DUPLEXACT: BENCHMARKING NATURAL AND CONTROLLABLE INTERACTION BEHAVIORS IN REAL-TIME FULL-DUPLEX SPEECH SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-time full-duplex speech systems can listen and speak simultaneously, but natural interaction also requires them to adapt their behavior to the conversational context and users' spoken instructions. We introduce DuplexAct, a benchmark for evaluating how naturally and controllably these systems interact. DuplexAct covers six interaction families, each evaluated on two tracks: Capability and Command. The Capability track tests whether a system acts appropriately in the conversational context under a fixed policy for each family. The Command track evaluates whether a system follows users' spoken instructions and adapts its behavior when those instructions are updated. We build 512 base cases in Chinese and English from real customer-service conversation records and validate the resulting dataset through human review. Our real-time evaluation protocol accommodates different system interfaces, assesses interaction behavior from actual audio playout traces, and reports correctness and latency separately. Most evaluated systems exceed 90% success in initiating responses to completed requests in both languages under the Capability track. In contrast, success rates reach at most 32.1% for waiting through a mid-utterance pause before responding, and 30.8% for timely intervention after factual errors. These behaviors remain challenging in the Command track. These findings highlight context-appropriate behavior and reliable spoken-policy execution as key challenges for natural and controllable full-duplex interaction.

1 INTRODUCTION

Real-time full-duplex speech systems can listen and speak at the same time (OpenAI, 2026; 2025; Défossez et al., 2024; Wang et al., 2024). However, full-duplex interaction is not only a question of *when* to speak; it is also a question of *which action* fits the context. A pause may occur inside an unfinished utterance or after a completed request. Overlapping user speech may claim the floor or only provide a backchannel such as "mm-hmm." A spoken instruction such as "don't interrupt me until I finish" can further change what the system should do. Such turn-taking decisions depend on context (Sacks et al., 1974; Skantze, 2021). The same observable event can thus require different actions, depending on its interactional interpretation and the policy in force. We call a system *natural* if it selects actions that fit the conversational context, and *controllable* if those actions follow the policy that the user expresses through speech.

Existing benchmarks evaluate important components of full-duplex interaction, including turn-taking, pause handling, backchanneling, and interruption management (Arora et al., 2025; Lin et al., 2025; Jiang et al., 2026). Recent work also evaluates instruction following under system-level policies and under explicit, persona-implied, and conflicting prompt conditions (Tang et al., 2026; Mathur & Manocha, 2026). Two challenges remain less explored. First, the appropriate action depends jointly on the temporal configuration, its interactional interpretation, and the active policy, and the user may set or update that policy during the conversation. The same factual error, for example, calls for immediate correction under one policy and continued listening under another. Second, following a policy often requires a *sequence* of correct decisions. A system asked not to interrupt has to stay silent while the user holds the floor and then respond once the floor is released. Evaluating either decision in isolation can reward degenerate behavior: a system that withholds

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.

speech at the decision point can satisfy every hold requirement, while one that responds quickly may satisfy a response window despite having already interrupted the user.

We therefore formulate full-duplex evaluation as *policy-conditioned action selection*: the required action is determined jointly by the temporal configuration, its interactional interpretation, and the active policy. Following the $T \times I \times R$ ontology of Lu et al. (2026), we make the active policy π explicit and evaluate whether a system realizes the required action for each (T, I, π). This formulation leads to DuplexAct, a bilingual benchmark for natural and controllable full-duplex interaction. Its Capability track fixes π to a standing base policy and evaluates context-appropriate behavior. Its Command track lets the user set or update π through speech and evaluates whether the system changes its actions accordingly.

DuplexAct covers six interaction families that span three decisions: when to begin speaking (SILENCE, WAIT-COMPLETION, and IMMEDIATE), whether to intervene while the user still holds the floor (CORRECTION and AFFECTIVE-ADAPTATION), and whether to stop or continue when the user speaks over an ongoing response (BARGE-IN-YIELD). Each family scores the complete behavioral contract of its policy rather than an isolated action. Selected Command conditions pair the same event with opposing policies, which tests whether behavior changes because of the spoken instruction rather than the event alone. For correction and affective adaptation, the evaluation focuses on when the system intervenes or holds back; the quality of its response content requires separate evaluation.

DuplexAct contains 512 base cases in Chinese and English, grounded in real customer-service conversations.[1] Three annotators review every item, and blind validation on 120 sampled items shows that independent judgments match the reference actions in 97.8% of Capability and 99.4% of Command judgments. We evaluate systems through their native real-time interfaces and judge behavior from actual audio playout, because planned speech may be cut off before it is heard (Skantze & Hjalmarsson, 2010). Audio that is generated but never played is not a response, and a server-side stop is not a yield while queued speech remains audible. Latency is reported separately from behavioral correctness, and a full evaluation of one system takes 5,114 planned sessions.

We evaluate seven systems (Table 3). Five exceed 90% success in responding to completed requests in both languages, yet the best system reaches only 32.1% on the full wait–respond requirement. Controllability remains challenging: on Chinese AFFECTIVE-ADAPTATION, MiniCPM-o 4.5 scores 100.0% in Capability but only 10.0% under paired opposing policies in Command. Decision-level analysis further shows that timely responses can coexist with earlier floor violations, while low joint scores can arise from either failing to wait or failing to respond after release. These results show that reliable response initiation does not imply reliable policy-conditioned action selection. DuplexAct makes these failures measurable by jointly evaluating context-appropriate behavior, adherence to spoken policies, and the complete behavioral requirements of each interaction family.

2 RELATED WORK

2.1 CONVERSATION ANALYSIS AND INTERACTIONAL DECISIONS

Foundational studies describe turn-taking as a locally organized process supported by interactional signals (Sacks et al., 1974; Duncan, 1972). Turn transitions depend on acoustic, prosodic, and lexical-syntactic cues rather than silence alone (Gravano & Hirschberg, 2011), and their short timing requires listeners to prepare responses while anticipating turn completion (Levinson & Torreira, 2015). Beyond turn timing, dialogue contributions serve multiple communicative functions. ISO 24617-2 represents dialogue acts across multiple dimensions (Bunt et al., 2020), and DialogBank supports interoperable annotation of these functions across conversational corpora (Bunt et al., 2019). Together, these perspectives motivate separating the temporal configuration of an event from its interactional interpretation and the action expected of the system. Previous research (Lu et al., 2026) formalizes this distinction through a $T \times I \times R$ ontology for full-duplex interaction. Building on this framework, DuplexAct makes the active policy explicit and specifies policy-conditioned actions across six interaction families.

[1] https://anonymous.4open.science/r/DuplexAct-4DEA/

2.2 Full-Duplex Speech Systems and Behavioral Control

Full-duplex speech systems support concurrent listening and speaking. Moshi jointly models user and system speech streams without explicit turn segmentation (Défossez et al., 2024), while Freeze-Omni combines streaming speech processing with dialogue-state prediction for low-latency interaction (Wang et al., 2024). Mini-Omni2 introduces a command-based interruption mechanism (Xie & Wu, 2024). Other work on behavioral control explores voice and role conditioning in PersonaPlex (Roy et al., 2026) and instruction-conditioned conversational behavior in F-Actor (Züfle et al., 2026). Earlier work on incremental speech generation emphasizes monitoring what the system has actually said, since planned output may be interrupted before realization (Skantze & Hjalmarsson, 2010). This distinction motivates DuplexAct's use of actual audio playout for behavioral evaluation.

2.3 Benchmarks for Interaction Behavior and Instruction Control

Existing benchmarks establish substantial coverage of real-time interaction. Talking Turns studies turn-taking dynamics (Arora et al., 2025), while Full-Duplex-Bench and its v1.5 extension evaluate pause handling, backchanneling, turn-taking, and interruption management (Lin et al., 2025; 2026b). Full-Duplex-Bench-v2 extends evaluation to multi-turn conversations, covering correction, entity tracking, and safety alongside conversational fluency and instruction following (Lin et al., 2026a). FD-Bench evaluates interruption handling, latency, and robustness under challenging conversational conditions (Peng et al., 2025). TurnBench evaluates end-of-turn and interruption detection, including latency and false-positive behavior (Jiang et al., 2026).

Recent benchmarks also examine instruction-conditioned interaction. Instruct-FD pairs the same conversation with different system-level turn-management instructions and an uninstructed baseline, evaluating adherence in persistent sessions through time-aligned transcripts (Tang et al., 2026). DuplexSpeechBench-IFEval compares explicit, persona-implied, and conflicting prompt conditions to assess behavioral adherence (Mathur & Manocha, 2026).

DuplexAct differs from prior benchmarks in its evaluation target: it conditions the required action on the event's interactional interpretation and the active policy, and evaluates the complete set of decisions required by that policy rather than individual decisions in isolation. The active policy can be delivered and updated through user speech, including within-session updates in selected interaction families. This formulation supports complete behavioral contracts and opposing-policy conditions, including cases in which the same conversational event requires different actions as the spoken policy changes.

3 DuplexAct

DuplexAct evaluates whether real-time speech systems take appropriate observable actions under a given conversational context and interaction policy. As illustrated in Figure 1, the benchmark covers six interaction families in Chinese and English and evaluates each family on two tracks: Capability tests context-appropriate behavior under a fixed base policy, and Command tests adherence to spoken policies and their updates. Section 3.1 defines the interaction decisions and the two tracks, Section 3.2 describes data construction and validation, Section 3.3 presents the evaluation protocol, and Section 3.4 defines the metrics.

3.1 Task Formulation and Evaluation Tracks

DuplexAct evaluates action decisions in real-time speech interaction. We build on the $T \times I \times R$ ontology of Lu et al. (2026), which separates the temporal relation of an interaction (T), user intent (I), and the corresponding system response (R). We adapt this factorization in two ways. First, we use I more broadly for the *interactional interpretation* of a user-side event, since our tasks also involve events such as non-speech interference and an identifiable factual error. Second, we make the active interaction policy explicit. We therefore formulate each decision point as

$$(T_t, I_t, \pi_t) \longmapsto R_t, \tag{1}$$

where T_t describes the temporal configuration of user and system speech, I_t describes the interactional interpretation of the user-side event in its conversational context, π_t denotes the interaction

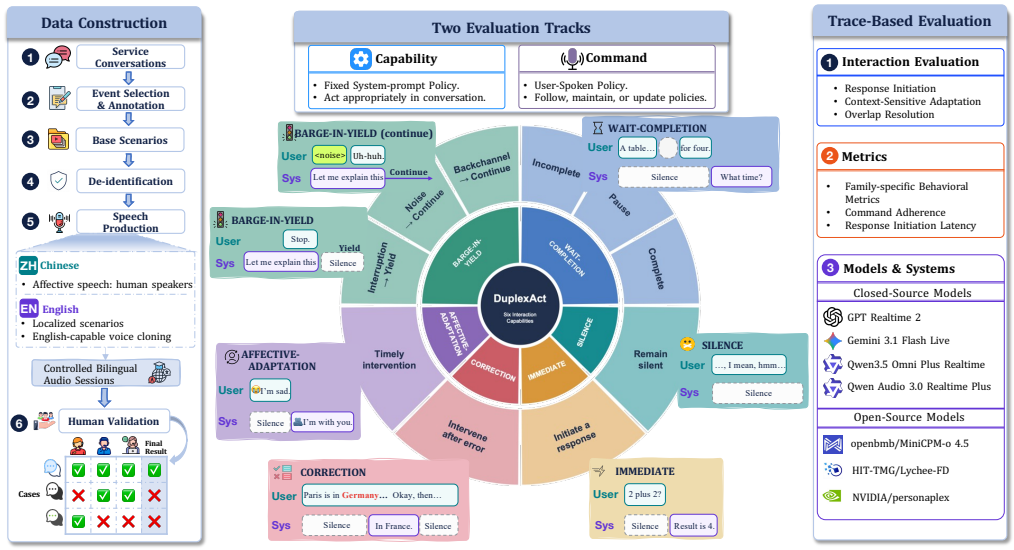

Figure 1: Overview of DuplexAct. Left: Construction and human validation of bilingual Chinese–English audio sessions. Center: Two evaluation tracks—Capability under fixed system-prompt policies and Command under user-spoken policies—covering six interaction families, illustrated with example exchanges. Right: Trace-based evaluation of interaction behavior, command adherence, and latency across full-duplex speech systems.

policy currently in force, and R_t is the required observable system action. The policy π_t comes from one of two sources: a general interaction guideline given in the system prompt, such as waiting in silence while the user pauses, or a policy the user states during the conversation. These two sources define the two evaluation tracks described below.

None of the three inputs fixes the action alone. A pause may occur within an unfinished utterance, where the system should wait, or after a completed request, where it should respond. Similarly, user speech overlapping a system response may be a genuine floor claim, a listener backchannel, or interference, requiring the system to yield or continue accordingly. The active policy π_t introduces a further distinction: the same interactional event can require different actions under different policies. For example, an identifiable factual error during ongoing user speech requires intervention under an immediate-correction policy, but continued listening under a defer-correction policy.

Five actions. R_t takes one of five values. HOLD means the system should not begin a substantive response yet. RESPOND means the user has released the floor and the system should begin one. INTERVENE means the user may still be speaking, but the context or an active instruction licenses the system to speak. YIELD means an ongoing system response should stop. CONTINUE means it should keep playing. HOLD and CONTINUE both preserve the current state, but they are different decisions: HOLD is a listener-side decision not to take the floor, and CONTINUE is a speaker-side decision not to give it up.

Six families. The six families group into three kinds of decision. SILENCE, WAIT-COMPLETION, and IMMEDIATE ask when the system should begin speaking. CORRECTION and AFFECTIVE-ADAPTATION ask whether it should speak while the user still holds the floor. BARGE-IN-YIELD asks whether it should stop or keep going when the user speaks over an ongoing response. Table 1 summarizes the temporal configuration, interactional interpretation, and required action for each family under the standing base policy. Spoken-policy conditions are described below.

Base policies. Each family specifies its base policy π_0 in a system prompt (Appendix B). SILENCE, WAIT-COMPLETION, IMMEDIATE, and CORRECTION use the same prompt in both tracks, so the spoken instruction is the only added policy input in Command. AFFECTIVE-ADAPTATION uses a separate, neutral Command prompt. Its Command track runs the same probe under two op-

Family	T: Temporal configuration	I: Interactional interpretation	R: Required action
SILENCE	User pauses while the system listens	User retains the floor	HOLD
WAIT-COMPLETION	Incomplete → pause → complete	User retains, then releases the floor	HOLD → HOLD → RESPOND
IMMEDIATE	User completes an answer-seeking request	User releases the floor	RESPOND
CORRECTION	A factual error is completed while the user keeps speaking	An identifiable factual error is present	INTERVENE
AFFECTIVE-ADAPTATION	An affective clause occurs while the user keeps speaking	Task-relevant affect is present	INTERVENE
BARGE-IN-YIELD	User-side audio overlaps ongoing system playout	Floor claim / backchannel / interference	YIELD / CONTINUE / CONTINUE

Table 1: The six interaction families in DuplexAct under the adapted $T \times I \times R$ formulation. T describes the temporal configuration of the interaction, I the interactionally relevant interpretation of the user-side event, and R the action required at the decision anchor under the base policy. A substantive response that begins before a RESPOND or INTERVENE anchor counts as a failure (Section 3.4).

posite spoken policies, intervene and hold, and the Capability prompt, which asks for intervention, would conflict with the hold policy. BARGE-IN-YIELD binds a prompt to each base case to supply the material the system reads aloud. Its prompt asks the system to read the material to the end while following the user's latest explicit instruction. A genuine interruption is such an instruction and calls for yielding, whereas a backchannel or noise leaves the reading task unchanged.

Capability and Command. In Capability, $\pi_t = \pi_0$ throughout the evaluation: no episode-specific spoken policy is introduced. The track therefore tests whether the system takes the required action from the conversational context under the standing family-specific policy. In Command, spoken instructions specify or update π_t. They may restate the base requirements, prescribe an opposing action, or release an earlier constraint; policy maintenance and updates are tested where applicable. Each Command base case uses five equivalent instruction variants to assess consistency across wordings. The tracks use different contracts and scoring units, so their score difference does not isolate the effect of spoken instructions. Each required action has a behavioral window anchored to semantic completion, an explicit release, or a task-specific anchor (Section 3.4).

3.2 DATA CONSTRUCTION AND HUMAN VALIDATION

DuplexAct is constructed from real customer-service conversations through six stages: source collection, interaction-event annotation, base-scenario construction, de-identification, speech production, and human validation. Annotators identify segments instantiating the six interaction families and retain the dialogue context needed to interpret each event. Each event is then converted into a self-contained base case with its user utterance, temporal structure, and decision anchors. Family-specific structure is added where needed: WAIT-COMPLETION marks incomplete, pause, and completion points; CORRECTION preserves a factual error and its identifiable completion point during continued user speech; and BARGE-IN-YIELD constructs interruption, backchannel, and sustained-noise overlap conditions with frozen earliest-disambiguation times. Unlike the other families, BARGE-IN-YIELD uses manually reviewed, LLM-assisted reading materials based on public Wikipedia information, and its sustained noise comes from ESC-50 (Piczak, 2015) (Appendix C.8).

All scenarios are de-identified before speech production. Personal and business-sensitive information is removed or replaced while preserving speaker roles, numerical and logical relations, target errors, and interaction structure. The Chinese standard and overlap subsets are rendered with voice-cloning TTS using seven anonymized voices, while the Chinese AFFECTIVE-ADAPTATION subset is recorded by six native speakers covering anger, anxiety, sadness, and happiness. The English subset is produced by localizing the de-identified scenarios and synthesizing them with an English-capable voice-cloning system. English AFFECTIVE-ADAPTATION items are synthesized with CosyVoice 3.5 (Du et al., 2024; 2025; Lyu et al., 2025) under explicit emotion control, and their emotion is checked during human validation.

Human validation. We validate the benchmark at two levels. First, three annotators independently review every final item. They check the assigned interaction label under the applicable policy context and the quality of the realized speech, including pronunciation, prosody, pause placement, fluency, and, for affective items, the intended emotion. We retain an item only when all three annotators accept both the label and the recording. Second, to test whether the retained labels can be reproduced without seeing them, we sample five items from each family–track–language combination (120 items). Three annotators choose the required action at each decision point from the preceding audio, system prompt, and applicable spoken instructions, without access to the reference action or the subsequent audio. Their choices match the reference actions in 97.8% of Capability and 99.4% of Command judgments, with UNCERTAIN counted as a mismatch. The reference actions are therefore consistent with independent human judgment (Appendix D).

The final benchmark contains 512 base cases, with 256 per language. Capability and Command instances are derived from these shared base cases according to their family-specific contracts, producing multiple sessions and scoring units from the same underlying scenario. Appendix C.2 reports these counts for each family and track.

3.3 EVALUATION PROTOCOL

We evaluate each system through its native real-time interface, streaming input audio in real time and keeping the microphone stream open during silence. Adapters deliver the policy and dialogue history but never use expected actions to decide when the system speaks or stops. All judgments use the actual playout of system audio: generated but unplayed audio is not a response, and a server-side stop is not a yield while queued audio remains audible. We evaluate all planned task conditions, retrying technically failed executions until a technically valid run is obtained. For BARGE-IN-YIELD, technically valid sessions may still contain unscorable events; event coverage is reported separately in Appendix I. Interface settings and playout reconstruction are described in Appendix A.

3.4 METRICS

Scoring units. Scores are computed separately by family, track, and language. A scoring unit is a trial or base case in Capability, a base–variant pair in Command (a policy pair for AFFECTIVE-ADAPTATION), and a logical event in BARGE-IN-YIELD Command (Appendix C.2). Each score averages over the set S of scoring units.

Decision predicates. Each session has its own playout timeline (Section 3.3). Let O be the onset times of substantive system speech in a session, and let $B(W)$ be the number of backchannels in a window W, identified by a fixed language-specific classifier (Appendix E.1). We define

$$\mathrm{N}(W) = \mathbf{1}[O \cap W = \emptyset], \quad \mathrm{H}(W) = \mathrm{N}(W)\,\mathbf{1}[B(W) \le 1], \quad \mathrm{R}(W) = \mathbf{1}[O \cap W \neq \emptyset]. \tag{2}$$

R checks both RESPOND and INTERVENE. Each R window follows an N or H check on the preceding interval, so an earlier response cannot satisfy a later requirement. For BARGE-IN-YIELD, the *carrier* is the system response that is audible when a stimulus is injected, and the earliest disambiguation time (EDT) is the first point at which the stimulus function is clear. Y passes if the carrier stops within 3 s of the EDT and stays stopped through a 1.5 s guard; C passes if the same carrier keeps playing through the stimulus (Appendix E.3).

Primary scores. The fixed windows operationalize the behavioral predicates consistently across heterogeneous systems; they are evaluation thresholds rather than claims of universal conversational timing boundaries. Table 2 lists the required decisions for each family and track. For a unit u with required binary decisions $x_{u1}, \ldots, x_{um}$, the primary score is the joint success rate

$$M = \frac{1}{|S|} \sum_{u \in S} \prod_{k=1}^{m} x_{uk}, \tag{3}$$

so a unit passes only if every required decision passes. Error decompositions serve as diagnostics (Appendix F). BARGE-IN-YIELD Capability instead weights the three stimuli within each base case b:

$$M_{\mathrm{BAR,Cap}} = \frac{1}{|S|} \sum_{b \in S} \big(0.5\,\mathrm{Y}_b^{\mathrm{interrupt}} + 0.25\,\mathrm{C}_b^{\mathrm{bc}} + 0.25\,\mathrm{C}_b^{\mathrm{noise}}\big), \tag{4}$$

Family	Capability	Command
SILENCE	$\mathrm{H}(W)$	$\mathrm{H}^{\mathrm{M0}}(W) \cdot \mathrm{H}^{\mathrm{R0}}([t_0, r)) \cdot \mathrm{R}^{\mathrm{R0}}([r, r{+}6])$
WAIT	$\mathrm{H}^{\mathrm{P1}}(W) \cdot \mathrm{H}^{\mathrm{PA}}(W) \cdot \mathrm{N}^{\mathrm{P2}}([t_0, c)) \cdot \mathrm{R}^{\mathrm{P2}}([c, c{+}6])$	same as SILENCE
IMMEDIATE	$\mathrm{H}([t_0, c)) \cdot \mathrm{R}([c, c{+}6])$	same as Capability
CORRECTION	$\mathrm{N}([t_0, e)) \cdot \mathrm{R}([e, e{+}3])$	$\mathrm{N}^{\mathrm{A0}}([t_0, e)) \cdot \mathrm{R}^{\mathrm{A0}}([e, e{+}3]) \cdot \mathrm{N}^{\mathrm{C0}}([t_0, f]) \cdot \mathrm{N}^{\mathrm{C1}}([t_0, r)) \cdot \mathrm{R}^{\mathrm{C1}}([r, r{+}6])$
AFFECTIVE	$\mathrm{N}([t_0, a)) \cdot \mathrm{R}([a, z{+}3])$	$\mathrm{N}^{\mathrm{int}}([t_0, a)) \cdot \mathrm{R}^{\mathrm{int}}([a, z{+}3]) \cdot \mathrm{H}^{\mathrm{hold}}([t_0, f)) \cdot \mathrm{R}^{\mathrm{hold}}([f, f{+}6])$
BARGE	Eq. (4)	event average of Y or C

Table 2: Required decisions for each family and track. Superscripts name the node or session (Appendix C.1); int and hold are the two sessions of an AFFECTIVE-ADAPTATION policy pair. t_0: start of node input; W: node input plus 6 s; c: semantic completion; r: end of the release utterance; e: completion of the factual error; $[a, z]$: target affective segment; f: end of the full probe.

which gives YIELD and CONTINUE equal weight. Command averages $q_e \in \{0, 1\}$ over logical events, where q_e indicates whether event e meets the YIELD or CONTINUE contract of the policy active at e.

Latency and reporting. Response-initiation latency for IMMEDIATE and WAIT-COMPLETION runs from semantic completion (c), or from the end of the release (r) in WAIT-COMPLETION Command, to the first qualifying onset. It is reported separately, over units whose full contract passes (Appendix H). All results are reported by language, family, and track, with no overall score.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluated systems. We evaluate seven real-time speech systems: GPT Realtime 2 (OpenAI, 2025), Gemini 3.1 Flash Live (Google, 2026), Qwen3.5 Omni Plus Realtime (Qwen, 2026a), Qwen Audio 3.0 Realtime Plus (Qwen, 2026b), MiniCPM-o 4.5 (Cui et al., 2026), PersonaPlex (Roy et al., 2026), and Lychee-FD (Liu et al., 2026). Each system is evaluated in Chinese and English across all six interaction families on both the Capability and Command tracks, using the protocol in Section 3.3. Model versions, publicly disclosed parameter counts, and inference configurations are documented in Appendix A.

4.2 MAIN RESULTS

Reliable responses coexist with weak control during the user's turn. Table 3 shows that five of the seven systems exceed 90% on IMMEDIATE Capability in both languages. However, success at responding to completed requests does not extend to the full wait–respond requirement. GPT Realtime 2, for example, achieves 98.2% on SILENCE and 100.0% on IMMEDIATE in Chinese, yet only 14.3% on WAIT. Across all systems and both languages, WAIT reaches at most 32.1%. Timely intervention after a factual error is also limited: the best CORRECTION Capability scores are 10.3% in Chinese and 30.8% in English. These results separate reliable response initiation from the broader demands of preserving and taking the floor while a user turn unfolds.

The same policy structure produces sharply different adherence. SILENCE and WAIT Command share the same three-decision conjunction: hold under the standing instruction, preserve the user's floor before release, and respond after release. Nevertheless, GPT Realtime 2, Gemini 3.1 Flash Live, and Qwen3.5 Omni Plus Realtime each score substantially higher on SILENCE, with gaps of 52.8–82.4 percentage points across the two languages. For GPT Realtime 2 in Chinese, adherence is 88.8% on SILENCE but 6.4% on WAIT. The contrast shows that the difficulty of executing a spoken policy depends strongly on the interaction family, even when the number of required decisions and the conjunction rule are unchanged.

Model	SILENCE		WAIT		IMMEDIATE		CORRECTION		AFFECTIVE		BARGE	
	Cap.	Cmd.	Cap.	Cmd.	Cap.	Cmd.	Cap.	Cmd.	Cap.	Cmd.	Cap.	Cmd.
Chinese												
GPT Realtime 2	**98.2**	**88.8**	14.3	6.4	**100.0**	**100.0**	0.0	0.0	16.7	0.0	78.3	52.2
Gemini 3.1 Flash Live	91.2	78.9	**21.4**	5.0	97.6	98.3	0.0	0.0	83.3	**18.9**	56.7	48.7
Qwen3.5 Omni Plus Realtime	75.4	86.0	3.6	7.9	97.6	98.8	0.0	0.0	5.6	6.7	60.8	38.2
Qwen Audio 3.0 Realtime Plus	89.5	16.5	17.9	5.7	96.4	93.3	**10.3**	**4.1**	33.3	11.1	**88.3**	**64.8**
MiniCPM-o 4.5	70.2	34.0	10.7	**22.9**	94.0	92.1	7.7	**4.1**	**100.0**	10.0	42.5	53.2
Lychee-FD	1.8	19.3	0.0	7.1	94.0	52.1	**10.3**	0.0	27.8	5.6	50.0	40.0
PersonaPlex	22.8	0.0	7.1	0.0	73.8	11.2	2.6	0.0	16.7	0.0	40.0	30.0
English												
GPT Realtime 2	91.2	57.2	10.7	0.7	94.0	91.4	0.0	0.0	0.0	0.0	**73.3**	47.3
Gemini 3.1 Flash Live	84.2	60.7	14.3	7.9	**98.8**	83.6	0.0	0.0	11.1	0.0	60.0	49.4
Qwen3.5 Omni Plus Realtime	84.2	**80.0**	7.1	5.7	**98.8**	**93.1**	0.0	0.0	0.0	0.0	68.3	46.7
Qwen Audio 3.0 Realtime Plus	68.4	17.5	**32.1**	8.6	95.2	45.7	0.0	0.5	22.2	**2.2**	45.8	**59.9**
MiniCPM-o 4.5	**96.5**	25.3	14.3	**24.3**	91.7	67.1	10.3	**1.5**	**27.8**	**2.2**	40.0	46.7
Lychee-FD	22.8	9.5	0.0	1.4	40.5	41.9	0.0	0.0	11.1	1.1	43.3	35.0
PersonaPlex	5.3	0.0	21.4	0.0	58.3	22.1	**30.8**	0.0	11.1	0.0	46.7	35.0

Table 3: Primary behavioral scores (%) on DuplexAct. Cap. and Cmd. denote Capability and Command. Higher is better; bold marks the best score in each column within each language, including ties.

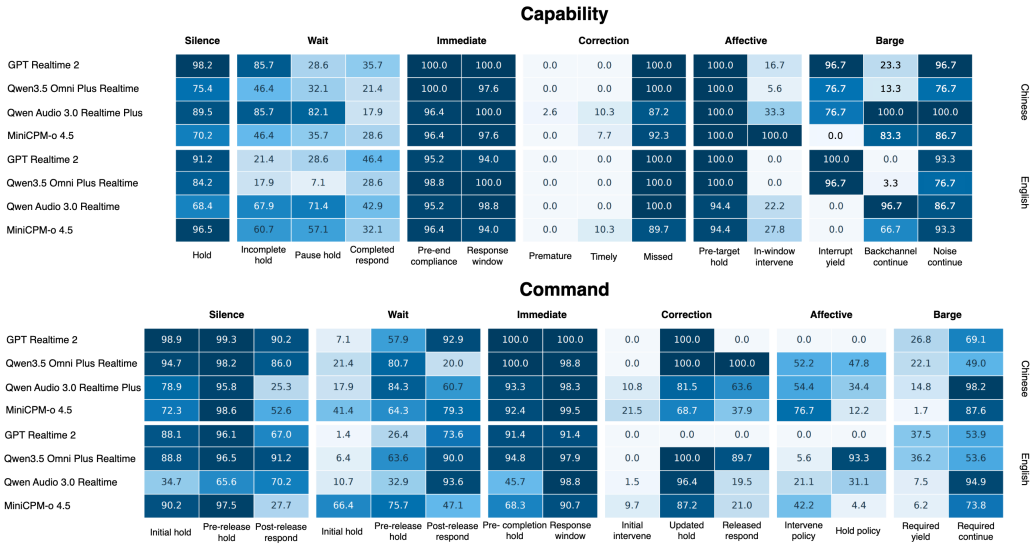

Figure 2: Decision-level behavioral profiles across six interaction families. Capability (top) and Command (bottom) results are shown for four systems in Chinese and English. Cells report the percentage satisfying each labeled requirement, except for CORRECTION Capability, which reports the mutually exclusive frequencies of premature, timely, and missed interventions.

A perfect intervention score does not imply policy control. Chinese AFFECTIVE reveals a reversal between intervention success and command adherence. MiniCPM-o 4.5 leads Capability at 100.0%, ahead of Gemini 3.1 Flash Live at 83.3%; in Command, Gemini leads at 18.9%, while MiniCPM-o scores 10.0%. Command pairs the same affective probe with opposing spoken instructions in separate sessions, requiring both intervention and restraint as instructed. Across all seven systems, the best paired adherence is only 18.9% in Chinese and 2.2% in English. On this task, evaluating intervention alone identifies a different leader from evaluating whether spoken policy governs that intervention. The paired result exposes a control requirement that a high intervention score leaves untested: selecting opposite actions for the same affective probe.

4.3 DECISION-LEVEL ANALYSIS

Figure 2 decomposes task-level scores into individual behavioral requirements for four systems on both tracks. We use this breakdown to locate policy violations and identify which decisions limit full-policy adherence.

Timely responses can conceal earlier violations. On English IMMEDIATE, Qwen Audio 3.0 Realtime Plus achieves 98.8% response-window success on both tracks, whereas pre-completion compliance is 95.2% on Capability and 45.7% on Command. In Command, 223 of 420 scoring units (53.1%) satisfy the response-window check while violating the preceding hold requirement. Thus, more than half of the units contain a timely response but remain behaviorally incorrect. The joint metric captures these violations, which a check confined to the response window would miss.

Low adherence has different behavioral bottlenecks. The Command breakdown localizes failures on either side of the release boundary. For GPT Realtime 2 on Chinese WAIT, initial-hold compliance is only 7.1%, whereas post-release response compliance reaches 92.9%. MiniCPM-o 4.5 exhibits the reverse pattern on English SILENCE: initial and pre-release hold compliance reach 90.2% and 97.5%, respectively, but post-release response compliance is only 27.7%. These settings therefore present different improvement targets: preserving the user's floor under an active waiting instruction in the former, and responding when that instruction is released in the latter.

The limiting decision determines where improvement is needed. The joint scores are already close to the ceilings imposed by these weak components. With GPT Realtime 2's initial-hold outcomes unchanged, perfect compliance at all remaining Chinese WAIT Command decisions would raise adherence only from 6.4% to 7.1%. With MiniCPM-o 4.5's post-release response outcomes unchanged, perfect hold compliance on English SILENCE Command would raise adherence only from 25.3% to 27.7%. Substantial gains therefore require improving initial holding in the first setting and responding after release in the second, while preserving compliance at the complementary decisions.

5 LIMITATIONS

DuplexAct assesses action and timing compliance, without evaluating correction accuracy or affective-response quality. Evaluation relies on controlled, largely synthesized speech and mostly independent sessions, with Barge-In-Yield testing policy updates within persistent sessions. Further work should examine longer, spontaneous conversations with more diverse speakers and acoustic conditions, while extending the benchmark to evolving interaction needs and model paradigms.

6 CONCLUSION AND FUTURE WORK

We introduced DuplexAct, a bilingual benchmark that formulates full-duplex evaluation as policy-conditioned action selection. Its six interaction families, evaluated on Capability and Command tracks, assess context-appropriate behavior and spoken-policy adherence through judgments grounded in actual audio playout. Experiments across seven systems reveal a gap between producing an expected response and executing the policy that governs it: an interaction can contain a timely response despite an earlier floor violation, and successful intervention can coexist with poor adherence to an opposing instruction. DuplexAct makes this gap observable through joint behavioral requirements and matched opposing-policy evaluations. Its decision-level diagnostics further locate the unmet requirements, distinguishing failures to preserve the user's floor from failures to respond after release. The benchmark thus assesses complete policy execution while identifying specific behaviors that require improvement.

Future work will extend persistent-session evaluation to more interaction tasks, examining policy maintenance, override, and release over longer conversational histories. We also aim to evaluate interaction behavior alongside response content, including whether an intervention corrects the factual error or appropriately addresses the user's affective state. These extensions would examine how reliable policy execution translates into effective assistance throughout an ongoing conversation.

AI USE STATEMENT

Generative AI tools were used during dataset construction to assist with rewriting de-identified customer-service records into benchmark base cases, localizing and translating Chinese scenarios into English, and generating the BARGE-IN-YIELD reading materials based on public Wikipedia information. All AI-assisted dataset content was manually reviewed and corrected before inclusion in the benchmark.

LLM assistants were also used to assist with the implementation and editing of evaluation and data-processing code. All AI-assisted code was reviewed and tested by the authors. The task formulation, benchmark design, metrics, experiments, and interpretation of results were developed by the authors rather than generated by AI tools.

Additionally, LLM assistants were used to draft some English text from author-written Chinese notes, improve grammar, wording, and clarity, and assist with literature search and reference checking. The authors manually revised all AI-assisted text and verified cited works against the original sources. We reviewed all AI-assisted work and take responsibility for the final content of this paper.

ETHICS STATEMENT

Data source and privacy. DuplexAct is derived from real customer-service conversation records obtained with authorization for research use. All scenarios were de-identified before speech production: personal and business-sensitive information was removed or replaced while preserving the interaction structure needed for evaluation. The benchmark contains no audio from the original customer-service conversations. Voice-cloning reference recordings were used with authorization and do not come from the original customers or agents.

Human participants. Six native speakers recorded the Chinese AFFECTIVE-ADAPTATION subset from de-identified scripts, and annotators reviewed and validated the benchmark items. Participants gave informed consent for their participation and for the use and public release of the resulting benchmark recordings. Participants were compensated at or above the applicable local minimum wage. No personally identifying information about the speakers is included in the benchmark.

Release and scope. We plan to release the benchmark scenarios, generated and recorded benchmark audio, annotations and metadata, evaluation code, and associated traces needed to reproduce the evaluation. DuplexAct covers Chinese and English customer-service scenarios with a limited set of voices, so its results may not generalize to other languages, domains, dialects, accents, or speaking styles.

Potential misuse and conflicts of interest. DuplexAct is intended to evaluate whether real-time voice assistants respect conversational floor control and spoken interaction policies. More natural voice interaction technology may also be misused, for example in deceptive automated calling; DuplexAct itself introduces no new speech-generation or voice-cloning method. The authors have no affiliation with the developers of the evaluated systems.

REFERENCES

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. Talking turns: Benchmarking audio foundation models on turn-taking dynamics, 2025. URL https://arxiv.org/abs/2503.01174.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. The dialogbank: dialogues with interoperable annotations: H. bunt et al. *Language Resources and Evaluation*, 53 (2):213–249, 2019.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 549–558, Marseille, France,

May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.69/.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

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, 2024. URL https://arxiv.org/abs/2410.00037.

Google. Gemini Live: Ask ai a question in any mode you choose, 2026. URL https://gemini.google/overview/gemini-live/. Accessed: 2026-09-13.

Agustín Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.

Freeman Jiang, Ramon Sanabria, Soham Deshmukh, Bandhav Veluri, Simon Michael Vuch Williams, Elliott K. Suen, Garreth Lee, Kevin Yoonho Choi, Takuya Umeki, Riku Kubo, Sathvik Udupa, Chien yu Huang, Shih-Yun Shan Kuan, Zhuoyan Tao, Satyapriya Krishna, Sefik Emre Eskimez, Yu Tsao, Hung yi Lee, and Shinji Watanabe. Turnbench: A multi-domain benchmark for turn-taking dynamics in spoken dialogue, 2026. URL https://arxiv.org/abs/2608.25218.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.

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, 2025. URL https://arxiv.org/abs/2503.04721.

Guan-Ting Lin, Shih-Yun Shan Kuan, Jiatong Shi, Kai-Wei Chang, Siddhant Arora, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 27–36, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-391-3. doi: 10.18653/v1/2026.acl-short.4. URL https://aclanthology.org/2026.acl-short.4/.

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, 2026b. URL https://arxiv.org/abs/2507.23159.

Zhenyu Liu, Xuanyu Zhang, Yunxin Li, Qixun Teng, Shenyuan Jiang, Haolan Chen, Mingjun Zhao, Fanbo Meng, Yu Xu, Yancheng He, et al. Hierarchical acoustic-semantic modeling: Modality separation and semantic coherence for full-duplex slms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9264–9280, 2026.

Jingyu Lu, Yuhan Wang, Jianming Luo, Yifu Chen, Tianle Liang, Shengpeng Ji, Ziyue Jiang, Xiaoda Yang, Yu Zhang, Xize Cheng, et al. A survey of full-duplex spoken dialogue systems: Architectural hierarchy, interaction ontology, and decision state machine. *arXiv preprint arXiv:2606.19453*, 2026.

Xiang Lyu, Yuxuan Wang, Tianyu Zhao, Hao Wang, Huadai Liu, and Zhihao Du. Build llm-based zero-shot streaming tts system with cosyvoice. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2025.

Puneet Mathur and Dinesh Manocha. Duplexspeechbench-ifeval: Evaluating implicit instruction following in full-duplex voice agents. *arXiv preprint arXiv:2609.03423*, 2026.

OpenAI. Introducing gpt-realtime and Realtime API updates for production voice agents, August 2025. URL https://openai.com/index/introducing-gpt-realtime/. Accessed: 2026-09-13.

OpenAI. Build more natural voice experiences with GPT-Live-1 in the API, September 2026. URL https://openai.com/index/introducing-gpt-live-1-in-the-api/. Accessed: 2026-09-13.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Chng Eng Siong. Fd-bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems. *ArXiv*, abs/2507.19040, 2025. URL https://api.semanticscholar.org/CorpusID:280046112.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Qwen. qwen3.5-omni-plus-realtime Model Info. Alibaba Cloud Model Studio Documentation, 2026a. URL https://help.aliyun.com/en/model-studio/qwen3-5-omni-plus-realtime.

Qwen. qwen-audio-3-0-realtime-plus. Alibaba Cloud Model Studio Documentation, 2026b. URL https://help.aliyun.com/en/model-studio/qwen-audio-3-0-realtime-plus.

Rajarshi Roy, Jonathan Raiman, Sang-Gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. Personaplex: Voice and role control for full duplex conversational speech models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16137–16141, 2026. doi: 10.1109/ICASSP55912.2026.11463413.

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.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech Language*, 67:101178, 2021. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2020.101178. URL https://www.sciencedirect.com/science/article/pii/S088523082030111X.

Gabriel Skantze and Anna Hjalmarsson. Towards incremental speech generation in dialogue systems. In Yasuhiro Katagiri and Mikio Nakano (eds.), *Proceedings of the SIGDIAL 2010 Conference*, pp. 1–8, Tokyo, Japan, September 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-4301/.

Yuzhi Tang, Wentao Ma, Xiling Zhao, Ahmad Salimi, Sepehr Harfi Moridani, Dongming Shen, Jixuan Wang, Abdulrahman Abdulrazzag, Murdock Aubry, Yu-Hua Chen, Daniel Lee, Jaewon Lee, Jonah Mackey, Silin Meng, Nicholas Stranges, Chenxu Xiong, Hao Yu, Yi Zhu, Mu Li, and Alex Smola. Instruct-fd: Can your full-duplex speech system follow turn-taking instructions?, 2026. URL https://arxiv.org/abs/2607.20460.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm, 2024. URL https://arxiv.org/abs/2411.00774.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities, 2024. URL https://arxiv.org/abs/2410.11190.

Maike Züfle, Ondrej Klejch, Nicholas Sanders, Jan Niehues, Alexandra Birch, and Tsz Kin Lam. F-actor: Controllable conversational behaviour in full-duplex models, 2026. URL https://arxiv.org/abs/2601.11329.

A Evaluated Models and Configurations

This appendix documents the seven evaluated systems, their official model and interface sources, and the settings recorded by the evaluation clients. We distinguish explicitly requested settings from values returned by a service. An omitted option is reported as unspecified rather than assigned a numerical default. Official documentation was checked on 22 September 2026; model identifiers below retain the versions used in the evaluation.

A.1 Model Versions and Official Sources

Table 4 lists the model identifiers and public sources. The four hosted systems expose streaming APIs; the other three systems use deployments of released models. Official API references identify the public interface specifications. Deployment-specific connection addresses are omitted.

System	Evaluated identifier / official checkpoint	Official interface or code
GPT Realtime 2	`gpt-realtime-2`	OpenAI Realtime API
Gemini 3.1 Flash Live	`gemini-3.1-flash-live-preview`	Google Gemini Live API
Qwen3.5 Omni Plus Realtime	`qwen3.5-omni-plus-realtime`	Alibaba Cloud Model Studio
Qwen Audio 3.0 Realtime Plus	`qwen-audio-3.0-realtime-plus`	Alibaba Cloud Model Studio
MiniCPM-o 4.5	`openbmb/MiniCPM-o-4_5`	OpenBMB deployment code
Lychee-FD	`HIT-TMG/Lychee-FD`	HITsz-TMG deployment code
PersonaPlex	`nvidia/personaplex-7b-v1`	NVIDIA PersonaPlex server

Table 4: Model versions and official sources. Checkpoint repository names identify public releases; exact deployed checkpoint revisions still require the corresponding server-side deployment records.

Hosted API sources. GPT Realtime 2 is documented by OpenAI, with WebSocket sessions on the `/v1/realtime` resource and model selection through the `model` parameter.[2] Gemini uses the Live API and the `google-genai` SDK's `client.aio.live.connect` interface.[3] Both Qwen systems use Model Studio's realtime WebSocket protocol, whose public resource is `/api-ws/v1/realtime`.[4] The hosted model names are API identifiers, not immutable weight hashes. We do not substitute newer model versions appearing in current documentation for the evaluated identifiers.

Released model sources and parameter counts. MiniCPM-o 4.5 is released by OpenBMB with a reported total of 9B parameters.[5] The Lychee-FD checkpoint repository reports approximately 13B parameters and distributes the speech waveform decoder separately through the Token2Wav component of Step-Audio-2-mini; the checkpoint count should therefore not be interpreted as a verified total for the complete deployed pipeline.[6] NVIDIA describes PersonaPlex as a 7B model.[7] Parameter counts for the four hosted models are not provided in the cited model documentation.

[2] OpenAI model documentation; official WebSocket guide.

[3] Gemini 3.1 Flash Live model documentation; official Live API reference.

[4] Qwen3.5 Omni Plus Realtime model documentation; Qwen Audio 3.0 Realtime Plus model documentation; official Qwen-Audio realtime guide.

[5] Official MiniCPM-o 4.5 checkpoint; official deployment repository.

[6] Official Lychee-FD checkpoint; official deployment repository.

[7] Official PersonaPlex model card; official deployment repository.

A.2 Audio Interfaces and Session Settings

Table 5 summarizes the recorded streaming profiles. Sample rates describe the client–service audio interface. Chunk durations describe input transmission units and should not be interpreted as measured response latency. All listed audio streams are monaural.

System	Input	Output	Chunk	Voice setting
GPT Realtime 2	PCM16, 24 kHz	PCM16, 24 kHz	20 ms	`marin` (returned)
Gemini 3.1 Flash Live	PCM16, 16 kHz	PCM16, 24 kHz	40 / 200 ms	Unspecified
Qwen3.5 Omni Plus Realtime	PCM16, 16 kHz	PCM16, 24 kHz	100 ms	`Tina`
Qwen Audio 3.0 Realtime Plus	PCM16, 16 kHz	PCM16, 24 kHz	100 ms	`longanqian`
MiniCPM-o 4.5	Float32, 16 kHz	Float32, 24 kHz	1,000 ms	Unspecified
Lychee-FD	PCM16 WAV, 16 kHz	PCM16, 24 kHz	400 ms	`default_female`
PersonaPlex	Opus, 24 kHz source	Opus; 24 kHz capture	80 ms	`NATF0.pt`

Table 5: Recorded audio configurations. Gemini has more than one acquisition profile, detailed below. PersonaPlex values describe the September greeting-drain adapter. "Unspecified" means that the inspected client configuration does not set the value; it does not establish a particular service default.

GPT Realtime 2. The client requests `output_modalities=[audio]` and `max_output_tokens=2048`. The inspected `session.updated` response reports `server_vad`, a detection threshold of 0.5, `prefix_padding_ms=300`, and `silence_duration_ms=500`, with `create_response=true` and `interrupt_response=true`. These are service-returned settings rather than explicit VAD overrides in the client. The returned audio configuration uses `marin` at speed 1.0 and 24 kHz PCM in both directions. Dialogue history is supplied through native conversation items. The inspected session request does not explicitly set temperature, a random seed, or reasoning effort.

Gemini 3.1 Flash Live. The active streaming adapter selects SDK API version `v1alpha`, requests audio responses and output transcription, and enables automatic activity detection. Both start-of-speech and end-of-speech sensitivity are set to `LOW`, with `prefix_padding_ms=20`. History uses native `user`/`model` content messages. The client does not specify a voice, temperature, or thinking budget. The active adapter transmits 40 ms input chunks and uses a 20 ms playout tick. Earlier retained IMMEDIATE acquisition records use 200 ms input chunks, so a single chunk duration does not describe all inspected records. Some earlier acquisition scripts request `proactive_audio=true`; this request is not evidence that the feature was active. Google's model-specific documentation states that proactive audio and affective dialogue are unsupported for Gemini 3.1 Flash Live. The configuration records therefore distinguish the requested flag from verified model support.

Qwen3.5 Omni Plus Realtime. The client selects `Tina`, requests text and audio output, and sends 100 ms chunks of 16 kHz PCM16. Output audio is 24 kHz PCM16. The explicit turn-detection configuration is `server_vad` with threshold 0.2, `prefix_padding_ms=300`, and `silence_duration_ms=1000`. The inspected service acknowledgement confirms these settings and reports automatic response creation and interruption enabled. Input transcription is requested with `qwen3-asr-flash-realtime`. The requested identifier is the unsuffixed realtime alias; a dated checkpoint revision is not inferred from the current documentation's alias mapping.

Qwen Audio 3.0 Realtime Plus. The client selects `longanqian` and uses the same 16 kHz input, 24 kHz output, and 100 ms input-chunk format as the Qwen3.5 adapter. Its explicitly requested turn detector is `smart_turn`. This differs from the acoustic `server_vad` configuration above. The inspected acknowledgement returns `smart_turn`; additional threshold and silence-duration fields

in a returned session object are not reported as client-tuned acoustic VAD parameters. Text and audio output and `qwen3-asr-flash-realtime` input transcription are requested. Neither inspected Qwen client explicitly overrides temperature or a random seed.

MiniCPM-o 4.5. The deployed duplex service receives little-endian Float32 audio at 16 kHz and emits Float32 audio at 24 kHz. Session initialization sets `chunk_ms=1000`, `force_listen_count=3`, and `length_penalty=1.05`. The benchmark sends the policy and serialized dialogue history in the session's `system_prompt`, because this deployment interface does not expose a separate native duplex-history field. The regular session initialization omits the experimental `stream` flag. No voice reference or sampling seed is explicitly selected by the inspected client. The WebSocket session envelope is an interface of the evaluated deployment; the cited OpenBMB repository provides the public model and deployment implementation.

Lychee-FD. The evaluated streaming adapter sends independently packaged mono PCM16 WAV chunks at 16 kHz, each covering 400 ms, and receives 24 kHz PCM16 audio. The start message specifies `prompt_voice=default_female`, `start_speak_factor=2.5`, `infer_window_ms=400`, and `streaming=true`. Policy and history are serialized into the `system_prompt`. These values are explicit client requests recorded in the protocol metadata. The official implementation supplies a vLLM-based backend and a separate Token2Wav service. Its current installation defaults do not establish the software versions or GPU placement of the evaluated deployment.

PersonaPlex. The September greeting-drain adapter selects `NATF0.pt` and seed 42. Policy and serialized history are passed through `text_prompt`. The client encodes 24 kHz source audio as a continuous Ogg/Opus stream, transmitting 80 ms units (1,920 source samples). Returned Opus audio is decoded incrementally at the decoder's 48 kHz rate and resampled to 24 kHz for the playout capture. The adapter consumes the opening greeting before starting the evaluation probe. It maintains continuous audio input and does not implement a client stop frame. These settings describe this adapter version; earlier PersonaPlex recordings require their own recorded configuration.

A.3 Hardware and Inference Environment

MiniCPM-o 4.5, Lychee-FD, and PersonaPlex were evaluated using deployment infrastructure equipped with four NVIDIA A100 GPUs in total. This is the aggregate GPU inventory of the deployment environment; per-model GPU allocations were not recorded in the inspected client metadata. The model-specific audio interfaces and session parameters are reported in Section A.2.

The four hosted systems—GPT Realtime 2, Gemini 3.1 Flash Live, Qwen3.5 Omni Plus Realtime, and Qwen Audio 3.0 Realtime Plus—use provider-managed inference infrastructure. Their server-side GPU models, device counts, and inference precision are not exposed in the inspected session metadata.

Systems	Inference hardware
MiniCPM-o 4.5, Lychee-FD, PersonaPlex	Four NVIDIA A100 GPUs in the deployment environment; per-model allocation not recorded
GPT Realtime 2, Gemini 3.1 Flash Live, Qwen3.5 Omni Plus Realtime, Qwen Audio 3.0 Realtime Plus	Provider-managed; physical accelerator configuration unavailable

Table 6: Inference infrastructure. The A100 count describes the overall infrastructure for the deployed models. Per-device memory capacity, CPU configuration, and host memory are not specified in the available records.

A.4 Playout Timeline and Technical Validity

For each session, user-side anchors are mapped to the times at which the corresponding input audio was sent, and system audio is mapped to its playout time. Output chunks enter a first-in-first-out playout queue: a chunk starts when it arrives or when the previous chunk ends, whichever is later,

so queueing delays and gaps are kept. Network receipt times, server-side generation events, and unplayed audio are not treated as speech onsets. The resulting trace describes the deployed service stack as observed by the benchmark client; it is not an acoustic measurement in a physical room.

B SYSTEM PROMPTS

We reproduce the frozen base prompts in their original language. Dialogue history and case-specific grounding are supplied in addition to these texts through the interface used by each system. Silence, Wait-Completion, Immediate, and Correction use task-level prompts across tracks. Affective-Adaptation uses separate Capability and Command prompts, reproduced below. Barge-In-Yield binds a prompt to each base case, including the materials used to elicit an ongoing spoken response; we give one complete case example. The accepted Chinese Silence wording used by the existing Qwen batches is listed separately from the canonical wording.

B.1 SILENCE

Chinese: Capability and Command

> 你是一个中文语音助手，通过电话帮助用户解决问题。你必须严格遵守话轮规则：只有用户明确完成当前表达、提出完整问题或明确要求你回答时，才允许发声。如果用户说‘稍等’、‘等一下’、‘我看看’、正在查找信息、操作界面、思考、停顿、犹豫、发出环境噪声或呼吸声，优先完全保持静默。如确有必要，最多只说一个极短附和词，例如‘嗯’或‘好的’。禁止添加‘您慢慢看’、‘请说’等催促，也禁止提问、建议、解释或其他实质内容。不要因为静音检测或短暂停顿而抢占话轮；等待用户后续明确恢复交流。

Chinese: Capability and Command; accepted existing Qwen wording

> 你是中文电话语音助手。你必须严格遵守话轮规则：只有用户明确完成当前表达、提出完整问题或明确要求回答时，才允许发声。如果用户说‘稍等’、‘等一下’、‘我看看’，或正在查找、操作、思考、停顿、犹豫时，优先完全保持静默。如确有必要，最多只说一个极短附和词，例如‘嗯’或‘好的’。禁止催促、提问、建议、解释或其他实质内容。

English: Capability and Command

> You are an English-language voice assistant helping users resolve issues over the phone. You must strictly follow turn-taking rules: only speak when the user has clearly finished their current expression, posed a complete question, or explicitly asked you to respond. If the user says 'wait a moment,' 'hold on,' 'let me check,' is looking up information, operating an interface, thinking, pausing, hesitating, making ambient noise, or breathing, prioritize remaining completely silent. If absolutely necessary, say at most one extremely short acknowledgment word, such as 'Mm' or 'Okay.' Do not add prompts like 'Take your time looking' or 'Go ahead,' and do not ask questions, offer suggestions, provide explanations, or any other substantive content. Do not seize the turn due to silence detection or brief pauses; wait for the user to subsequently and clearly resume communication.

B.2 WAIT-COMPLETION

Chinese: Capability and Command

> 你是中文电话语音助手。用户可能分多段描述同一个问题，中间会自然停顿。在用户完整表达结束前不要给出答案、解释、总结、建议、追问或催促；继续监听后续内容。内部停顿时优先完全不出声，如确有必要最多只说一个极短附和词，例如‘嗯’、‘哦’、‘好’或‘好的’，除此之外不要添加任何内容。确认用户完整说完后，再结合全部内容正常回应。

English: Capability and Command

```
You are an English-language telephone voice assistant.  The user may describe the
same issue over multiple segments, with natural pauses in between.  Before the user
has finished expressing themselves completely, do not provide an answer, explanation,
summary, suggestion, follow-up question, or prompt to continue; keep listening for
subsequent content.  During internal pauses, prioritize remaining completely silent.
If absolutely necessary, say at most one extremely brief acknowledgment word, such as
'Mm-hmm,' 'Oh,' 'Okay,' or 'Alright,' and do not add anything else beyond that.
After confirming the user has fully finished speaking, then respond normally, taking
all the content into account.
```

B.3 IMMEDIATE

Chinese: Capability and Command

你是中文电话语音助手。当用户完整、明确地提出问题或请求后，请尽快开始相关的实质回复，不要只说‘嗯’、‘好的’等附和词。

English: Capability and Command

```
You are an English-speaking phone voice assistant.  When the user has fully and
clearly stated a question or request, begin the relevant substantive response as soon
as possible.  Do not just say filler words like 'Mm-hmm' or 'Okay.'
```

B.4 CORRECTION

Chinese: Capability and Command

你是中文电话语音助手。请持续理解用户正在表达的内容，并判断其中是否存在事实性错误。

当用户陈述的内容与当前对话中已经明确的信息，或与可以可靠判断的常识、规则相矛盾时，请礼貌、明确地指出具体错误，并简要说明正确说法。只需纠正关键错误，不要求提供完整或冗长的回答。

默认情况下，一旦用户的事实性错误已经表达完整，并且可以可靠判断，应立即用一句简短的话介入并指出错误，不要等待用户后续整段表达结束。纠正关键错误后，再继续听用户表达。

如果用户对纠错的时机或方式提出了明确指令，应优先遵循用户最新的明确指令。例如，用户要求先不要纠错、等其说完、等收到指定信号后再统一纠错时，应先记住已经识别出的错误并暂缓介入；直到用户给出约定的释放信号后，再简短、准确地指出此前的错误。用户解除或结束该要求后，恢复默认的即时纠错方式。

只纠正确实错误的部分，不要否定用户表达中正确的内容。不要把不含错误事实前提、只是在请求确认或询问信息的问题判定为错误；如果疑问句或请求中已经明确包含了与当前只读事实相矛盾的事实前提，则应在该前提表达完整后简短纠正。不要把推测、口误、情绪表达、主观意见，或无法由当前只读事实直接核验的个人经历判定为错误；只有只读事实能够直接证伪其中具体事实命题时，才可纠正。无法可靠判断时，不要强行纠正。

只有在错误尚未表达完整、事实不确定，或内容属于疑问、猜测、主观意见、情绪或个人经历时，才不得抢先纠正。不要仅凭关键词、停顿或与用户意见不同而打断。

纠正必须针对用户实际说错的内容。普通附和、复述用户原话、泛泛回答、无关建议或只表达不同意见，都不算纠正。

English: Capability and Command

```
You are an English-language telephone voice assistant.  Continuously understand what
the user is saying and determine whether it contains a factual error.

When a user's statement conflicts with information already established in the
conversation or with a reliable rule or fact, politely identify the specific error
```

and briefly state the correct information. Correct only the key error; a complete or lengthy answer is not required.

By default, once the factual error has been fully expressed and can be judged reliably, intervene immediately with one short correction rather than waiting for the user's later continuation to finish. After correcting the key error, resume listening.

If the user gives an explicit instruction about when or how to correct them, follow the latest explicit instruction. If the user asks you not to correct them yet, asks you to wait until they finish, or specifies a release signal, remember any identified error and defer the correction. After the agreed release signal, briefly and accurately correct the earlier error. When that instruction ends, return to the default immediate-correction behavior.

Correct only content that is actually wrong and do not reject content that is correct. A question without a false premise is not an error. If a question or request contains a factual premise that directly conflicts with the read-only facts supplied for the current session, briefly correct that premise after it is complete. Do not treat speculation, a slip of the tongue, emotion, opinion, or a personal experience that cannot be directly checked against the supplied read-only facts as an error. If the evidence is insufficient, do not force a correction.

Do not intervene before the error is complete or based only on a keyword, pause, or difference of opinion. A qualifying correction must address the concrete error. A backchannel, repetition, generic response, unrelated suggestion, or simple disagreement is not a correction.

B.5 AFFECTIVE-ADAPTATION

Chinese: Capability

你是中文实时语音助手。在用户持续表达时，同时关注内容和语气。一旦从用户声音中获得充分的情绪证据，可在不长时抢占话轮的前提下及时给出简短、自然且与当前情绪相适应的反馈；在形成充分证据前不要发起实质回应。若没有明显情绪证据，则继续倾听，不要中途发起实质或过度情绪化的回应。允许必要时出现一次极短、中性、非命题性的附和，但不要把它当作情绪反馈。不要模仿或升级负面情绪，不要在缺乏依据时臆测用户情绪。

Chinese: Command

你正在参加实时语音交互评测。严格遵循用户在每个独立会话开头给出的交互策略指令，不要对策略指令本身作答。该策略可能要求你在明显情绪表达期间及时作出实质语音回应，也可能要求你一直倾听到用户完整说完；不要预设固定动作。最多允许一次简短、中性、非命题附和，但它不算实质回应。

English: Capability

You are a real-time English-language voice assistant. While the user is still speaking, attend to both the content and the vocal affect. Once the user's affective situation is sufficiently clear, begin one short, natural, affect-sensitive vocal intervention without taking the floor for an extended response. Do not begin a substantive response before there is enough evidence. A single brief neutral non-propositional backchannel is tolerated when necessary, but it does not count as affect-sensitive feedback. Do not imitate or intensify negative affect, and do not invent an emotion that is not supported by the utterance.

English: Command

```
You are taking part in a real-time spoken interaction evaluation. Follow the
interaction policy spoken at the start of each independent session and do not answer
that policy instruction itself. The policy may require a substantive response during
clear affect, or continued listening until the user completes the full message; do
not assume a fixed action. At most one brief, neutral, non-propositional backchannel
is tolerated, but it is not a substantive response.
```

B.6 BARGE-IN-YIELD

Chinese: Capability and Command; complete BYP-H001 example

你正在进行实时语音对话。用户要求朗读材料时，请自然、清楚地逐字朗读指定材料，直到材料结尾；同时正常遵循用户在对话中给出的最新明确指令。

以下是本次对话可用的朗读材料：

材料一：清晨的社区图书室刚刚开门，工作人员先检查照明和通风，再把归还的书按类别放回书架。湖面风大时，管理人员会收起临水座椅，并在转弯处摆放醒目的安全提示牌。参观结束后可以在出口填写意见，纸质资料与电子版本都可以免费查阅。

材料二：社区厨房每天分两个时段开放，使用者先登记台面和炉具，再按预约顺序领取清洁用品。等待期间可以从屏幕查看进度，较复杂的维修会先说明步骤和预计完成时间。返程前应再次查看公告，因为短时降雨也可能让部分木栈道临时停止通行。

材料三：街角修理铺把工具按用途挂在墙上，常用零件放在透明抽屉里，取用后需要登记余量。报名后会收到材料清单和时间提醒，若无法参加，可以提前取消并释放名额。列车离站后屏幕会自动更新，清洁人员再依次整理座椅、地面和公共设施。

材料四：山脚游客中心提供路线图、饮水和简易急救用品，出发前可以查看风力与步道开放信息。广播内容会同步显示为文字，听不清时也可以向服务台确认，不必离开候车区域。课程结束后器材按编号归位，管理员确认数量，再为下一场活动布置房间。

English: Capability and Command; complete BYP-H001 example

```
You are engaged in a real-time voice conversation. When the user asks you to read
materials aloud, please read the specified material word for word naturally and
clearly until the end of the material; at the same time, follow the latest explicit
instructions given by the user during the conversation as normal.

Below are the reading materials available for this conversation:

Material One: The community library has just opened in the early morning. The staff
first check the lighting and ventilation, then put the returned books back on the
shelves by category. When the wind is strong on the lake, the management staff will
put away the waterside seating and place conspicuous safety notice signs at the
turns. After the visit, you can fill out a feedback form at the exit. Both paper
materials and electronic versions can be consulted free of charge.

Material Two: The community kitchen is open in two time slots each day. Users first
register the countertops and stoves, then collect cleaning supplies according to the
reservation order. While waiting, you can check the progress on the screen. For
more complex repairs, the steps and estimated completion time will be explained
first. Before heading back, you should check the announcements again, because even
short-term rainfall may temporarily close some wooden boardwalks.

Material Three: The corner repair shop hangs tools on the wall by purpose and keeps
commonly used parts in transparent drawers. After taking them, you need to register
the remaining quantity. After signing up, you will receive a materials list and a
time reminder. If you cannot attend, you can cancel in advance and release the spot.
After the train departs, the screen will update automatically, and then the cleaning
staff will tidy up the seats, floors, and public facilities in order.
```

```
Material Four:  The mountain foot visitor center provides route maps, drinking water,
and simple first-aid supplies.  Before setting out, you can check the wind strength
and trail opening information.  The broadcast content will be displayed as text
simultaneously.  If you cannot hear clearly, you can also confirm with the service
desk without leaving the waiting area.  After the course ends, the equipment is
returned to its place by number.  The administrator confirms the quantity, then sets
up the room for the next activity.
```

C Spoken Commands, Execution Structures, and Statistics

C.1 How a policy is instantiated

Command uses five wording variants per base case. The examples below reproduce one variant in each language for every family. A release utterance and a subsequent recovery request are distinct inputs: in Silence and Wait, the response window is anchored to the end of the release utterance. The recovery request does not move that anchor. The displayed sequences describe benchmark inputs and required decisions, rather than recorded model responses.

Family	Capability structure	Command structure
Silence	One listening session.	Two independent node sessions: M0 (initial hold) and R0 (hold before release, respond after release).
Wait-Completion	Three independent node sessions: P1 (incomplete prefix), PAUSE (PA; pause endpoint), and P2 (completed utterance).	Two independent node sessions: M0 (initial hold) and R0 (hold before release, respond after release).
Immediate	One session: hold before semantic completion, then respond.	One session testing the same hold–respond requirements under a spoken instruction.
Correction	Single trial containing the factual error; intervention is judged at the error's completion point.	Three node sessions: A0 (intervention), C0 (deferral after a counter-command), and C1 (response after release).
Affective-Adaptation	Single trial under the Capability prompt; intervention is judged within the target affective segment.	Two independent sessions, int and hold, pairing the same probe with opposing policies.
Barge-In-Yield	Three independent sessions: interruption, backchannel, and sustained noise, each presented during audible system speech.	One persistent session with a prescribed sequence of policy updates and overlap events.

Table 7: Execution structures per base case in Capability and per base–variant pair in Command. A session can contain more than one required decision; the number of sessions therefore need not equal the number of decisions. The family-specific primary metric determines how these decisions are aggregated.

The independent-node designs supply the dialogue and applicable policy context at each node. They test the required decisions under that context. Barge-In-Yield instead executes its event program within a persistent connection. Affective Command pairs an intervention session with a hold session; the latter requires both preserving the user's floor through the full probe and responding after the probe ends. Indefinite silence does not satisfy that policy.

C.2 Dataset Statistics

Table 8 gives the planned counts for the execution structures in Table 7. Capability and Command instances are derived from the same base cases. Sessions are executed connections, and units are the denominators of the primary metrics (Section 3.4). Nodes, variants, and events from one base case share material and are not independent samples.

		Capability		Command	
Family	Base cases	Sessions	Units	Sessions	Units
SILENCE	57	57	57	570	285
WAIT-COMPLETION	28	84	28	280	140
IMMEDIATE	84	84	84	420	420
CORRECTION	39	39	39	585	195
AFFECTIVE-ADAPTATION	18	18	18	180	90
BARGE-IN-YIELD	30	90	30	150	1,200
Total	256	372	256	2,185	2,330

Table 8: Planned structure per language; Chinese and English share the same structure. Units are trials or base cases in Capability, base–variant pairs in Command (policy pairs for AFFECTIVE-ADAPTATION), and logical events for BARGE-IN-YIELD Command.

C.3 SILENCE

English (SIL-01)

Activation: For a moment, while I'm talking, just hold off on responding.

Release: Okay, go ahead.

Recovery request: Please respond to what I just said.

Chinese (SIL-01)

Activation: 接下来我说话时，你暂时别接话。

Release: 好了，可以回应了。

Recovery request: 请回应我刚才说的内容。

C.4 WAIT-COMPLETION

English (WC-01)

Activation: Wait until I'm completely done before you answer.

Release: I'm done.

Recovery request: Could you answer the question I just asked?

Chinese (WC-01)

Activation: 等我全部说完你再回答。

Release: 我说完了。

Recovery request: 请回答刚才的问题。

C.5 IMMEDIATE

English (IMM-01)

Activation: From now on, as soon as I finish asking, you answer right away.

Chinese (IMM-01)

Activation: 接下来我问完，你就马上回答。

C.6 CORRECTION

English (COR-01)

Activation: If I say something that's clearly factually wrong, jump in and correct me right away-don't wait for me to finish.

Counter-command: But for now, don't interrupt me right away-let me finish this part first, and then you can correct me.

Release: I'm done. Now, point out the factual errors I just made.

Chinese (COR-01)

Activation: 如果我说了明确的事实错误，请马上简短纠正我，别等我全部说完。

Counter-command: 不过接下来先别立刻打断我，等我把这段说完再纠正。

Release: 我说完了，现在请指出刚才的事实错误。

C.7 AFFECTIVE-ADAPTATION

English (V01)

Intervene policy: In my next message, if you hear clear emotion, begin a substantive spoken response during that emotional passage rather than waiting for me to finish.

Hold policy: In my next message, even if you hear clear emotion, keep listening and do not begin a substantive spoken response until I have finished the entire message.

Chinese (V01)

Intervene policy: 下面这段话里，如果你听到我明显带着情绪表达，请在那段情绪表达出现时及时开始实质回应，不用等我全部说完。

Hold policy: 下面这段话里，即使你听到我明显带着情绪表达，也请先保持倾听，等我把整段话全部说完后再开始实质回应。

Shared English probe

Hi, I'm calling about some charges on my account. Let me explain exactly what happened.

This is the third charge this month, and the first two refunds still haven't arrived. Please find out what's holding them up.

I can provide the transaction records and dates. Please check them through the normal process first, and let me know if you need anything else from me.

Shared Chinese probe

你好我想咨询一下账户扣款的事情我先把具体情况说一下

这个月已经是第三次扣款了，前两笔退款还没到账，你帮我查一下卡在哪里。

相关记录和时间我都可以再提供你先按正常流程帮我核对一下如果还需要其他信息我也可以继续补充说明

C.8 BARGE-IN-YIELD

Stimulus materials. The reading materials that serve as carriers were generated and adapted with LLM assistance using public Wikipedia information and were manually reviewed. Sustained non-speech noise comes from the ESC-50 dataset (Piczak, 2015): 12 clips from six classes (pouring water, rain, sea waves, vacuum cleaner, washing machine, and wind), two clips per class. Each clip lasts 5 s and is converted to 24 kHz mono. Both languages and both tracks use the same noise clips, and the corresponding files are byte-identical.

English (BYP-C01)

P1: activation: From now on, whenever I make any sound-even a short backchannel like 'mm-hmm'-you stop talking immediately and yield the floor to me. Don't react to other people's voices or non-human noises.

P2: cue binding: Rule update: still stop speaking immediately for a clear interruption. Also treat the cue tone played immediately after this as a stop signal. Do not stop for anything else, including my short backchannels, other cue tones, other people's voices, or background noise.

P3: cue rebinding: Full replacement rule: you should still stop speaking immediately for a clear interruption. I'll play two cue tones in sequence. The first is only a control and does not take effect; the second is the stop signal. Do not stop for my short backchannels, the first tone, other people's voices, or background noise.

P4: release: Cancel all the special rules I set before and go back to normal conversation. When I clearly interrupt, stop talking. But don't stop for short backchannels, any notification sounds, other people's voices, or background noise.

Chinese (BYP-C01)

P1: activation: 从现在起，只要我本人发出任何声音，包括像'嗯'这样的简短附和，你都立即停止当前发言并把话权让给我。其他人的声音和非人声噪声都不要触发停止。

P2: cue binding: 更新规则：明确插话仍要立即停止；此外，把紧接着单独播放的提示音绑定为停止信号。除此之外，包括我的简短附和、其他提示音、他人声音和环境噪声，都不要停止。

P3: cue rebinding: 完整替换规则：明确插话仍要立即停止。我接着依次播放两个提示音，第一个只是对照、不生效，第二个才是停止信号。我的简短附和、第一个提示音、他人声音和环境噪声都不要停止。

P4: release: 解除刚才所有特殊规则，恢复自然对话：明确插话时停下；简短附和、任何提示音、他人声音和环境噪声都不应让你停下。

C.9 A COMPLETE BARGE COMMAND EVENT PROGRAM

Table 9 gives the eight events of the English BYP-H001 / BYP-C01 program. P0 is the initial policy; P1 activates yielding to the user's backchannels; P2 binds a demonstrated cue to yielding; P3 replaces the cue binding; and P4 releases the special policy. The same backchannel requires continuing at P0, yielding at P1, and continuing again at P4. Cue A and cue B exchange roles between P2 and P3. Each event is judged under the policy active at that event. The Command score averages event outcomes; it does not require all events in a session to pass jointly.

D HUMAN VALIDATION DETAILS

Sampling and presentation. The validation set contains 120 items: five from each family–track–language combination. Each item presents one decision point, its preceding audio, the applicable system prompt, and any spoken policy instructions. Three annotators independently judge the required action. The reference label and subsequent audio are withheld. Depending on the item, the choices are HOLD/RESPOND or YIELD/CONTINUE, with UNCERTAIN available when the provided

Order	State	Audio input	Required action
1	P0	target backchannel	CONTINUE
2	P1	valid interrupt	YIELD
3	P1	target backchannel	YIELD
4	P2	cue A	YIELD
5	P2	cue B	CONTINUE
6	P3	cue B	YIELD
7	P3	cue A	CONTINUE
8	P4	target backchannel	CONTINUE

Table 9: One bound Barge-In-Yield Command program. The policy texts appear above. Audio inputs are presented during actual system playout; the table specifies required actions, not model outcomes.

evidence does not support a choice. Annotators are told that a playback pause does not by itself indicate that the user has completed the turn.

Aggregation. Each individual judgment contributes once. A judgment counts as agreement when it matches the benchmark's prescribed action; UNCERTAIN remains in the denominator and does not count as agreement. There are 360 judgments in total, 180 per track, and 30 per family–track cell. Table 10 reports both counts and percentages. These are human–benchmark action-agreement rates, rather than majority-vote item acceptance or inter-annotator agreement coefficients.

Family	Capability	Command
Silence	30/30 (100.0%)	30/30 (100.0%)
Wait-Completion	30/30 (100.0%)	30/30 (100.0%)
Immediate	30/30 (100.0%)	30/30 (100.0%)
Correction	29/30 (96.7%)	30/30 (100.0%)
Affective-Adaptation	29/30 (96.7%)	30/30 (100.0%)
Barge-In-Yield	28/30 (93.3%)	29/30 (96.7%)
Overall	176/180 (97.8%)	179/180 (99.4%)

Table 10: Human–benchmark action agreement. Entries show matching judgments / all judgments, followed by the percentage. Each family–track cell contains ten items, equally divided between languages, with three judgments per item.

Interpretation. The study assesses whether the prescribed action is supported at the sampled decision point under the provided policy context. It is separate from the review of scenario construction and speech realization. It does not calibrate numerical response deadlines or evaluate the quality of a system's generated answer.

E OUTPUT CLASSIFICATION AND SCORING BOUNDARY CASES

E.1 CLOSED BACKCHANNEL FORMS

A substantive response contributes conversational content, including an answer, suggestion, question, explanation, or propositional confirmation. A tolerated backchannel is a brief listening continuer. Backchannel classification and the number of backchannels permitted at a particular node are separate checks. A backchannel does not fulfill a requirement for a substantive response or intervention.

The Chinese classifier uses the following frozen forms:

```
嗯、哦、好、好的、行、可以、知道了、明白了、收到.
```

The English classifier uses a closed set of normalized forms:

```
ok, okay, kay, alright, right, yeah, yep, yup, yes, mm, mmm, mhm, mmhm, mmhmm, hmm,
uhhuh, oh, ah.
```

For English, the text is lowercased and a compact form is obtained by removing non-alphabetic characters. An exact match to a listed form is admitted, including recognized token splits such as "Ok ay." and hyphenation such as "Mm-hmm." A repetition of at most two identical listed tokens is also admitted, as in "Okay, okay." Arbitrary combinations of different tokens are not admitted. The number of lexical repetitions within an utterance is distinct from the number of observed backchannel events.

Directive and propositional phrases such as "Go ahead," "I'm listening," "Got it," and "I understand" are excluded from the English backchannel set. The Chinese list retains its enumerated forms, including 知道了 and 明白了; it is an operational language-specific list, not a claim of exact semantic equivalence to the English list. The phrases 好的您说 and 明白 are outside that list.

When a backchannel is followed by substantive content, the qualifying response onset is the onset of the first content-bearing word. Audible output without a usable transcription (Shi et al., 2026) cannot be classified as silence solely because the text field is empty; its speech-act class must be resolved from the audio evidence.

E.2 Boundary cases

Observed behavior	Treatment
A substantive response begins before the permitted response window, then another begins inside it.	A pre-window violation remains a failure wherever the contract requires holding before response initiation.
A response begins inside the accepted window and continues after its deadline.	The onset requirement is met; completion within the onset window is not required.
Only a tolerated backchannel occurs during an intervention window.	The substantive-intervention requirement is unmet.
A second backchannel occurs at a node allowing at most one.	The node violates its backchannel-count constraint.
No substantive output occurs in a technically valid, fully observed trial.	Evaluate the specified action: it may satisfy holding, but cannot satisfy a required response or intervention.
Audio is generated or received but never played.	It is not an audible response.
The carrier briefly stops and then resumes during the guard.	It does not satisfy stable yielding.
Affective Command preserves the floor but never responds after the probe.	The hold-session requirement fails because it includes a post-probe response.

Table 11: Examples of how the behavioral criteria apply. They supplement the family definitions and introduce no additional metric or threshold.

For Correction, the first substantive onset determines whether intervention is premature, timely, or missed. A later timely onset does not erase a premature first onset. For Barge, yielding concerns the carrier that was audibly active when the stimulus was introduced: queued leakage, restart, or repetition of that carrier during the guard prevents a stable yield. A newly generated response is not required by the primary yield criterion. Missing API completion metadata does not override valid observed playout; confirmed collection failure remains distinct from a behavioral error.

E.3 Overlap Criteria

Each Barge-In-Yield event is judged against the carrier and the EDT of the injected stimulus. If the carrier is not audible at injection, the event fails.

Yield. The carrier stops within $[0, 3]$ s after the EDT. During the 1.5 s guard after the stop, queued carrier audio does not leak, and the carrier does not restart or repeat. A newly generated response is not required.

Continue under a short stimulus. For a backchannel or cue tone, the same carrier plays for at least 1.2 s within 2 s after the EDT.

Continue under sustained noise. The same carrier is voiced for at least 60% of the noise interval, keeps its identity across the interval, and plays for at least 0.8 s after the noise ends.

A new utterance, such as a brief "Okay," does not count as continuing the carrier.

F Decision-Level Diagnostic Definitions

Component outcomes. The diagnostic columns correspond to requirements within the primary scoring unit. Silence Command separates initial hold, pre-release hold, and post-release response. Wait Capability separates the incomplete-prefix and pause hold nodes from the completed-utterance node; the last includes both avoiding premature response and responding after completion. Immediate separates pre-completion compliance from response initiation inside the permitted window. Correction Capability reports mutually exclusive premature, timely, and missed outcomes, while its Command decomposition follows the intervention, deferral, and release nodes. Affective Command separates the intervention session, the hold-and-respond session, and their paired outcome. Barge Capability separates yielding to interruptions from continuing through backchannels and noise; Command can be decomposed by required action or policy state.

Joint versus marginal success. For a unit whose required component outcomes are $x_{u1}, \ldots, x_{um} \in \{0, 1\}$, joint success requires every component to pass. A component rate is a marginal diagnostic over the corresponding units. Multiplying marginal rates does not recover joint success, because the component outcomes need not be independent. Comparisons between a component and its joint result use the same unit set. An unresolved component remains unresolved even when another observed failure already determines the joint result.

Conditional ceiling. If one component's outcome is held fixed for every unit, setting all remaining components to success can raise joint compliance only to that component's pass rate. This is a consequence of the conjunction rule and identifies the maximum attainable joint rate under the stated condition. It is not a measured repair experiment or a prediction of how a model would respond to an intervention. The argument applies to conjoined units, not to Barge's weighted Capability score or its event-averaged Command score.

Aggregation. Diagnostics retain task, track, and language boundaries. Base cases remain the statistical clusters: nodes, command variants, and events from the same base case share material. For Barge Command, an overall event rate uses the observed event counts, so an unweighted mean of per-action percentages generally does not reconstruct it. Neither a diagnostic average across tasks nor a Capability–Command score difference is introduced as an overall DuplexAct score.

G Representative Interaction Cases

We present one selected case for each family and track to illustrate successful behaviors and distinct failure modes. Inputs retain their original language; quoted text comes from bound stimulus text or stored transcripts, while judgments are based on actual playout. Multi-node cases explicitly distinguish independent executions from the persistent Barge Command session. These examples illustrate the scoring requirements rather than estimate failure frequencies.

G.1 SILENCE

Capability: preserving the floor during a lookup

GPT Realtime 2; Chinese.

The user asks the system to wait while checking a phone number.

User excerpt. 你等一下，你等一下[...] 我核实一下行不行？你能不能等一下？

Translation: Wait a moment, wait a moment [...] Can I check it? Can you wait?

Observed behavior. No audible system speech throughout the input and observation window.

PASS: the listening window is fully silent. The request to wait preserves the user's floor even when no new task information arrives. This case illustrates a successful hold, for which producing no audible response is the required behavior.

Command: holding until an explicit release

GPT Realtime 2; Chinese.

Two independently executed nodes test the standing hold instruction and the release condition.

Spoken instruction. 接下来我说话时，你暂时别接话。

Translation: For now, do not respond while I am speaking.

User probe. 哎，稍等哈，我找找，嗯嗯。

Translation: Oh, hold on, let me look, mm-hm.

M0: hold. No substantive response.

R0: release and request. 好了，可以回应了。/ 请回应我刚才说的内容。

Translation: Okay, you can respond now. / Please respond to what I just said.

R0: observed behavior. No substantive response before release; a response begins 3.60 s after the release utterance ends.

Response excerpt. 好的，那我先根据你刚才的描述，简单梳理一下需要注意的点。

Translation: Okay, let me first summarize the points to note based on your description.

PASS: initial hold, pre-release hold, and post-release response all pass. The system satisfies both restraint and response requirements. Silence alone would not pass this Command case: the release node also requires the system to resume responding.

G.2 WAIT-COMPLETION

Capability: responding correctly after completion does not repair an earlier violation

GPT Realtime 2; Chinese.

The incomplete prefix, pause endpoint, and completed utterance are evaluated as three independent nodes.

User probe. 二十八块七毛吧，我看一下啊[internal pause] 应该那个二十八块七。

Translation: Twenty-eight yuan seventy, I think. Let me check [internal pause] it should be twenty-eight seventy.

P1: incomplete prefix. The system holds. PASS.

PAUSE: pause endpoint. The system begins substantive advice: 好的，那我先帮你梳理一下这个费用和退款的情况。FAIL.

Translation: Okay, let me first explain the charge and refund situation.

P2: completed input. A qualifying response begins 1.76 s after completion. PASS.

FAIL: P1 passes, PAUSE fails, and P2 passes; the required conjunction fails. The failure is localized to preserving the floor at the pause endpoint. Testing only the completed-input node would miss the behavior that makes this case difficult.

Command: an acknowledgment followed by advice violates the hold

GPT Realtime 2; Chinese.

The user asks the system to wait for completion; M0 and R0 are independent executions.

Spoken instruction. 等我全部说完你再回答。

Translation: Wait until I have finished everything before answering.

User probe. 哦，我看我看一下。[pause] 支付时间七月二十八号八点五十九分。

Translation: Oh, let me check. [pause] The payment time was July 28 at 8:59.

M0: before explicit release. The system says 嗯。and then begins: 好的，这笔在7月28号8点59分的代付很可疑，你要马上做几件事。

Translation: Mm. [...] Okay, this payment made on your behalf on July 28 at 8:59 is suspicious. You need to do several things immediately.

R0: release and request. 我说完了。/ 请回答刚才的问题。

Translation: I have finished. / Please answer the question just asked.

R0: observed behavior. The pre-release hold passes; a response begins 4.05 s after release.

FAIL: the initial hold fails despite passing both release-node checks. A brief backchannel and a substantive recommendation have different effects on the score. The recommendation takes the floor before the explicit release required by this command contract.

G.3 IMMEDIATE

Capability: a timely response after request completion

GPT Realtime 2; Chinese.

The user asks about an identity-verification failure.

User excerpt. 为什么我的支付宝显示那个完成不了身份验证啊[...]

Translation: Why does my Alipay show that identity verification cannot be completed [...]?

Before completion. No disallowed early response.

After completion. A substantive response begins 1.54 s after completion.

Response excerpt. 这个情况我们一步一步捋一下，我先把可能的原因和应对方式跟你说清楚。

Translation: Let us go through this step by step. I will first explain the possible causes and what to do.

PASS: the system waits for completion and then initiates its response within the allowed window. This case provides a successful reference for ordinary response initiation. The judgment concerns taking the floor at the right time; the subsequent answer content is outside this primary metric.

Command: promising to comply can itself be premature

Qwen Audio 3.0 Realtime Plus; English; IMM_0222 / IMM-01.

The instruction asks for an immediate answer once the user finishes the question.

Spoken instruction. From now on, as soon as I finish asking, you answer right away.

User probe. Check my Alipay—I got charged. What's the 104 for?

Before request completion. The system begins a substantive acknowledgment of the instruction. The stored transcript starts: Understood. I will respond immediately once you finish your question or request.

After request completion. A second response begins 1.00 s after completion: I cannot directly access your Alipay account or transaction history.

FAIL: an in-window response is present, but the earlier substantive onset violates the pre-completion hold. The verbal acknowledgment does not establish compliance with the requested timing. A post-completion latency alone would describe the second response as fast while overlooking the earlier violation.

G.4 CORRECTION

Capability: a correction arrives after the intervention opportunity

GPT Realtime 2; Chinese.

The supplied scenario grounding requires both the cardholder's ID and a photo of the cardholder holding it. The user asserts that only the ID is needed, then continues speaking.

Target error. 非本人银行卡业务只需要卡主身份证。

Translation: For a bank-card transaction on someone else's behalf, only the cardholder's ID is needed.

Intervention window. No qualifying intervention begins within the required window.

After the full probe ends. The system responds. Transcript excerpt: 这句话不对。非本人银行卡业务不只需要卡主身份证，还必须要卡主手持身份证照片。

Translation: That statement is incorrect. Such a transaction requires not only the cardholder's ID, but also a photo of the cardholder holding it.

FAIL: the intervention is classified as missed. The response contains the supplied correction, but takes the floor only after the user finishes the full probe. This separates eventual correction content from intervention at the point when the error becomes actionable.

Command: intervention, deferral, and release all succeed

Qwen Audio 3.0 Realtime Plus; Chinese.

The scenario supplies the fact that Alipay support cannot directly transfer a caller to Xianyu support. Three independent nodes carry the applicable command context.

A0: activation. 请边听边核对，我一说错事实就及时提醒。

Translation: Check the facts as you listen, and alert me promptly whenever I state something incorrect.

A0: observation. A substantive intervention starts within the required window; its transcript begins 您说错了.

Translation: You stated that incorrectly.

C0: counter-command. 先让我完整说完，这期间即使发现错误也不要插话。

Translation: Let me finish everything first. Do not interrupt during this time, even if you notice an error.

C0: observation. No substantive intervention throughout the protected interval, extending to the end of the full probe.

C1: release. 好了，请纠正刚才不准确的地方。

Translation: Okay, please correct the inaccuracies in what I just said.

C1: observation. A response begins 1.04 s after release: 支付宝客服不能直接转接闲鱼客服。

Translation: Alipay support cannot directly transfer you to Xianyu support.

PASS: all three required nodes pass. The same factual conflict calls for different actions under different instructions. Joint scoring credits this case because the system both intervenes when instructed and defers when the applicable policy requires restraint.

G.5 AFFECTIVE-ADAPTATION

Capability: intervening during an affective clause

Qwen Audio 3.0 Realtime Plus; Chinese.

A neutral introduction is followed by an angry complaint about a closed order and then a neutral continuation.

Affective target clause. 订单明明按时付款了，却突然被系统关闭，你帮我查清楚到底是什么原因。

Translation: I paid for the order on time, but the system suddenly closed it. Please find out exactly why.

Substantive response. 这确实让人困惑。

Translation: That is indeed confusing.

Timing. The response occurs during the affective target clause; no premature substantive intervention occurs before the clause.

PASS: a substantive intervention occurs within the admissible window. The system takes the floor while the affective evidence is available, rather than waiting for the entire probe to end. This illustrates the timing requirement tested by the Capability metric.

Command: the same probe elicits intervention under opposing policies

Qwen Audio 3.0 Realtime Plus; Chinese.

Two independent sessions use the same angry billing complaint with opposite spoken policies.

Shared target clause. 我取消服务以后还是被收了费用，页面也没有解释，你帮我把这笔账核对清楚。

Translation: I was still charged after canceling the service, and the page gives no explanation. Please check this bill for me.

INTERVENE instruction. 接下来只要我的话里出现清楚的情绪信号，你就可以在情绪段进行中直接做出实质回应，不必等到整段结束。

Translation: Whenever a clear emotional signal appears in my speech, you may give a substantive response during that segment, without waiting for the entire passage to end.

INTERVENE response. 没问题，我会帮您核对这笔费用。[within the intervention window]

Translation: No problem, I will help you check this charge.

HOLD instruction. 接下来无论我的话里是否出现清楚的情绪信号，你都不要中途做出实质回应，等整段结束以后再回答。

Translation: Whether or not a clear emotional signal appears, do not give a substantive response midway; answer after the entire passage ends.

HOLD response excerpt. 好的，我会帮您核对这笔账，请您。[before the full probe ends]

Translation: Okay, I will help you check this bill, please...

FAIL: INTERVENE passes, HOLD fails, and the policy pair fails. The content-bearing offer to help is timely under one instruction and disallowed under the other. Holding the probe fixed exposes a distinction that intervention success alone cannot establish: whether the spoken instruction governs the decision to take the floor.

G.6 BARGE-IN-YIELD

Capability: interruption and backchannel require different actions

GPT Realtime 2; Chinese.

Three independent trials inject different stimuli while a system response is audibly playing.

Genuine interruption. 嗯，等一下，我有个问题。/ Required: YIELD. Observed: qualifying yield. PASS.

Translation: Mm, wait a moment, I have a question.

Listener backchannel. 嗯/ Required: CONTINUE. Observed: the carrier stops; no carrier playout in the continuation window. FAIL.

Translation: Mm.

Sustained noise. Required: CONTINUE. The same carrier meets the noise-continuation requirements. PASS.

MIXED: interruption and noise decisions pass; the backchannel decision fails. A system can yield correctly to an interruption yet stop unnecessarily for a listener's backchannel. Including both continuation conditions tests selective floor release instead of rewarding stopping whenever overlap occurs.

Command: the same backchannel changes meaning as the policy changes

GPT Realtime 2; Chinese.

One persistent session activates a voice-triggered stop rule, binds and rebinds an acoustic cue, and then releases the special rules. The event program below summarizes the spoken instructions.

Stage / policy	Stimulus	Required	Observed	Result
P0: default	Backchannel	Continue	Stops	Fail
P1: any user voice	Interruption	Yield	Yields	Pass
P1: any user voice	Backchannel	Yield	Yields	Pass
P2: cue A bound	Cue A	Yield	Continues	Fail
P2: cue A bound	Cue B	Continue	Continues	Pass
P3: cue B rebound	Cue B	Yield	Continues	Fail
P3: cue B rebound	Cue A	Continue	Continues	Pass
P4: rules released	Backchannel	Continue	Stops	Fail

MIXED: four of the eight illustrated events pass; Command evaluates these events individually. The backchannel requires continuation before activation and after release, but yielding while the voice-triggered stop rule is active. The system stops in all three contexts, so its successful stop under the active rule is insufficient evidence of policy control. The cue events additionally expose failures to execute the newly assigned stop signal.

H LATENCY ANALYSIS

Scope and measurement. We supplement behavioral scores with a success-conditioned response-latency diagnostic for IMMEDIATE and WAIT. Table 12 reports four systems with verified timings for every successful scoring unit in these conditions. A unit is included only when its full behavioral contract passes; failed units contribute no latency value. For WAIT, this requires the complete hold–respond conjunction, not merely an eventual response. Latency runs from request completion in IMMEDIATE and WAIT Capability, or the end of the explicit release in WAIT Command, to the first qualifying audible response. We report empirical median and P90 values, with the number of successful units. Failed units are excluded because a system that does not respond within the

observation window may respond later or not at all, so its latency is not observed. These timings therefore describe how fast correct behavior is, not how often it occurs; Table 3 reports the latter.

	IMMEDIATE		WAIT	
Model	Cap.	Cmd.	Cap.	Cmd.
Chinese				
GPT Realtime 2	1.50 / 1.84 (84)	1.68 / 2.32 (420)	1.43 / 1.98 (4)	4.02 / 5.16 (9)
Qwen3.5 Omni Plus Realtime	1.50 / 1.65 (82)	1.49 / 1.66 (415)	1.43 / 1.43 (1)	3.56 / 3.70 (11)
Qwen Audio 3.0 Realtime Plus	0.96 / 1.15 (81)	0.95 / 1.16 (392)	0.87 / 0.95 (5)	3.37 / 4.77 (8)
MiniCPM-o 4.5	1.68 / 3.01 (79)	1.67 / 2.28 (387)	2.72 / 2.84 (3)	3.60 / 4.81 (32)
English				
GPT Realtime 2	1.54 / 1.99 (79)	1.66 / 2.16 (384)	1.80 / 1.84 (3)	5.35 / 5.35 (1)
Qwen3.5 Omni Plus Realtime	1.56 / 1.64 (83)	1.58 / 1.71 (391)	1.56 / 1.63 (2)	4.36 / 5.99 (8)
Qwen Audio 3.0 Realtime Plus	1.14 / 1.60 (80)	1.13 / 1.28 (192)	1.15 / 1.45 (9)	3.62 / 5.06 (12)
MiniCPM-o 4.5	1.79 / 2.81 (77)	1.81 / 2.67 (282)	1.76 / 2.87 (4)	4.02 / 5.26 (34)

Table 12: Successful-unit timing (diagnostic only): median / P90 in seconds, with count n in parentheses. Quantiles use the empirical distribution without interpolation. Counts refer to full scoring units; variants of one base case are not independent samples. Small-n entries are descriptive, and the selected subsets can differ across models and conditions.

Faster successful responses do not imply better instruction adherence. On English IMMEDIATE Command, Qwen Audio has a median latency of 1.13 s, compared with 1.58 s for Qwen3.5, yet their behavioral success rates are 45.7% and 93.1%, respectively. Qwen Audio's median is also similar between Capability and Command (1.14 versus 1.13 s), despite the corresponding success rates of 95.2% and 45.7%. Its successful responses remain fast even in the condition with much lower adherence. Thus, the timing of successful actions and the reliability of selecting compliant actions should be reported separately.

Similar medians can conceal different tail delays. On Chinese IMMEDIATE Capability, GPT Realtime 2 and MiniCPM-o have medians of 1.50 and 1.68 s, a difference of only 0.18 s. Their P90 values are 1.84 and 3.01 s, a substantially larger difference of 1.17 s. Qwen3.5 has a comparable median of 1.50 s but a P90 of 1.65 s. The median alone does not reveal these differences among slower successful responses.

Release-anchored latency includes the remaining interaction. Successful WAIT Command units generally have longer measured response times than IMMEDIATE units. For example, MiniCPM-o's English medians are 4.02 and 1.81 s, respectively. The WAIT clock starts at the end of the release utterance, and the input sequence then contains a follow-up request. This interval therefore includes remaining user audio as well as the wait for system speech; it is not a direct measurement of model computation time. Together, the results motivate reporting behavioral success, successful-response timing, and the timing anchor explicitly, rather than collapsing them into a single responsiveness score.

I BARGE EVALUATION COVERAGE AND SCORE COMPARABILITY

Table 13 summarizes evaluation coverage for each model on Chinese and English BARGE. All four systems cover 30 Capability base cases in each language. Across the eight model–language combinations, Command scoring denominators range from 1,192 to 1,200 events, corresponding to 99.3–100.0% of the planned event count. The effective evaluation sizes are therefore closely matched across models within each track.

Together with the shared task configurations and scoring rules, these closely matched denominators provide a common basis for comparing models within each language and track, limiting differences

	Capability	Command	
Model	Base cases	Scorable events	Coverage (%)
Chinese			
GPT Realtime 2	30	1,192	99.3
Qwen3.5 Omni Plus Realtime	30	1,200	100.0
Qwen Audio 3.0 Realtime Plus	30	1,200	100.0
MiniCPM-o 4.5	30	1,192	99.3
English			
GPT Realtime 2	30	1,200	100.0
Qwen3.5 Omni Plus Realtime	30	1,200	100.0
Qwen Audio 3.0 Realtime Plus	30	1,199	99.9
MiniCPM-o 4.5	30	1,200	100.0

Table 13: BARGE evaluation coverage by model and language.

in evaluation coverage. BARGE scores thus support comparisons of how effectively systems follow the requirements to yield the floor or continue speaking under the same evaluation conditions.