suggest a trend towards lower answer accuracy after longer conversation histories for some models, as observed for MiniCPM-o (App. A3.6).

4.5 Sensitivity and uncertainty

To check whether the silence-preservation ranking depends on brief speech detections, we repeat the analysis using only detected assistant speech duration, without the semantic exemption for brief acknowledgements. A silence-requiring window counts as a violation only when the total detected assistant speech within it exceeds a threshold of 0, 100, 200, 300 or 500 ms. At 0 ms, any detected speech counts; at 500 ms, only speech totalling more than half a second counts. The model ordering within each addressing condition is unchanged across these thresholds and matches the reported silence-preservation ordering (App. A3.5).

The 95% confidence intervals support the reported silence-preservation ordering for all five models under both addressing conditions. For MiniCPM-o and Voila, the closest pair under explicit addressing, this interval places MiniCPM-o ahead by 8.25 to 9.58 percentage points (App. A7.1).

All models receive the same $2{,}000$ T requests in each addressing condition, but a model may remain silent on a request that another model responds to. To compare answer correctness on the same questions, each two-model test uses only T requests on which neither model remained silent. On these requests, MiniCPM-o has significantly higher conditional answer accuracy than every other model, and Moshi exceeds FLM-Audio under both addressing conditions. The remaining model pairs show no statistically significant difference in conditional answer accuracy (App. A7.1).

5 Conclusion

Duplex-MPE turns addressee inference into an observable speech-control test: an end-to-end model receives continuous multi-party audio and must decide whether to begin, what to say, when to remain silent and when to release the floor. Duplex-MPE contains $2{,}000$ addressing-inverted scenario pairs and reports fresh-onset response rate, conditional answer accuracy, silence preservation and answering-window yield on separate denominators. Across five systems, speech presence repeatedly fails to determine those capabilities, so a single response rate would conceal the behaviour that produced it. The separate scores expose missed requests, low-content responding, intrusion and near-continuous speech as distinct failure profiles. The paired intervention produces a large response-rate effect for one transcript-conditioned reference, while no significant paired response presence effect is detected for the evaluated speech systems. The same taxonomy and automated generation and evaluation pipeline support larger or fresh evaluation draws rather than relying on a fixed hand-labelled set.

AI USE STATEMENT

Generative models are part of both the benchmark and the research process. Claude Opus 5 generates scenario scripts and paired rewrites and supplies the semantic verdicts used to score the speech systems' conditional answer accuracy and silence preservation. Qwen3-TTS synthesises human turns, and Qwen3-ASR-1.7B transcribes assistant output. We evaluate Gemini 3.1 Pro on the scenarios' speaker-attributed text, measuring its decisions to respond or remain silent and its answers to T requests. The evaluated systems are themselves generative models.

We used an AI coding assistant to develop and debug generation, evaluation, scoring and table scripts, execute experiments, verify bibliographic records, and draft or revise manuscript text. Reported values are computed from model outputs and scoring decisions. The authors reviewed the code, artifacts and claims and take responsibility for the final submission.

ETHICS STATEMENT

All benchmark conversations are generated and all speech is synthesised under a deterministic speaker-to-voice map. No participant was recorded and no real conversation or personal record was collected for the dataset.

Duplex-MPE's synthetic English scenes do not represent real populations, room acoustics or social norms, and they may inherit biases from the generating and speech-synthesis models. Results should therefore be interpreted only for the benchmark scenarios and should not support claims about particular speaker groups.

REPRODUCIBILITY STATEMENT

All reported values are computed by scripts from fixed manifests, model outputs and verdict files. We will release the scenario manifests, per-turn audio and construction boundaries, the duty preamble, scheduler and detector settings, per-scenario seeds, and scoring code.

REFERENCES

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/681fe4ec554beabdc9c84a1780cd5a8a-Abstract-Datasets_and_Benchmarks_Track.html.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=2e4ECh0ikn.

Kratika Bhagtani, Mrinal Anand, Yu Chen Xu, and Amit Kumar Singh Yadav. Speak or stay silent: Context-aware turn-taking in multi-party dialogue. *CoRR*, abs/2603.11409, 2026. doi: 10.48550/ARXIV.2603.11409. URL https://doi.org/10.48550/arXiv.2603.11409.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried M. Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio (eds.), *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pp. 28–39. Springer, 2005. doi: 10.1007/11677482_3. URL https://doi.org/10.1007/11677482_3.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *Trans. Assoc. Comput. Linguistics*, 14:378–398, 2026. doi: 10.1162/TACL.A.628. URL https://doi.org/10.1162/tacl.a.628.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037, 2024. doi: 10.48550/ARXIV.2410.00037. URL https://doi.org/10.48550/arXiv.2410.00037.

Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972. doi: 10.1037/h0033031.

Erik Ekstedt and Gabriel Skantze. Turngpt: a transformer-based language model for predicting turn-taking in spoken dialog. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 2981–2990. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.268. URL https://doi.org/10.18653/v1/2020.findings-emnlp.268.

Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 5190–5194. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-10955. URL https://doi.org/10.21437/Interspeech.2022-10955.

Ryo Fukuda, Takatomo Kano, Siddhant Arora, Marc Delcroix, Naohiro Tawara, Atsunori Ogawa, Yuya Chiba, Atsushi Ando, William Chen, and Shinji Watanabe. Evaluating large language models abilities for addressee, turn-change, and next speaker prediction in meetings. *CoRR*, abs/2606.17542, 2026. doi: 10.48550/ARXIV.2606.17542. URL https://doi.org/10.48550/arXiv.2606.17542.

Kuofeng Gao, Shutao Xia, Ke Xu, Philip Torr, and Jindong Gu. Benchmarking open-ended audio dialogue understanding for large audio-language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 4763–4784. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.237. URL https://doi.org/10.18653/v1/2025.acl-long.237.

Vineet Garg, Ognjen Rudovic, Pranay Dighe, Ahmed Hussen Abdelaziz, Erik Marchi, Saurabh Adya, Chandra Dhir, and Ahmed H. Tewfik. Device-directed speech detection: Regularization via distillation for weakly-supervised models. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 1258–1262. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-11228. URL https://doi.org/10.21437/Interspeech.2022-11228.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 3682–3692. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.285. URL https://doi.org/10.18653/v1/2021.acl-long.285.

Zhang He, Wenqian Cui, Haoning Xu, Xiao-Hui Li, Lei Zhu, Haoli Bai, Shaohua Ma, and Irwin King. Mtr-duplexbench: Towards a comprehensive evaluation of multi-round conversations for full-duplex speech language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 5334–5351. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.263. URL https://doi.org/10.18653/v1/2026.findings-acl.263.

Chien-yu Huang, Wei-Chih Chen, Shu-Wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Alejandro Ritter Gutierrez, and et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=s7lzZpAW7T.

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. An LLM benchmark for addressee recognition in multi-modal multi-party dialogue. In María Inés Torres, Yuki Matsuda, Zoraida Callejas, Arantza del Pozo, and Luis Fernando D'Haro (eds.), *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, IWSDS 2025, Bilbao, Spain, May 27-30, 2025*, pp. 330–334. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.iwsds-1.36/.

Natasa Jovanovic and Rieks op den Akker. Towards automatic addressee identification in multi-party dialogues. In Michael Strube and Candy L. Sidner (eds.), *Proceedings of the SIGDIAL 2004 Workshop, The 5th Annual Meeting of the Special Interest Group on Discourse and Dialogue, April 30 - May 1, 2004, Cambridge, Massachusetts, USA*, pp. 89–92. The Association for Computer Linguistics, 2004. URL https://aclanthology.org/W04-2317/.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. A corpus for studying addressing behaviour in multi-party dialogues. *Lang. Resour. Evaluation*, 40(1):5–23, 2006. doi: 10.1007/S10579-006-9006-4. URL https://doi.org/10.1007/s10579-006-9006-4.

David Joohun Kim, Daniyal Anjum, Bonny Banerjee, and Omar Abbasi. Selective attention system (sas): Device-addressed speech detection for real-time on-device voice ai, 2026. URL https://arxiv.org/abs/2604.08412.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-Yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, pp. 1–8. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11433838. URL https://doi.org/10.1109/ASRU65441.2025.11433838.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-Yi Lee. Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19447–19451. IEEE, May 2026a. doi: 10.1109/icassp55912.2026.11463576. URL http://dx.doi.org/10.1109/icassp55912.2026.11463576.

Zhaojiang Lin, Yong Xu, Kai Sun, Jing Zheng, Yin Huang, Surya Teja Appini, Krish Narang, Renjie Tao, Ishan Kapil Jain, Siddhant Arora, Ruizhi Li, Yiteng Huang, Kaushik Patnaik, Wenfang Xu, Suwon Shon, Yue Liu, Ahmed A Aly, Anuj Kumar, Florian Metze, and Xin Luna Dong. Wearvox: An egocentric multichannel voice assistant benchmark for wearables, 2026b. URL https://arxiv.org/abs/2601.02391.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. Proactive conversational agents with inner thoughts. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas (eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pp. 184:1–184:19. ACM, 2025. doi: 10.1145/3706598.3713760. URL https://doi.org/10.1145/3706598.3713760.

Chengqian Ma, Wei Tao, and Steven Y. Guo. C3: A bilingual benchmark for spoken dialogue models exploring challenges in complex conversations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 22778–22796. Association for Computational Linguistics, 2025.

doi: 10.18653/V1/2025.EMNLP-MAIN.1160. URL https://doi.org/10.18653/v1/2025.emnlp-main.1160.

Gallil Maimon, Amit Roth, and Yossi Adi. Salmon: A suite for acoustic language model evaluation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10888561. URL https://doi.org/10.1109/ICASSP49660.2025.10888561.

Vihaan Nama, Shreya Mendi, Zian Ye, and Brinnae Bent. When2speak: A dataset for temporal participation and turn-taking in multi-party conversations for large language models. *CoRR*, abs/2605.05626, 2026. doi: 10.48550/ARXIV.2605.05626. URL https://doi.org/10.48550/arXiv.2605.05626.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266, 2023. doi: 10.1162/TACL_A_00545. URL https://doi.org/10.1162/tacl_a_00545.

Hiroki Ouchi and Yuta Tsuboi. Addressee and response selection for multi-party conversation. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2133–2143. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1231. URL https://doi.org/10.18653/v1/d16-1231.

Shruti Palaskar, Ognjen Rudovic, Sameer Dharur, Florian Pesce, Gautam Krishna, Aswin Sivaraman, Jack Berkowitz, Ahmed Hussen Abdelaziz, Saurabh Adya, and Ahmed H. Tewfik. Multimodal large language models with fusion low rank adaptation for device directed speech detection. In Itshak Lapidot and Sharon Gannot (eds.), *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*. ISCA, 2024. doi: 10.21437/INTERSPEECH.2024-1361. URL https://doi.org/10.21437/Interspeech.2024-1361.

Deep Anil Patel, Iain Melvin, Christopher Malon, and Martin Renqiang Min. Discussllm: Teaching large language models when to speak. *CoRR*, abs/2508.18167, 2025. doi: 10.48550/ARXIV.2508.18167. URL https://doi.org/10.48550/arXiv.2508.18167.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. Fd-bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems. In Odette Scharenborg, Catharine Oertel, and Khiet Truong (eds.), *26th Annual Conference of the International Speech Communication Association, Interspeech 2025, Rotterdam, The Netherlands, 17-21 August 2025*. ISCA, 2025. doi: 10.21437/INTERSPEECH.2025-739. URL https://doi.org/10.21437/Interspeech.2025-739.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. Multimodal continuous turn-taking prediction using multiscale rnns. In Sidney K. D'Mello, Panayiotis G. Georgiou, Stefan Scherer, Emily Mower Provost, Mohammad Soleymani, and Marcelo Worsley (eds.), *Proceedings of the 2018 on International Conference on Multimodal Interaction, ICMI 2018, Boulder, CO, USA, October 16-20, 2018*, pp. 186–190. ACM, 2018. doi: 10.1145/3242969.3242997. URL https://doi.org/10.1145/3242969.3242997.

Ognjen Rudovic, Pranay Dighe, Yi Su, Vineet Garg, Sameer Dharur, Xiaochuan Niu, Ahmed Hussen Abdelaziz, Saurabh Adya, and Ahmed H. Tewfik. Device-directed speech detection for follow-up conversations using large language models. *CoRR*, abs/2411.00023, 2024. doi: 10.48550/ARXIV.2411.00023. URL https://doi.org/10.48550/arXiv.2411.00023.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. doi: 10.2307/412243.

Gabriel Skantze. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie

Louis (eds.), *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pp. 220–230. Association for Computational Linguistics, 2017. doi: 10.18653/V1/W17-5527. URL https://doi.org/10.18653/v1/w17-5527.

Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.*, 67:101178, 2021. doi: 10.1016/J.CSL.2020.101178. URL https://doi.org/10.1016/j.csl.2020.101178.

Zhaokai Sun, Shuai Wang, Zhennan Lin, Chengyou Wang, Dehui Gao, Yuang Cao, Chunjiang He, Pan Zhou, and Lei Xie. Msu-bench: Towards speaker-centric understanding in conversational multi-speaker scenarios, 2026. URL https://arxiv.org/abs/2606.22868.

Dominik Wagner, Alexander W. Churchill, Siddharth Sigtia, Panayiotis G. Georgiou, Matt Mirsamadi, Aarshee Mishra, and Erik Marchi. A multimodal approach to device-directed speech detection with large language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 10451–10455. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10446224. URL https://doi.org/10.1109/ICASSP48485.2024.10446224.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. Audiobench: A universal benchmark for audio large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4297–4316. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.218. URL https://doi.org/10.18653/v1/2025.naacl-long.218.

Chengyou Wang, Hongfei Xue, Guojian Li, Zhixian Zhao, Shuiyuan Wang, Shuai Wang, Xin Xu, Hui Bu, and Lei Xie. Full-duplex interaction in spoken dialogue systems: A comprehensive study from the icassp 2026 humdial challenge, 2026. URL https://arxiv.org/abs/2604.21406.

Ke Xu, Yuhao Wang, and Yu Wang. From reactive to proactive: Assessing the proactivity of voice agents via provoice-bench, 2026. URL https://arxiv.org/abs/2604.15037.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. Air-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1979–1998. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.109. URL https://doi.org/10.18653/v1/2024.acl-long.109.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: speech processing universal performance benchmark. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek (eds.), *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pp. 1194–1198. ISCA, 2021. doi: 10.21437/INTERSPEECH.2021-1775. URL https://doi.org/10.21437/Interspeech.2021-1775.

A1 Related work and evaluated systems

A1.1 Full-duplex dialogue and turn-taking

Turn-taking research studies when a speaker can take or retain the floor (Sacks et al., 1974; Duncan, 1972; Skantze, 2021). Continuous prediction, lexical completeness and voice-activity projection provide cues for these decisions (Skantze, 2017; Roddy et al., 2018; Ekstedt & Skantze,

2020; 2022). Audio dialogue models such as dGSLM and Moshi generate parallel conversational streams (Nguyen et al., 2023; Défossez et al., 2024). Addressee recognition separately identifies whom an utterance addresses, using meeting recordings, text conversations or multimodal inputs (Jovanovic & op den Akker, 2004; Ouchi & Tsuboi, 2016; Gu et al., 2021; Inoue et al., 2025; Fukuda et al., 2026). Duplex-MPE combines these decisions in an audio-only interaction where the model's own speech determines the outcome.

Benchmark task details. Most audio benchmarks score understanding or generated answer quality on isolated inputs. They cover speech-task generalisation under instructions (Yang et al., 2021; Huang et al., 2025), audio-language comprehension (Yang et al., 2024; Wang et al., 2025), voice-assistant behaviour and acoustic conditions (Chen et al., 2026; Maimon et al., 2025), and spoken dialogue understanding beyond the literal words and in more than one language (Ao et al., 2024; Gao et al., 2025; Ma et al., 2025). In all of them a response is expected on every item, so silence is never a correct output. Recent full-duplex benchmarks instead evaluate interactive behaviour on continuous or multi-turn exchanges.

Addressee recognition interfaces. Recent benchmarks extend this setting to LLM and multimodal inputs. Inoue et al. (2025) gives GPT-4o five manually transcribed context turns and asks for one of four labels, A, B, C or O; Fukuda et al. (2026) supplies ground-truth speaker IDs and utterance-aligned transcripts, audio and video from AMI, then scores predictions over four participants, GROUP and NONE with accuracy and macro-F1. Taken together, prior work establishes full-duplex interaction, addressee recognition and response rejection as related but distinct evaluation problems. Duplex-MPE combines them in a full-duplex multi-party protocol in which every speaker contributes to the dialogue state, the assistant is one of several possible addressees, and the observed output is the waveform the model chooses to emit. The benchmark evaluates responding and withholding across every turn and pairs each scenario across two addressing forms while holding the scenario and gold answer fixed.

A1.2 Adjacent evaluation tasks

ProVoice-Bench evaluates proactive speech triggered by implicit intent, user-defined conditions, contextual contradictions and acoustic events (Xu et al., 2026). Its single-user setting differs from deciding whom to answer in a multi-party conversation. MSU-Bench evaluates speaker-centric understanding across 16 tasks, including who spoke and what was meant (Sun et al., 2026). These understanding and proactivity tasks complement the speech-control behaviour evaluated here.

A1.3 Evaluated systems

The five systems were selected for publicly available code and weights, continuous audio input, and an inference interface that lets the model decide when to emit speech while continuing to receive input. The evaluated model weights are hosted in the following Hugging Face repositories:

- MiniCPM-o 4.5: `https://huggingface.co/openbmb/MiniCPM-o-4_5`.
- Moshi: `https://huggingface.co/kyutai/moshiko-pytorch-bf16`.
- FLM-Audio: `https://huggingface.co/CofeAI/FLM-Audio`.
- Voila: `https://huggingface.co/maitrix-org/Voila-autonomous-preview`.
- Freeze-Omni: `https://huggingface.co/VITA-MLLM/Freeze-Omni`.

Why non-full-duplex models are not evaluated. Duplex-MPE requires a model to receive continuous room audio, decide when to speak, and keep listening while speaking. Adapting a non-full-duplex model would require evaluator-selected segmentation or invocation points, making fresh-onset response rate partly dependent on the harness and leaving answering-window yield undefined. Scored evaluation is therefore restricted to full-duplex systems; instruction sensitivity under a segmented interface is reported only as a non-scored probe.

Moshi uses the male-voice moshiko checkpoint. Throughout the paper, Voila denotes the autonomous-preview checkpoint listed above.

Table 4: A complete scenario with explicit T addressing: four classmates choosing a venue, 13 turns, 99.2 s.

#	Speaker	Turn type	Aria must	Utterance
1	A	N2	stay silent	Okay, option one: the taco place on Fifth. Twelve minute walk, and it's like nine bucks a head if you skip the fancy drinks.
2	B	N2	stay silent	Option two, the karaoke bar by Ninth. That's twenty-five minutes on the bus and fifteen each, but they throw in free chips, so morally it's cheaper.
3	C	N2	stay silent	Or Luigi's, which is basically downstairs, four minutes, eleven a person, and zero karaoke, tragically.
4	D	T	**speak**	Alright, somebody settle this before we all starve ; Aria, out of those three, which one's the cheapest per person?
5	A	N1	stay silent	See, this is exactly why we keep Aria in here. B has never once done the math right.
6	B	N2	stay silent	Rude. Hey Siri, how long is the bus from campus to Ninth and Pine right now?
7	C	N3	stay silent	Nine bucks... I've got eleven until Friday, so that's, uh, two dollars of personality left.
8	D	N2	stay silent	C, be honest, do you actually care about karaoke or are you just in it for the free chips?
9	C	N2	stay silent	Chips. Entirely chips. I sing like a smoke alarm and everyone knows it.
10	A	$N4_Q$	**answer**	Hold on, Aria, does the taco place take card?
11	A	$N4_R$	**yield**	Ugh, never mind, I paid card there Tuesday, we're fine.
12	B	N1	stay silent	Fine, tacos win. But Aria gets absolutely no vote in the karaoke rematch next week.
13	D	N2	stay silent	Deal. Lobby at seven, and whoever shows up late is buying the guacamole for the table.

Table 5: Explicit and implicit versions of the T request in Table 4.

Turn 4, the only turn whose wording differs in this example:	
Explicit	Alright, somebody settle this before we all starve ; Aria, out of those three, which one's the cheapest per person?
Implicit	Alright, somebody settle this before we all starve ; you've been logging all three numbers, so run it: which one's cheapest per person?

A2 Dataset construction and accounting

A2.1 A complete scenario and its paired request

Table 4 shows a complete scenario, and Table 5 shows its alternative T wording. The T request in turn 4 requires comparing prices from turns 1 to 3; the gold answer is the taco place at nine dollars per person. Turns 1 to 3 introduce options to the other participants, and turn 13 closes the discussion with a meeting arrangement. All four are N2 turns: they contribute to the conversation but do not address Aria, so it should remain silent. Turns 5 and 12 mention Aria without requesting a response; turns 10 and 11 are $N4_Q$ and $N4_R$, respectively. Only T's text differs in this pair, and the gold answer is identical.

A2.2 Scenario composition and request placement

The 2,000 scenarios comprise 1,500 three-human and 500 four-human conversations, each with one assistant. Generation assigns T to early-middle, middle or late-middle position buckets in 667, 667 and 666 scenarios, respectively. Every scenario contains exactly one T, which is never first and is followed by further conversation. There are 968 scenarios with an N4 event after T. An N4 question is addressed to Aria by design but is followed by a human answer or instruction that the assistant need not answer; it is not a second T. The position constraint and the explicit/implicit addressing assignment are separate controls.

A2.3 SPEECH SYNTHESIS

Qwen3-TTS synthesises human utterances separately using a fixed speaker-to-voice map. A, B, C and D are identifiers assigned to the speakers within each scenario, not personal names. They select the voices `ryan`, `aiden`, `eric` and `dylan`, respectively, with clips stored at 24 kHz. Names mentioned in an utterance do not introduce additional speakers or voices. Each manifest record contains the scenario and pair IDs, speaker, turn text, event label, audio path, duration and T gold answer.

Conversation duration. The 4,000 constructed conversations total 110.93 hours, including reused clips and inter-turn gaps; the spoken duty preamble and model-dependent waits are additional. An evaluated conversation lasts 99.8 s on average; App. A2.5 provides the duration distributions.

A2.4 PAIRED EDITS AND AUDIO ACCOUNTING

The addressing rewrite changes only T in 1,958 pairs and also changes the preceding turn in 42 pairs. The final scripts, including N4 edits, differ only at T in 1,661 pairs; 339 pairs have additional text differences, including 16 with unequal turn counts. All pairs retain the same T gold answer.

The two addressing conditions contain 2,000 scenarios each. The final human clips total 105.55 h across all 4,000 conversations, counting clips reused between versions each time they occur. Adding 5.37 h of 0.4 s inter-turn gaps gives 110.93 h of constructed conversations. Of the clip duration, 44.45 h is reused between versions, leaving 61.10 h of distinct clips. These durations exclude the spoken duty preamble and model-dependent waits during evaluation.

A2.5 DURATION STATISTICS

Tables 6 and 7 report scene and utterance durations by T addressing condition; Table 8 focuses on the T requests. The "All turns" rows in Table 7 sum T, N1–N3, $N4_Q$ and $N4_R$, excluding silence inserted between clips. Totals are computed before rounding the displayed values. Named and unnamed T requests average 8.18 s and 8.58 s, respectively. These are descriptive duration comparisons, not a separate causal estimate of the effect of wording.

Scene-level durations in Table 6 include 0.4 s gaps between consecutive human clips. Tables 7 and 8 use only the individual clip durations, with no added inter-turn silence.

Table 6: Scene duration by T addressing condition, in seconds except total hours. p5/p95 denote the fifth/ninety-fifth percentiles.

Addressing	Scenarios	Total (h)	Mean	Median	p5	p95	Max
Explicit	2,000	55.27	99.5	98.2	79.4	123.0	297.6
Implicit	2,000	55.66	100.2	98.9	79.6	123.8	298.0

A3 RUNTIME PROTOCOL AND SCORING

A3.1 SHARED SCHEDULER AND DUTY INSTRUCTION

Each episode starts from fresh model state with the 31.9 s spoken duty preamble, followed by 0.5 s of silence before the first human turn. Between ordinary human turns, the scheduler plays 0.5 s of silence after one clip ends and before the next begins. After T, the model has 5 s to start speaking; once speech is present, playback waits for 2 s of confirmed silence or a 60 s cap. The cap starts at a new onset, or at T's end if the model is already speaking. After $N4_Q$, the scheduler always waits 3 s and then plays $N4_R$, regardless of whether the model has finished. The same settings apply to all five systems.

All timing is measured on the decoded assistant waveform with a causal Silero voice-activity detector (VAD). The detector requires at least 120 ms of speech, merges pauses shorter than 800 ms within

Table 7: Turn duration by type and T addressing condition, in seconds except total minutes. Totals exclude inter-turn silence.

Addressing	Turn type	n	Total (min)	Mean	Median	Max
Explicit	T (request)	2,000	272.6	8.18	7.84	24.16
	N1	3,041	387.0	7.64	7.28	188.72
	N2	14,052	1754.1	7.49	7.12	49.60
	N3	3,267	427.5	7.85	7.44	41.92
	$N4_Q$	1,911	128.8	4.04	3.68	12.64
	$N4_R$	1,911	184.8	5.80	5.36	27.60
	All turns	26,182	3154.9	7.23	6.96	188.72
Implicit	T (request)	2,000	285.8	8.58	8.24	24.40
	N1	3,039	387.6	7.65	7.28	188.72
	N2	14,054	1760.1	7.51	7.12	49.60
	N3	3,267	428.1	7.86	7.44	41.92
	$N4_Q$	1,911	129.8	4.08	3.68	22.96
	$N4_R$	1,911	186.8	5.86	5.44	27.60
	All turns	26,182	3178.3	7.28	6.96	188.72

Table 8: T-utterance duration by addressing condition. Mean and median are in seconds; chars denotes mean character count.

Addressing	n	Mean	Median	Chars
Explicit	2,000	8.18	7.84	111.6
Implicit	2,000	8.58	8.24	123.3

an episode, and applies the separate 2 s endpoint rule only when deciding whether a T response has finished. Token timestamps are not used because vocoding precedes audible output.

The spoken instruction, also used verbatim by the text reference, is:

> You are the voice assistant Aria, that is A R I A, Aria. You are in a room with several people. Only respond when someone speaks directly to you, Aria, and asks something you can answer. When people are talking to each other or to another device, or thinking out loud, or when the question they asked you has already been answered by someone else, stay silent and do not speak.

A3.2 Fresh onset and response presence

Table 9 partitions T requests into fresh onset, speech already active at request completion, and no response within the deadline. Response presence is the sum of the first two rates. A response exceeding the 60 s cap still belongs to its original onset category; duration and initiation are separate outcomes.

Metric-dependent rankings. The choice between fresh-onset response rate and response presence does change the ranking. Pooling the five models' explicit and implicit results gives ten entries; the rankings of these entries under the two metrics have Spearman correlation -0.4303. Under explicit addressing, none of the five models retains the same rank. Freeze-Omni moves from first under response presence to last under fresh-onset response rate because 0.7430 of its T turns contain continuation speech, whereas MiniCPM-o's response presence and fresh-onset response rate differ by 0.0045 (App. A3.2).

Table 9: Fresh speech, continuation and silence across the 2,000 T requests per addressing condition.

Model	Addressing	Fresh onset	Already speaking	Miss	Fresh-onset response rate	Response presence
MiniCPM-o 4.5	Explicit	1,900	9	91	0.9500	0.9545
	Implicit	1,887	10	103	0.9435	0.9485
Moshi	Explicit	1,268	363	369	0.6340	0.8155
	Implicit	1,340	304	356	0.6700	0.8220
FLM-Audio	Explicit	1,076	825	99	0.5380	0.9505
	Implicit	1,085	823	92	0.5425	0.9540
Voila	Explicit	1,293	10	697	0.6465	0.6515
	Implicit	1,334	4	662	0.6670	0.6690
Freeze-Omni	Explicit	505	1,486	9	0.2525	0.9955
	Implicit	516	1,479	5	0.2580	0.9975

A3.3 SEMANTIC SCORING

For conditional answer accuracy, Qwen3-ASR-1.7B transcribes the response and Claude Opus 5 compares it with the gold answer, ignoring filler and wording differences. An empty response transcript is scored as incorrect.

For silence preservation, a window with no model speech detected by VAD passes directly, without ASR or an Opus call. Only windows with detected model speech require transcription and semantic classification. All detected model-speech fragments inside one window are joined and transcribed together. Claude Opus 5 receives that text, up to three human turns ending at the current turn, the total model speech duration and its overlap with human speech.

For silence preservation, a brief acknowledgement that adds no information and answers no question is allowed (`BACKCHANNEL`). Answering, explaining, advising, asking a new question, or repeating acknowledgements until they hold the floor fails this metric (`SUBSTANTIVE_INTRUSION`). Duration alone does not determine whether speech is allowed: even a short answer such as "Nine dollars" counts as an intrusion.

An empty transcript after detected speech does not establish that the model was silent or produced an allowed acknowledgement. The window therefore fails silence preservation; it also fails if Opus provides no classification or returns an unrecognised label. These are scoring decisions in the absence of a content classification, not confirmed substantive intrusions.

Table 10 counts only N1–N3 windows. Its columns report windows where VAD detects model speech, the subset with an empty transcript, and the subset with a non-empty transcript but no valid Opus label. For example, in the explicit condition Voila has detected speech in $4{,}765$ windows; 171 have no transcript, and all remaining windows receive an Opus label. Treating those 171 windows as passes would add $171/20{,}360 \approx 0.00840$ to Voila's silence preservation, or 0.84 percentage points, raising it from 83.51% to 84.35% without changing the model ordering.

A3.4 HUMAN VALIDATION OF DATA AND SEMANTIC SCORING

Six students who understand spoken and written English evaluate the sampled materials. Each item is assessed by two reviewers; when their judgments differ, those same two reviewers discuss the item and determine its final label. We compare these final labels with the original Opus 5 verdicts to assess semantic scoring reliability.

For construction validation, we randomly sample 50 scenario pairs: 38 with three human speakers and 12 with four, including both addressing versions of each pair. These provide 100 conversation items containing $1{,}282$ human-turn occurrences and 95 N4 question–resolution pairs. Reviewers check whether T and $N4_Q$ address Aria, whether T has a unique answer supported by the preceding dialogue, whether the supplied gold answer is supported, whether $N4_R$ answers or cancels its request, and whether N1–N3 turns require silence rather than a substantive response. T answerability is judged using only the request and preceding dialogue, before revealing the gold answer or subsequent turns. $N4_Q$ is checked for its addressee, without requiring an answer inferable from the

Table 10: Window counts at each stage of silence-preservation scoring.

Model	Addressing	VAD detects speech	Empty ASR transcript	No valid Opus label
MiniCPM-o 4.5	Explicit	2,172	7	0
	Implicit	2,071	6	0
Moshi	Explicit	18,632	127	3
	Implicit	18,653	118	3
FLM-Audio	Explicit	12,492	37	5
	Implicit	12,411	43	5
Voila	Explicit	4,765	171	0
	Implicit	4,825	189	0
Freeze-Omni	Explicit	20,356	0	3
	Implicit	20,356	0	8

preceding dialogue. Reviewers also check whether Qwen3-TTS speech is intelligible and preserves the script's meaning.

To validate model evaluation, we separately sample outputs from the full benchmark runs. Each of the five models is evaluated on $2{,}000$ scenarios per addressing condition; from its available outputs in each condition, we select ten T responses and ten N1–N3 speech outputs. This gives $5 \times 2 \times 10 = 100$ outputs for each semantic scoring task, sampled independently of the 50 conversation pairs above. Within each model and condition, we randomly draw five T responses labelled correct and five labelled incorrect by Opus 5; for N1–N3 outputs, we draw five labelled acknowledgement and five labelled intrusion. When a label has fewer than five examples, we include all of them and randomly fill the remaining slots from the other label. Reviewers check whether Qwen3-ASR-1.7B transcripts preserve the meaning of these 200 output clips and assess answer correctness or acknowledgement versus intrusion for comparison with Opus 5. Together with the 100 conversation items, this gives 300 review items, each assessed by two reviewers, for 600 initial assessments.

Table 11: Human validation of sampled conversations and model outputs.

Check	**Sample size**	**Rate**
Construction: Opus 5 and Qwen3-TTS		
T addresses Aria: explicit	50	100%
T addresses Aria: implicit	50	100%
T has a unique answer supported by preceding dialogue	100	100%
Gold answer is correct and supported	100	98%
$N4_Q$ addresses Aria	95	100%
$N4_R$ answers or cancels its request	95	100%
N1–N3 turn requires no substantive response	992	99.70%
TTS is intelligible and preserves script meaning	1,282	100%
Evaluation: Qwen3-ASR-1.7B and Opus 5		
ASR preserves model output meaning	200	100%
Agreement with Opus 5: T answer correctness	100	99%
Agreement with Opus 5: acknowledgement versus intrusion	100	98%

Table 11 summarises validation of benchmark construction and model evaluation. Construction checks achieve 98%–100% pass rates, including intelligible and meaning-preserving TTS for all checked turns. For evaluation, all checked ASR transcripts preserve meaning, and reviewers' final labels agree with Opus 5 on 99% of answer-correctness judgments and 98% of acknowledgement-versus-intrusion judgments. These results support the reliability of both stages on the sampled items.

A3.5 SENSITIVITY TO DETECTED-SPEECH DURATION

For each silence-requiring window, the sweep sums the overlap of stored model-speech segments with that window and counts occupancy when the sum exceeds $\tau \in \{0, 100, 200, 300, 500\}$ ms. Thus $\tau = 0$ counts any detected speech, while $\tau = 500$ ignores up to half a second. This check applies no semantic backchannel exemption. Table 12 shows that the model ordering is unchanged at every threshold under both addressing conditions. At $\tau = 0$, occupied-window counts match the number of semantic-judge records for every run. At larger thresholds, duration-only and semantic rules differ: the former can ignore short answers, whereas the latter can exempt longer non-substantive acknowledgements.

Table 12: Silence preservation under speech-duration thresholds, without semantic exemptions, alongside the reported semantic scores.

		Acoustic silence preservation (no semantic exemption)					Reported
Model	Addressing	0 ms	100	200	300	500	(semantic)
MiniCPM-o 4.5	Explicit	0.8933	0.8935	0.8939	0.8948	0.9125	0.9242
	Implicit	0.8983	0.8985	0.8988	0.8991	0.9166	0.9289
Moshi	Explicit	0.0849	0.0899	0.0946	0.1102	0.2122	0.2734
	Implicit	0.0838	0.0880	0.0930	0.1090	0.2128	0.2676
FLM-Audio	Explicit	0.3864	0.3877	0.3897	0.3917	0.3966	0.3957
	Implicit	0.3904	0.3921	0.3934	0.3955	0.3997	0.3991
Voila	Explicit	0.7660	0.7676	0.7934	0.8222	0.8690	0.8351
	Implicit	0.7630	0.7650	0.7916	0.8192	0.8661	0.8307
Freeze-Omni	Explicit	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	Implicit	0.0002	0.0002	0.0002	0.0002	0.0003	0.0002

A3.6 REQUEST TIMING AND RESPONSE QUALITY

Table 13 gives the complete results for the analysis in Sec. 4.4. Elapsed time t runs from the first human turn to T onset, including intervening speech and scheduled gaps but excluding the duty preamble; the observed range is 14.50–263.97 s. We sort the 4,000 requests by this time and divide them into three groups with approximately equal sample sizes. The one-third and two-thirds quantiles give boundaries of 40.224 and 54.656 s, placing 1,330, 1,335 and 1,335 requests in the earlier, middle and later groups. All models use the same grouping, determined from input timing rather than model scores. Conditional answer accuracy uses response-present requests as its denominator; the two response rates use all requests in the group.

A4 N4 STOPPING ANALYSIS

A4.1 ELIGIBILITY AND STOPPING RULE

Window response covers any detected model speech from $N4_Q$ onset through its 3 s answering window. Table 14 separates three cases: the model is already speaking when $N4_Q$ begins, starts speaking during $N4_Q$, or first speaks in the 3 s interval after $N4_Q$ ends. Only the third group can enter answering-window yield, and only if the model is still speaking when $N4_R$ begins. Models that finish speaking before $N4_R$ begins are excluded from the stopping denominator.

Let t_R be the time when the $N4_R$ human audio clip starts playing, and t_E the time when that clip ends plus 0.5 s of inter-turn silence. An eligible event passes if the model is silent at $t_R + 2$ s. If t_E is later than this deadline, the model must also remain silent from the deadline until t_E. The deadline stays at $t_R + 2$ s even when the human clip and its following silence end earlier. For example, if $N4_R$ lasts 1 s, then $t_E = t_R + 1.5$ s; an eligible model that stops at $t_R + 1.8$ s and is silent at the deadline passes, although it stops after the human clip ends. A model still speaking at $t_R + 2$ s fails. Stopping is computed from the detector's speech segments, which merge gaps shorter than 800 ms. Answer content is not part of this stopping test.

Freeze-Omni is already speaking when $N4_Q$ begins in 1,784 explicit-condition events and 1,789 implicit-condition events. It starts during $N4_Q$ in another 126 and 120 events, respectively, leaving

Table 13: Performance by request time; all rates are percentages.

Model	Addressing	T time	Requests	Fresh onset	Response presence	Answer accuracy
MiniCPM-o 4.5	Explicit	Earlier	670	96.57	97.46	50.69
		Middle	665	94.29	94.44	44.27
		Later	665	94.14	94.44	46.82
	Implicit	Earlier	660	94.70	95.45	51.27
		Middle	670	94.03	94.33	44.30
		Later	670	94.33	94.78	44.41
Moshi	Explicit	Earlier	670	62.39	81.04	1.66
		Middle	665	64.06	80.30	1.12
		Later	665	63.76	83.31	0.90
	Implicit	Earlier	660	66.82	80.76	1.88
		Middle	670	69.10	84.18	1.42
		Later	670	65.07	81.64	1.83
FLM-Audio	Explicit	Earlier	670	67.61	94.03	0.16
		Middle	665	46.02	96.39	0.00
		Later	665	47.67	94.74	0.16
	Implicit	Earlier	660	66.97	95.30	0.00
		Middle	670	47.01	95.82	0.16
		Later	670	48.96	95.07	0.00
Voila	Explicit	Earlier	670	71.04	71.64	0.21
		Middle	665	63.91	64.21	0.23
		Later	665	58.95	59.55	0.51
	Implicit	Earlier	660	75.91	76.21	0.99
		Middle	670	68.66	68.81	0.65
		Later	670	55.67	55.82	0.53
Freeze-Omni	Explicit	Earlier	670	23.43	99.55	0.45
		Middle	665	26.32	99.85	0.45
		Later	665	26.02	99.25	0.15
	Implicit	Earlier	660	28.48	99.70	0.15
		Middle	670	23.58	99.85	0.30
		Later	670	25.37	99.70	0.00

Table 14: Events entering answering-window yield. Q and R denote the N4 question and resolution; the answering interval is the 3 s gap between them.

			Model speech relative to Q			Among starts after Q	
Model	Addressing	Window response hits	Already speaking at Q start	Starts during Q	Starts after Q within 3 s	Silent at R start	Still speaking at R start
MiniCPM-o 4.5	Explicit	1,818	9	73	1,736	103	1,633
	Implicit	1,808	6	62	1,740	76	1,664
Moshi	Explicit	1,763	913	424	426	207	219
	Implicit	1,779	906	425	448	206	242
FLM-Audio	Explicit	1,897	618	283	996	19	977
	Implicit	1,899	598	269	1,032	16	1,016
Voila	Explicit	1,188	39	236	913	293	620
	Implicit	1,184	41	223	920	270	650
Freeze-Omni	Explicit	1,911	1,784	126	1	0	1
	Implicit	1,910	1,789	120	1	0	1

one eligible event per condition. All four other models have substantially larger denominators. Conditional stopping rates must therefore be read together with the counts in Table 14.

A4.2 Stopping outcomes

Table 15 accounts for every scored event: successful stopping, speech spanning the deadline, or silence at the deadline followed by later speech. In the explicit condition, FLM-Audio is speaking at the deadline in 968 of 977 events. Moshi is silent at the deadline but speaks again before observation ends in 97 of its 177 failures. These events fail the subsequent-silence check even though the model briefly paused.

Table 15: Stopping outcomes among scored N4 events. *Still voicing*: speaking at the 2 s deadline; *resumed*: silent then, but speaking again before observation ends.

Model	Addressing	Scored	Yielded	Still voicing	Resumed
MiniCPM-o 4.5	Explicit	1,633	974	654	5
	Implicit	1,664	986	672	6
Moshi	Explicit	219	42	80	97
	Implicit	242	51	91	100
FLM-Audio	Explicit	977	8	968	1
	Implicit	1,016	6	1,010	0
Voila	Explicit	620	526	44	50
	Implicit	650	523	47	80
Freeze-Omni	Explicit	1	0	1	0
	Implicit	1	0	1	0

A5 Addressing and transcript-reference controls

A5.1 Paired addressing comparison

Each of the 2,000 scenarios is evaluated in both an explicit and an implicit addressing version, giving two observations of the same model on the corresponding T request. Here, a response means detected model speech in T's response window, including speech already underway; answer correctness is evaluated separately. Table 16 places each pair in one of four groups: speech in both versions, only in the explicit version, only in the implicit version, or in neither. These four counts sum to 2,000 for each model.

The explicit-minus-implicit difference measures the net change in how often the model speaks between the paired versions. It is the explicit-only count minus the implicit-only count, divided by

2,000; the table multiplies this value by 100 to express it in percentage points. For MiniCPM-o, 90 pairs favour the explicit version and 78 favour the implicit version, giving 12 more responses overall, or 0.60 percentage points. This net difference can be small even when the model changes its response on many individual pairs.

The exact McNemar test asks whether the explicit-only and implicit-only response counts are consistent with equal response probabilities. Under this assumption, its p value is the probability of an imbalance at least as large as the observed one, given the number of pairs with different responses. The exact two-sided McNemar test evaluates whether the two directions of change are equally likely. The significance level $\alpha = 0.05$ is the threshold applied to the table's p column: only $p < 0.05$ is called statistically significant. MiniCPM-o's $p = 0.3961$ exceeds this threshold, so its 90 versus 78 split does not provide sufficient evidence of unequal response probabilities. The same holds for all five models; insufficient evidence of a difference does not establish that the conditions are equivalent.

Table 16: Paired addressing comparisons over 2,000 scenario pairs.

	Model speech after T in each pair					
Model	Both versions	Explicit only	Implicit only	Neither	Explicit − implicit (percentage points)	p
MiniCPM-o 4.5	1,819	90	78	13	+0.60	0.3961
Moshi	1,355	276	289	80	−0.65	0.6137
FLM-Audio	1,815	86	93	6	−0.35	0.6539
Voila	908	395	430	267	−1.75	0.2365
Freeze-Omni	1,986	5	9	0	−0.20	0.4240

A5.2 TEXT-BASED EVALUATION OF RESPONSE DECISIONS AND ANSWERS

We evaluate whether Gemini 3.1 Pro chooses to respond to the right utterances and answers T requests correctly when given text instead of audio. For each evaluated utterance, it receives that utterance, the duty instruction in App. A3.1 and preceding speaker-attributed dialogue, then chooses `RESPOND` or `SILENT`. Decisions are queried independently using the script history, without future utterances or previous model-generated answers.

The first four result columns in Table 17 correspond to the following tasks:

- **T requests**: response presence is the fraction receiving a `RESPOND` decision. For these requests only, Gemini generates an answer, and a separate Gemini call compares it with the gold answer to compute conditional answer accuracy.
- **N1, N2 and N3 utterances**: silence preservation is the fraction receiving a `SILENT` decision.
- **$N4_Q$ questions**: window response is the fraction receiving a `RESPOND` decision. No answer is generated for these questions, and this text decision does not measure whether speech begins within three seconds.

The text evaluation produces no speech timeline, so it does not measure fresh-onset response rate or stopping after $N4_R$ begins.

The final column is a separate diagnostic: Gemini is asked whether each $N4_Q$ question addresses Aria, regardless of whether it can answer. All these questions are intended for Aria; the column reports how often Gemini identifies that addressee, and does not enter any of the four preceding metrics.

The explicit and implicit rows group scenarios by T's addressing form; their N4 and silence-preservation columns do not imply that every turn in those scenarios has that addressing form. Across all 2,000 matched pairs, explicit and implicit response rates are 0.9470 and 0.3040, giving a difference of 0.6430.

To examine consistency across runs, we compare two text evaluations of Gemini 3.1 Pro on the same scenario versions with identical duty instructions. The runs agree on 97.2% of the 48,542

Table 17: Transcript-conditioned results for Gemini 3.1 Pro. Bold marks the higher value across conditions in columns marked ↑.

Addressing	Response presence (T)	Conditional answer accuracy↑	Silence preservation↑	Window response ($N4_Q$)	$N4_Q$: Gemini identifies Aria
Explicit	0.9470	**0.9039**	0.9537	0.7425	0.9236
Implicit	0.3040	0.7911	**0.9934**	0.7609	0.9299

decisions to respond or remain silent. For Gemini, response presence, silence preservation and window response differ by at most 0.95 percentage points between runs. Its conditional answer accuracy changes from 90.39% to 97.02% on explicit requests, a 6.63-percentage-point increase. On implicit requests it changes from 79.11% to 81.03%, a 1.92-percentage-point increase. The paired addressing differences are 0.6430 and 0.6455, so the large addressing contrast is reproduced despite the answer-accuracy variation.

A5.3 Silence preservation by turn type

Table 18 breaks down text-reference silence preservation by taxonomy label.

Table 18: Transcript-conditioned silence preservation by turn type.

Addressing	N1	N2	N3	All
Explicit	0.9231	0.9715	0.9057	0.9537
Implicit	0.9865	0.9964	0.9865	0.9934

Table 19 groups the N1, N2 and N3 windows for all five speech systems by utterance type and addressee. Requests and questions are split into device-directed and other requests; the latter are listed as requests to a human. Parentheses in the column headings identify the N categories represented in each group; a column need not cover a whole category. For example, a question to another human that mentions Aria belongs to N1, so the human-directed request column contains both N1 and N2. Statements span N1, N2 and N3, while self-talk is a subset of N3. Failures include both substantive intrusions and judge errors, as in the main silence-preservation score. For MiniCPM-o under explicit addressing, the device-request column contains 140 failures among 2,570 windows and the human-request column contains 76 among 2,134. Combining their counts gives $(140 + 76)/(2{,}570 + 2{,}134) \approx 4.59\%$, reported as 4.6% in the main text.

Table 19: Speech-system silence-preservation failure rates by utterance type and addressee. Lower is better; bold marks the lowest value per addressing condition and column.

		Carries a request…			
Model	Addressing	to a device (N2)	to a human (N1/N2)	statement (N1/N2/N3)	self-talk (N3)
MiniCPM-o 4.5	Explicit	**5.4%**	**3.6%**	**9.3%**	**5.3%**
	Implicit	**4.7%**	**2.5%**	**9.1%**	**4.4%**
Moshi	Explicit	73.5%	79.5%	70.0%	77.7%
	Implicit	76.0%	79.8%	71.0%	75.3%
FLM-Audio	Explicit	49.9%	56.8%	63.5%	59.5%
	Implicit	50.3%	56.4%	63.1%	58.9%
Voila	Explicit	11.6%	11.9%	16.0%	25.1%
	Implicit	11.3%	12.7%	16.2%	26.9%
Freeze-Omni	Explicit	100.0%	100.0%	100.0%	100.0%
	Implicit	100.0%	100.0%	100.0%	100.0%
Window count	Explicit	2,570	2,134	12,437	3,219
Window count	Implicit	2,570	2,134	12,437	3,219

A6 QUALITATIVE EXAMPLES

A6.1 ALLOWED ACKNOWLEDGEMENTS AND PENALISED SPEECH

Table 20 shows selected model outputs and the judge's reasons for allowing or penalising them in silence-requiring windows.

Table 20: Examples of exempted acknowledgements and penalised speech in silence-requiring windows.

Model	Label (duration)	What the assistant said	Judge's reason
MiniCPM-o	exempt (0.67 s)	"Understood."	Brief phatic acknowledgement adding no information and not ta…
	penalised (2.30 s)	"Yeah, I'm free for all three."	Adds new content, claiming availability, beyond phatic acknow…
Moshi	exempt (0.38 s)	"I'm sorry."	Brief phatic sympathetic reaction; adds no information and an…
	penalised (2.18 s)	"It's a valid point, but."	Evaluates and begins to counter a human's claim, starting an…
FLM-Audio	exempt (0.67 s)	"Good luck."	Brief phatic reaction; adds no information, answers nothing,…
	penalised (5.60 s)	"Siri is developed by Apple Inc. and is part o…"	Introduces unrelated factual information in multiple clauses,…
Voila	exempt (0.42 s)	"Thank you."	Brief phatic politeness; adds no information and doesn't take…
	penalised (0.45 s)	"The ants."	Content-bearing utterance introducing new information, not a…
Freeze-Omni	penalised (7.30 s)	"True. Nonverbal cues. Use nonverbing or makin…"	Prolonged multi-clause advice introducing unrelated new conte…

A7 STATISTICAL COMPARISONS

A7.1 CONFIDENCE INTERVALS AND PAIRED TESTS

To estimate uncertainty in silence preservation, we randomly select $2{,}000$ scenarios from each addressing condition's stored evaluation results, allowing a scenario to be selected more than once, and recalculate the differences between models. Each selected scenario contributes all its scored turns; the models are not run again. We repeat this calculation $10{,}000$ times and use the middle 95% of the recalculated differences as the confidence interval. These analyses compare models on the same evaluation items to determine which score differences have statistical support. For silence preservation, all model pairs differ significantly under both addressing conditions, supporting the ordering MiniCPM-o, Voila, FLM-Audio, Moshi, then Freeze-Omni from highest to lowest. The scenario-level bootstrap supports the same ordering: every 95% interval for a model-pair difference lies entirely above or below zero. For the closest pair under explicit addressing, MiniCPM-o and Voila, the interval places MiniCPM-o ahead by 8.25 to 9.58 percentage points. For each resampled set of scenarios, we subtract Voila's silence-preservation rate from MiniCPM-o's rate and multiply by 100. The 2.5th and 97.5th percentiles of these $10{,}000$ differences are 8.2497 and 9.5753 percentage points, rounded to the reported bounds.

On their shared response-present requests, Moshi answers $28/1{,}640$ implicit requests correctly and Freeze-Omni answers $2/1{,}640$ ($p_{\text{corr}} = 1.74 \times 10^{-5}$). For explicit requests the corresponding counts are $20/1{,}623$ and $5/1{,}623$ ($p_{\text{corr}} = 0.0815$), so only the implicit comparison is statistically significant; both models have low absolute accuracy. For conditional answer accuracy, MiniCPM-o exceeds all other models, and Moshi exceeds FLM-Audio under both addressing conditions. For

Moshi versus Freeze-Omni, corrected p values are 0.0815 under explicit addressing and 1.74×10^{-5} under implicit addressing. Only the latter is statistically significant, although both models answer fewer than 2% of their shared requests correctly. The remaining comparisons are not statistically significant, so the tests do not establish a complete ranking for this metric. In particular, Moshi's higher point estimate than Voila does not establish higher answer accuracy on their shared requests.

Table 21 reports exact two-sided McNemar tests; corrected $p < 0.05$ indicates a significant difference. Bonferroni correction covers 20 tests per addressing condition: ten model pairs for each of the two metrics. Silence comparisons use all common windows; answer-accuracy comparisons use only T requests where both models produced speech, so their rates can differ from the model-specific rates in Table 3. For the silence-preservation intervals, we sample $2{,}000$ scenarios with replacement per addressing condition, keeping each scenario's windows together, and repeat this $10{,}000$ times with seed 20260902. The interval bounds are the 2.5th and 97.5th percentiles of the resulting score differences.

In a row labelled "A vs B", r_1 is model A's score and r_2 is model B's score on the same compared items; higher is better for both metrics. "disc." counts items where one model passes and the other fails: one preserves silence and the other does not, or one answers correctly and the other incorrectly. "shared" counts T requests on which both models produced speech; it is the denominator for both answer-accuracy scores in that row. The corrected value p_{corr} is the raw McNemar p value multiplied by 20, capped at 1; values below 0.05 indicate a statistically significant difference, while larger values do not establish equal performance. For example, under explicit addressing FLM-Audio and Freeze-Omni both speak on $1{,}892$ T requests, answer 2 and 7 correctly, respectively, and differ in correctness on 9 requests. Their corrected $p = 1.000$ does not support an answer-accuracy difference.

Table 21: Pairwise comparisons with Bonferroni-corrected McNemar tests.

	Silence preservation (all silence windows)				Conditional answer accuracy (shared items)				
Pair	r_1	r_2	disc.	p_{corr}	r_1	r_2	shared	disc.	p_{corr}
Explicit									
FLM-Audio vs Freeze-Omni	0.3957	0.0002	8,053	$< 10^{-300}$	0.0011	0.0037	1,892	9	1.000
FLM-Audio vs MiniCPM-o 4.5	0.3957	0.9242	11,529	$< 10^{-300}$	0.0011	0.4706	1,817	853	6.7×10^{-256}
FLM-Audio vs Moshi	0.3957	0.2734	8,967	7.9×10^{-153}	0.0000	0.0117	1,545	18	1.5×10^{-4}
FLM-Audio vs Voila	0.3957	0.8351	10,756	$< 10^{-300}$	0.0016	0.0032	1,237	6	1.000
Freeze-Omni vs MiniCPM-o 4.5	0.0002	0.9242	18,812	$< 10^{-300}$	0.0037	0.4729	1,901	894	2.7×10^{-265}
Freeze-Omni vs Moshi	0.0002	0.2734	5,568	$< 10^{-300}$	0.0031	0.0123	1,623	25	0.082
Freeze-Omni vs Voila	0.0002	0.8351	16,997	$< 10^{-300}$	0.0031	0.0031	1,298	8	1.000
MiniCPM-o 4.5 vs Moshi	0.9242	0.2734	13,958	$< 10^{-300}$	0.4631	0.0122	1,559	711	3.9×10^{-203}
MiniCPM-o 4.5 vs Voila	0.9242	0.8351	4,139	1.3×10^{-179}	0.4683	0.0032	1,247	582	1.5×10^{-171}
Moshi vs Voila	0.2734	0.8351	13,017	$< 10^{-300}$	0.0085	0.0028	1,062	10	1.000
Implicit									
FLM-Audio vs Freeze-Omni	0.3991	0.0002	8,120	$< 10^{-300}$	0.0005	0.0016	1,903	4	1.000
FLM-Audio vs MiniCPM-o 4.5	0.3991	0.9289	11,485	$< 10^{-300}$	0.0006	0.4669	1,814	848	1.8×10^{-251}
FLM-Audio vs Moshi	0.3991	0.2676	8,988	1.7×10^{-176}	0.0006	0.0152	1,576	25	3.1×10^{-5}
FLM-Audio vs Voila	0.3991	0.8307	10,671	$< 10^{-300}$	0.0008	0.0063	1,276	9	0.781
Freeze-Omni vs MiniCPM-o 4.5	0.0002	0.9289	18,907	$< 10^{-300}$	0.0016	0.4662	1,892	879	9.9×10^{-264}
Freeze-Omni vs Moshi	0.0002	0.2676	5,448	$< 10^{-300}$	0.0012	0.0171	1,640	30	1.7×10^{-5}
Freeze-Omni vs Voila	0.0002	0.8307	16,909	$< 10^{-300}$	0.0007	0.0075	1,335	11	0.234
MiniCPM-o 4.5 vs Moshi	0.9289	0.2676	14,105	$< 10^{-300}$	0.4650	0.0173	1,557	713	1.5×10^{-195}
MiniCPM-o 4.5 vs Voila	0.9289	0.8307	4,144	6.1×10^{-219}	0.4722	0.0072	1,258	589	3.4×10^{-171}
Moshi vs Voila	0.2676	0.8307	13,091	$< 10^{-300}$	0.0189	0.0081	1,114	30	0.855