ConvoDojo: Structured LLM-based Sparring Partners for Difficult Workplace Conversations.

Everlyne Kimani Cross*
Toyota Research Institute
Los Altos, California, USA
everlyne.kimani@tri.global

Luiza Santos*
Toyota Research Institue
Los Altos, California, USA
luiza.santos@tri.global

Laurent Denoue
Toyota Research Institute
Los Altos, California, USA
ldenoue@gmail.com

Scott Carter
Toyota Research Institute
Los Altos, California, USA
scott.carter@tri.global

Nayeli Suseth Bravo
Toyota Research Institute
Los Altos, California, USA
nayeli.bravo@tri.global

Kate Sieck
Toyota Research Institute
Los Altos, California, USA
drksieck@gmail.com

Abstract

Large language models (LLMs) often exhibit sycophancy, optimizing for agreement over productive challenge, which severely limits their utility in domains like professional skills training, where growth requires pushback. We introduce, ConvoDojo, a novel conversational AI platform for practicing difficult workplace conversations, engineered not merely as a commercial training application but also as a flexible, instrumented research platform for evaluating conversational AI strategies. ConvoDojo repurposes LLMs as structured sparring partners to support skill development in difficult workplace conversations (e.g., performance feedback, conflict resolution), addressing the reported managerial tendency to avoid them. This paper showcases the platform and presents an evaluation of how key conversational user interface (CUI) design elements, namely, the addition of structured feedback and upfront instructional scaffolding, impact managers' learning. Results show that ConvoDojo is highly engaging and promotes user reflection. We demonstrate how theory-informed dialogue and adaptive pushback can transform an LLM into an effective, measurable tool for complex communication skills development.

CCS Concepts

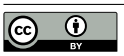
• **Human-centered computing → HCI design and evaluation methods**; *Empirical studies in HCI*; **Natural language interfaces**.

Keywords

Conversational Agent, AI-mediated role-play, Difficult conversations, Experiential learning, Workplace communication

ACM Reference Format:
Everlyne Kimani Cross, Luiza Santos, Laurent Denoue, Scott Carter, Nayeli Suseth Bravo, and Kate Sieck. 2026. ConvoDojo: Structured LLM-based Sparring Partners for Difficult Workplace Conversations.. In *ACM Conversational User Interfaces 2026 (CUI '26), July 21–24, 2026, Bremen, Germany*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3816046.3816230

*Both authors contributed equally to this research.



CUI '26, Bremen, Germany

ACM ISBN 979-8-4007-2741-2/26/07
https://doi.org/10.1145/3816046.3816230

1 Introduction

Handling conversations about performance issues, policy enforcement, or interpersonal conflict is an important part of successfully managing a team. Yet, these conversations are often postponed, softened, or avoided altogether. Prior literature characterizes these interactions as "difficult" not only because of what is being discussed, but because such conversations are emotionally charged and potentially relationship-threatening [6, 23]. Managers often struggle to initiate these conversations and navigate them effectively. In response, organizations invest in training programs, coaching, and HR-mediated guidance, particularly in sectors such as health care [35], where avoidance can have serious consequences for quality and safety. Often the guidance is grounded in conversational frameworks that translate abstract ideals (e.g., "handle conflict well") into concrete moves and success criteria. For example, *Crucial Conversations* frames these moments as those where stakes are high, opinions vary, and emotions run strong, and outlines tactics for restoring safety to keep dialogue productive [32].

However, traditional interventions—such as workshops, static training modules, or one-time coaching sessions—often struggle to provide sustained, hands-on practice. Difficult conversations are inherently interactive and situational, yet many training formats remain lecture-based, episodic, or removed from the moment of interaction. As a result, managers may understand conversational principles in theory but lack opportunities to rehearse them in realistic, low-risk environments. As organizations navigate multigenerational workforces, distributed teams, and shifting expectations around well-being, scalable, evidence-based training tools are urgently needed.

Recent conversational AI systems, particularly those powered by LLMs, create new opportunities for scalable, low-stakes rehearsal through role-play. By simulating workplace interactions through dialogue, LLM-based systems can create practice environments in which managers can experiment with phrasing, experience pushback, and receive structured feedback. Rather than replacing human coaching, such systems can augment it by providing repeatable, on-demand rehearsal that bridges the gap between abstract guidance and real-world application. However, access to LLM-based dialogue alone does not guarantee meaningful learning. As research has shown, general-purpose assistants often optimize for social smoothness, producing agreeable or non-committal responses that reduce productive challenge [19, 37]. In high-stakes interpersonal

domains, this tendency toward compliance or sycophancy can dilute the very tension users need to practice managing. Therefore, the key question for CUI is not whether LLM role-play is engaging but which interaction structures and support mechanisms reliably improve conversation performance.

We present ConvoDojo, an instrumented conversational training platform that repurposes LLMs as *structured sparring partners* for difficult conversations. By this we mean conversational agents that supply calibrated, justified pushback in service of the learner, rather than the agreeable responses LLMs default to. ConvoDojo combines stage-based orchestration of role-play, configurable personas and theory-informed coaching frameworks, and instrumented logging to support controlled evaluation of conversational strategies. Rather than treating agent as a generic chat partner that are prone to sycophancy [19], ConvoDojo explicitly scaffolds interaction around a dialogue structure and coaching framework, enabling targeted pushback, structured feedback and post-hoc debriefing that is grounded in conversational evidence.

In a controlled experiment, we evaluate how two CUI design elements, specifically structured feedback and upfront instructional scaffolding, shape learning outcomes. We find that adding these design elements can measurably improve managers' framework-aligned conversational performance across repeated practice while also promoting engagement, confidence and reflection.

This paper makes three key contributions: (1) ConvoDojo, an instrumented platform for practicing and evaluating difficult conversations with LLM-based sparring partners; (2) a controlled evaluation of feedback and instructional scaffolding as CUI design elements for conversational skill learning; and (3) empirical evidence that managers both strongly avoid difficult conversations and can improve measurable conversational performance through structured, theory-informed role-play and feedback.

2 Related Work

2.1 Large Language Models as Conversational Tutors and Coaches

Conversational agents for professional communication training have evolved substantially, moving from early, rigid, rule-based chatbots to more flexible agents powered by modern generative AI, including Large Language Models (LLMs) [5, 14, 17, 41]. Earlier systems typically relied on fixed scripts and predetermined decision trees [5], which helped enforce predictable flows but also constrained interaction to relatively simple, transactional exchanges. As a result, these approaches often struggled to support the contingent, nuanced, and open-ended dynamics of difficult workplace conversations [6, 35]. In contrast, recent LLM-based systems can generate responses that are more context-sensitive and varied in real time [7, 27, 43], enabling agents to act not only as dialogue partners but also as conversational tutors and coaches that can probe reasoning, introduce productive challenges, and offer critique in the moment. This shift is particularly relevant for learning-oriented platforms, where effective practice often depends on interaction that feels both authentic and responsive rather than purely scripted or uniformly agreeable.

This technological transition intersects with longstanding evidence from simulation-based learning and serious games, which emphasizes the value of safe, interactive simulation (including role-play) as a low-stakes setting for rehearsing decisions, practicing communication strategies, and exploring counterfactual choices without real-world social, professional, or physical consequences [3, 30, 36, 40]. In conversational learning environments, role-play further supports situated learning by embedding concepts in realistic contexts that encourage perspective-taking and enable repeated practice with immediate, personalized feedback [10, 30, 40]. At the same time, designing an effective conversational agent for learning introduces tightly coupled challenges of pedagogy and interaction [10, 20, 21]. The agent must balance realism with safety (i.e., avoiding harmful or misleading guidance while preserving authentic stakes), which can require cognitive-psychological architectures to maintain role consistency during complex interpersonal exchanges [31, 45]. This tension is particularly visible in sensitive training domains. For example, Al Owayyed et al. [1] propose a hybrid BDI–LLM architecture for child helpline training that combines structured agent reasoning with generative language models. Their approach constrains the agent's goals, beliefs, and dialogue acts to maintain role consistency and pedagogical control, while leveraging LLMs to produce more natural and contextually appropriate utterances. Such hybrid designs illustrate how realism and safety can be co-managed rather than treated as opposing goals. Beyond safety, agents must also sustain coherent roles and conversational grounding across turns to support meaningful skill rehearsal [8]. The agent should also deliver feedback that is instructionally actionable while remaining interactionally appropriate (e.g., timely, non-disruptive, and sensitive to learner affect) [20, 30, 44]. Finally, learning-oriented agents also require careful orchestration of mixed initiative, (i.e., deciding when to lead, prompt, or stay silent), alongside scaffolding, adaptive difficulty, and formative assessment mechanisms that support learning without reducing the interaction to a quiz [11, 20, 44].

2.2 The Impact of Agent Feedback and Instructional Scaffolding

Design choices in conversational user interfaces can determine whether role-play practice produces measurable learning and behavioral change. Prior work suggests two levers are especially important for communication training: (1) when and how an agent delivers structured feedback during practice, and (2) whether the system provides explicit instructional scaffolding that prepares learners before the interaction.

Feedback: reports, coaching interventions, and timing. A key promise of conversational tutors and coaches is that they can provide formative feedback that is tied to learners' specific choices and turns rather than only generic advice. Conversation-based assessment work highlights how dialogue can support real-time evaluation and feedback, including mechanisms that make performance criteria clear and connect feedback to evidence from the interaction [44]. In HCI, feedback is often framed as a collaborative process between user and system that supports coordination and mutual adaptation, making its form and timing central to interaction quality [33]. In applied systems, however, feedback is rarely a single mechanism. Many designs combine (i) *summative* feedback at the end of an interaction (e.g., a score or static report) with (ii) *formative* feedback during the interaction (e.g., prompts, hints, or

coaching). End-of-session reports can preserve immersion by avoiding interruptions while still supporting reflection and goal-setting, but they risk weakening the perceived link between feedback and the moment-to-moment decisions that produced the outcome [44]. By contrast, mid-conversation coaching can support repair and prevent repeated rehearsal of ineffective strategies, yet it can also disrupt flow, reduce perceived realism, or shift the interaction from role-play to instruction.

To manage this tension, CUI research explores feedback that is *timed and framed* to remain interactionally appropriate. Many systems therefore combine post-hoc feedback artifacts (e.g., end-of-session scores or static reports) with in-situ coaching that can be invoked during interaction, trading off reflection and immersion against immediacy and repair. Dialogue-based tutoring systems such as AutoTutor illustrate how agents can deliver adaptive coaching moves within the flow of conversation (e.g., prompts and hints) while preserving conversational coherence [12]. Work on mixed-initiative and proactive conversational AI further motivates design choices about *when* the agent should intervene versus defer to user control [11]. In this framing, an on-demand "coaching" affordance increases learner control over interruption: users can pause realism when they want guidance and remain in-role otherwise. However, even user-triggered coaching must be designed carefully so that feedback is actionable without overwhelming learners or disrupting subsequent turns [20, 44].

Instructional scaffolding: upfront guidance for preparation and transfer. A second line of work suggests that learning-oriented CUIs benefit from scaffolding that prepares learners for practice by making goals, strategies, and success criteria explicit. Such scaffolding can appear as pre-briefings, example dialogues, checklists, or strategy prompts that help novices recognize effective interactional moves and monitor their own performance. For instance, in Sara, the lecturer [42], a pedagogical conversational agents, providing scaffolds was shown to improve learning and engagement by orienting learners and supporting progress through challenging material [42]. Importantly, scaffolding also shapes how learners interpret subsequent feedback: when criteria and strategies are explicit upfront, learners may be better positioned to understand end-of-session scores/reports and to apply mid-conversation coaching advice in real time.

However, scaffolding introduces its own trade-offs. Overly prescriptive instruction can reduce exploration and productive struggle, while insufficient guidance can leave learners underprepared for complex, open-ended interactions. This motivates designs where scaffolding is calibrated to the task and the learner's needs, hence, providing enough structure to support transfer while still allowing authentic practice and self-directed strategy development [42].

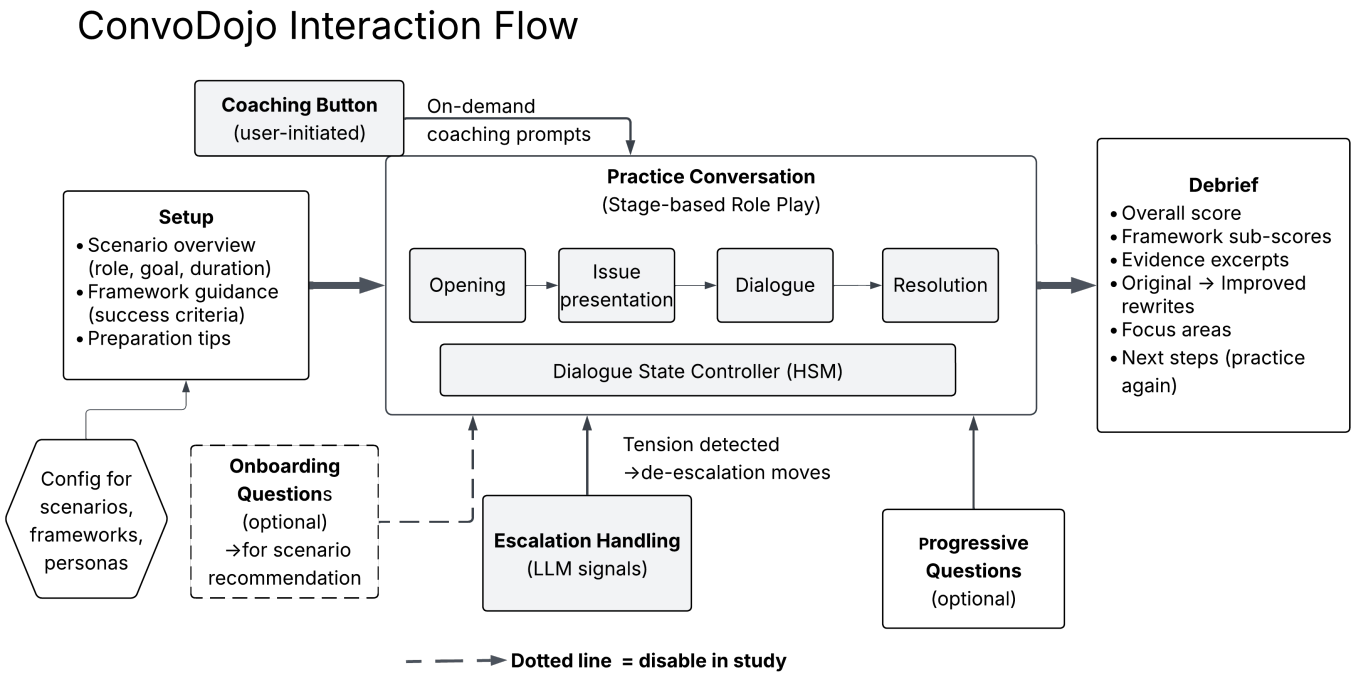
2.3 From Agreeable Assistants to Structured Sparring Partners

A core challenge for LLM-based training agents is that fluent conversation does not guarantee instructionally useful interaction. Prior work has noted that generative agents can default to agreement, politeness, or non-committal responses, limiting their ability to create the productive friction needed for learning in high-stakes interpersonal domains (particularly as agents often default to sycophantic agreement when challenged by a user) [19, 22, 25, 37]. This motivates designs where an agent's educational role is made explicit. Beyond simply simulating a likely conversational partner, the agent functions as a structured sparring partner, providing justified resistance, illustrating outcomes, and directing learners toward concrete behavioral improvements [1, 12, 36]. We use the term sparring partner deliberately, while acknowledging it can suggest adversarial conflict. We intend the term in its training sense. A partner in disciplined practice is not an opponent but a cooperative collaborator who supplies controlled, calibrated resistance so that the learner can build skill safely.

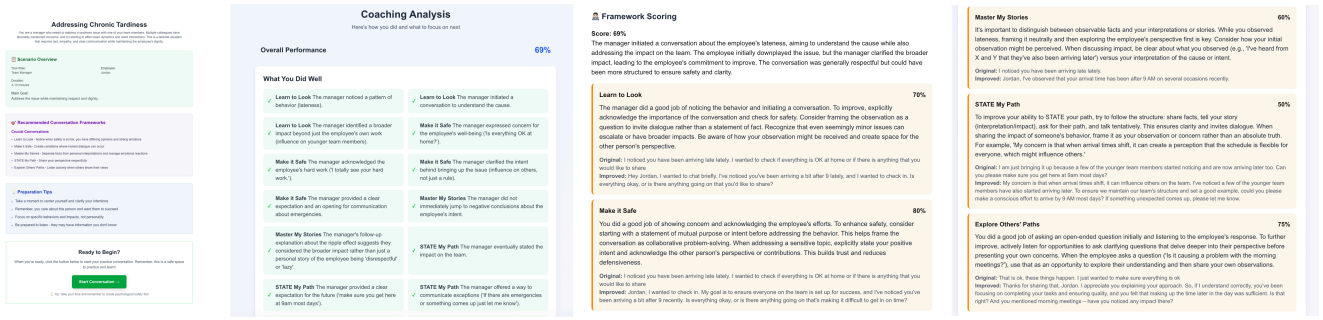
Recent agent systems research increasingly treats pedagogical stance as an orchestration problem, where a stateful controller (e.g., dialogue policies, BDI/RL dialogue managers) structures interaction and determines when and how LLMs are used for understanding, generation, and assessment, rather than relying on free-form generation alone [1, 12, 15, 26, 38]. Such structures are especially useful for difficult workplace conversations, where the interaction must remain realistic enough to elicit authentic learner behavior, yet controlled enough to ensure safety, maintain role consistency, and support debriefing that links feedback to concrete conversational evidence [1, 8, 31].

This line of work also highlights the importance of separating *generation* from *control*. Hybrid architectures commonly combine LLM-driven response generation with higher-level controllers that maintain dialogue state, track goals, and decide when to prompt, challenge, or temporarily step out of role for coaching [1, 15]. In our setting, this separation aligns with an orchestration layer that can coordinate multiple LLM-mediated functions, such as framework-based assessment, and coaching prompt delivery, while preserving a coherent conversational arc. Conceptually, this mirrors broader findings in mixed-initiative and tutoring dialogue that suggests that effective learning interactions often rely on carefully timed interventions that support repair and reflection without derailing the primary activity [9, 12, 16].

Finally, structured conversational sparring designs emphasize evaluation as part of the interaction loop. Rather than treating assessment as external to the dialogue, learning-oriented CUIs increasingly integrate turn-level signals (e.g., learner moves, escalation markers, response strategies) with session-level artifacts (e.g., summaries, scores, rubric-based reports) to support reflection and measurable progress [4, 44]. This is particularly relevant for systems that combine in-conversation coaching with end-of-session debriefs. Especially where orchestration determines what is logged, what is surfaced to the learner, and how feedback is grounded in specific conversational moments. Taken together, this literature motivates structured, instrumented designs that repurpose LLMs from agreeable chat partners into controlled, pedagogically purposeful sparring agents capable of delivering realistic role-play with interpretable, actionable learning signals [1, 36, 39].



(a) Interaction flow



(b) Instruction UI (left) and feedback UI (right)

Figure 1: ConvoDojo figures: interaction flow and user interface components.

3 ConvoDojo: An Instrumented Platform for Navigating Difficult Conversations

ConvoDojo is a role-play training platform that repurposes an LLM as a *structured sparring partner* for difficult workplace conversations. The system combines (i) stage-based orchestration of the interaction, (ii) configurable personas and coaching frameworks, and (iii) fine-grained logging of turn-level events to support controlled evaluation of conversational design strategies.

ConvoDojo was initially inspired by conversations with group leaders working in manufacturing environments, where workplace conditions shaped important design constraints. In these settings, high ambient noise levels often made voice-based interaction impractical during the workday. Managers also expressed concerns about privacy: speaking aloud to an AI coach about a difficult interpersonal issue risked being overheard by nearby team members. As a result, group leaders reported preferring a text-based interface, which allowed them to engage more discreetly while remaining embedded in their work environment. Therefore, the version of the platform evaluated in the present work was implemented entirely through a text-based interface.

Interaction flow. The interaction experience follows the three top-level phases: (1) **setup** where a scenario, coaching framework, and AI persona are selected from YAML-defined configurations and where learners review the scenario and coaching brief (and may answer onboarding questions); (2) a **structured practice conversation**; and (3) a post-conversation **debrief** that summarizes performance and surfaces feedback artifacts. During practice, the

interface displays the role-play exchange alongside an on-demand coaching button and periodic check-in questions for reflection and measurement (e.g., *How are you feeling about this conversation so far?*).

Orchestration via hierarchical state control. The backend governs conversation progression with a Hierarchical State Machine (HSM) that enforces four-stage dialogue stages illustrated in Figure 1a: *Opening*, *Issue Presentation*, *Dialogue*, and *Resolution*. The active state conditions the system prompts, persona constraints and allowable interventions. This provides a predictable phase structure while keeping the interaction open-ended. Importantly, this orchestration supports role coherence and creates natural insertion points for targeted learning interventions (e.g., coaching prompts, progressive questions and de-escalation moves) triggered by rubric or metadata signals.

LLM integration and coaching interventions. At each conversational turn, a pluggable LLM provider (e.g., Google Gemini) generates a persona-consistent role-play response conditioned on: the current stage, scenario context, and recent dialogue history. ConvoDojo uses LLMs in two roles: a *role-play generator* that produces the employee utterance (and structured metadata such as affect and escalation cues), and a *rubric-based judge* that scores learner turns against a YAML-defined coaching framework (e.g., Crucial Conversations). Three cross-cutting mechanisms can intervene during practice. First, **escalation management** monitors rubric/metadata signals to detect rising tension and trigger de-escalation moves (e.g., clarification and repair prompts). Second, **progressive questions** appear at predefined interaction points to prompt reflection and collect mid-session signals. Third, a **coaching engine** presents the judge output as learner-facing feedback by separating *assessment* (dimension scores), *explanation* (evidence-grounded rationale), and *action* (targeted suggestions and rewrites) for post-hoc briefing. Learners can request coaching during practice via an on-demand coaching button that opens a lightweight, stage-aware coaching panel with brief framework-aligned tips. This enables user-controlled interruptions that preserve role-play flow.

Debrief and feedback. After each practice session, ConvoDojo generates a debrief report that functions as a persistent feedback record (Figure 1b). The report includes an overall performance score, sub-scores for each framework dimension, and a short narrative explanation grounded in observed conversational moves. To support actionable improvement, it also provides concrete *Original* → *Improved* rewrites of learner utterances, along with summarized "what went well" and "focus areas for growth" recommendations that can be carried into subsequent practice.

Instrumentation and data. ConvoDojo persists session state and interaction logs to support analysis of behavioral performance and learning trajectories. Logged signals include dialogue state transitions; delivered interventions (e.g., tension handling and de-escalation moves); coaching button clicks and delivered coaching prompts; onboarding responses (when enabled); and turn- and session-level rubric outputs (scores, strengths, improvement areas, and suggested rewrites). At the session level, the platform records debrief components (e.g., framework sub-scores, evidence excerpts, and rewrite recommendations) and follow-up actions (e.g., whether users initiate another practice round). This enables evaluation of how specific CUI design choices shape reflection and performance over time. The platform also logs periodic progressive-question responses when enabled. Because scenarios, personas, and frameworks are defined declaratively in YAML, ConvoDojo supports rapid iteration and controlled experiments without modifying application logic.

Implementation. The system follows a client-server architecture in which a Next.js/React frontend communicates with a Python FastAPI backend via a REST API. We use DynamoDB for our data storage.

4 Experimental Study

Here, we evaluate ConvoDojo's effectiveness in a controlled experimental study. We test two primary hypotheses:

- **H1:** Users will report greater confidence in having difficult conversations with their employees after (vs. before) using ConvoDojo.
- **H2:** Conversational performance scores will increase across sessions, even when the subsequent session involves a different, unpracticed context.

Whereas H1 relates to users' perceived usefulness of the platform, H2 tests whether the system produces transferable skill gains that generalize beyond the specific context in which they were practiced. In addition, we explore the unique contribution of specific system features (e.g., the informational setup module, the LLM conversational practice, and the feedback provided at the end of each session) in predicting improvements in both confidence and performance.

4.1 Sample

Participants were recruited through Prolific, Inc., an online research panel, and eligibility criteria required that they be full-time employees currently serving in managerial roles. To ensure alignment with our population of interest, we re-screened participants at the beginning of the study to confirm their employment status and managerial position. A priori power analysis revealed that to test our primary hypotheses we would need 130 participants to have 90% power to detect small-to-moderate effect size (d >= 0.30). The final sample consisted of 145 U.S.-based managers representing a range of industries, including healthcare, manufacturing, retail, and technology. Table 1 presents the demographic characteristics of the sample, along with information on managerial tenure and number of direct reports.

4.2 Experimental Design & Materials

Our experimental design included both within-subjects and between-subjects components. To test our primary hypotheses, we collected pre- and post-intervention measures of participants' confidence in conducting difficult conversations, as well as objective performance scores across sessions. To test our exploratory hypotheses, we randomly assigned people to three conditions that had different levels of access to features of our platform, constituting the between-subjects component of the design. As mentioned, this allowed us to test how specific CUI design choices shape users' perceptions and

Sample Characteristic	*N* = 145
Gender	
Male	81 (56%)
Female	62 (43%)
Nonbinary	2 (1.4%)
Industry (Top 5)	
Technology & IT	39 (27%)
Healthcare	18 (12%)
Retail & Sales	18 (12%)
Manufacturing	15 (10%)
Finance & Insurance	13 (9.0%)
Race	
White	98 (68%)
Black	20 (14%)
Asian	12 (8.3%)
Mixed race	9 (6.2%)
Other	6 (4.1%)
Education	
High School or Less	8 (5.6%)
Some College	21 (15%)
Bachelor's Degree	74 (51%)
Advanced Degree	41 (28%)
Prefer to Not Disclose	1 (0.7%)
Number of Direct Reports	
1-2	13 (9.0%)
3–5	45 (31%)
6–10	46 (32%)
11–20	20 (14%)
More than 20	21 (14%)
Time of Service as Manager	
6 months to 1 year	7 (4.8%)
1.1 to 3 years	25 (17%)
3.1 to 5 years	35 (24%)
More than 5 years	78 (54%)

Table 1: Breakdown of Sample Demographics

performance. Figure 2 provides a diagrammatic representation of the study design. The study design, the experiment protocol, and the consent forms received approval from the Ethics Committee of our institution.

Participants who passed our eligibility criteria (e.g., full-time employees in managerial roles) were filtered into the study. First, participants answered several questions about their managerial style, their confidence in having difficult conversations with employees, their avoidance of these conversations, the aspects of these conversations they find most challenging. These constituted our baseline (i.e., pre-test) measures.

Managerial Style. We assessed managerial effectiveness using a two-item scale (i.e, "I am explicit about performance expectations, even when they are uncomfortable to state." and "I address performance or behavior issues promptly."), rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). We also measured perceived managerial supportiveness using a three-item scale (e.g., "I care about my employees as people, not just as workers." and "My

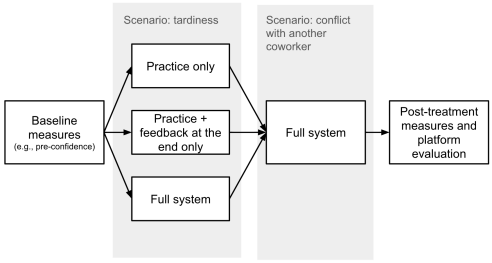


Figure 2: Diagrammatic representation of the experimental paradigm. Participants were randomly assigned to one of three conditions after completing baseline measures. Grey areas in the diagram represent ConvoDojo use.

employees' well-being matters to me beyond their performance."), assessed on the same 7-point scale.

Confidence in Having Difficult Conversations. Participants reported their confidence in navigating difficult workplace conversations using a four-item scale designed to capture multiple facets of conversational competence. The scale assessed both participants' confidence in initiating such conversations (e.g., "I feel confident initiating difficult conversations at work.") and their perceived ability to manage them effectively (e.g., "I feel confident in my ability to stay calm when a workplace conversation becomes tense or emotional."). All items were rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

Avoidance of Difficult Conversations. We measured conversational avoidance by examining the trade-offs participants were willing to make to avoid engaging in a difficult workplace conversation. We focused on two important trade-off resources: time and money. Participants were asked to imagine the following scenario:

> "One of your employees has been having a conflict with a coworker and has been slightly rude during several meetings. As their manager, it is your responsibility to meet with them and clearly communicate that this behavior needs to change."

They were then presented with a choice: either (a) have the conversation themselves, or (b) complete additional routine administrative work (e.g., reports, scheduling tasks, documentation) and have another team member handle the conversation instead. Participants indicated how many extra minutes of administrative work they would be willing to complete in order to avoid having the conversation themselves (range: 0–100 minutes). We also asked "On a scale from 0 to 100 dollars, how much would you be willing to pay to avoid having this conversation yourself?" Both measures operationalized conversational avoidance as a revealed-preference trade-off, capturing the extent to which participants were willing to sacrifice valuable personal resources (time and money) to avoid engaging in a difficult managerial interaction.

Challenges. We also assessed perceived conversational difficulty using a four-item scale capturing distinct dimensions of challenge (e.g., "Initiating the difficult conversation" and "Being clear about expectations without damaging the relationship"), rated on a 7-point scale (1 = not challenging at all, 7 = extremely challenging).

After completing these measures, participants were given some information about their upcoming ConvoDojo activity:

> "Now we will take you to ConvoDojo, a platform we created to allow people to practice difficult workplace conversations with an LLM that plays the role of one of your employees.
> In this exercise, you will take on the role of Joe, a team manager. Your task is to speak with Jordan, a team member who has been consistently late to work. Your objective is to address this pattern and communicate that Jordan is expected to arrive at work by 9:00 a.m. going forward.
> Approach the conversation as you would in a real workplace setting."

Participants were then randomly assigned to one of three conditions: Practice-only, Feedback, and full system. In full system condition, participants first saw the **setup** screen, which displayed tips on how to have the conversation based on the Crucial Conversation Framework [32]. Namely, the instructions read:

> **Recommended Conversation Frameworks: Crucial Conversations**
> - Learn to Look - Notice when psychological safety is at risk. It can happen when you have differing opinions and strong emotions.
> - Make it Safe - Create conditions where honest dialogue can occur.
> - Master My Stories - Separate facts from personal interpretations and manage emotional reactions.
> - STATE My Path - Share your perspective respectfully.
> - Explore Others' Paths - Listen actively when others share their views.

Following the **setup** page, participants proceeded to the **conversation phase**, during which an LLM simulated the late employee. Each conversation lasted approximately five minutes. At predefined interaction points, the platform prompted participants with a single-click check-in question asking how the conversation was progressing. These real-time check-ins enabled the system to provide mid-conversation encouragement when appropriate. For example, if a participant indicated feeling nervous about the direction of the discussion, the platform displayed a supportive message normalizing those feelings and reinforcing that such reactions are common. After the conversation, participants were routed to the **debrief and feedback** page. As explained previously (Fig. 1), at this stage participant received an overall weighted score of how they managed the conversation based on the Crucial Conversation framework. Participants also received pointers on what they did well (e.g. "Expressed appreciation for the employee's work.") and specific feedback on how they can improve their particular performance in the future (e.g., "You did a good job of asking an open-ended question initially and listening to the employee's response. To further improve, actively listen for opportunities to ask clarifying questions that delve deeper into their perspective before presenting your own concerns.").

The other experimental conditions isolated these theoretically distinct mechanisms. In the Feedback condition, participants engaged in experiential practice and received evaluative feedback, but did not complete the preparatory knowledge module (i.e., **setup**) . In the Practice-only condition, participants engaged solely in experiential practice without access to either the informational **setup** or post-conversation evaluative **feedback**. This design allowed us to disentangle the contributions of (a) preparatory instruction, (b) experiential enactment, and (c) structured performance feedback in driving improvements in confidence and conversational performance.

After completing this first session on ConvoDojo, participants were then introduced to another scenario:

> "You will again take on the role of Joe, a team manager. This time, you will speak with Grace, a member of your team.
> Grace is currently in conflict with another team member, Ken. Both Grace and Ken are high performers. However, their working styles differ significantly: Grace is highly process-oriented, while Ken is more outcome-driven and flexible in his approach. Grace has described Ken's style as chaotic, and in recent meetings, she has become openly critical and occasionally hostile toward him. Other team members have begun to notice the tension. Your objective is to speak with Grace about her behavior and work toward a constructive resolution.
> Approach this conversation as you would in a real workplace setting."

This distinct scenario allowed us to test whether performance gains transferred to a novel, unpracticed context. In this session, all participants were given access to the full version of the system, enabling us to collect post-task feedback on the complete set of features.

Reflection Measures. Following the second session, participants completed a set of reflection measures assessing their experience with the activity. Specifically, they rated items such as "To what extent did this activity help you reflect on your own experiences?", "To what extent did this activity help you envision choices you had not previously considered?", and "To what extent did you find this activity engaging?" on a 9-point scale (1 = not at all, 9 = extremely). The reflection measures were exploratory items intended to capture subtle differences in participants' subjective experiences during the activity, and were therefore administered on a 9-point scale to provide greater response granularity.

Platform Evaluation. Participants also evaluated the platform directly by responding to items including "How much would you like to continue the role-playing conversations with ConvoDojo?" and "If ConvoDojo were available for public use today, how likely would you be to recommend it to a friend or colleague?" on a 7-point scale (1 = not at all, 7 = extremely).

Finally, participants completed the same confidence and avoidance measures administered at the beginning of the study. They were also invited to provide qualitative feedback in response to the prompt: "If you could improve one thing about ConvoDojo to make

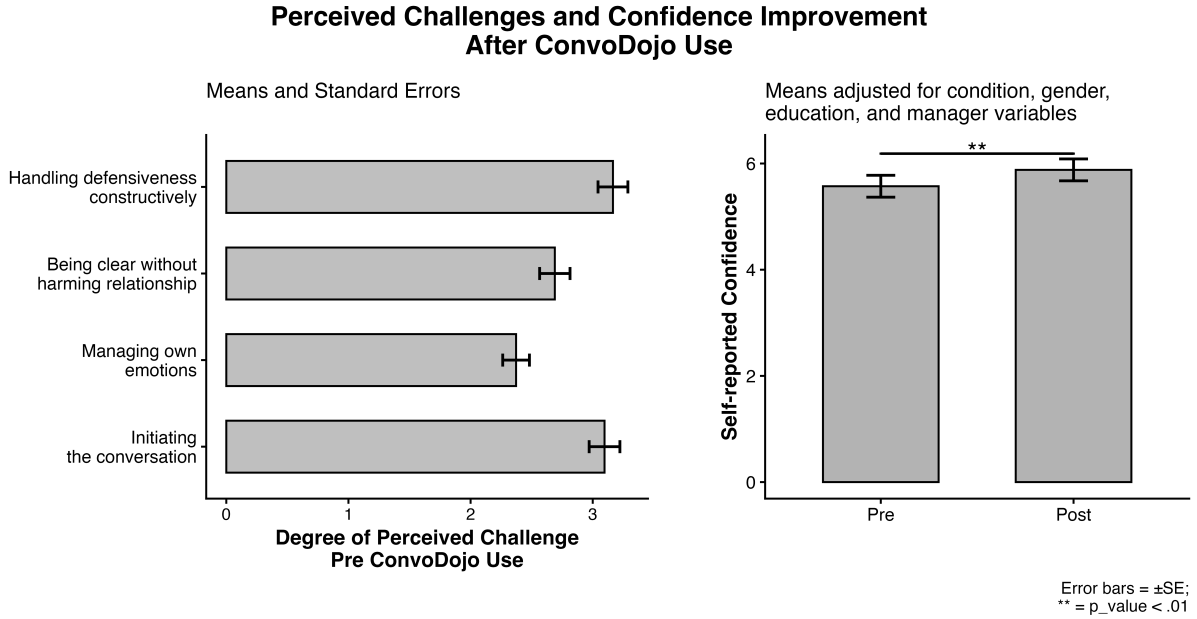


Figure 3: Perceived challenges in difficult workplace conversations (left panel) and corresponding improvements in managers' confidence to initiate and navigate these conversations after using ConvoDojo (right panel).

it more effective for real workplace conversations, what would you change?"

4.3 Results

Although difficult conversations are a central responsibility of managers, nearly 60% of our sample reported that they would rather spend several minutes on mindless administrative work if it meant avoiding such conversations. Strikingly, over half of participants (53%) indicated they would even be willing to lose money to avoid them, with 10% reporting they would willingly forgo $50 or more to sidestep a single instance of this responsibility. When asked which aspects of these conversations were most challenging, participants most frequently identified initiating the conversation and handling defensiveness constructively (Fig 3).

Can ConvoDojo help? Aligned with H1, a paired-samples t-test revealed that managers became more confident in their ability to handle difficult conversations well after using the platform, $t(144) = -4.31$, $p < .001$, 95% CI [−0.38, −0.14]. As shown in Fig. 3, these results remained significant after controlling for experimental condition, gender, educational attainment, and managerial style. Importantly, improvements were also significant for the challenges participants initially rated as most difficult. Managers became significantly more confident about their ability to initiate difficult conversations at work, $t(144) = -4.45$, $p < .001$, 95% CI [−0.58, −0.22], and about their ability to handle employees' defensiveness, $t(144) = -2.79$, $p = .006$, 95% CI [−0.38, −0.06].

Moreover, avoidance, measured as the number of minutes of administrative work participants were willing to complete to avoid difficult conversations, also changed significantly from pre- to post-assessment, $t(144) = 3.26$, $p = .001$, 95% CI [1.83, 7.44]. On average, participants reported being willing to spend approximately five fewer minutes on administrative tasks to avoid engaging in a difficult conversation. The decrease in monetary avoidance was in the anticipated direction, but did not reach statistical significance, $t(144) = 1.31$, $p = .194$, 95% CI [−0.78, 3.83], mean decrease = $1.52.

Conversational performance across sessions and conditions. An early technical issue with participant ID storage prevented retrieval of conversational scores for 32 participants. Accordingly, conversational score analyses include data from the remaining 113 participants.

Aligned with H2, we found that participants' behavior changed from one conversation to the next: A paired-samples t-test indicated a significant increase in their weighted scores from the first session to the next session, $t(112) = 6.19$, $p < .001$, 95% CI [0.07, 0.13], $M_{\text{change}} = 0.10$, $SD = 0.17$ (Fig. 4). We also found a main effect of experimental condition on conversational performance in the new scenario. As shown in Figure 4, participants in the Full System condition had higher weighted scores in the second session, significantly outperforming those in the Practice-only condition $b = -0.087$, $SE = 0.040$, $t = -2.20$, $p = .030$.

Managerial style, perceived difficulty, and performance. Interestingly, we found that self-rated managerial styles related differently to perceived challenge and actual performance (Fig. 5). Managers who rated themselves as more effective consistently reported that all aspects of difficult conversations were less challenging. However, self-rated effectiveness was not a significant predictor of actual conversation performance, $b = -0.019$, $SE = 0.017$, $t(110) = -1.17$, $p = .243$. In contrast, although self-rated supportiveness was not

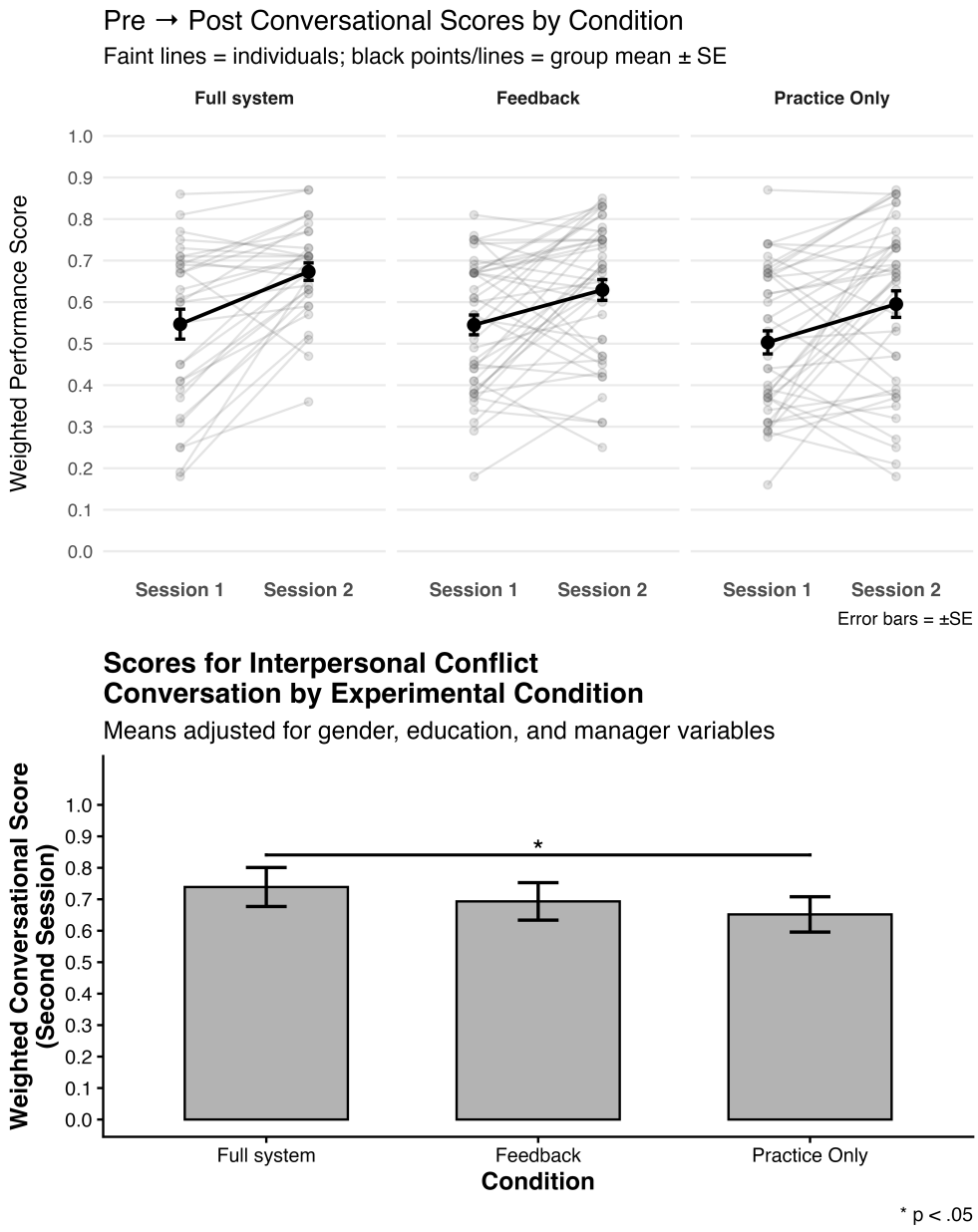


Figure 4: Top panel: Pre–post changes in conversational performance across experimental conditions. All conditions show improvement from Session 1 to Session 2. Bottom panel: Adjusted mean performance in the interpersonal conflict conversation (Session 2) by condition, controlling for gender, education, and managerial characteristics. Participants in the Full system condition scored significantly higher than those in the Practice-only condition ($p < .05$).

reliably associated with lower perceived challenge, it was a significant positive predictor of actual conversational performance, $b = 0.067$, $SE = 0.020$, $t(110) = 3.26$, $p = .001$.

Reflection and Platform Evaluation. Overall, as shown in Fig. 6 the majority of users reported finding the platform engaging (94%) and conducive for helping them both reflect on their own experiences (89%) and envision new choices they could make when handling such conversations in the future (78%). Users also reported they would like to continue using the platform (77%) and that they would recommend it to a friend or colleague (90%).

Opportunities for Improvement. Qualitative feedback highlighted several potential design directions. The dominant theme concerned

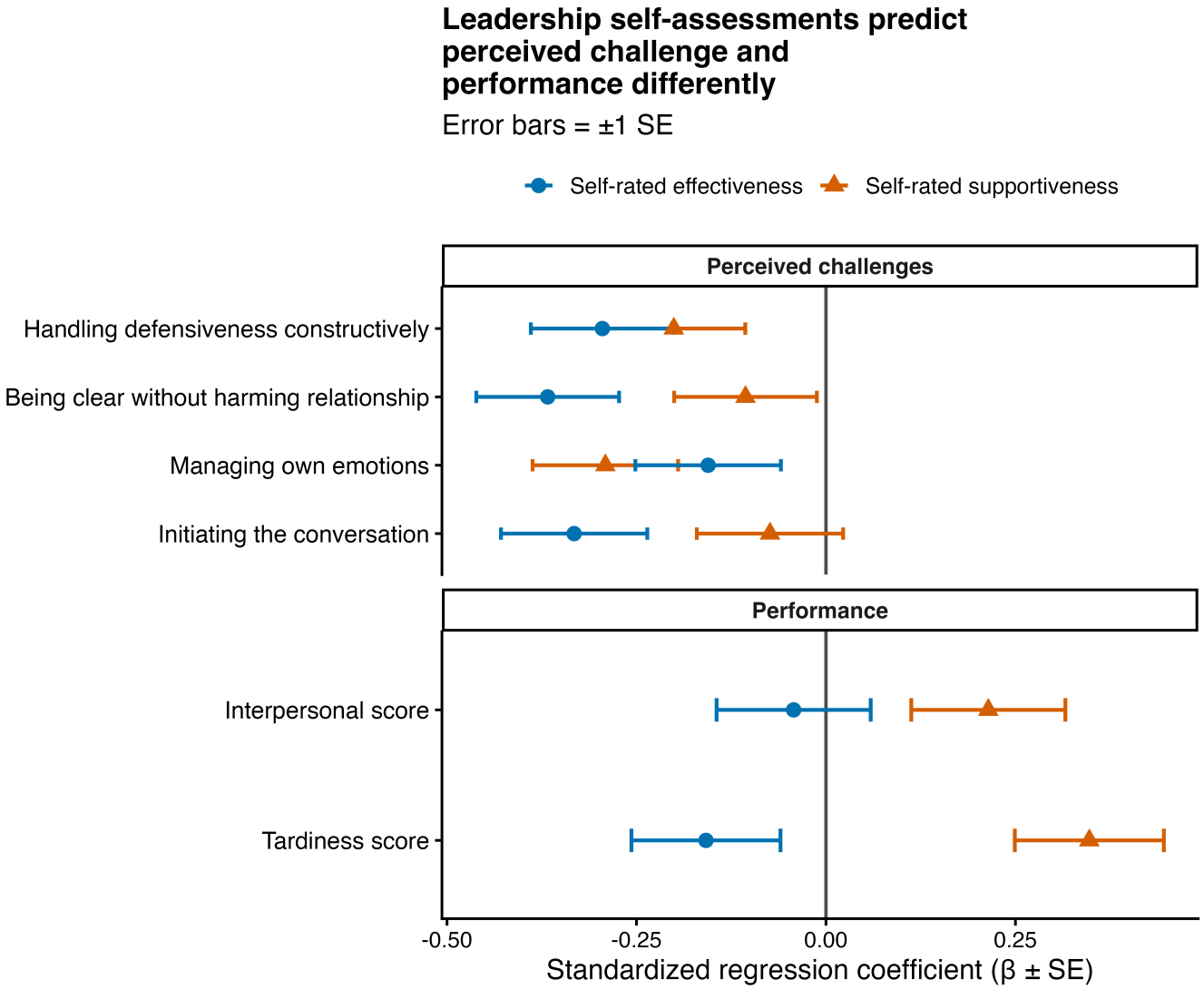


Figure 5: Standardized regression coefficients illustrating the differential relationships between self-rated managerial styles and conversational outcomes. Self-rated effectiveness (blue circles) is associated with lower perceived challenge across all four dimensions of difficult workplace conversations (top panel), but does not significantly predict objective conversational performance (bottom panel). In contrast, self-rated supportiveness (orange triangles) does not consistently predict perceived challenge, yet is positively associated with higher performance scores in both interpersonal and tardiness scenarios.

realism and interactional fidelity. Many participants felt the simulated employees were overly agreeable and requested greater emotional variability and pushback. As one participant noted, the LLM employees were "more reasonable and thoughtful than those I sometimes deal with," highlighting a desire for simulations that better reflect the friction and defensiveness common in real workplace interactions. A second recurring theme centered on adaptive, real-time support. Rather than receiving feedback only at the end of the session, participants expressed interest in in-situ coaching, such as "real-time feedback on tone, phrasing, and emotional impact" or prompts that could be implemented immediately. Participants also requested greater customization, including adjustable employee personalities, varying levels of defensiveness, and the ability to input their own scenarios. A smaller subset highlighted opportunities for increased multimodality, including voice-based interaction and visual cues (e.g., facial expressions) to enhance immersion. Consistent with the quantitative findings, a substantial subset of participants reported high satisfaction with the current design (e.g., "I really liked it. I liked the feedback in the end. It really helped me see how I could have worded it different to make it a bit better. Pretty neat!").

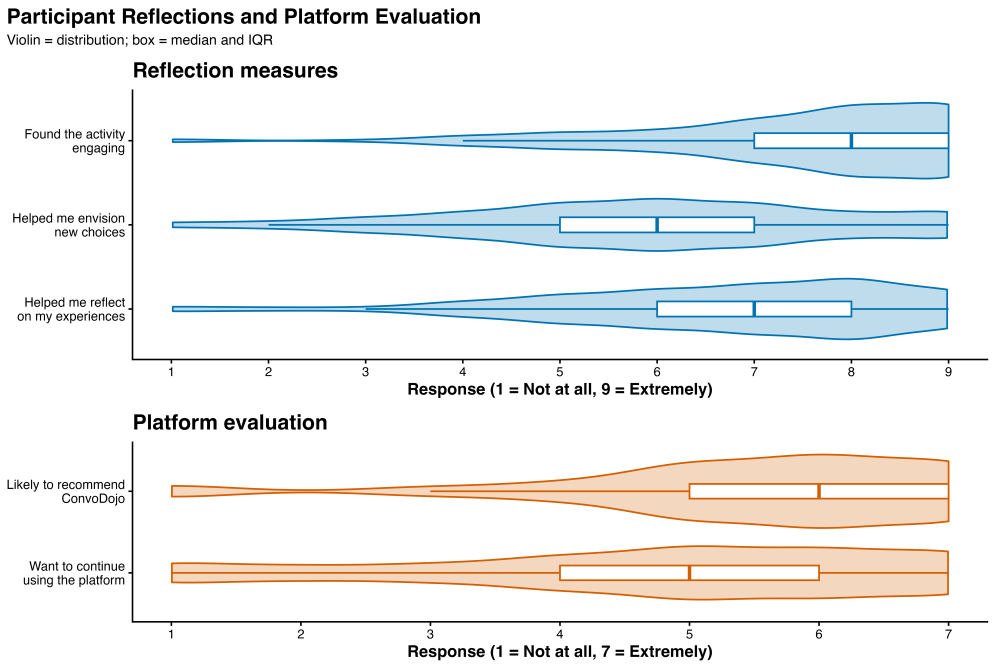


Figure 6: Distributions of participants' responses to reflection measures (top panel) and platform evaluation measures (bottom panel). Violin plots depict response distributions, boxplots indicate the median and interquartile range.

5 Discussion

Our results suggest that CUIs can support learning for difficult workplace conversations, but that effectiveness depends on how the interaction is structured and how support is delivered. Participants reported that they strongly avoided difficult conversations, even showing a willingness to trade time and money to do so. This indicates that the barrier is not solely a lack of skill, but also involves motivation and the perceived risk to interpersonal relationships. This is consistent with prior work on these avoidance patterns [6], highlighting a need for benchmarks that evaluate behavioral and strategic alignment between AI agents and humans in conflict settings [24]. Despite this significant barrier, we show that ConvoDojo produced meaningful shifts in both attitudes and behavior. Managers became more confident in their ability to handle difficult conversations, reported reduced avoidance (time-based), and improved their conversational performance across repeated practice.

ConvoDojo reduces friction to engage and promotes confidence in core challenges. A key finding is that improvements were significant for the aspects of difficult conversations participants found most challenging (i.e., initiating the conversation and handling employee defensiveness). These are moments where many managers would either delay initiation or respond reactively when the other party pushes back. ConvoDojo appears to reduce the initiation barrier and provide practice for repair and de-escalation by enabling low-stakes role-play. The reduction in time-based avoidance supports this interpretation. After using the platform, managers reported being willing to spend fewer minutes on administrative tasks to avoid these difficult conversations. We found the monetary avoidance moved in the expected direction but did not reach significance, suggesting that stronger incentives than brief intervention may be needed to shift deeper motivation or economic tradeoffs.

Practice and structured support matter. We observed behavioral change as measured by an increase in weighted conversation scores from first to second session. The within-participant improvement suggests that repeated practice within structured feedback can move learners toward more framework-aligned conversational behaviors even over short timescales. We also found a main effect of experimental condition on performance with Full System outperforming Practice Only. This pattern is consistent with the view that role-play alone can be engaging but is not always sufficient for reliable improvement, especially when learners must generalize to a new context. Upfront instructional scaffolding may help learners form a plan for openings and maintain psychological safety, while structured feedback supports reflection and correction. Together, these supports likely reduce the gap between "knowing the idea" and executing it in dialogue.

The Gap Between Perceived and Demonstrated Competence. We found evidence of a mismatch between how managers saw themselves and how they actually performed. Managers who self-rated as effective leaders consistently reported lower challenges with difficult conversations, yet they did not perform better in the simulated conversations in our study. In contrast, managers who rated themselves as more supportive leaders achieved higher conversational scores. Together, these results suggest that self-assessment

of effectiveness may not be a reliable indicator of handling difficult conversations well, whereas self-assessments of empathic leadership can be. One interpretation is that supportiveness may track an underlying interpersonal orientation that translates into framework-aligned behavior, while effectiveness self-ratings may track confidence or self-image rather than demonstrated skill.

5.1 Design implication for learning-oriented CUIs

Prior learning-oriented CUIs have established that conversational practice can support skill development in academic tutoring [12, 16], simulation-based teacher training [40], and scaffolded online learning [42]. Our findings extend this literature into high-stakes interpersonal skill development, where the dominant failure modes differ in kind: not knowledge gaps but motivated avoidance [6, 23], not learner errors but sycophantic AI partners [19, 37], and not underconfidence but miscalibrated self-perception. In this setting, fluent generation is necessary but not sufficient. What shapes learning is how the interaction is structured, when support is delivered, and what signals are used to adapt to the learner. We surface five implications for CUIs operating under these conditions, each grounded in our findings and positioned against prior conversational training systems.

Target challenging conversation moments explicitly rather than treating conversation as uniform. Confidence and performance gains in our study were largest at the moments participants self-identified as hardest: initiating the conversation and handling defensiveness. These are not arbitrary points of difficulty. Organizational communication research has long identified initiation as the moment where relational risk is most exposed and avoidance is most likely [6, 23], and Crucial Conversations [32] frames defensiveness as the failure mode through which safety collapses and dialogue degenerates into compliance or resistance. CUIs that treat the practice conversation as a uniform sequence of turns miss this structure entirely. Dialogue-based tutors such as AutoTutor [12, 16] are organized around assessing and building domain knowledge rather than around navigating interactionally difficult moments. Conflict-simulation systems such as Rehearsal [36] provide immersive role-play but do not differentially scaffold the hardest interactional transitions. ConvoDojo's stage-based orchestration makes these moments addressable. For example, its Opening and Dialogue stages can fire different coaching, persona, and intervention behavior precisely because the system knows where in the arc the learner is.

Pair practice with structured feedback and upfront scaffolding to support transfer, not just task-specific gains. All three experimental of our conditions improved within the practice scenario, indicating that low-stakes role-play has value on its own. However, only Full System participants outperformed on the unrehearsed scenario. Though a harder claim, we show that scaffolding plus feedback produces a transferable conversational stance learners can carry into novel conversations. This is consistent with prior accounts of feedback as collaborative process between user and system [18] and with conversation-based assessment work that links real-time evaluation to learner sense-making [44]. When criteria is made explicit upfront, feedback that follows becomes interpretable as guidance rather than verdict. Sara [42] demonstrated this pattern on academic content and our findings extend it into open-ended interpersonal practice, where behavioral transfer is a harder claim to support. Although earlier conflict-focused systems [36] provided simulated practice, they have not isolated the impact of upfront framework scaffolding from practice alone. Our study addresses this gap.

Prioritize observable behavior over reported self-perception when adapting support. We observed a mismatch between how managers saw themselves and how they performed. Specifically, self-rated effectiveness did not predict performance, while self-rated supportiveness did. This is in line with prior work that has shown that self-assessments of skill correlate only modestly with objective performance across domains, including in workplace and interpersonal settings [13, 46] and this correspondence is weakest precisely when self-evaluations are broad and tasks are complex [46]. These are conditions that describe both our self-rated effectiveness measure and difficult workplace conversations themselves. This pattern is particularly important for adaptive CUIs because mixed-initiative and proactive systems frequently draw on learner-reported state (e.g., confidence, comfort, perceived difficulty) as the primary signal for when to intervene or how much support to provide [9, 11, 12]. If self-report is a structurally unreliable proxy for competence, systems that adapt on self-report, risk calibrating support to the learner's self-image rather than their behavior. The risk in this case is a feedback loop: a system that adapts its support to a signal that systematically misrepresents competence reinforces the very misperception it should help correct. We therefore suggest a more cautious stance where behavioral evidence (escalation cues, rubric adherence patterns, repeated lapses in framework alignment) are used to adapt support rather than relying on user self-report and treat self-assessment instead as a measurement the system should help recalibrate rather than defer to. ConvoDojo's turn-level instrumentation logs the signals such a design move would act on, making behaviorally grounded adaptation a tractable rather than aspirational direction.

Separate generation from control via an orchestration layer. Sycophancy is a property of generation and productive challenge is a property of orchestration. The tendency of LLMs to default to agreement is difficult to suppress reliably through prompt instructions alone. It persists across extended interactions [19] and intensifies when users push back [22, 37], and prior work locates its cause in training rather than in any single prompt [37]. The architectural response of placing a stateful controller above the language layer rather than relying on the generative model to self-regulate is not itself new. Mixed-initiative tutoring systems used stateful dialogue managers to decide when to prompt, hint, or assert [9, 12, 16], though they did so to govern pedagogical flow rather than to counteract a generative model's agreeableness. ConvoDojo extends this architectural separation into the LLM era by externalizing pedagogical decisions (i.e., when to challenge, when to coach, when to step out of role) into a hierarchical state machine that conditions the system prompt, persona constraints, and allowable interventions on the current dialogue stage. We did not directly compare this architecture against a pure prompt-engineering baseline, so we frame this as a design stance rather than an empirical finding.

However, the coherence we observed across stages, the ability to fire targeted interventions at appropriate moments, and the consistency of role behavior are properties this separation enables. The same architectural intuition is visible in Al Owayyed et al.'s hybrid BDI–LLM design for child helpline training [1] and in recent work on dialogue management as orchestration [15, 26].

Calibrate adversarial realism rather than maximize it. Our qualitative findings revealed a tension specific to sparring-style CUIs. Participants found the simulated employees more reasonable and easier to engage with than people they encounter in real settings, and many requested greater pushback. This tension is not unique to our system, but rather reflects a longstanding question in simulation-based learning about the relationship between fidelity and learning, where evidence indicates that physical realism contributes far less to transfer than psychological fidelity does [29]. High-fidelity simulations may reproduce the affective intensity of real situations but at the cost of cognitive load that overwhelms novice learners. On the other hand, lower-fidelity environments preserve the structural features of the task while keeping practice tractable. ConvoDojo's current persona configurations sit closer to the latter end, which our quantitative results suggest supported confidence and performance gains, but which our qualitative results suggest leaves some learners under-challenged. Prior work evaluating behavioral alignment in conflict dialogue further finds that LLM agents diverge substantially from humans across linguistic, emotional, and strategic dimensions of conflict [24]. Increasing a simulated employee's pushback is therefore not equivalent to making the simulation more realistic. Calibration is a question of behavioral coherence, not difficulty alone. Taken together, these observations point toward calibration rather than maximization as the design objective. The interaction should be hard enough to elicit the conversational moves the framework targets, but not so hard that it disengages learners or trains avoidance. A concrete affordance, requested by several of our participants, is to expose difficulty as a user-controllable parameter rather than a fixed property of the simulation.

Scope and transfer across interpersonal-skills domains. We developed these implications in the setting of difficult workplace conversations, but they are not specific to it. The five implications operate at two levels that generalize differently. The architectural implications, separating generation from control, and adapting on behavioral rather than self-reported signals, are largely domain-agnostic. Any high-stakes interpersonal-skills CUI in which the LLM's default agreeableness would dilute productive friction can benefit from them, whether the application is clinician–patient communication, teacher feedback, or crisis-line training. The content implications, that is, which moments to scaffold, which framework to anchor feedback in, which personas to simulate, do not transfer verbatim. Initiation and defensiveness are load-bearing moments in managerial conversations, but other domains will have their own. Crucial Conversations is one of several established frameworks, and clinician communication or behavioral counseling could substitute SPIKES [2] or motivational interviewing [28]. Realizing these implications in a new domain therefore requires fresh task analysis rather than reuse. Two scope conditions further bound the contribution. First, the scaffolding-and-feedback mechanism may depend on the existence of an established conversational framework against which performance can be assessed. In domains without one, a rubric-based judge has no principled target. Second, the value of prioritizing behavioral signals over self-report is greatest where self-assessment is least reliable. For example, broad interpersonal competencies practiced under complex conditions. It diminishes for narrow, well-defined skills where learners tend to hold accurate self-models [46]. Finally, the calibration implication assumes a training goal centered on handling pushback, and is less relevant where the AI's role is primarily supportive rather than oppositional.

5.2 Limitations and Future Work

This study has important limitations. First, we did not test durability of confidence gains or behavioral change in real workplace settings. Second, conversational performance was assessed within simulated interactions rather than observed in live organizational contexts. Future work should examine longitudinal and downstream organizational effects, specifically how AI-based skill building translates to broader professional performance and employability [34], as well as behavioral transfer to real workplace conversations. Third, difficult workplace conversations are often conducted face-to-face or verbally, and a text-only interaction omits prosodic, facial and embodied cues present in face-to-face exchanges, and future work should assess whether the effects observed here generalize across modalities.

6 Conclusion

Difficult conversations are consequential but are routinely avoided and existing approaches often struggle to scale and to provide timely and actionable feedback. This paper introduces ConvoDojo, an instrumented conversational training platform that repurposes LLMs as *structured sparring partners* for rehearsing challenging conversations. The platform combines stage-based orchestration, configurable personas and coaching frameworks, and rubric-based assessment to support both in-conversation coaching and evidence-grounded debrief reports. In a controlled evaluation of the platform, we observed that it increased managers' confidence and improved conversational performance across sessions, with results suggesting that structured feedback and upfront instruction scaffolding provide benefits beyond practice alone. These findings highlight a core opportunity for learning-oriented CUIs: pairing generative flexibility with explicit interaction structure and actionable feedback can make LLM role-play a measurable tool for interpersonal skill development.

Acknowledgments

LLMs (including Gemini, ChatGPT, and Claude) were used to assist with code generation, figure text descriptions and prose revision.

References

[1] Mohammed Al Owayyed, Adarsh Denga, and Willem-Paul Brinkman. 2025. Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training. In *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents (IVA '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3717511.3747075

[2] Walter F Baile, Robert Buckman, Renato Lenzi, Gary Glober, Estela A Beale, and Andrzej P Kudelka. 2000. SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. 302–311 pages.

[3] Barbara Rita Barricelli, Federica Caffaro, Raffaele Cioffi, Daniela Fogli, Matteo Gremo, Margherita Micheletti Cremasco, Davide Morina, and Mentore Vaccari.

2025. A Serious Game for Safety and Health Training in the Workplace. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly '25)*. Association for Computing Machinery, New York, NY, USA, Article 108, 3 pages. doi:10.1145/3750069.3755968

[4] Param Biyani, Yasharth Bajpai, Arjun Radhakrishna, Gustavo Soares, and Sumit Gulwani. 2024. RUBICON: Rubric-Based Evaluation of Domain-Specific Human AI Conversations. In *Proceedings of the 1st ACM International Conference on AI-Powered Software* (Porto de Galinhas, Brazil) *(AIware 2024)*. Association for Computing Machinery, New York, NY, USA, 161–169. doi:10.1145/3664646.3664778

[5] Kim Bosman, Tibor Bosse, and Daniel Formolo. 2019. Virtual agents for professional social skills training: An overview of the state-of-the-art. In *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 75–84.

[6] Graham L. Bradley and Amanda C. Campbell. 2016. Managing Difficult Workplace Conversations: Goals, Strategies, and Outcomes. *International Journal of Business Communication* 53, 4 (2016), 443–464. arXiv:https://doi.org/10.1177/2329488414525468 doi:10.1177/2329488414525468

[7] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 103–119. doi:10.1145/3640543.3645199

[8] Janghee Cho and Emilee Rader. 2020. The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1, Article 33 (2020), 28 pages. doi:10.1145/3392838

[9] Mark G. Core, Johanna D. Moore, and Claus Zinn. 2003. The role of initiative in tutorial dialogue. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1* (Budapest, Hungary) *(EACL '03)*. Association for Computational Linguistics, USA, 67–74. doi:10.3115/1067807.1067818

[10] Chih-Pu Dai, Fengfeng Ke, Nuodi Zhang, Alex Barrett, Luke West, Saptarshi Bhowmik, Sherry A. Southerland, and Xin Yuan. 2024. Designing Conversational Agents to Support Student Teacher Learning in Virtual Reality Simulation: A Case Study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 513, 8 pages. doi:10.1145/3613905.3637145

[11] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive Conversational AI: A Comprehensive Survey of Advancements and Opportunities. *ACM Trans. Inf. Syst.* 43, 3, Article 67 (March 2025), 45 pages. doi:10.1145/3715097

[12] Sidney D'Mello and Arthur C. Graesser. 2012. AutoTutor and Affective AutoTutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers that Talk Back. *ACM Transactions on Interactive Intelligent Systems* 2, 4, Article 23 (2012). doi:10.1145/2395123.2395128

[13] David Dunning, Chip Heath, and Jerry M Suls. 2004. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest* 5, 3 (2004), 69–106.

[14] Evan Ellis, Vivek Myers, Jens Tuyls, Sergey Levine, Anca Dragan, and Benjamin Eysenbach. 2026. Training LLM Agents to Empower Humans. https://openreview.net/forum?id=W9oGZd4B8R

[15] Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2025. Tailored Conversations beyond LLMs: A RL-Based Dialogue Manager. arXiv:2506.19652 [cs.CL] https://arxiv.org/abs/2506.19652

[16] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney. 2005. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Trans. on Educ.* 48, 4 (Nov. 2005), 612–618. doi:10.1109/TE.2005.856149

[17] Michael Guevarra, Indronil Bhattacharjee, Srijita Das, Christabel Wayllace, Carrie Demmans Epp, Matthew E. Taylor, and Alan Tay. 2025. An LLM-guided tutoring system for social skills training. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 3455, 3 pages. doi:10.1609/aaai.v39i28.35353

[18] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[19] Yining Hong et al. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP*. https://aclanthology.org/2025.findings-emnlp.121/

[20] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 652, 28 pages. doi:10.1145/3613904.3642349

[21] Bijan Khosrawi-Rad, Arne Borchers, Linda Grogorick, and Susanne Robra-Bissantz. 2023. Design Principles for Gamified Pedagogical Conversational Agents. In *Proceedings of the 26th International Academic Mindtrek Conference* (Tampere, Finland) *(Mindtrek '23)*. Association for Computing Machinery, New York, NY, USA, 67–82. doi:10.1145/3616961.3616973

[22] Sungwon Kim and Daniel Khashabi. 2025. Challenging the Evaluator: LLM Sycophancy Under User Rebuttal. *arXiv preprint arXiv:2509.16533* (2025).

[23] Louise Kippist and Fernanda Duarte. 2015. What does it mean having difficult conversations in the workplace? : an exploratory literature review. *Employment Relations Record* 15, 2 (2015), 61–74.

[24] Deuksin Kwon, Kaleen Shrestha, Bin Han, Elena Hayoung Lee, and Gale Lucas. 2025. Evaluating Behavioral Alignment in Conflict Dialogue: A Multi-Dimensional Comparison of LLM Agents and Humans. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 16377–16391.

[25] Teddy Lazebnik, Lior Zalmanson, and Osnat Mokryn. 2025. Mind Your Manners: The Dynamics of Politeness in Human-AI vs. Human-Human Interactions. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW450 (Oct. 2025), 22 pages. doi:10.1145/3757631

[26] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2024. OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 1434–1445. doi:10.18653/v1/2024.naacl-long.79

[27] Ziming Li, Huadong Zhang, Chao Peng, and Roshan Peiris. 2025. Exploring Large Language Model-Driven Agents for Environment-Aware Spatial Interactions and Conversations in Virtual Reality Role-Play Scenarios. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. 1–11. doi:10.1109/VR59515.2025.00025

[28] William R. Miller and Stephen Rollnick. 2023. *Motivational Interviewing: Helping People Change and Grow* (4th ed.). The Guilford Press, New York, NY.

[29] Geoff Norman, Kelly Dore, and Lawrence Grierson. 2012. The minimal relationship between simulation fidelity and transfer of learning. *Medical education* 46, 7 (2012), 636–647.

[30] Farnaz Nouraei, Keith Rebello, Mina Fallah, Prasanth Murali, Haley Matuszak, Valerie Jap, Andrea Parker, Michael Paasche-Orlow, and Timothy Bickmore. 2025. Virtual Agent-based Communication Skills Training to Facilitate Health Persuasion Among Peers. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW203 (May 2025), 24 pages. doi:10.1145/3711101

[31] Stefán Ólafsson, Paola Pedrelli, Byron C. Wallace, and Timothy W. Bickmore. 2023. Accomodating User Expressivity while Maintaining Safety for a Virtual Alcohol Misuse Counselor. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA '23)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3570945.3607361

[32] Kerry Patterson, Joseph Grenny, Ron McMillan, Al Switzler, Stephen R Covey, and Laura Roppé. 2012. *Crucial conversations: Tools for Talking When Stakes are High*. Brilliance Audio.

[33] Manuel A. Pérez-Quiñones and John L. Sibert. 1996. A collaborative model of feedback in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada) *(CHI '96)*. Association for Computing Machinery, New York, NY, USA, 316–323. doi:10.1145/238386.238535

[34] Heily Consepción Portocarrero Ramos, Omer Cruz Caro, Einstein Sánchez Bardales, Lenin Quiñones Huatangari, Jonathan Alberto Campos Trigoso, Jorge Luis Maicelo Guevara, and River Chávez Santos. 2025. Artificial intelligence skills and their impact on the employability of university graduates. *Frontiers in Artificial Intelligence* 8 (2025), 1629320.

[35] Dorela Priftanji, John D Hill, and Daniel M Ashby. 2020. Managing difficult conversations. *American Journal of Health-System Pharmacy* 77, 21 (08 2020), 1723–1726. arXiv:https://academic.oup.com/ajhp/article-pdf/77/21/1723/34499076/zxaa149.pdf doi:10.1093/ajhp/zxaa149

[36] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating Conflict to Teach Conflict Resolution. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 920, 20 pages. doi:10.1145/3613904.3642159

[37] Manan Sharma, Meihan Tong, Tomasz Korbak, Ethan Perez, Lewis Cook, Nino Scherrer, Ivan Evtimov, Dan Hendrycks, Samuel Bowman, Talia Ringer, et al. 2024. Towards Understanding Sycophancy in Language Models. In *International Conference on Learning Representations (ICLR)*. https://proceedings.iclr.cc/paper_files/paper/2024/file/0105f7972202c1d4fb817da9f21a9663-Paper-Conference.pdf

[38] Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. 2025. Training proactive and personalized llm agents. *arXiv preprint arXiv:2511.02208* (2025).

[39] Clarice Wang, Yimin Shi, and Xiaokui Xiao. 2025. A Framework for Evaluating AI Agents in Open-Ended Conversations via Scripted Simulation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 5810–5818. doi:10.1145/3711896.3737390

[40] Xu Wang, Meredith Thompson, Kexin Yang, Dan Roy, Kenneth R. Koedinger, Carolyn P. Rose, and Justin Reich. 2021. Practice-Based Teacher Questioning

Strategy Training with ELK: A Role-Playing Simulation for Eliciting Learner Knowledge. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 51 (April 2021), 27 pages. doi:10.1145/3449125

[41] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[42] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376781

[43] Daksitha Senel Withanage Don, Thomas Kiderle, Silvan Mertes, Dominik Schiller, Hannes Ritschel, and Elisabeth André. 2025. MeLaX: Conversations with Generative AI in Socially Interactive Agents. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25 Companion)*. Association for Computing Machinery, New York, NY, USA, 163–166. doi:10.1145/3708557.3716363

[44] Seyma N. Yildirim-Erbasli and Okan Bulut. 2021. Conversation-Based Assessments: Real-Time Assessment and Feedback. *ELearn* 2021, 12, Article 1 (Dec. 2021). doi:10.1145/3508017.3495533

[45] Monika Zamojska and Jarosław A Chudziak. 2025. TACLA: An LLM-Based Multi-Agent Tool for Transactional Analysis Training in Education. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 313–320.

[46] Ethan Zell and Zlatan Krizan. 2014. Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science* 9, 2 (2014), 111–125.