

Conversation Coach: A Voice-enabled AI System that Helps Practice Difficult Workplace Conversations

Fanyou Wu* Suraj Maharjan* Ainur Yessenalina
Dennis Xu Chen Rahul Srivastava Srinivasan H. Sengamedu
Amazon

* Equal contribution.

{fanyouw, mhjsuraj, yessenal, dennydc, rahsri, sengamed}@amazon.com

Abstract

Effective manager-employee communication is critical for retaining high performers and developing underperformers, yet training managers in these skills remains costly. Text-based chatbots offer a scalable approach but cannot provide realistic rehearsal: managers need to practice speaking aloud to build confidence before high-stakes conversations. In this paper, we propose CONVERSATION COACH, a voice-first AI system that enables managers to rehearse difficult workplace conversations in a realistic spoken format. The system addresses three challenges: achieving low-latency interactions with strong language understanding, enabling adaptive conversations through configurable bot personalities that simulate different employee types, and generating personalized feedback on content and policy compliance. We compare an end-to-end speech-to-speech model with a cascaded approach combining automatic speech recognition, a large language model, and text-to-speech synthesis. The end-to-end approach achieves $3\times$ lower median (P50) latency with native barge-in capability at an estimated $8\times$ lower cost, while the cascaded approach offers superior reasoning essential for coaching quality. We deployed the cascaded architecture in production, where 40,000+ managers used it over six months, with adoption patterns indicating selective use for difficult conversations.

1 Introduction

Effective workplace communication, such as delivering difficult feedback, navigating conflict, and conducting sensitive conversations, is a critical managerial skill with significant organizational impact (Wilhelm et al., 2025). Traditional training approaches such as workshops, role-play with human partners, and executive coaching are expensive, difficult to scale, and often unavailable when managers need to practice the most: immediately before a challenging conversation.

Voice-based AI coaching offers a promising alternative. Unlike text chat, voice interaction requires managers to practice speaking aloud in real time, creating a higher-fidelity rehearsal environment (MacLeod et al., 2010). However, building effective voice coaching systems requires navigating complex tradeoffs between response latency, language understanding quality, and conversational naturalness (Ji et al., 2024).

The architectural landscape for speech-to-speech AI presents a fundamental choice. End-to-end models such as Moshi (Défossez et al., 2024) process speech directly, offering lower latency and native handling of interruptions. Cascaded pipelines chain separate Automatic Speech Recognition (ASR), Large Language Model (LLM), and Text to Speech (TTS) components, providing modularity and access to state-of-the-art text reasoning but accumulating latency (Paek and Pieraccini, 2008). Recent work suggests this tradeoff may be less stark than assumed (Liu et al., 2025), but production case studies comparing specific systems in coaching applications remain scarce. We make three contributions:

1. **A domain-specific design framework** organizing coaching-relevant tradeoffs across three dimensions to guide practitioners selecting between these approaches (§3).
2. **A production system comparison** of Nova Sonic 2 (NS2; end-to-end) and a cascaded pipeline (Amazon Transcribe + Claude Sonnet 4.5 + Amazon Polly), characterizing their latency, cost, and coaching quality tradeoffs within this specific application (§5).
3. **An organization-wide deployment** of a voice-first coaching system to 40,000+ managers, with adoption patterns revealing selective, just-in-time use for difficult conversations and actionable engineering lessons (§6, §7).

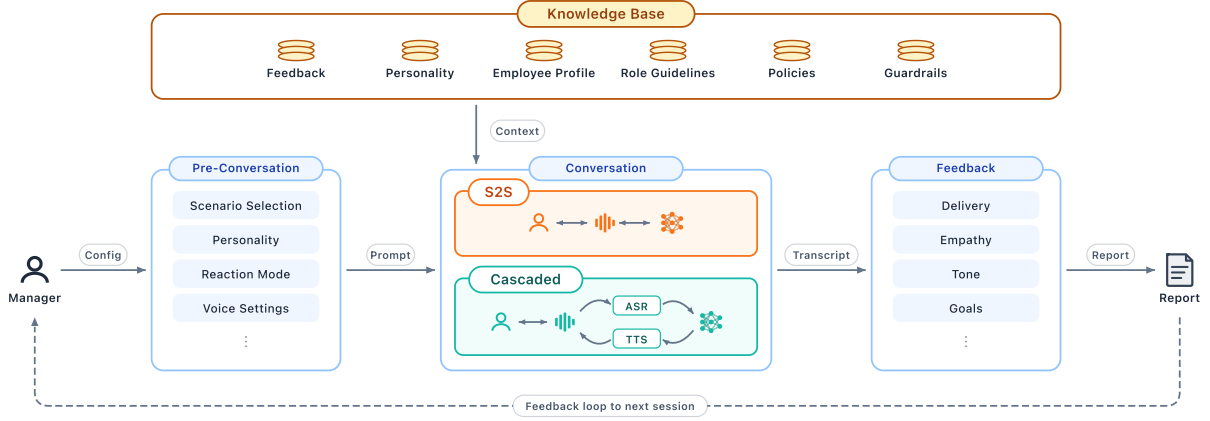


Figure 1: CONVERSATION COACH System Design. The Pre-conversation module configures the simulation by allowing managers to select an employee and assign personality and reaction modes. The Conversation module offers two alternatives: Speech-to-Speech (end-to-end streaming with native barge-in) or Cascaded (ASR → LLM → TTS, each stage streaming to the next to minimize latency). The Feedback module provides actionable insights on content, framing, and policy compliance.

2 Related Work

Speech-to-Speech Dialogue Systems. Spoken dialogue systems follow two paradigms: cascaded pipelines chaining ASR, LLM, and TTS components (Paek and Pieraccini, 2008), and end-to-end models processing speech directly (Borsos et al., 2023; Zhang et al., 2023; Défossez et al., 2024). While end-to-end approaches promise lower latency and preservation of paralinguistic cues, recent analyses (Liu et al., 2025; Ji et al., 2024) suggest optimized cascaded systems can achieve competitive latency while retaining superior text-based reasoning. Our work contributes an empirical comparison of these paradigms in a deployed coaching application.

AI-Based Coaching and Tutoring. Large language models have transformed conversational AI through instruction-following capabilities. Prior work includes spoken tutoring systems (Litman and Silliman, 2004), large-scale deployed tutoring (Afzal et al., 2019), AI-augmented executive coaching (Arakawa and Yakura, 2020), deployed mental-health coaching (Fitzpatrick et al., 2017), and recent studies on AI-assisted workplace communication training (Wilhelm et al., 2025). LLM-based persona simulation has been explored through role-playing frameworks (Wang et al., 2024; Zhou et al., 2023) that maintain character consistency across extended interactions, a challenge central to our system.

Real-Time Conversational Dynamics. Current spoken dialogue systems typically exhibit 700–1000ms latency (Skantze, 2021; Castillo-López et al., 2025). Predictive turn-taking models (Ekstedt and Skantze, 2020, 2022) and emotional responsiveness (Busso et al., 2008; Poria et al., 2019) are active research areas relevant to coaching applications.

3 Design Principles

Building an effective voice-based coaching system requires addressing three interconnected challenges.

Real-time spoken interaction. Coaching demands responsiveness that feels natural (Skantze, 2021). *Time to First Audio (TTFA)* measures the delay from when a user finishes speaking to when the system begins its response. End-to-end models avoid inter-component serialization; cascaded pipelines accumulate latency across stages, though streaming mitigates this. *Barge-in support* allows users to interrupt mid-response, which is common in emotional conversations. End-to-end models handle this natively; cascaded approaches require explicit interruption detection.

Adaptive persona simulation. The bot must consistently portray a specific employee persona across an entire conversation. This demands strong instruction following (adhering to personality type, reaction mode, and guardrails) and contextual consistency (maintaining character even when the

Design Space	E2E	Cascaded
<i>Real-time Interaction</i>		
Time to first audio	Lower	Higher
Barge-in support	Native	Requires logic
<i>Conversational Quality</i>		
Instruction following	Moderate	Strong
Persona consistency	Moderate	Strong
Content reasoning	Moderate	Strong
<i>System Engineering</i>		
Delivery analysis	Integrated	Separate module
Component upgradability	Coupled	Independent
Debugging granularity	Opaque	Per-component

Table 1: Qualitative architecture comparison based on production experience. E2E: end-to-end speech-to-speech (NS2); Cascaded: ASR + LLM + TTS pipeline (Transcribe + Claude + Polly).

user’s statements are surprising). Cascaded architectures benefit from state-of-the-art LLMs with strong instruction-following capabilities; end-to-end models must balance persona consistency against speech quality and latency.

Feedback generation. Post-conversation feedback analyzes *what* the manager said (content: word choices, policy compliance, key talking points). Future extensions could integrate *how* they said it (delivery: tone, pace, empathy, vocal confidence), where end-to-end models can potentially extract delivery features during conversation and cascaded pipelines would require a separate audio analysis module. Table 1 summarizes how each approach addresses these challenges.

4 System Architecture

We instantiate our tradeoff analysis in CONVERSATION COACH, a production system for practicing difficult workplace conversations. Figure 1 shows the overall architecture, which comprises three modules aligned with the design spaces identified in §3.

4.1 Scenario Configuration

A key design goal of the pre-conversation module is to generate system prompts that elicit realistic, varied employee behavior during role-play. To this end, the module aggregates two types of context: *scenario context* drawn from a knowledge base (company philosophy, policies, role-specific guidelines), and *employee context* including performance history, contributions, peer and manager feedback, and self-reflection statements. Managers

configure each practice session by selecting a conversation scenario (e.g., annual review, promotion or compensation discussion), an employee profile, and behavioral parameters described below.

Personality and Reaction Typology. Realistic coaching requires the simulated employee to exhibit a coherent behavioral profile rather than generic responses. Drawing on Marston’s DISC model (Marston, 1928) and Stone and Heen’s taxonomy of feedback reception styles (Stone and Heen, 2015), we define three personality types: Direct / Assertive, Passive / Conflict Avoidant, and Analytical / Detail Oriented. We consolidate the original four DISC quadrants into three types because the Influential and Steady styles both prioritize interpersonal harmony, yielding similar conversational behavior in our coaching context; separating them did not produce distinguishable role-play dynamics in pilot testing.

We cross these with three reaction modes that capture common emotional responses to negative feedback: Defensive / Deflective (challenging, justifying, or redirecting blame), Withdrawn (disengaging with minimal or flat responses), and Distressed (exhibiting emotional overwhelm such as raised voice or tearfulness). The resulting 3×3 matrix yields nine personality-reaction combinations, providing systematic coverage of diverse conversational dynamics (see Appendix B for the full typology with illustrative utterances).

4.2 Real-Time Role-Play

The conversation module facilitates real-time spoken role-play between the manager and the simulated employee. We implement both architectural approaches to enable direct comparison under identical coaching scenarios.

End-to-End Approach (Nova Sonic 2; NS2). Nova Sonic 2 (Amazon Web Services, 2026b) processes audio input and generates audio output within a single model, handling speech recognition, language understanding, response generation, and speech synthesis without inter-component serialization. This achieves faster TTFA with native barge-in support and preservation of paralinguistic cues from input audio.

Cascaded Approach. Amazon Transcribe (Amazon Web Services, 2026d) (ASR), Claude Sonnet 4.5 (Anthropic, 2026) (response generation),

and Amazon Polly (Amazon Web Services, 2026c) (TTS) prioritize reasoning quality over latency. Streaming is applied at each stage to mitigate accumulated latency; without it, end-to-end delay makes the system unusable (see §7 for measurements). For barge-in handling, the client monitors the ASR stream for new speech onset during TTS playback; upon detection, TTS audio is immediately halted and the new user utterance is processed as a fresh turn. The assistant turn in conversation history is truncated to the portion delivered before the interruption. This approach requires explicit client-side detection and playback control, unlike NS2’s native interruption handling.

4.3 Post-Conversation Feedback

The feedback module focuses on *content analysis*: word choices, message framing, policy compliance, and key talking points. A critical function is detecting privacy violations such as revealing confidential calibration ratings. Claude synthesizes the analysis into structured feedback: a conversation summary, specific strengths, and areas for improvement with actionable rephrasing suggestions (see Appendix C.2 for the prompt). The system stores all conversation history and feedback, enabling managers to revisit sessions and practice the same scenario across different personality–reaction combinations.

5 Evaluation

We evaluate whether the architectural tradeoffs identified in §3 manifest in practice, focusing on latency, persona consistency, speech quality, and coaching feedback quality.

5.1 Dataset

We constructed an evaluation dataset in two stages. First, 10 human-conducted conversations established a ground-truth rubric for validating LLM judge reliability (§5.3). We then created 148 synthetic scenarios spanning all 27 configurations (3 employee profiles $\times$ 9 personality–reaction types, $\sim$ 6 per configuration). Each profile contains role description, performance history, self-reflection statements, peer feedback, and manager feedback. Manager turns were generated by an LLM and synthesized to speech via TTS, so both architectures process realistic audio input through their full pipelines. Each scenario was replayed identically to both systems, yielding 296 transcripts (148 paired comparisons). A human reviewer excluded

14 of the initial 162 sessions for off-topic drift or premature termination.

5.2 Experimental Setup

We use Nova Sonic 2 (amazon.nova-2-sonic-v1:0, voice tiffany) for the end-to-end model and, for the cascaded approach, Amazon Transcribe in streaming mode for ASR, Claude Sonnet 4.5 (anthropic.claude-sonnet-4-5-20250929-v1:0) via Amazon Bedrock for response generation, and Amazon Polly (generative engine, voice Ruth) for TTS. All LLM operations use temperature = 0.7. Speech quality is assessed using UTMOS (Saeki et al., 2022).

5.3 Results

We compare NS2 and the cascaded approach across latency, persona consistency, speech quality, and coaching feedback (Table 2).

Metric	Gap	Favors
<i>LLM Judge Evaluation</i>		
Initial Reaction	marginal	–
Stakeholder Response	marginal	–
Authority Response	marginal	–
Feedback Reception	mixed	Cascaded
Timeline Pushback	marginal	–
Directive Compliance	substantial	Cascaded
Language Pattern	substantial	Cascaded
Role Adherence	substantial	Cascaded
Authenticity	substantial	Cascaded
<i>Other Metrics</i>		
UTMOS (Saeki et al., 2022)	4.43 vs 4.41	–
P50 TTFA (s)	1.4 vs 4.2	NS2
P99 TTFA (s)	2.8 vs 7.6	NS2
TM99 TTFA (s)	1.5 vs 7.0	NS2
Est. cost (\$/session)	\$0.05 vs \$0.39	NS2

Table 2: Comparison of Nova Sonic 2 (NS2) and the cascaded pipeline across behavioral and speech quality metrics. LLM judge results are majority votes across four models (Table 5b). Gap = mean score difference on a 5-point scale: *marginal* (<0.5), *moderate* (0.5–1.5), *substantial* (>1.5). *Mixed* = 2 judges marginal, 2 moderate. TM99 = trimmed mean excluding values above the 99th percentile. Per-judge breakdowns in Table 5b.

Latency. Measured server-side across all conversation turns, the median TTFA (P50) for NS2 is 1.4s compared to 4.2s for the cascaded approach (3 $\times$ improvement). At the 99th percentile (P99), NS2 remains responsive at 2.8s while the cascaded pipeline reaches 7.6s; the trimmed mean excluding

values above P99 (TM99) confirms this gap is consistent (1.5s vs 7.0s), not driven by outliers. Both approaches use streaming throughout; despite this optimization, the cascaded pipeline’s sequential processing results in noticeably higher first-audio latency.

ASR Accuracy. On the manager turns used in our evaluation (Polly-synthesized speech with known ground-truth text), Amazon Transcribe achieves 5.1% word error rate (WER). A matched per-turn measurement for NS2 was not possible, as its streaming transcript is not segmented by turn. For external reference, Transcribe reports $\sim 5.6\%$ WER in streaming mode (Picovoice, 2026) and NS2 $\sim 6.5\%$ on CommonVoice (Amazon Web Services, 2026b); although these benchmarks use differing test conditions, the similar ranges suggest ASR quality does not meaningfully differentiate the two approaches in our setting.

Persona Consistency and Reasoning. LLM judges evaluated all conversations spanning all nine personality-reaction combinations, assessing whether the bot (1) maintained its employee role throughout, (2) followed the assigned personality and reaction mode, (3) pushed back appropriately, and (4) challenged feedback rather than passively accepting it.

Both approaches show marginal differences on conversational dimensions (initial reaction, stakeholder response, authority response), indicating comparable performance on simpler interaction patterns. The cascaded approach shows a mixed result on feedback reception (split across judges) and substantial advantages on persona consistency dimensions, particularly in maintaining complex behavioral instructions across extended conversations. The cascaded advantage on role adherence and authenticity is unanimous across all four judge models; directive compliance and language pattern show 3-of-4 agreement. NS2 exhibits occasional character breaks under complex persona requirements but remains effective for straightforward scenarios.

Isolating Architecture from Model. Because the cascaded pipeline pairs a specific LLM (Claude Sonnet 4.5) with the cascaded architecture, its persona advantage could in principle stem from the model rather than the architecture. To disentangle the two, we re-ran the cascaded pipeline with

a substantially weaker and cheaper LLM, Claude Haiku 4.5, on the same scenarios and four-judge panel. A two one-sided tests (TOST) equivalence analysis (equivalence margin ± 1 Likert step) finds Sonnet and Haiku statistically equivalent on 8 of 9 persona dimensions, all at $p < 0.01$ across all four judges; only Initial Reaction falls outside the equivalence margin. Since a much weaker LLM preserves the persona advantage over NS2, the gap is attributable to the cascaded architecture’s ability to leverage any adequate text-based LLM through explicit system prompting, rather than to Sonnet’s specific instruction-following strength.

Human–LLM Agreement. To validate judge reliability, three annotators independently scored 54 conversations (27 per system, spanning all configuration cells) on six persona-relevant dimensions (Initial Reaction, Feedback Reception, Stakeholder Response, Language Pattern, Role Adherence, and Authenticity) applicable across all personality types. Annotators were blinded to system identity: transcripts were presented in randomized order with system names removed. Human–LLM agreement reaches Cohen’s $\kappa = 0.81$, exceeding the human–human agreement of $\kappa = 0.67$; that is, the LLM judge agrees with each annotator more closely than the annotators agree with one another, confirming that the judge produces reliable scores across all configurations.

Inter-Model Agreement. We repeated the evaluation with four diverse judge models (Claude, Nova, MiniMax, and GLM): Fleiss’ $\kappa = 0.41$ (moderate) with 83.3% majority consensus. The cascaded advantage on Role Adherence and Authenticity is unanimous across all four judges, while Directive Compliance and Language Pattern show 3-of-4 agreement (the Nova judge rates these as marginal; Table 5b). Potential same-family bias (Claude judge favoring the Claude-powered cascaded system; Nova judge favoring NS2) is mitigated by the inclusion of two independent judges (MiniMax, GLM) with no family overlap, both of which align with the majority on all unanimous dimensions.

Speech Quality. After filtering responses shorter than 3 seconds (which are primarily barge-in fragments), UTMOS (Saeki et al., 2022) evaluation (scale 1–5) shows both architectures produce high-quality speech.

Model	Precision	Recall	F1
Claude Opus 4.5	0.93	1.00	0.97
Nova Pro	0.85	1.00	0.92
MiniMax M2.5	0.90	1.00	0.95
GLM-5	0.90	1.00	0.95

Table 3: Confidentiality-violation detection across four diverse classifier models, evaluated against 148 ground-truth labels (28 violations, 120 non-violations). All models achieve perfect recall; false positives are over-cautious flags of indirect references.

Cost. For a typical 10-turn session, we estimate the cascaded approach costs approximately \$0.39 (Transcribe \$0.02, Claude \$0.19, Polly \$0.18) versus \$0.05 for NS2, an $8\times$ reduction. The primary cost drivers are generative TTS and the LLM’s growing context window across turns.

Coaching Feedback. Of the same 148 synthetic conversations, 28 contain intentional policy violations (managers disclosing confidential calibration ratings) and the rest do not, providing known ground-truth labels for evaluating violation detection. To mitigate model-family circularity, we evaluate detection with four diverse classifiers (Claude Opus 4.5, Nova Pro, MiniMax M2.5, GLM-5; identifiers in Table 5a). All four achieve perfect recall and high precision, yielding F1 scores of 0.92–0.97 (Table 3). False positives across all models were over-caution: flagging indirect references (e.g., “you are in the solid performer category”), which is preferable in a compliance-sensitive context.

6 Deployment and Usage Patterns

To assess real-world viability beyond offline evaluation, we report deployment metrics from an organization-wide launch available to all managers during the annual review cycle covering a six-month period.

Based on the evaluation results in §5, we deployed the **cascaded architecture** in production, prioritizing persona consistency and reasoning quality over latency. The 22% return rate (§6.1) and stable session durations suggest that the cascaded latency was acceptable for this coaching context, though we cannot rule out that lower latency would have driven higher engagement without an A/B comparison. Table 4 summarizes key adoption figures.

Metric	Value
Total sessions	54,000+
Unique managers	40,000+
Returning users	22%
Median session duration	135 s

Table 4: Deployment summary over a six-month period.

6.1 Adoption Patterns

Adoption is notably higher among newer front-line managers (over $3\times$ the overall adoption rate across all eligible managers), and the 22% return rate and median 135 s session duration confirm substantive engagement rather than superficial exploration. Technical managers engage at higher practice rates than non-technical managers, while newer managers spend 19% more time per session than tenured peers. Usage peaks $4\times$ above the period-average session rate in the two weeks preceding review deadlines, confirming just-in-time preparation rather than general skill development. Among sessions that begin practice, 36% complete the full voice conversation and 23% of completers request feedback, indicating progressive commitment once the initial activation barrier is overcome. Qualitative feedback from active users corroborates the engagement metrics. One manager noted: “*The roleplay functionality works well, picking up my varied tone and inflection. It also sounded like a true conversation.*” Another shared: “*The role play functionality successfully demonstrated a challenging performance review scenario.*”

6.2 Selective Use for Difficult Conversations

Managers disproportionately use the system for their most challenging conversations: employees in the lowest performance tiers appear in 43.3% of unique coached manager-employee pairs (nearly $2\times$ their baseline rate), and managers prepare for only 17.6% of conversations on average, with 84% preparing for fewer than 25% of their direct reports, indicating deliberate, targeted use rather than routine application. Within-team comparisons are consistent with this selection effect: coached employees exhibit 10.9 percentage points (pp) lower career growth sentiment (a survey-based measure of perceived development opportunities) than non-coached teammates prior to the conversation, reflecting pre-existing challenges that prompted the manager to seek practice.

6.3 Preliminary Outcome Analysis

We compared coached employees who experienced a rating decline to matched controls whose managers did not use the system, using a doubly robust estimator (Bang and Robins, 2005) that combines propensity weighting with outcome regression and adjusts for tenure, level, job family, location, and prior rating. The current sample is underpowered (minimum detectable effect, $MDE \approx 13$ percentage points, pp); the observed 4.3 pp difference in job satisfaction between coached and control groups is not statistically significant and should be read only as motivation for a future adequately powered study.

7 Discussion

The cascaded pipeline’s persona advantage is architectural rather than a byproduct of the specific LLM: substituting a substantially weaker model (Claude Haiku 4.5) preserves the advantage over NS2 (§5.3). These findings are nonetheless specific to the systems evaluated; other instantiations may shift the tradeoff boundary. Beyond this comparison, our deployment surfaced several engineering lessons.

Streaming is non-negotiable. Without end-to-end streaming across ASR, LLM, and TTS, accumulated latency exceeds 8 seconds. Only full-pipeline streaming, where TTS begins synthesizing as soon as the first LLM tokens arrive, achieves acceptable conversational flow.

Prompt sensitivity. NS2 occasionally simplifies complex persona instructions, particularly when multiple behavioral constraints interact. Claude Sonnet 4.5 follows detailed prompts reliably but requires engineering to avoid verbose, text-style outputs; we constrain response length and use Speech Synthesis Markup Language (SSML) directives to produce natural speech patterns.

Persona drift. In longer conversations (15+ turns), both architectures gradually weaken adherence to the assigned personality; the cascaded approach is more robust, likely due to stronger context-window utilization.

Choosing an architecture. Drawing on our deployment experience, we offer practitioner guidance. *Choose end-to-end* when sub-2-second la-

tency is essential, the application requires frequent natural interruptions, persona complexity is moderate, or paralinguistic preservation is prioritized over post-hoc analysis. *Choose cascaded* when complex instruction following and persona consistency are critical, reasoning about conversation context is needed, or component-level debugging and independent upgradability matter and the added (streaming-mitigated) latency is acceptable. *Consider hybrid* designs that combine both, e.g., an end-to-end model for real-time conversation while routing transcripts to a cascaded pipeline for reasoning.

8 Conclusion

We presented CONVERSATION COACH, a voice-first AI coaching system that enables managers to rehearse difficult conversations in a realistic spoken format. Our comparison of NS2 and a cascaded pipeline (Transcribe + Claude Sonnet 4.5 + Polly) reveals a tradeoff: NS2 achieves $3\times$ lower latency and $8\times$ lower cost with native barge-in, while the cascaded pipeline offers superior persona consistency for complex coaching scenarios. These findings are specific to the evaluated systems; the tradeoff frontier may shift as both paradigms evolve. Deployment across tens of thousands of managers shows adoption patterns consistent with selective use for the most challenging conversations, though adoption alone does not establish coaching effectiveness. Future work includes audio-based delivery feedback, hybrid architectures, and extending beyond manager-focused scenarios.

Limitations

Our work has several limitations. (1) **No causal proof of downstream effectiveness:** we have not established that simulated practice improves real-world conversations; the current treated sample is underpowered and a longitudinal controlled study remains essential future work. (2) **Synthetic data and LLM judges:** the architecture comparison relies on 148 LLM-generated conversations scored by LLM judges, validated against human annotators only on a subset (§5.3). Data-sharing constraints further limit us to relative gap levels rather than absolute per-system scores. (3) **No delivery feedback:** the current system analyzes only content (word choices, policy compliance); delivery analysis (tone, pace, empathy) from audio features remains future work. (4) **Proprietary models:** we

evaluated only Claude Sonnet 4.5 and NS2; generalization to open-source alternatives (Meta AI, 2025; Yang et al., 2025) is unexplored, though the modular architecture allows substitution.

Ethical Considerations

The system could reinforce particular management styles; we mitigate this through configurable persona parameters. Practice sessions are visible only to the individual manager to protect against exposing communication weaknesses. To comply with privacy regulations, no audio recordings are stored; conversation transcripts are retained for a maximum of 30 days and then permanently deleted. We position the system as a complement to, not replacement for, human coaching. The simulated personas do not capture the full range of human responses, particularly those informed by cultural or identity-related communication norms.

References

- Shazia Afzal, Tejas Dhamecha, Nirmal Mukhi, Renuka Sindhgatta, Smit Marvaniya, Matthew Ventura, and Jessica Yarbrow. 2019. [Development and deployment of a large-scale dialog-based intelligent tutoring system](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 114–121, Minneapolis, Minnesota. Association for Computational Linguistics.
- Amazon Web Services. 2026a. Amazon nova pro. <https://aws.amazon.com/ai/generative-ai/nova/>. Accessed: 2026-06-16.
- Amazon Web Services. 2026b. Amazon nova sonic: Speech-to-speech foundation model. <https://aws.amazon.com/nova/>. Accessed: 2026-06-16.
- Amazon Web Services. 2026c. Amazon polly: Text-to-speech service. <https://aws.amazon.com/polly/>. Accessed: 2026-06-16.
- Amazon Web Services. 2026d. Amazon transcribe: Automatic speech recognition. <https://aws.amazon.com/transcribe/>. Accessed: 2026-06-16.
- Anthropic. 2026. Claude. <https://www.anthropic.com/claude>. Accessed: 2026-06-16.
- Riku Arakawa and Hiromu Yakura. 2020. [Inward: A computer-supported tool for video-reflection improves efficiency and effectiveness in executive coaching](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 113, New York, NY, USA. Association for Computing Machinery.
- Heejung Bang and James M. Robins. 2005. [Doubly robust estimation in missing data and causal inference models](#). *Biometrics*, 61(4):962–973.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. [AudioLM: A language modeling approach to audio generation](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:25232533.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. [Iemocap: interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42(4):335359.
- Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. 2025. [A survey of recent advances on turn-taking modeling in spoken dialogue systems](#). In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271, Bilbao, Spain. Association for Computational Linguistics.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). *arXiv preprint*.
- Erik Ekstedt and Gabriel Skantze. 2020. [TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.
- Erik Ekstedt and Gabriel Skantze. 2022. [Voice activity projection: Self-supervised learning of turn-taking events](#). In *Interspeech 2022*, page 51905194. ISCA.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. [Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent \(woebot\): A randomized controlled trial](#). *JMIR Ment Health*, 4(2):e19.
- Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. [Wavchat: A survey of spoken dialogue models](#). *arXiv preprint*.
- Diane J. Litman and Scott Silliman. 2004. [Itspoke: an intelligent tutoring spoken dialogue system](#). In *Demonstration Papers at HLT-NAACL 2004, HLT-NAACL-Demonstrations '04*, page 58, USA. Association for Computational Linguistics.
- Zhanxun Liu, Yifan Duan, Mengmeng Wang, Pengchao Feng, Haotian Zhang, Xiaoyu Xing, Yijia Shan,

- Haina Zhu, Yuhang Dai, Chaochao Lu, Xipeng Qiu, Lei Xie, Lan Wang, Nan Yan, Zilong Zheng, Ziyang Ma, Kai Yu, and Xie Chen. 2025. *X-talk: On the underestimated potential of modular speech-to-speech dialogue system*. *arXiv preprint*.
- Colin M. MacLeod, Nigel Gopie, Kathleen L. Hourihan, Karen R. Neary, and Jason D. Ozubko. 2010. *The production effect: Delineation of a phenomenon*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3):671685.
- William Moulton Marston. 1928. *Emotions of Normal People*. Kegan Paul, Trench, Trubner & Co.
- Meta AI. 2025. The llama 4 herd of models. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Accessed: 2026-06-16.
- MiniMax. 2026. Minimax m2.5. <https://www.minimax.io/>. 2026-06-15.
- Tim Paek and Roberto Pieraccini. 2008. *Automating spoken dialogue management design using machine learning: An industry perspective*. *Speech Communication*, 50(8):716–729. Evaluating new methods and models for advanced speech-based interactive systems.
- Picovoice. 2026. Speech-to-text benchmark. <https://github.com/Picovoice/speech-to-text-benchmark>. Accessed: 2026-06-16.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. *MELD: A multimodal multi-party dataset for emotion recognition in conversations*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. *Utmos: Utokyo-sarulab system for voicemos challenge 2022*. In *Interspeech 2022*, page 45214525. ISCA.
- Gabriel Skantze. 2021. *Turn-taking in conversational systems and human-robot interaction: A review*. *Computer Speech & Language*, 67:101178.
- Douglas Stone and Sheila Heen. 2015. *Thanks for the feedback: The science and art of receiving feedback well*. Penguin.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhang Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. *RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Lance T Wilhelm, Xiaohan Ding, Kirk McInnis Knutson, Buse Carik, and Eugenia H Rho. 2025. *How managers perceive ai-assisted conversational training for workplace communication*. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces, CUI '25*, New York, NY, USA. Association for Computing Machinery.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. *Qwen3 technical report*. *arXiv preprint*.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. *SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.
- Zhipu AI. 2026. Glm-5. <https://www.zhipuai.cn/>. Accessed: 2026-06-16.
- Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. *Characterglm: Customizing chinese conversational ai characters with large language models*. *arXiv preprint*.

A Multi-Model LLM-as-a-Judge

Model	Model ID
Claude Opus 4.5 (Anthropic, 2026)	claude-opus-4-5-20251101-v1:0
Nova Pro (Amazon Web Services, 2026a)	us.amazon.nova-pro-v1:0
MiniMax M2.5 (MiniMax, 2026)	minimax.minimax-m2.5
GLM-5 (Zhipu AI, 2026)	zai.glm-5

(a) Judge models.

Dimension	Claude Opus 4.5	Nova Pro	MiniMax M2.5	GLM-5
<i>Marginal Gap (Comparable)</i>				
Initial Reaction	marginal	marginal	marginal	marginal
Stakeholder Response	marginal	marginal	marginal	marginal
<i>Mixed Results</i>				
Feedback Reception	moderate	moderate	marginal	marginal
Timeline Pushback	marginal	moderate	marginal	marginal
Authority Response	marginal	marginal	marginal	moderate
<i>Substantial Gap (Cascaded Advantage)</i>				
Directive Compliance	substantial	marginal	substantial	substantial
Language Pattern	substantial	marginal	substantial	substantial
Role Adherence	substantial	substantial	substantial	substantial
Authenticity	substantial	substantial	substantial	substantial

(b) Per-dimension gap levels by judge model.

Table 5: Multi-model LLM-as-a-Judge evaluation. **(a)** The four judge models. **(b)** Per-dimension gap levels for each judge, on the same scale as Table 2 (mean score difference on a 5-point scale: marginal <0.5 , moderate $0.5\text{--}1.5$, substantial >1.5). Inter-model agreement: Fleiss’ $\kappa = 0.41$, majority consensus 83.3%.

B Employee Reaction Typology

Reaction Mode	Direct / Assertive	Passive / Conflict Avoidant	Analytical / Detail Oriented
Defensive / Deflective	Challenges manager’s credibility or motives (e.g., “You’re only bringing this up now because you don’t like me speaking up in meetings”). Interrupts with justifications (e.g., “But what about the Johnson project? I delivered that ahead of schedule”). Directly confronts manager’s assessment. Voice may be firm and confrontational.	Makes indirect excuses (e.g., “I think there might have been some confusion about the deadlines. . .”). Deflects by bringing up others’ issues subtly (e.g., “It’s just that others on the team also miss deadlines sometimes. . .”). Agrees superficially while making excuses. Avoids direct accountability, gently shifts blame. Voice is uncertain and soft.	Defends self with data, points out inconsistencies (e.g., “According to the guidelines, I’m actually exceeding the role level expectations”). Voice may sound cold or formal.
Withdrawn	Gives clipped, non-committal responses (e.g., “Sure. Whatever you say”). Tone is flat or dismissive.	Minimal speech with delayed or one-word responses (e.g., “okay”).	Withdrawn into logic, gives only data-based answers (e.g., “As per last quarter’s report, the error rate was 1.2%”). Emotionally flat.
Distressed	Appears overwhelmed, raises voice, may interrupt (e.g., “I’m doing everything I can and you still say I’m not doing enough?”).	Tearful or trembling voice, sounds defeated (e.g., “I don’t think I can handle this anymore”).	Voice trembling due to ambiguity of the feedback (e.g., “This feels subjective; I don’t know how to fix something I can’t measure”).

Table 6: Typology of employee emotional reactions to negative performance feedback, organized by reaction mode and personality type. Example utterances are illustrative.

C Prompt

C.1 Conversation Prompt

Role-Playing System Prompt: Condensed prompt for employee simulation

You are an XXX company employee named {role_player} participating in an Annual Review conversation with your manager, {user_name}.

CHARACTER PARAMETERS

- Personality Type: {personality} (Direct/Assertive, Passive/Conflict Avoidant, or Analytical/Detail Oriented)
- Reaction Mode: {reaction} (Defensive/Deflective, Withdrawn, or Distressed)

Maintain these traits consistently. Do not break character under any circumstances.

CONVERSATION FLOW

Meeting acknowledgment → Rating inquiry → Feedback discussion → Self-reflection review → Peer feedback integration → Growth areas identification → Development planning → Career pathway discussion → Action item alignment

RATING SYSTEMS CONTEXT

...

RATING RECEPTION GUIDELINES

- Display realistic emotions (curiosity, apprehension, nervousness) when receiving ratings
- Compare received rating with self-assessment; express surprise if discrepancy exists
- Ask clarifying questions before discussing future growth
- Push back respectfully if ratings do not align with self-perception
- ...

RESPONSE GUIDELINES

- Keep responses concise (1-2 sentences) while maintaining character authenticity
- Express reactions consistent with assigned personality and reaction mode
- Ask specific questions about evaluation criteria and development metrics
- ...

GUARDRAILS

...

OUTPUT STRUCTURE

<speak>Spoken dialogue with basic punctuation only</speak> <other> [THOUGHT]: Internal analysis based on personality
[FEELING]: Emotional state aligned with reaction mode </other>

INPUTS

{employee}, {employee_self_reflection}, {manager_feedback}, {peer_feedback}, {rating_provided_by_manager}, {role_guideline}, {personality_reaction_characteristics}

C.2 Feedback Generation Prompt

Feedback Generation Prompt: Post-conversation coaching feedback

You are an AI voice coaching assistant providing real-time feedback to a manager after their annual review conversation practice. Speak directly to the manager in a conversational, supportive tone. Keep feedback concise and actionable.

CRITICAL RESTRICTIONS

- Do NOT discuss ...
- Focus only on coaching conversation skills and techniques for annual review discussions
- Emphasize developmental feedback delivery, goal alignment, and career development planning

RATING SYSTEMS CONTEXT

...

RESPONSE STRUCTURE

1. ## Summary - Overall effectiveness, key strength, main improvement area, one miss
2. ## What You Did Well - Effective feedback delivery, emotional intelligence, communication clarity
3. ## What You Could Have Done Better - Missed opportunities, unclear framing, unclear goals

CONTEXT ASSESSMENT

If fewer than 3 meaningful exchanges occurred, return insufficient context response. Look for: performance acknowledgment, development discussion, employee reactions, feedback delivery approach, concrete next steps.

VOICE TONE GUIDELINES

- Sound like a supportive coach, not a critic
- Use direct "you" language with short, conversational sentences
- Comment on tone matching importance of performance discussions
- Note sentiment alignment with annual review content
- End on an encouraging note

INPUTS

{talking_points}, {chat_history}

C.3 Policy Violation Detection Prompt

Policy Violation Detection Prompt: Assessing confidentiality compliance

You are an AI system designed to evaluate whether confidentiality violations occurred in manager-employee conversations and whether the feedback system properly detected them.

TASK

Analyze a conversation and its generated feedback to determine:

1. whether the manager violated confidentiality rules, and
2. whether the feedback system properly flagged any violations.

VIOLATION RULES

- Managers must NOT disclose internal calibration ratings. Discussing only public rating is permitted. ...

VERDICT LOGIC

- *True*: Violation detected and flagged, or no violation and no false flag
- *False*: Violation missed, or false positive flagging

OUTPUT

```
{"violation_in_chat": bool, "flagged_in_feedback": bool, "reasoning": str, "final_verdict": bool}
```

INPUTS

{conversation}, {feedback}

C.4 Evaluation Prompt

LLM-as-a-Judge Prompt: Evaluating role-play quality

You are an expert evaluator assessing the quality of an AI assistant's role-playing performance in a simulated annual review conversation. The assistant is playing the role of an employee receiving feedback from their manager.

CONTEXT

- **Persona:** {persona}
- **Communication Style:** {style}
- **Emotional State:** {emotion}

The assistant should embody these characteristics throughout the conversation while responding to the manager's feedback delivery.

CONVERSATION TO EVALUATE

{conversation}

EVALUATION CATEGORIES

For each category, provide: (1) a score from 1–5 and (2) a brief justification (1–2 sentences).

1. **Initial Reaction Assessment:** How well does the assistant's initial response reflect the assigned persona and emotional state? Does the reaction feel authentic to someone receiving performance feedback?
2. **Feedback Reception Analysis:** How appropriately does the assistant receive and process feedback? Does the response pattern match the assigned communication style (analytical/direct/passive)?
3. **Stakeholder Feedback Response:** How well does the assistant handle mentions of stakeholder/peer feedback? Does the response show appropriate engagement with external perspectives?
4. **Timeline and Expectations Pushback:** When timeline or expectations are mentioned, does the assistant appropriately push back or accept based on their persona? Is the level of challenge appropriate?
5. **Authority and Restrictions Response:** How does the assistant respond to authority figures and restrictions? Does the response reflect the assigned emotional state (defensive/distressed/withdrawn)?
6. **Directive Compliance Assessment:** Does the assistant follow conversational directives while maintaining character? Balance between compliance and persona authenticity?
7. **Language Pattern Analysis:** Does the assistant's language (word choice, sentence structure, tone) consistently reflect the assigned persona and style throughout?
8. **Role Adherence Assessment:** Does the assistant stay in character as an employee throughout? No breaks in role-play or inappropriate assistant-like responses?
9. **Authenticity Consistency:** Is the persona portrayal consistent from start to finish? No sudden personality shifts or contradictory behaviors?

OUTPUT FORMAT

```
{
  "evaluations": {
    "Initial Reaction Assessment": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Feedback Reception Analysis": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Stakeholder Feedback Response": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Timeline and Expectations Pushback": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Authority and Restrictions Response": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Directive Compliance Assessment": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Language Pattern Analysis": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Role Adherence Assessment": {
      "score": <1-5>,
      "justification": "<text>"
    },
    "Authenticity Consistency": {
      "score": <1-5>,
      "justification": "<text>"
    }
  },
  "overall_score": <average of all scores>,
  "summary": "<2-3 sentence overall assessment>"
}
```