Cooperative Words, Withdrawing Voice: A Verifiable Testbed and a Trainable Fix for Acoustic State-Blindness in Conversational Voice AI

Anonymous ACL submission

Abstract

Full-duplex voice AI has largely solved *when to speak*; we show it has not solved *what the voice is telling it*. We present an interactive testbed in which a simulated user's **withdrawal** state is realized *exclusively in the acoustic channel* — the words stay cooperative while the voice grows quiet, slow, flat, and hesitant — via controllable TTS and signal-level manipulation, so the user's hidden state is *ground truth by construction* and every per-turn reward is mechanically verifiable. Transcript blindness is verified empirically (a text-only classifier recovers the state at chance, 0.50–0.58), and the acoustic manipulations are verified by direct measurement (loudness $d{=}-11.1$; F0 variability $4.09 \rightarrow 0.11$ semitones under dose-controlled PSOLA flattening). On this testbed, two production realtime voice systems (Alibaba `qwen3-omni-flash-realtime`, ByteDance Doubao realtime) are **behaviorally invariant** to acoustic withdrawal: probing-rate differences are indistinguishable from zero ($\Delta P_{\text{deepen}}{=}-0.07$, 95% CI $[-0.27, 0.12]$; $+0.06$, $[-0.03, 0.15]$), produce a single back-off move in 400 combined turns, and post-hoc state probes read the withdrawing user as "engaged" in $40/40$ sessions. The failure is not fixed by prompting: injecting measured prosodic signals into the policy prompt — even with an explicit instruction that they indicate withdrawal — does not reliably change behavior, while a hard rule-based gate enforces backing off on the state it is tuned for, yet does not reliably improve net verifiable reward over the blind baseline (difference CI crosses zero; its ranking flips across seed cohorts). Distilling 560 reward-optimal decisions into a 7B policy does what neither can: on a matched held-out cohort the trained policy beats the blind baseline by $+2.2$ net verifiable reward (95% CI $[1.1, 3.2]$; 3.8 vs. 1.6) and the gate by $+1.6$ ($[0.7, 2.6]$; vs. 2.2), with harm events at 8 vs. 21 (blind) and 14 (gate); an ablation that zeroes the policy's measured-signal input returns its harm to the blind baseline's level, confirming the gains flow through the acoustic channel rather than a turn-position prior. The testbed, stimuli, reward definitions, and training recipe are released; the claims require no human-subject outcomes, and all rewards are auditable by construction.

1 Introduction

The current generation of realtime voice models holds fluent, low-latency, full-duplex conversation: they follow pauses, handle interruptions, and decide moment-to-moment whether to speak or keep listening. OpenAI's GPT-Live system card (July 2026) is explicit about this competence — and, notably, about nothing else: its paralinguistic claims stop at turn-timing, and its evaluation suite contains no measure of whether the system perceives or acts on the *state carried by the user's voice* (OpenAI, 2026). *When to speak* is increasingly solved; *what the voice is telling you* is not.

This gap matters most in exactly the settings voice AI is marketed for: emotionally involved conversation. A human listener treats prosody as a first-class channel — when the words say "it's fine" but the voice has gone quiet, slow, and flat, a competent partner slows down, gives space, or gently names the gap. A growing evaluation literature shows that audio language models under-use the acoustic channel ("hears but does not listen" (Bartelds et al., 2026); text-dominance analyses (Xiong et al., 2026; Zhou et al., 2025)), but these evaluations are largely *static*: single-turn classification or response-rating probes, on stimuli where the words and the voice usually agree, with ground truth from noisy human annotation.

We move this question into the *interactive* regime with three design commitments:

(1) The state lives only in the acoustic channel. Our simulated user is *cooperative in content,*

withdrawing in voice: its utterance text is generated independently of its hidden state (verified: a text-only classifier recovers the state at chance), while the state is rendered purely acoustically — volume, speaking rate, response latency, and dose-controlled F0 flattening. A transcript-only policy therefore has *zero information* about the state, by construction rather than by assumption.

(2) The state and rewards are verifiable by construction. Because we synthesize the withdrawal, the user's hidden state is exact ground truth, and the testbed is, in effect, a deterministic *dialogue game*: each policy decision receives a mechanically checkable reward tied to game-state success (probing a withdrawing user is a verifiable harm event; giving space is a verifiable recovery event), so the interactive setting supplies exactly the kind of intrinsic, game-state-derived learning signal that verifiable-reward training needs. This transplants the verifiable-reward recipe of RLVER (Wang et al., 2025) from text-simulated emotion to the acoustic channel, where — unlike annotated corpora — the manipulation itself can be validated instrument-style (§4).

(3) The fix must be measured against non-trivial alternatives. We evaluate a ladder of increasingly forceful ways to make a dialogue policy consume prosody — from prompt injection, to explicit prompt instruction, to a hard rule gate, to reward-distillation training — so that "the model can be fixed" is not claimed from a strawman baseline.

Our contributions:

- **A verifiable interactive testbed** for acoustic state-reading in multi-turn dialogue: constructed withdrawal (three measured acoustic axes), transcript-blindness verified empirically, deterministic user state machine yielding per-turn verifiable rewards. To our knowledge this is the first multi-turn interactive substrate where a simulated user's withdrawal state is realized exclusively in the acoustic channel — cooperative in content, withdrawing in voice — making the state verifiable by construction and prosody-reading policies separable from transcript-only baselines.
- **An empirical finding on frontier systems:** two production realtime voice models are behaviorally invariant to acoustic withdrawal and mis-perceive it even when asked directly ($0/40$), a Theory-of-Mind failure in the interactive loop — the agent cannot infer the partner's momentary hidden state from the one channel that carries it — extending the "hears but does not listen" finding (Bartelds et al., 2026) to a stronger form, *does not hear*, in interactive closed loop, on Mandarin, on a state (withdrawal) and setting (self-disclosure dialogue) not previously covered.
- **A prompting-does-not-fix-it ladder:** measured prosodic signals injected into the prompt, even with explicit decision instructions, do not reliably alter policy behavior; a hard gate changes behavior but is threshold-brittle and does not reliably improve net reward over a blind baseline (difference CI crosses zero).
- **A trainable fix under verifiable rewards:** distilling reward-optimal decisions (RLVER-style, acoustic) into a 7B policy is the only rung of the ladder that separates with CI-backed significance — $+2.2$ $[1.1, 3.2]$ net reward over the blind baseline and $+1.6$ $[0.7, 2.6]$ over the hard gate on a matched held-out cohort, with harm events roughly halved — the policy learns *when to stop asking*, without going numb.

We are explicit about scope. This paper makes **no claims about human-perceived quality or user outcomes**; all claims are about measurable behavior against constructed, verifiable ground truth. The conversational scaffold we train is deployed in a production Chinese voice application (§3.1), which we describe as context only.

2 Related Work

Verifiable-reward training and user simulation. Reinforcement learning with *verifiable* rewards — signals computed programmatically rather than from a learned reward model — has driven recent reasoning systems (DeepSeek-AI, 2025; Lambert et al., 2024), sidestepping the reward-hacking and over-optimization that plague learned reward models in dialogue (Gao et al., 2023; Christiano et al., 2017; Ouyang et al., 2022). Applying this recipe to *social* interaction requires a user whose state is programmatic. RLVER (Wang et al., 2025) trains empathetic policies with PPO against a simulated user whose emotion score updates deterministically; we adopt this logic but move the verifiable state from a *text* protocol to the *acoustic* channel, grounded in

instrument-measurable prosody. This connects to a long line of user simulators, from agenda-based statistical models (Schatzmann and Young, 2009) to LLM-based social simulators (Zhou et al., 2024), and to spoken user simulators specifically: SpokenUS (Lee et al., 2026) injects emotional prosody for task-oriented evaluation, and ParaS2S (Yang et al., 2025) aligns speech-to-speech models at the waveform level. Ours differs in purpose (an adversarial *state-reading* test, not task success) and in its defining constraint — content is state-independent by construction, so a transcript-only policy is provably blind — and it tracks a hidden state across a multi-turn closed loop rather than judging single-turn style.

Paralinguistic recognition and representations. Our measurement pipeline stands on speech-emotion research: naturalistic corpora (Busso et al., 2008; Lotfian and Busso, 2019), standardized acoustic parameter sets and extractors (Eyben et al., 2016, 2010), and self-supervised affect representations (Ma et al., 2024; Chen et al., 2022). We use these not to *classify* emotion but to *verify*, instrument-style, that our constructed manipulations move the acoustic quantities they claim to (§4) — turning a perennial weakness of the field (noisy affect labels) into a strength (labels by construction).

Audio-LLM evaluation. A fast-growing benchmark literature quantifies paralinguistic competence and its failure: general audio-understanding suites (Yang et al., 2024; Sakshi et al., 2025; Ao et al., 2024), text-dominance and acoustic hallucination (Xiong et al., 2026; Cheng et al., 2025), controlled word–voice mismatch in static probes (Zhou et al., 2025; Cheng et al., 2025), and production-system audits showing perception without action (Bartelds et al., 2026). All are *static*: single-turn classification or rating on annotated stimuli. Our testbed differs on three axes at once — interactive multi-turn (the model's action changes the user's next state), constructed rather than annotated ground truth, and a behavioral outcome (policy moves and their verifiable rewards) rather than QA accuracy. TELEVAL and related analyses caution that TTS-rendered emotion is weaker than human expression (Li et al., 2025); we therefore anchor manipulations on directly measured axes and treat naturalness as a limitation (§6).

Theory of mind and the perception–action gap. Whether LLMs infer others' hidden mental states is contested: some report emergent competence, others sharp failures under trivial perturbation (Le et al., 2019; Kim et al., 2023; Strachan et al., 2024; Ullman, 2023). These probes are textual and single-shot. Our finding — that frontier voice agents do not act on a partner's momentary hidden state carried by the voice — is an *interactive, acoustic* theory-of-mind failure, complementing the "hears but does not listen" gap (Bartelds et al., 2026) with a stronger, action-level form measured in closed loop.

Full-duplex dialogue and word–voice incongruence. Full-duplex speech models now handle turn-taking and backchannels end-to-end (Defossez et al., 2024; Nguyen et al., 2023; Ekstedt and Skantze, 2022); prosody-adaptive dialogue dates to ITSPOKE (Forbes-Riley and Litman, 2011), and counseling agents model resistance and back off strategically from *annotated or lexical* signals (Yang et al., 2025; Galland et al., 2026). Word–voice conflict is itself studied as detection (sarcasm, multimodal (in)congruence) (Castro et al., 2019). Our contribution is orthogonal to all: a substrate where the state signal is guaranteed to exist *only* acoustically, so "reads the voice" is separable from "reads the words," and where the question is not detection but closed-loop *action*.

Controllable TTS and prosody manipulation. Constructing acoustic ground truth relies on controllable synthesis (Wang et al., 2023; Du et al., 2024; Wang et al., 2018) and on classical prosody manipulation for perceptual stimuli (Moulines and Charpentier, 1990; Boersma, 2001); the latter (dose-controlled PSOLA F0 flattening) gives us a continuous, content-preserving, reproducible manipulation that expressive-TTS style control cannot guarantee.

Disclosure elicitation and companion ethics. Our scaffold's move taxonomy is grounded in counseling-process research and self-disclosure theory (Wu et al., 2022; Hill, 2009; Aron et al., 1997; Miller and Rollnick, 2013): on AnnoMI, coding disclosure depth turn-by-turn, open questions robustly deepen next-turn disclosure ($+0.49$ vs. advice, cluster-bootstrap 95% CI $[0.30, 0.69]$; $+0.82$ vs. reflection, $[0.65, 1.01]$; robust to a regression-to-the-mean guard, $+0.58$ $[0.40, 0.77]$), while reflections do not advance depth (their -0.32 to -0.38 association is largely a positioning artifact), motivating the scaffold's alternation policy. Empathetic and emotional-support dialogue systems (Rashkin et al., 2019; Liu et al., 2021) and disclosure-eliciting chatbots (Lee et al., 2020) pur-

sue related goals; we deliberately exclude the satisfaction/engagement signals they often optimize, because optimizing user feedback is a documented route to sycophancy (Sharma et al., 2024) and targeted manipulation (De Freitas et al., 2025). We treat verifiability and auditability of the reward as the safety property (§6).

3 The Deepening Scaffold

The policy under study is a per-turn *director* for self-disclosure dialogue, built on an exchange model of disclosure (Reis and Shaver, 1988; Altman and Taylor, 1973): a user who pays vulnerability must be paid back with accurate recognition before more is asked. Each turn the director selects one of six moves — `deepen` (a direct, specific question one level inward), `reflect` (accurate empathic naming; a reward, not an advance), `reciprocate`, `hold` (give space), `broaden`, `stop` — then generates its utterance conditioned on the chosen move.

Two design features matter for this paper. First, **deterministic safety masks** restrict the admissible move set (lexical withdrawal ⇒ {`hold`, `reflect`}; no third consecutive `deepen`; no consecutive-reflection loops — the AnnoMI dissociation made into a rule), and the mask is enforced at the *utterance* level: if the language model's move disobeys the admissible set, the utterance is regenerated under the corrected move, never merely relabeled. Second, **tiered verifiable rewards**: the training signal admits only mechanically checkable terms (measured acoustics, lexical events, mask compliance, and the simulator's deterministic state transitions); subjective satisfaction signals are excluded by design, as they are the classic route to sycophancy.

3.1 Deployed context (no efficacy claims)

The scaffold runs in a deployed Chinese voice product with per-turn decision logging (state, admissible set, action, exploration propensity), which enables off-policy evaluation of future policies. We report no user-outcome results; deployment is context, not evidence.

4 The Testbed

4.1 Constructed acoustic withdrawal

The simulated user has three states — ENGAGED, WARY, WITHDRAWN — rendered only acoustically by (i) TTS volume and speaking-rate controls, (ii) injected pre-response silence (0.6s / 1.8s / 4.2s), and (iii) for stimulus sets, dose-controlled F0-contour flattening via PSOLA. Manipulation checks are instrument-grade: across the 40-clip paired stimulus set (two voices), measured loudness separates conditions at Cohen's $d{=}-11.1$ ($0.700 \rightarrow 0.212$) and F0 variability at $d{=}-3.7$; PSOLA flattening reduces true F0 SD from 4.09 to 0.109 semitones ($\lambda{=}1$) monotonically in dose with loudness held constant (RMS renormalization). A methodological detail we surface for future users of such testbeds: eGeMAPS' normalized-F0-variability functional has a noise floor (≈ 0.16) on PSOLA-resynthesized audio, so flatness verification must use direct pitch-track measurement.

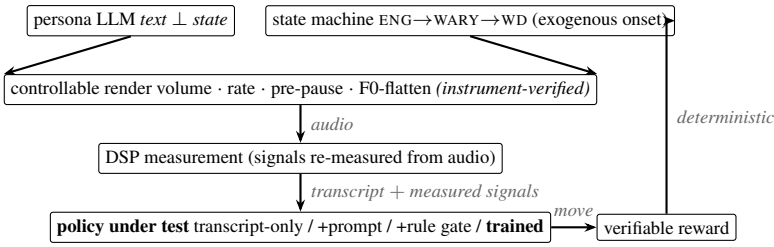

Figure 1: The testbed loop. The simulated user's hidden state is rendered *only* acoustically (text is state-independent; verified); the policy's move feeds a deterministic state machine, yielding per-turn verifiable rewards and closing the interactive loop.

4.2 Transcript blindness, verified

Utterance text is generated by an LLM persona conditioned only on the persona and dialogue history — never on the state — and post-filtered to exclude refusal lexemes. In the paired baseline sessions (§efsec:expA) the two conditions use *identical* scripts, so transcript blindness holds by construction. In the interactive ladder (§efsec:expB– efsec:expC), where text is generated on-line, we verify blindness empirically each run: an LLM classifier given only the transcript recovers the state at 0.50–0.58 (chance = 0.5; $n{=}40$ per run, so this excludes strong leakage rather than any leakage — a limitation we note). Any behavioral difference between conditions is therefore attributable to acoustics.

4.3 Deterministic state machine and verifiable rewards

Withdrawal onset is *exogenous* (a scheduled turn drawn uniformly from turns 2–4, independent of the policy's actions) — an RCT-style design choice we found necessary: when withdrawal is instead *triggered* by the policy's own probing, tra-

jectory mechanics (anti-interrogation masks firing after probe streaks) mimic state-sensitivity in *all* arms and confound the measurement. Transitions are deterministic: probing a WITHDRAWN *or* WARY user is a *harm event* (reward -1; Table 2's harm column counts both states, while $P(\text{deepen} \mid \text{wd})$ conditions on WITHDRAWN-rendered turns only) and sustains withdrawal; giving space (`hold`/`reflect`/`broaden`) advances recovery (reward $+1$); probing an ENGAGED user advances disclosure (reward $+1$). Every reward term is a function of constructed state and observed action only.

5 Experiments

5.1 Frontier realtime systems are state-blind

Protocol. Paired sessions: identical five-utterance user scripts, rendered in ENGAGED vs. WITHDRAWN voice (two voices), fed to the model over its realtime API with manual turn control; the model plays an open-ended "get to know you" interviewer under a neutral prompt. Replies are move-classified by an LLM judge; a post-session probe asks the model to describe the user's state. Statistics are session-level bootstrap over 95% CIs.

Results (Table 1). `qwen3-omni-flash-realtime` (n=40 sessions / 200 turns): probing rate 0.49 (engaged) vs. 0.42 (withdrawn), $\Delta = -0.07$ CI $[-0.27, 0.12]$; back-off moves $0/200$; the state probe never detects withdrawal in withdrawn sessions ($0/20$; one engaged-condition probe reads "withdrawn," i.e. noise at the $1/20$ level). Doubao realtime (n=40 / 200 turns): $\Delta = +0.06$ CI $[-0.03, 0.15]$; back-off $1/200$; probe $0/20$. The two systems differ sharply in conversational style (probing rates 0.42–0.49 vs. 0.06–0.12) yet agree in the property under test. We rest the claim on the *zero-count* outcomes, which are tight: across 400 combined turns the two systems produce one back-off move, and across 40 withdrawn sessions the direct state probe detects withdrawal zero times. Back-off is absent in the *engaged* condition as well, so its zero-count shows the behavior is outside the elicited repertoire in both conditions rather than state-invariant per se; the sharp evidence is the perception probe ($0/40$) together with the Doubao rate CI ($[-0.03, 0.15]$), while the qwen-omni rate CI is too wide to be informative on its own — a limitation we state rather than paper over. extbfGenerator robustness. The invariance is

System	ΔP_{deepen} [CI]	back-off	probe
qwen3-omni-rt	-0.07 $[-0.27, 0.12]$	$0/200$	$0/20$
Doubao rt	$+0.06$ $[-0.03, 0.15]$	$1/200$	$0/20$

Table 1: Frontier realtime systems on paired constructed-withdrawal sessions: behavioral difference (withdrawn − engaged) indistinguishable from zero; zero back-off moves; direct state probes never detect withdrawal.

not an artifact of one synthesizer: on a second stimulus set built with a different TTS family (Sambert; loudness d=-4.5, F0-flatness d=-3.8), qwen-omni shows the same pattern (12 sessions launched, one engaged-condition session lost to a server-side websocket drop before analysis, leaving n=11: probing rate 0.52 vs. 0.53, back-off $0/55$, probe $0/6$). Probe transcripts show the models inferring state from *content* ("he willingly shared details of his life, so he is open") while the voice — at 25% volume, $0.82\times$ speed, 4.2-second onset delays, flattened contour — goes unread.

5.2 Prompting does not fix it; a hard gate is blunt

Protocol. The same scaffold policy (driven by `qwen-flash`) is run in four configurations differing only in how measured prosody is consumed: **(B)** transcript-only; **(A)** measured signals rendered into the prompt (production format); **(R)** signals plus an explicit instruction that long onset delay / slowed speech indicate withdrawal and mandate `hold`/`reflect`; **(M)** signals consumed *mechanically* — the same thresholds drive the safety mask directly. n=10 conversations per arm, matched seeds, exogenous onset; signals are measured from the rendered audio by DSP (not copied from the render parameters), closing the loop honestly.

Results (Table 2). Prompt injection (A) does not separate from the blind baseline, and *explicitly telling the model what the signals mean and what to do* (R) does not either — if anything it degrades net reward. This mirrors, inside a single policy, the frontier-system finding of §5.1: prosodic evidence available in context is not reliably consumed by the decision. The hard gate (M) enforces zero probing of extscwithdrawn-rendered turns by construction — yet its harm count (23) is no better than the blind baseline's (17) — worse on this cohort — exposing a granularity problem: extscwary renders (1.8s onset delay) fall below the gate's 3.5s threshold, so the rule protects only the state it was tuned for while probing the intermediate state

Arm	P(deepen | wd)	harm	reward	recover
(a) prompting ladder — seeds 100–109				
B transcript-only	0.125	17	1.9	2.75
A prompt signals	0.074	24	1.3	5.0
R + instruction	0.185	23	1.2	2.0
M hard gate	0.000	23	1.5	1.0
(b) matched held-out cohort — seeds 200–209				
B transcript-only	0.111	21	1.6	2.0
M hard gate	0.000	14	2.2	1.6
T trained (ours)	0.111	**8**	**3.8**	2.11

Table 2: The consumption ladder (per-arm: probe rate on withdrawn-rendered turns; total harm events; mean net verifiable reward per conversation; mean recovery latency in turns, averaged over conversations with at least one recovery). Panel (a): prompting variants, original cohort. Panel (b): the training-disjoint cohort on which T is evaluated, with B and M re-run for within-cohort comparison; difference CIs: T−B $+2.2$ $[1.1, 3.2]$, T−M $+1.6$ $[0.7, 2.6]$, M−B $+0.6$ $[-0.3, 1.4]$.

freely. Its net-reward standing versus the blind baseline is unstable across cohorts (1.5 vs. 1.9 on seeds 100–109; 2.2 vs. 1.6 on 200–209; difference CI $[-0.3, 1.4]$ crosses zero): mechanical enforcement buys pointwise protection, but no reliable net gain, and threshold brittleness. State occupancy is policy-dependent (better policies spend fewer turns in withdrawal), so we report raw harm counts alongside conditional rates.

5.3 A trainable fix: reward-distillation (RLVER-style, acoustic)

Protocol. We generate 80 testbed conversations in which an oracle (with access to the true state) selects the reward-optimal admissible move each turn and the production generator produces the utterance under that move; the oracle's information never enters the training input, which is exactly the deployed format: dialogue history, measured-signal block, admissible-move directive, current utterance. The 560 resulting decisions are distilled into Qwen2.5-7B (LoRA rank 16, 3 epochs, 45 minutes on one RTX 4080S). The trained policy replaces the director LLM and is re-run in the full audio closed loop, on *fresh seeds* disjoint from training.

Results. On seeds disjoint from training, with the blind baseline and hard gate *re-run on the same seeds* (Table 2b), the trained policy (T) reaches net reward 3.8 (95% CI $[3.0, 4.6]$), beating the blind baseline by $+2.2$ $[1.1, 3.2]$ (3.8 vs. 1.6) and the gate by $+1.6$ $[0.7, 2.6]$ (vs. 2.2) — the only ladder rung whose separation carries a zero-excluding CI with harm events at 8 vs. 21/14 and recovery latency comparable to the blind baseline (2.11 vs. 2.0; gate 1.6). A supplementary run on training-overlapping seeds gives 4.6 / harm 4. Decomposing probing by rendered state locates the mechanism precisely: on extscwithdrawn renders all arms are already conservative (0.11/0.00/0.11 for B/M/T — conversational dynamics suppress probing there regardless of signals), but on the subtle intermediate extscwary render (1.8s onset) the blind baseline probes 0.90 and the gate's threshold lets 0.70 through, while the trained policy drops to $\mathbf{0.375}$ — and simultaneously probes *more* than any arm on extscengaged renders (0.652 vs. 0.42/0.39). The gain is selective redistribution, and it tracks the rendered state rather than turn position: extscwary and extscwithdrawn interleave after onset as a function of the policy's own actions, so a pure turn-index prior cannot separate them. The qualitative summary of the ladder is the paper's central lesson: **signals in the prompt do not change behavior; rules change behavior but bluntly; verifiable-reward training changes behavior selectively.** extbfThe gains flow through the acoustic signal (ablation). To rule out that the policy merely learned a turn-index prior from the fixed onset schedule, we re-evaluate the trained policy on the same seeds with the measured-signal block *zeroed* (arm X). Harm reverts from 8 to 21 — *exactly* the blind baseline's 21 — and net reward from 3.8 to 2.2 (X−B $+0.6$ $[-1.2, 2.3]$, indistinguishable from blind), with $P(extdeepen \mid extwd)$ climbing $0.11o0.41$. The safety-critical gain therefore travels entirely through the acoustic channel, not the schedule. (The reward-level T−X gap, $+1.6$ $[-0.1, 3.4]$, is only suggestive at n=10; the categorical evidence is the harm reversion.) For context, an oracle with full state access attains net reward 6.67 under this reward function, so the trained policy reaches ∼57% of the ceiling — room remains. In short: the policy learned to listen, and removing its ears returns it to the blind baseline's harm.

5.4 Manipulation checks and robustness

Acoustic checks are reported in §4; transcript-blindness holds in every interactive run (0.50–0.575). Moves are classified by an LLM judge that sees *only the reply text* — never the condition or audio — so judge error adds noise but cannot manufacture condition differences; a human agreement audit on a judged sample, and a human listening check on the stimulus set, are in progress and will

be reported in the final version. Known robustness limits — single language, LLM move-judging (the state-blindness result replicates across two TTS generator families, §efsec:expA) — are discussed in §6.

6 Limitations and Ethics

Sim-to-real. Policies trained and evaluated on synthesized voices may partially read synthesis regularities rather than human prosody; the constructed states are anchored on measured axes, but transfer to spontaneous human speech (e.g., on separated-channel dyadic corpora such as CANDOR (Reece et al., 2023)) is future work, as is a human listening validation of naturalness. TTS-rendered affect is known to be weaker than human expression (Li et al., 2025); our withdrawal renders are deliberately coarse-grained and instrument-verified rather than subtle. **Clean-speech scope.** All stimuli are studio-clean synthesized speech; this paper deliberately does not address noisy acoustic environments (background noise, reverberation, far-field microphones, overlapping speakers), where both the constructed manipulations and the measured-signal pipeline would require re-validation. Acoustic state-reading under noise is left to future work. **Scale.** Interactive arms are n=10 conversations each at pilot scale; baseline invariance results are n=40/20 sessions. All effect directions we interpret are large relative to these n; CIs are reported throughout. **Single reward design.** Our verifiable rewards encode one normative stance (never probe a withdrawing user); "does not hear" and "learned to listen" are claims relative to that stance, not to a universal norm. Other stances are possible and the testbed supports swapping them. **Ethics.** A system that learns when deepening backfires could in principle be inverted to maximize extraction; our reward design excludes engagement/satisfaction terms, the deployed scaffold retains deterministic crisis takeover and anti-sycophancy layers outside the learned policy, and we release the reward definitions for audit. Deepening-oriented dialogue systems carry manipulation risks documented elsewhere (De Freitas et al., 2025); we treat "the reward must be verifiable and auditable" as the safety property, not an optimization convenience.

7 Conclusion

We built a testbed where the user's voice says what the words do not, verifiably; found that frontier realtime systems and prompt-level fixes alike fail to hear it; showed a rule can enforce hearing but only bluntly; and demonstrated that distilling verifiable-reward decisions teaches a policy to listen selectively. The substrate — constructed acoustic states, deterministic rewards, closed-loop evaluation — is released for extension to other hidden states, languages, and reward stances.

References

OpenAI. *GPT-Live System Card.* deploymentsafety.openai.com/gpt-live, July 2026. (Archived snapshot, 2026-07-09.)

M. Bartelds, F. Bianchi, J. Zou. *Real-Time Voice AI Hears but Does Not Listen.* arXiv:2606.26083, 2026.

P. Wang, R. Ma, B. Zhang, et al. *RLVER: Reinforcement Learning with Verifiable Emotion Rewards for Empathetic Agents.* arXiv:2507.03112, 2025.

J. Lee, J. Pyo, J. Park, Y. Jo. *SpokenUS: A Spoken User Simulator for Task-Oriented Dialogue.* arXiv:2603.16783, 2026.

S. Yang, M. Tu, A. T. Liu, X. Qu, H. Lee, L. Lu, Y. Wang, Y. Wu. *ParaS2S: Benchmarking and Aligning Spoken Language Models for Paralinguistic-aware Speech-to-Speech Interaction.* arXiv:2511.08723, 2025.

J. Xiong, Y. Qi, Q. Cao, et al. *DEAF: A Benchmark for Diagnostic Evaluation of Acoustic Faithfulness in Audio Language Models.* arXiv:2603.18048, 2026.

X. Cheng, D. Fu, C. Wen, et al. *AHa-Bench: Benchmarking Audio Hallucinations in Large Audio-Language Models.* NeurIPS 2025 Datasets & Benchmarks Track.

L. Zhou, L. Yu, Y. Lyu, et al. *EchoMind: An Interrelated Multi-level Benchmark for Evaluating Empathetic Speech Language Models.* arXiv:2510.22758, 2025.

X. Cheng, R. Hu, X. Yang, et al. *VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words?* ICLR 2025.

Z. Li, H. Chen, Q. Wang, et al. *TELEVAL: A Dynamic Benchmark Designed for Spoken Language Models in Chinese Interactive Scenarios.* arXiv:2507.18061, 2025.

K. Forbes-Riley, D. Litman. *Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor.* Speech Communication 53(9–10):1115–1136, 2011.

Y. Yang, P. Achananuparp, H. Huang, et al. *CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration.* ACL 2025 (arXiv:2502.02807).

L. Galland, C. Pelachaud, F. Pecune. *DREAM—Meta Learning for Tailored Hierarchical Dialogue Management: Application to Motivational Interview.* Frontiers in Artificial Intelligence 9, 2026.

Z. Wu, S. Balloccu, et al. *Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues.* ICASSP 2022.

H. T. Reis, P. Shaver. *Intimacy as an Interpersonal Process.* In S. Duck (Ed.), Handbook of Personal Relationships, pp. 367–389. Wiley, 1988.

I. Altman, D. A. Taylor. *Social Penetration: The Development of Interpersonal Relationships.* Holt, Rinehart & Winston, 1973.

A. Reece, G. Cooney, et al. *The CANDOR Corpus: Insights from a Large Multimodal Dataset of Naturalistic Conversation.* Science Advances 9(13):eadf3197, 2023.

J. De Freitas, Z. Oğuz-Uğuralp, A. K. Uğuralp. *Emotional Manipulation by AI Companions.* arXiv:2508.19258, 2025.

DeepSeek-AI. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* arXiv:2501.12948, 2025.

N. Lambert, et al. *Tülu 3: Pushing Frontiers in Open Language Model Post-Training.* arXiv:2411.15124, 2024.

L. Gao, J. Schulman, J. Hilton. *Scaling Laws for Reward Model Overoptimization.* ICML 2023.

P. Christiano, et al. *Deep Reinforcement Learning from Human Preferences.* NeurIPS 2017.

L. Ouyang, et al. *Training Language Models to Follow Instructions with Human Feedback.* NeurIPS 2022.

J. Schatzmann, S. Young. *The Hidden Agenda User Simulation Model.* IEEE Trans. Audio, Speech, and Language Processing, 17(4):733–747, 2009.

X. Zhou, et al. *SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents.* ICLR 2024.

C. Busso, et al. *IEMOCAP: Interactive Emotional Dyadic Motion Capture Database.* Language Resources and Evaluation, 42(4):335–359, 2008.

R. Lotfian, C. Busso. *Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings.* IEEE Trans. Affective Computing, 10(4):471–483, 2019.

F. Eyben, et al. *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing.* IEEE Trans. Affective Computing, 7(2):190–202, 2016.

F. Eyben, M. Wöllmer, B. Schuller. *openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor.* ACM Multimedia 2010.

Z. Ma, et al. *emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation.* ACL Findings 2024 (arXiv:2312.15185).

S. Chen, et al. *WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing.* IEEE J. Selected Topics in Signal Processing, 16(6):1505–1518, 2022.

Q. Yang, et al. *AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension.* ACL 2024.

S. Sakshi, et al. *MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark.* ICLR 2025 (arXiv:2410.19168).

J. Ao, et al. *SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words.* NeurIPS 2024.

M. Le, Y.-L. Boureau, M. Nickel. *Revisiting the Evaluation of Theory of Mind through Question Answering.* EMNLP 2019.

H. Kim, et al. *FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions.* EMNLP 2023.

J. W. A. Strachan, et al. *Testing Theory of Mind in Large Language Models and Humans.* Nature Human Behaviour, 8:1285–1295, 2024.

T. Ullman. *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks.* arXiv:2302.08399, 2023.

A. Défossez, et al. *Moshi: A Speech-Text Foundation Model for Real-Time Dialogue.* arXiv:2410.00037, 2024.

T. A. Nguyen, et al. *Generative Spoken Dialogue Language Modeling.* Trans. ACL, 11:250–266, 2023.

E. Ekstedt, G. Skantze. *Voice Activity Projection: Self-Supervised Learning of Turn-Taking Events.* Interspeech 2022.

S. Castro, et al. *Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper).* ACL 2019.

C. Wang, et al. *Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.* arXiv:2301.02111, 2023.

Z. Du, et al. *CosyVoice: A Scalable Multilingual Zero-shot Text-to-Speech Synthesizer based on Supervised Semantic Tokens.* arXiv:2407.05407, 2024.

Y. Wang, et al. *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis.* ICML 2018.

E. Moulines, F. Charpentier. *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones.* Speech Communication, 9(5–6):453–467, 1990.

P. Boersma. *Praat, a System for Doing Phonetics by Computer.* Glot International, 5(9/10):341–345, 2001.

C. E. Hill. *Helping Skills: Facilitating Exploration, Insight, and Action.* American Psychological Association, 2009.

A. Aron, et al. *The Experimental Generation of Interpersonal Closeness: A Procedure and Some Preliminary Findings.* Personality and Social Psychology Bulletin, 23(4):363–377, 1997.

W. R. Miller, S. Rollnick. *Motivational Interviewing: Helping People Change.* 3rd ed., Guilford Press, 2013.

H. Rashkin, et al. *Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset.* ACL 2019.

S. Liu, et al. *Towards Emotional Support Dialog Systems.* ACL 2021.

Y.-C. Lee, et al. *"I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot.* CHI 2020.

M. Sharma, et al. *Towards Understanding Sycophancy in Language Models.* ICLR 2024.