Cao, L., Etemadi, A., Wheeler, M., & Dede, C. (2025). Keeping the "glass box" transparent: Comparing expert and AI-generated ratings and feedback in stealth assessment for judgement-focused negotiation simulations. *Journal of Research on Technology in Education*, 1–23. https://doi.org/10.1080/15391523.2025.2568526

Keeping the "Glass Box" Transparent:

Comparing Expert and AI-generated Ratings and Feedback in Stealth Assessment for Judgement-focused Negotiation Simulations

Authors

Lydia Cao, Ontario Institute for Studies in Education, University of Toronto

Ashley Etemadi, Motiva Education

Mike Wheeler, Harvard Business School

Chris Dede, Harvard Graduate School of Education

Details of the corresponding author

Lydia Cao ly.cao@utoronto.ca

Funding

This research has not received any external funding.

Declaration of interest

The authors report there are no competing interests to declare.

Word count 7586 (exluding references).

Keeping the "Glass Box" Transparent: Comparing Expert and AI-generated Ratings and Feedback in Stealth Assessment for Judgement-focused Negotiation Simulations

Negotiation is a hard-to-measure competency, involving a dynamic balance of relational and outcome-oriented dimensions. Generative AI has opened avenues for delivering real-time assessment and feedback after a negotiation simulation. However, there is a tension between the "black box" architecture of GenAI and the "glass box" approach of stealth assessment. This case study uses a mixed-method approach to compare ratings and feedback given by GenAI and a human expert on seven negotiation transcripts. The results illustrate that with predetermined criteria, GenAI provides more formulaic feedback, while an expert focuses on the context of exchanges. Implications for stealth assessment and negotiation feedback are discussed.

Keywords: stealth assessment; generative AI; feedback; negotiation

Subject classification codes: include these here if the journal requires them

Introduction

Negotiation represents a hard-to-measure competency. Inherently complex and multifaceted, negotiation requires a delicate balance of assertiveness and warmth as well as the orchestration of communication, strategy, emotional intelligence, contextual awareness, and improvisation. This requires in-the-moment improvisation and cannot be reduced to a set of procedures and recipes (Wheeler, 2013). As a result, negotiation training often adopts a 'learning-by-doing' approach through using simulations — whether through in-person role play, simulations with digital puppeteering (e.g., Mursion), or interactions with AI agents (Dinnar et al., 2021; Movius, 2008; Susskind & Corburn, 2000; Wheeler, 2006).

Stealth assessment offers a promising framework for evaluating complex competencies such as negotiation. Unlike traditional outcome-focused assessment, stealth assessment is *process-oriented* and is seamlessly embedded into the dynamics of learning. Stealth assessment uses rich data from technologically enhanced environments (e.g., digital games, VR, digital simulations) to make real-time inferences that simultaneously assess and foster improvement in hard-to-measure competencies and skills, such as creativity, persistence, and problem-solving (Rahimi & Shute, 2024). These competencies require sophisticated metrics that often manifest in the process and interactions rather than solely in the outcome. Critically, stealth assessment operates as a "glass box" that uses evidence-based methodology to ensure validity, reliability, and fairness (Mislevy et al., 2003; Rahimi & Shute, 2024; Shute, 2009).

Recent advances in generative AI (GenAI), especially in natural language understanding (NLU) and processing (NLP), offer new possibilities for negotiation training and assessment. GenAI can potentially play the role of a negotiation counterpart and provide immediate evaluation and personalized feedback based on rich sources of multimodal data collected in the process, such as learner response, tone, and outcome.

However, the potential offered by GenAI comes with a fundamental tension: while stealth assessment requires transparency and evidence-based methodology (the "glass box"), generative AI operates within an inherently "black box" architecture. This tension raises crucial questions about how to harness AI's capabilities while maintaining the glass box of stealth assessment to ensure its psychometric rigor and transparency.

The use of GenAI in stealth assessment is still at a nascent stage, and little is known about the affordances and limitations of GenAI in assessing and producing feedback for hard-to-measure skills. Interpretability for GenAIs is challenging due to its

technological architecture that relies on billions of parameters. According to Rose (2025), AI models can appear to understand complexity when they are actually just memorizing specific examples seen before in their training dataset. This creates the illusion of intelligence and expertise when the AI might just be recalling specific answers. It is thus crucial to understand the nature of AI-generated assessment and feedback in terms of validity, reliability, and fairness. Otherwise, we risk reducing complex human skills to what AI is capable of grasping, simplifying negotiation into gaming the AI system and perpetuating systematic algorithmic bias.

This study is an initial step towards understanding the validity of AI-generated assessment and feedback for complex human skills, such as negotiation. By quantitatively and qualitatively comparing feedback generated by GenAI and human experts, we delineate on the affordances and limitations of GenAI for complex skill assessment and feedback. We propose potential models for hybrid systems that leverage the unique strengths of both humans and AI —a form of intelligence augmentation (Dede et al., 2021). These fundamental insights are crucial for maintaining the glass box of stealth assessment as generative AI becomes increasingly integrated into technologically rich learning environments.

Literature review

Simulations have long been used in complex skill development (e.g., negotiation, teacher training) as they create rich contexts for learners to enact the skill in specific situations rather than simply following a one-size-fits-all recipe of procedures. To better understand the historic landscape of negotiation simulations and their relationship with GenAI-enhanced stealth assessment, we identify two distinct ways to categorize negotiation simulations based on: 1) interaction design (i.e., menu-based vs. free text) and 2) pedagogical focus (i.e., strategy-focused to foster routine expertise vs.

judgement-focused to foster adaptive expertise).

1. ***Interaction design: Menu-based vs free text***

Simulations can be categorized based on their interaction design. Menu-based simulations provide pre-determined options and actions for learners from which to choose. Examples are IAGO for negotiation and SimSchool for teacher training (Deale & Pastore, 2014; Mell & Gratch, 2016; Murawski et al., 2024). In contrast, free-text simulations allow natural language interaction through either open-ended writing or speech. Examples include Simulation Labs, ACE for negotiation training, Wonda, and Mursion (Dinnar et al., 2021; Mikeska et al., 2023; Shea et al., 2024). To offer learners experiences that better approximate real-world interactions, this type of simulation can leverage machine learning, digital puppeteering that involves human in the loop, and recently large language models that use AI-based agents.

2. ***Instructional focus: Strategy-focused vs judgement-focused simulations***

Simulations can also be categorized based on their instructional focus and the type of expertise they aim to foster. Strategy-focused negotiation simulations emphasize developing fluency with proven strategies and tactics that have clear benchmarks for success—what can be characterized as "routine expertise" (Hatano & Oura, 2003; Murawski et al., 2024; Shea et al., 2024). For instance, Monahan et al. (2018) developed a virtual agent to engage in negotiation with learners using a set of pre-determined principles quantified through automated methods, including 1) make high initial offers, 2) use more of available time, and 3) maintain strong offers throughout. Though such training is helpful for learners to establish fluency with strategies and tactics, predetermined principles can overlook crucial contextual factors, as real-world scenarios demand adaptability and nuanced responses to changing situations.

Judgement-focused simulations, in contrast, aim to support learners in developing the ability to flexibly apply knowledge and skills across varied contexts, which can be characterized as "adaptive expertise" (Hatano & Oura, 2003). Unlike strategy-focused simulations, which often have clear benchmarks for success, judgement-focused simulations have no single correct approach and require learners to make multiple types of judgements:

1) Judgement of the outcome: Beyond simple metrics like getting the best price, learners must judge what will be a favourable outcome given a specific context. This requires learners to understand how to prioritize their goals and the interests of the parties involved. For instance, they might prioritize a long-term partnership more than maximizing short-term financial gain.
2) Judgement on sociocultural contexts: Learners need to navigate the complex interplay of culture, interpersonal relations, and power dynamics. For example, a subordinate formulating a statement in a particular way might be expected in negotiation in one culture while inappropriate in another. Thus, learners need to recognize how cultural differences can influence negotiation expectations as well as the management of power imbalances and positive relationships.
3) Judgement on strategy and tactics: Learners have to judge the appropriate strategy and tactic to use for a particular context rather than applying a predetermined recipe or a one-size-fits-all approach. A question could be relevant and appropriate in one point in the negotiation but, later on in the same conversation, such a question could be inappropriate.
4) Judgement in the moment: Learners need to adapt and improvise in the moment, given the uncertainty in a negotiation. For example, they must adapt to decide

when to assert themselves, when to take a step back, when to continue or pause a negotiation, and how to respond to unexpected situations (Wheeler, 2013).

While recognizing the importance of strategy-focused simulations as foundational for further learning, judgement-based simulations prepare learners to improvise and adapt their strategies in response to evolving dynamics, including being attentive to both the substantive outcomes and the relational aspects of the negotiation (Wheeler, 2013). This creates a tension in objectives. Ma et al. (2024) found that outcome-focused AI approaches in negotiation misalign with the process-oriented needs of humanitarian negotiators, who need contextual understanding and human rapport-building. Thus, rather than those offering direct recommendations of invariant negotiation strategies, they emphasized the need for flexibility that contextualizes cases and supports negotiators exploring alternative options with their associated benefits and risks.

3. GenAI for assessment and feedback in negotiation simulations

The distinction between strategy-focused and judgement-focused negotiation simulations has important implications for how GenAI can be harnessed in stealth assessment. Research shows that GenAI performs well at simple, closed-ended assessments but struggles with nuanced and open-ended ones. For example, GenAI was shown to effectively classify student comments in online courses but struggled with interpreting nuanced content in longer student forum posts (Chien et al., 2024; Misiejuk et al., 2024; Wang et al., 2023).

In strategy-focused simulations, research has found that GenAI can provide assessment and feedback based on clear benchmarks and well-defined criteria. For example, Shea et al. (2024) demonstrates the effectiveness of ACE (Assistant for Coaching Negotiation) in providing feedback and enhancing learners' negotiation skills.

ACE is an LLM-based system to coach learners' negotiation skills; it functions as both a negotiation partner and a coach by identifying learners' deviations from proven negotiation principles, such as setting strategic walk-away prices, breaking the ice to establishing social rapport, making ambitious first offers, and providing rationales for proposals, strong counteroffers, and strategic closing. The system demonstrated over 90% accuracy in error detection in all categories

In contrast, assessing judgement-focused simulations is much more challenging as there is not a pre-determined right or wrong strategy approach. These simulations thus require the evaluator to judge whether a strategy or tactic used is contextually appropriate within the dynamics of a given negotiation (Murawski et al., 2024; Shea et al., 2024). This demands a more sophisticated understanding of the relationship between context and strategy. While GenAI has shown the capability of identifying deviations from standard practices, it often struggles with nuanced situations where multiple valid approaches exist or where cultural, relational, power, and other situational factors require negotiators to improvise in the moment (Wheeler, 2021). Its inherent complexity and ambiguity make judgement-focused negotiation challenging to assess since success depends not only on following tactics but also on navigating various complexities, adapting to unexpected responses, and making contextual judgements in the moment (Wheeler, 2013). So far, little is known about the extent to which GenAI can evaluate and provide feedback on judgement-focused negotiation simulations.

Research questions:

1) To what extent does AI-generated rating on judgement-focused negotiation simulations align with human experts?
2) To what extent does AI-generated feedback on judgement-focused negotiation simulations align with human experts? What are the similarities and differences

in terms of *approach* and *content* between human expert and AI-generated feedback?

3) What process does a negotiation expert undertake when assessing judgement-focused negotiation simulations?

Methodology

Given current limited understanding of GenAI's capability for rating and producing feedback on judgement-focused negotiation, we used a case study method with a mixed-method approach that affords close examination of the alignment and discrepancies between GenAI and human experts, both qualitatively and quantitatively. We also conducted a think-aloud protocol with a human negotiation expert to shed light on experts' cognitive processes, illuminating what makes negotiation assessment particularly challenging (Ericsson & Simon, 1993). This small-scale study aligns with the principle outlined by Rose (2025): "In order to achieve the greatest synergy between human intelligence and AI, we must probe deeply into the nuances of what makes some tasks challenging." Using a mixed-methods approach, we identify areas where GenAI evaluations converge with or diverge from human expert assessments.

Negotiation scenario

This case study used seven transcripts of a judgement-focused negotiation with an AI agent. This simulation presents an open-ended scenario that involves an imbalance of power dynamics between two parties. The learner plays the role of an employee who has been performing significant additional responsibilities beyond their official positions. After demonstrating significant value to the company, they must navigate a high-stakes negotiation with the CEO, with whom they have had challenging and tense interactions in the past (for the full scenario, see Appendix 2).

This scenario represents a judgement-focused negotiation as it requires learners to:

1) Judgement on the outcome: Determine what constitutes a favourable outcome for them beyond just salary, considering other elements including position, transition period, and work-life balance.
2) Judgement on the sociocultural context: Navigate the unbalanced power dynamic between themselves and the CEO.
3) Judgement on strategies and tactics: Decide on an appropriate negotiation approach given their troubled history of prior attempts with the CEO.
4) Judgement in the moment: Manage unexpected emotional situations, such as when the CEO refuses to recognize the value of the employee with condescending remarks that require learners to respond effectively in the moment.

Data sources and analysis

Both a human expert of negotiation with 40+ years of experience and GenAI (GPT4o with structured prompting) assessed seven transcripts on a scale of 1-5 across four dimensions of criteria: persuasiveness, managing emotions, creativity, and managing the negotiation process. All dimensions require adaptive expertise rather than applying predetermined strategies. The human expert and the AI each provided qualitative feedback on five randomly selected transcripts. It is important to note that, to ensure consistency in the assessment framework, the AI assessments were generated using structured prompts designed by the same human expert who conducted the expert assessments. In addition, to better understand the process human experts undertake when assessing negotiation, we also conducted a think-aloud protocol which lasted 1.5

hours (Ericsson & Simon, 1993). The session was audio recorded and transcribed. A summary of quantitative and qualitative analyses is shown in Table 1.

Table 1. Summary of data sources and analysis methods

RQ	Data sources	Analysis methods
RQ1: To what extent does AI-generated rating on judgement-focused negotiation simulations align with human experts?	Expert and AI **Rating** on 7 simulation transcripts	Quantitative methods: - Descriptive statistics - Visualization (Heatmap and swarm plot)
RQ2: To what extent does AI-generated feedback on judgement-focused negotiation simulations align with human experts? What are the differences in terms of approach and content between human expert and AI-generated feedback?	Expert and AI **Feedback** on 5 simulation transcripts	Quantitative method: **Semantic similarity**: compare semantic similarity between expert and AI feedback using SBERT (e.g., all-MiniLM-L6-v2 model in our case) to compare cosine similarity. Qualitative methods: **Systematic coding of AI/human feedback**: two coders collaboratively conducted a combination and inductive and deductive coding sentence by sentence across 3 iterations to finalize a codebook that captures human and AI feedback in terms of feedback approach (Table 2) and feedback content (Table 3) (Saldaña, 2025; Strauss, 1990). One coder reviewed all transcripts systematically with the finalized codebook and the other coder conducted a second round of review to ensure accuracy and comprehensiveness. For full list of codes see Tables 5 and 6. This enriches the quantitative comparison and provides relevant examples to illustrate the differences
RQ3: What process does a negotiation expert undertake when assessing judgement-focused negotiation simulations?	Expert **think-aloud** on 2 transcripts	Qualitative method: 1. Reflexive thematic analysis of think-aloud transcript (Braun & and, 2019) to reveal verbal reasoning for decision criteria and patterns, identify implicit evaluation criteria not in the rubric, document human intuition and implicit knowledge.

Table 2. High-level codes for feedback approach

Feedback Approach	**Description**
Observe	Noticing specific actions, language choices, and interactions

Interpret	Sense-making of observation, including connecting observed behaviors to negotiation concepts, recognizing strategies used, making comparisons across simulations, and speculate why participants chose particular approaches.
Assess	Evaluating performance of the learner, what they did well, and what they could have done differently
Metacognitive reflection	Metacognitively reflecting on the feedback process itself, such as considering whether there is sufficient evidence to conduct assessment, the role of the AI agent.

Table 3. High-level codes for feedback content

Feedback content	Element	Description
Relational	Empathy	Understanding the motivations and feelings of other parties
	Assertiveness	Asserting your interests and point of view
Substantive	Creating Value	Recognizing and capitalizing on opportunities to create value
	Claiming Value	Getting the maximum possible in the agreement

Findings

RQ1: Comparing Expert vs AI-generated Rating

The descriptive statistics reveal substantial differences in ratings between the human expert and AI. The Heatmap in Figure 1 shows the expert and AI ratings on each of the seven simulation transcripts across four dimensions. Darker color indicates higher rating scores. The AI ratings are relatively homogenous in color, whereas the human rating shows clear color contrast. This indicates that the experts are distinguishing between high and low performance.

Furthermore, unlike the AI, which always provided a rating on every dimension in every transcript, the human expert exercised judgement on when evaluation was appropriate. For instance, for simulations 5 and 7, the expert deliberately chose to withhold evaluation for the "managing emotions" dimension as they noted that there was not enough evidence or emotional moments in the simulation to make an accurate assessment.

For persuasiveness, expert ratings (M = 3.43, SD = 1.99, Range = 1-5) showed considerably higher standard deviation and wider range compared to AI ratings (M =

3.43, *SD* = 0.53, Range = 3-4), though they had the same mean value. Similarly, the swarm plots illustrated that AI ratings clustered narrowly in the middle while expert ratings displayed a bimodal or hourglass pattern with concentrations at both low and high ends, indicating the expert was making a distinction between a poor and strong performance.

For emotion management, AI ratings (*M* = 4.29, *SD* = 0.49, Range = 4-5) skewed towards higher ends of the scores compared to more varied expert assessments (*M* = 3.60, *SD* = 1.52, Range = 2-5). This suggests the AI was systematically overestimating scores in the emotional dimension, potentially indicating a lack of understanding of what emotion dimensions entail. The swarm plot similarly shows a bimodal pattern with concentrations at both low and high ends. It is important to note that human experts have deliberately withheld assessment in two instances as discussed above.

For creativity, the human expert and AI showed more similar assessment patterns with similar means, expert ratings (*M* = 3.43, *SD* = 1.40, Range = 2-5) and AI ratings (*M* = 3.00, *SD* = 0.82, Range = 2-4). Nevertheless, expert ratings still had a greater standard deviation, indicating more discernment between high and average performance.

For process management, expert evaluations (*M* = 3.57, *SD* = 1.81, Range = 1-5) showed greater standard deviation and wider range compared to AI ratings (*M* = 3.14, *SD* = 0.90, Range = 2-4). Similar to persuasiveness, the expert ratings displayed an hourglass distribution in the swarm plot, indicating clear discrimination between strong and poor performance in managing negotiation processes, while AI ratings remained clustered in the middle ranges.

Figure 1. Heatmap of AI and Human Expert Rating of Negotiation

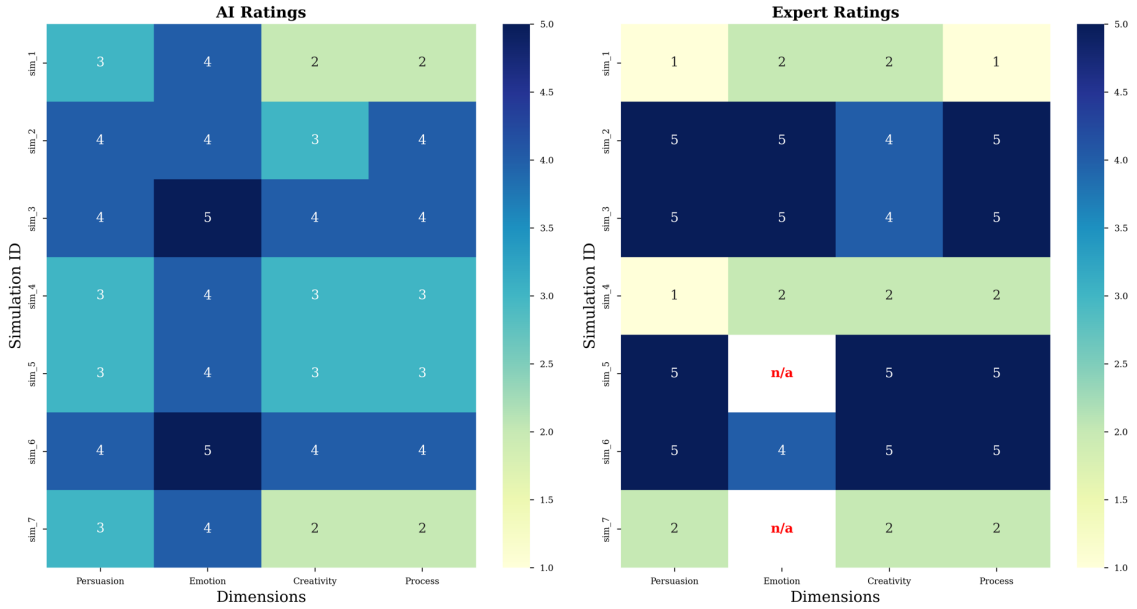
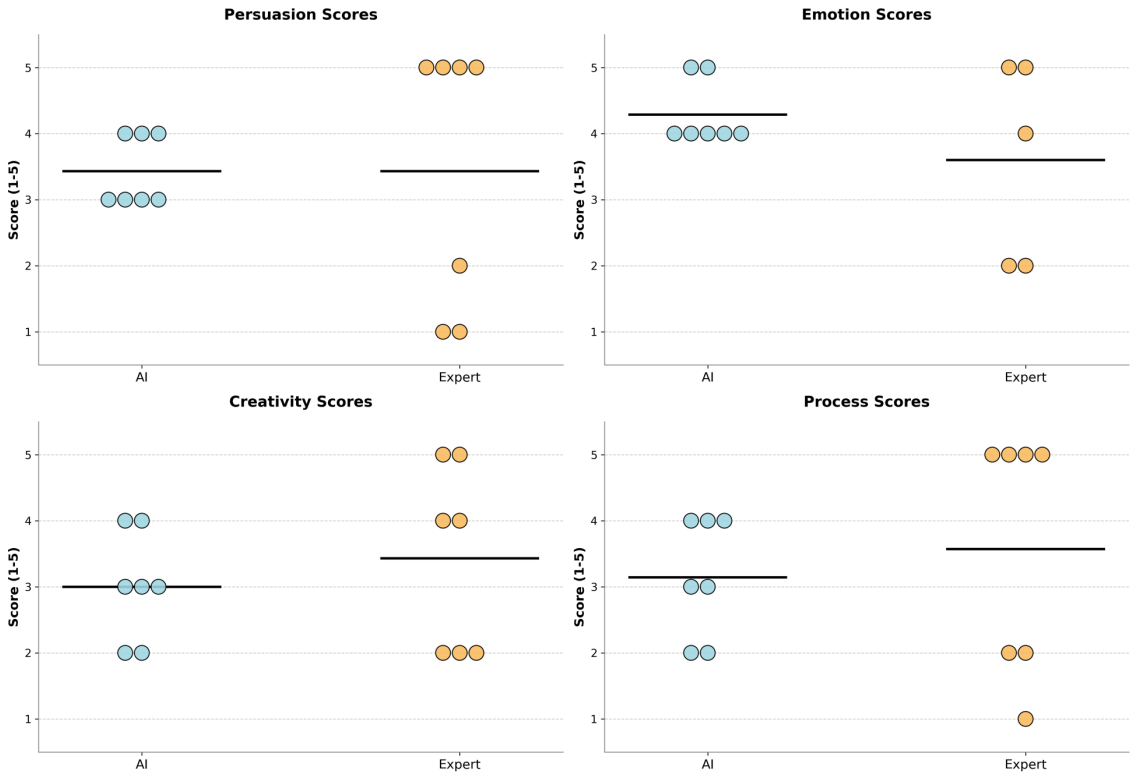


Table 4. Summary statistics of AI and Human Expert Rating of Negotiation

Dimension	Rater	Mean	Std. Dev.	Range
Persuasiveness	AI	3.43	0.53	1.0
	Expert	3.43	1.99	4.0
Emotion	AI	4.29	0.49	1.0
	Expert	3.60	1.52	3.0
Creativity	AI	3.00	0.82	2.0
	Expert	3.43	1.40	3.0
Process	AI	3.14	0.90	2.0
	Expert	3.57	1.81	4.0

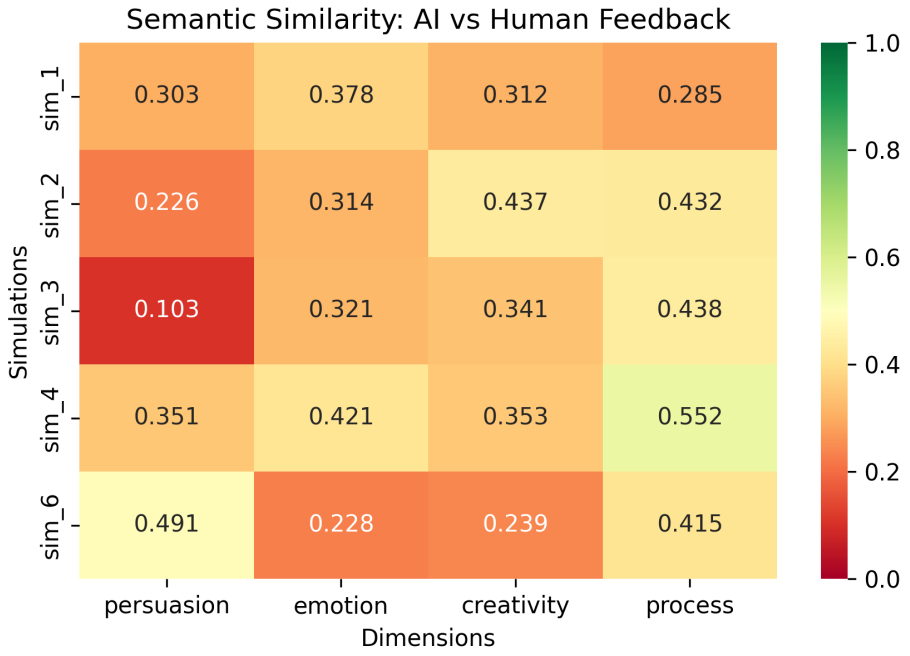
Figure 2. Swarm Plot of AI and Human Rating of Negotiation



RQ2: Comparing Expert vs AI-generated Feedback

1) Semantic similarity of Expert vs. AI feedback

Figure 3. Heat map of semantic similarity between AI and human feedback



The AI-generated feedback and human expert feedback shared little semantic

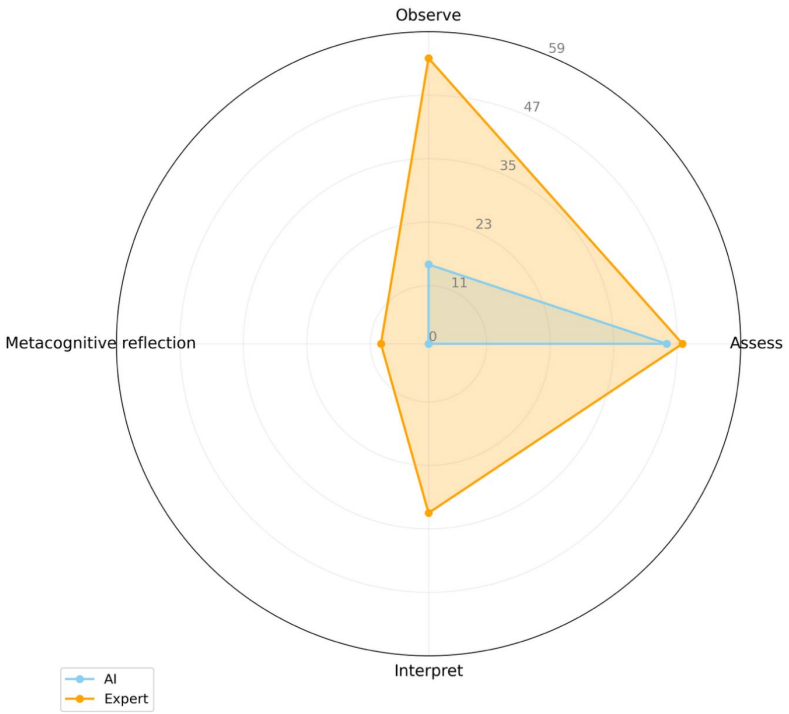
similarity, as evidenced by low overall cosine similarity (*Mean* = 0.347, Median = 0.346), considerable variability across dimensions (SD = 0.102; Range: 0.103 - 0.552), and all cases except one falling below the 0.5 similarity threshold. This quantitative finding was then triangulated with detailed qualitative analysis in the section below.

2) Qualitative analysis of Expert vs. AI feedback

Feedback approach

The qualitative analysis showed a considerable difference between AI and human experts in terms of their feedback approach, as shown in the radar chart in Figure 4. The human expert showed a balanced distribution of evaluative moves across observation (37.8%), interpretation (22.4%), assessment (33.6%), and metacognitive reflection (6.3%). The AI system had limited evaluative moves that were heavily focused on assessment (75%) and making observations (25%) without any interpretation or metacognitive reflection.

Figure 4. Radar chart of AI and human feedback approaches



Examining the codes within each larger category (see Summary in Table 5), we identified 16 distinct feedback approaches that were only observed in human experts. In terms of *observation*, the human expert observed the length of negotiation (4 instances) and wording choices (7 instances), pointed out factual mistakes (3 instances), and referred to specific lines or interactions in their feedback (15 instances). This suggests humans were able to identify and comment on granular textual elements and linguistic nuances within the context. In terms of *interpretation*, the human expert engaged in sense-making (16 instances) through: considering context (5 instances), making comparisons with other negotiation (3 instances), providing reasoning and justification for their ratings (3 instances), speculating to explain learners' approaches (3 instances), and teasing out specific strategies employed by learners (1 instance), and speculating about alternative options and outcomes (1 instance). This interpretive dimension represents a significant domain where AI feedback is currently lacking entirely. Finally, the human expert showed unique *metacognitive* awareness, with 9 instances of

metacognitive reflection on the task itself. This included acknowledging the absence of evidence for evaluation (6 instances) and specifically commenting on what the AI did during the interaction (3 instances). These findings reveal humans' distinctive ability to reflect on the evaluation process itself, which is currently missing in absent in current AI feedback.

Table 5. Comparison of codes in feedback approach

Feedback Approach	code	AI	human	difference	relationship
Assessment	Holistic assessment	9	14	5	Human more frequent
Assessment	What the learner did well	25	21	-4	AI more frequent
Assessment	What the learner didn't do well	11	13	2	Similar
Observation	Learner's disposition/attitude	4	4	0	Identical
Observation	Relational elements	3	2	-1	Similar (AI higher)
Observation	Delivery of rationale	1	1	0	Identical
Observation	Efficacy of the negotiation	0	1	1	Human only
Observation	Length of the negotiation	0	4	4	Human only
Observation	Outcome of the negotiation	2	3	1	Similar
Observation	Tone of the player	1	6	5	Human more frequent
Observation	Wording choice	0	7	7	Human only
Observation	What the AI did and learner response	4	8	4	Human more frequent
Observation	Pointing out factual mistake	0	3	3	Human only
Observation	Refer to specific line/interaction	0	15	15	Human only
Interpretation	Consider the context	0	5	5	Human only
Interpretation	Compare with other negotiations	0	3	3	Human only
Interpretation	Reasoning for rating/feedback	0	3	3	Human only
Interpretation	Speculate alternative options	0	1	1	Human only
Interpretation	Speculate learner's approach	0	3	3	Human only
Interpretation	Tease out specific strategy	0	1	1	Human only
Metacognition	Absence of evidence for evaluation	0	6	6	Human only
Metacognition	Comment on AI actions	0	3	3	Human only

Table 6. Comparison of codes in feedback content

Feedback Content	code	AI	human	difference	relationship
Relational	Advocate for oneself/being assertive	2	2	0	Identical
Relational	Create relationship with the counterpart	0	3	3	Human only
Relational	Understand others' needs and priorities	9	0	-9	AI only
Relational	Have a positive and professional tone	5	2	-3	AI more frequent
Relational	Introduce emotional dimension to the conversation	0	1	1	Human only
Relational	Maintain poise and balance	9	2	-7	AI more frequent
Relational	Show confidence and not arrogance	4	4	0	Identical
Relational	Start on the right footing	0	4	4	Human only
Relational	Use smart wording	0	1	1	Human only
Substantive	Ask strong, relevant, and polite questions	9	2	-7	AI more frequent
Substantive	Avoid haggling over number/number exchange	2	2	0	Identical
Substantive	Being specific with examples and proposals	6	1	-5	AI more frequent
Substantive	Come up with creative ideas	11	4	-7	AI more frequent
Substantive	Establish mutual benefits	5	7	2	Human more frequent
Substantive	Getting the maximum possible in the agreement	0	2	2	Human only
Substantive	Use evidence to justify	11	4	-7	AI more frequent

Feedback Content

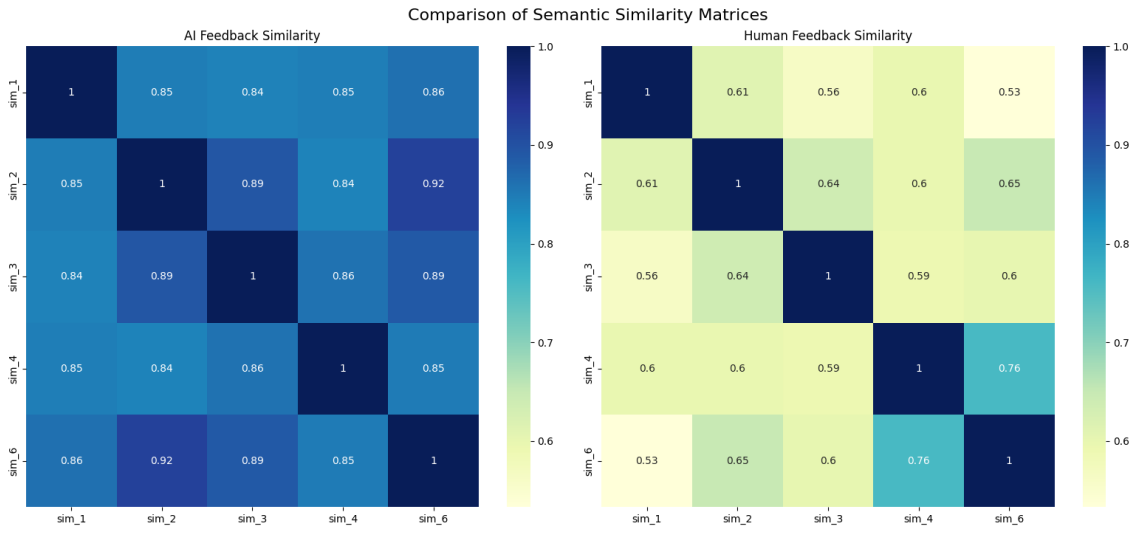
In terms of feedback content, we compared the similarities within AI-generated feedback across simulations, and separately we compared the similarities within human expert feedback across simulations (Table 6). The similarity matrices showed distinct patterns in feedback variability between humans and AI across simulations. A higher score and darker color indicate higher similarity. The AI feedback similarity matrix is predominantly in blue, which indicates that AI feedback shared high similarity scores across simulations ($M = 0.866$, $SD = 0.026$). This means that AI generates very similar feedback regardless of the simulation context. In contrast, the human feedback similarity matrix is lighter in color with substantially lower similarity ($M = 0.614$, $SD = 0.058$). This shows that human experts are more context-sensitive and adapt their feedback to the specific simulation.

The following example illustrates the human expert's contextual awareness, in which identical actions elicited different assessments and feedback depending on the situational context. In one simulation, a participant responded to the CEO's (played by the AI agent) opening question ("What's up? Is everything okay?") by saying "Everything is fine." The human expert noted: "Well, it's often pleasant to start on a positive note, but in this case things are definitely not fine. And the fact they aren't fine for [the player] or the company could set stage for making some important decisions." In another simulation, the same human expert identified a similar response from the participants as highly effective. As the human expert observed: "I raise my eyebrow over [the player's] statement, 'Things are great' (when they're not). But here it seems to set a very positive outlook that continues for the entire conversation." These contrasting evaluations of essentially identical communicative behaviors demonstrate that human expert feedback incorporates sophisticated contextual understanding that extends

beyond simple behavior/action recognition.

In addition to a lack of contextual awareness, AI also had difficulty understanding emotional nuances in interpersonal relationships. In one notable case, when a participant responded to a dismissive comment from the CEO:" I think that's a bit dismissive to call it luck," the human expert provided positive feedback noting: "The player stands their ground, keeps balanced, and chides [the CEO] firmly... 'a bit dismissive' nicely dials down some the potential heat. There are numerous such examples right up to the end." In contrast, the AI did not recognize the nuanced de-escalation of conflict with smart wording instead critiqued the player: "[the player] remained respectful and composed throughout the negotiation, even when [the CEO] was dismissive of her achievements. She did not retaliate and maintained a professional demeanor. However, she could have been more assertive in addressing George's dismissive comments about her contributions." This divergence showed that the AI did not grasp the emotional nuances in wording choices to de-escalate potential conflict that human experts deemed as effective.

Figure 5. Similarity matrices of AI-generated feedback (left) and human expert feedback (right)



Triangulating this quantitative finding with the qualitative analysis, we found

that AI followed the feedback criteria outlined in the prompt exclusively, which resulted in more formulaic feedback. In contrast, the human expert exhibited more flexibility and context-sensitivity and drew from their broader knowledge and experience to incorporate additional feedback beyond the criteria in response to the uniqueness of each transcript. The additional feedback from human experts included: creating relationships with counterparts (3 instances), starting negotiation on the right footing (4 instances), maximizing agreement value (2 instances), introducing emotional dimensions to the negotiation (1 instance), and using smart wording to de-escalate conflict (1 instance).

For example, when evaluating the same simulation where a player proposed a trial period, the AI provided structured feedback that adhered closely to feedback criteria outlined in the prompt: “[the player] showed creativity by proposing a trial period for the VP role, which balanced her aspirations with the company's need for proven performance. She also suggested a phased salary increase based on performance. However, she could have elaborated more on specific new initiatives or roles that could benefit both her and the company.” In contrast, the human expert showed sensitivity to tone and emotional implications: “The word, 'hey' is appropriately casual and implies 'I just thought of a good idea'. If the player instead said, 'Could you give me a trial period?', that might still work, but it that alone might prompt the boss to ponder pros and cons. By contrast, the player's formulation illuminates both the con (the chance it might not work) and the corresponding pro (that there's an exit option).” The human expert recognizes the strategic use of casual language to shape the interpersonal dynamics of the negotiation, which extends beyond the suggestions in the feedback criteria.

This tendency of AI's adherence to criteria producing formulaic and sometimes

contextually irrelevant feedback is shown in another example. Here, both human and AI evaluators noted creativity in the learner's approach. The AI stated: “Kim proposed a professional development plan and performance-based incentives, which showed creativity. However, she could have suggested more innovative solutions, such as a trial period for the VP role or a hybrid role that leverages her strengths while reducing her workload.” The AI seems to have mechanically included the suggestion of a trial period because it was listed as a creative idea in the prompt without necessarily considering whether it was contextually appropriate for this particular negotiation. In contrast, the human expert provided more contextually relevant feedback: “The player also does this well, expressing willingness to create a 'professional development plan with your guidance,' which potentially sets up a closer relationship between the two parties.” The human expert noticed the creative idea served the dual purpose of professional development and relationship building. This example further illustrates how human experts draw on their broader understanding of interpersonal relationships and negotiation expertise to provide feedback that is contextually relevant and goes beyond criteria that were not explicitly captured.

RQ3: An expert’s process for negotiation feedback

Our goal with the think aloud protocol was to understand how a negotiation expert approaches the process of providing feedback on a judgement-based negotiation. From the thematic analysis of the session, we deduced that our expert followed three high-level stages while reviewing the negotiation transcripts and providing feedback (Table 7).

Table 7. Three high-level stages of reviewing a negotiation transcript

Step 1: Opening lines analysis	Though the direction and tone of the negotiation can change over the course of the interactions, the

	expert dialled in on the first series of exchanges between the parties.
Step 2: Line-by-line deep dive	A highly contextualized review of each line or pairs of lines that dynamically leveraged steps such as comparison, anticipation, decoding, and speculation.
Stage 3: Holistic review	Brief synopsis of the conversation with high-level thoughts on how the negotiation was navigated

The first stage involved reading the opening statements from both the AI partner, who always started the dialogue, and the learner to gauge on what footing the learner started the negotiation. For the expert, the execution of the initial exchange set the tone or "framing" for the remainder of the discussion, even though a learner who "started off poorly could [still] do well." For example, in both transcripts, he points outs that the players started very broadly instead of being specific about their intentions of wanting a promotion and a raise. Despite the importance he placed on the opening statements, he also acknowledged and demonstrated that a learner could improve their footing even after generic language at the beginning, depending on the directions taken during the negotiation. These instances where there are shifts in the learner's positioning were explicitly highlighted. For instance, in one think aloud transcript that started with a broad opening, the expert pinpointed a shift at a specific point, "Kim is now being a very strong negotiator after setting the table."

The second stage involved a line-by-line reading of the remaining conversation followed by either a quick reaction after each quote (e.g., "I'm impressed here," "he doesn't challenge that," "good that she does that") or a longer contextual response.

From these longer responses, we gathered that the expert had a rough model of what a good conversation looked like in terms of the opening, language used, points raised, and outcomes negotiated. The model influenced the ideal responses not only for the learner but also for the AI agent. This model surfaced through the evaluative dance that consisted of the following dynamic moves:

1. Comparison: making comparisons to the approaches in other transcripts that were reviewed.
2. Anticipation: anticipating what will ensue or what the learner needs to do later on in order to have a successful negotiation within the context.
3. Decoding: attempting to make sense of the player's thought processes, implying meaning, and exploring alternatives.
4. Speculation: speculating the intention of the learner or AI agent based on the specific way a statement was delivered.

Comparison

Comparisons were made to other transcripts that were reviewed. These comparisons, sprinkled throughout the reviews, were either illustrating something that the learner did well or comparably to others or did poorly. For example, in one transcript, the learner suggests bringing in a new resource to assume the responsibilities of the old position, an idea that the boss approves and inquires as to what supports the learner will need to move forward with the plan. To this, the expert points out that some of the other transcripts also discussed interests and priorities early on, but in this negotiation, "more than in the other five we had looked at already, the boss was moving towards being collaborative, which was perceived favorably. On the contrary, in another transcript, the learner starts the conversation with a generic statement, and the expert mentions other learners being "a little more specific" (specificity is seen in a positive light).

Anticipation

Anticipating the moves a learner will make or need to make down the line in order to achieve a successful outcome typically occurs in the first half of the review. For instance, in response to one learner's request to be compensated at the same salary as the previous boss, the expert comments that the learner has set up a strong argument for the increase in pay, and yet he expects that there will be a counter from the AI agent. In another transcript, the learner is setting up the complication around which the negotiation revolves, and the expert identifies a point that was left out of the statement and assumes that it "will come up soon." The expert was aware that he was anticipating how the interaction would unfold. After the think aloud, he commented about reading the transcript with interest and in some instances, thinking about what was going to happen next.

Decoding

Decoding involved summarizing and making an attempt to understand the underlying meaning. When one learner exclaimed overcoming challenges in closing a deal and demonstrating their capabilities, the expert points out that the learner is standing up for themself and not accepting the boss' view that they took an ancillary role. In other cases, the expert expresses confusion as to why the learner made a certain statement and explains the source of the confusion.

Speculation

Speculation of intentions comes in the form of hypothesizing why a player in the negotiation might have made a particular decision. The speculation could be related to a particular word, such as when the expert deduces that the learner used the word 'fair' in hopes that the boss would be fair, or an overall scenario; for example, when the learner is outlining their responsibilities and timeline for the trial period, they mention needing

to continue working double time, which the expert presumes is an unintentional slip of the tongue. It could also come in the form of a question, as in the case when the learner offers to receive a bonus for taking on the role of the boss and defer the promotion. To the expert, it was not clear why the learner made that move since she may not be willing to take the meager bonus once push came to shove.

These evaluative moves did not follow a linear, rule-based progression. Instead, they were deployed dynamically, depending on the context surrounding the lines, the position of the lines within the larger negotiation, and the wording of the statement.

Interestingly, throughout the think aloud, the expert did not make many comments related to the explicit criteria outlined in the rubric. Only once did he bring up a provocative statement as relating to emotion. The connections to the rubric were not made until the third stage.

In the third stage, the expert did a brief holistic review of the conversation, sharing general thoughts on the learner's success navigating the relational and substantive parts of the negotiation. The review was short, broad, and did not explicitly address the criteria. For example, at the end of one transcript, the expert exclaimed that the two parties had "negotiated how to negotiate"--a crucial indicator of a successful negotiation and underscored at the beginning of the think aloud session—in the first half, as opposed to an offer-counteroffer approach common in strategy-based negotiation, but it had gone too far in that direction.

Overall, the expert's review of the transcripts looked both specifically and holistically at the conversations as a way to determine whether the negotiation as a whole was moving in an optimal or suboptimal direction. He did not strictly adhere to the criteria during the review process, but rather examined a combination of factors,

including the negotiation starting off on the right footing; the tone, attitude, and framing illustrated throughout; shifts in the dynamic over the course of the negotiation; the balancing of substantive vs. relational outcomes; the strength and relevance of questions to a particular context; and self-advocacy, among other factors. Each factor was taken in consideration dependant on the context. While speculating and anticipating how the negotiation might progress, the expert did not penalize the learner for taking the conversation in a different direction than what he might expect or do himself. This was explicitly illustrated when the expert gave a transcript two ratings of five and one six out of five, despite the learner starting with a general statement, making a confusing offer, and accepting a lower salary than what the optimal outcome would prescribe. For the expert, the learner's exceptional process and adaptive navigation appeared to hold more weight than the specific elements outlined in the criteria.

Discussion

This case study sheds light on the affordances and limitations of GenAI in assessing judgement-focused negotiation, as a hard-to-measure competency. We compared AI-generated ratings and feedback with those from a human expert with 40+ years of experience in negotiation. Our findings showed differences between the AI and the human expert in terms of ratings, feedback approach, and feedback content. These findings have implications for developing stealth assessment that leverages the affordances of GenAI while keeping its glass box transparent through evidence-centred methodology (ECD) (Mislevy et al., 2003).

Affordances and limitations of GenAI in judgement-focused negotiation assessment

Regarding ratings, our case study found that prompt-based GenAI has limited validity when used alone to assess judgement-focused negotiation. AI has the tendency to produce conservative ratings clustered in the middle (3-4), whereas a human expert was

able to clearly distinguish between strong and weak performance. This finding echoes previous research on the use of GenAI in other educational contexts, which found that GenAI was able to succeed in simple, closed-ended assessments with clear answers and benchmarks but struggles with nuanced and open-ended ones (Misiejuk et al., 2024; Shea et al., 2024; Wang et al., 2023). We found GenAI was able to identify surface-level behavior indicated in the prompt, but it does not have inherent expertise or contextual understanding for valid assessment of judgement-focused negotiation, which requires a holistic and nuanced understanding and awareness of the interplay among many human factors, such as culture, power dynamics, relationships, and context (Schneider et al., 2025). For instance, we found that AI was unable to assess the same behavior differently according to the context and struggled to grasp emotional nuances in wording choices.

In terms of the feedback approach, the AI primarily focused on assessment with limited moves to observation. In contrast, the human expert had a more balanced approach consisting of observation, interpretation, assessment, and metacognitive reflection on the task itself. For example, the human expert attempted to understand the rationale of learner behavior and deliberately withheld assessment in cases when they found insufficient evidence.

Regarding feedback content, we found that AI adhered rigidly to the criteria outlined in the prompt, which resulted in formulaic and sometimes contextually irrelevant feedback. On the other hand, the human expert exhibited flexibility to draw from their broader knowledge and expertise to provide feedback in response to the uniqueness of each simulation. This raises a tension between accuracy and creativity in setting the temperature of GenAI, which controls the randomness of AI-generated output. Low temperature can mitigate hallucination and enhance accuracy. However, for

assessment that requires contextual judgement, our study found that a low temperature, in this case, results in formulaic and decontextualized feedback that is not necessarily helpful. Increasing the temperature could potentially result in more flexible and creative feedback, but risks reducing accuracy and reliability. This tension highlights the nature of GenAI, which does not have inherent expertise and relies on pattern recognition from its training data. Future experiments are needed to locate a sweet spot for the temperature setting while recognizing the fundamental limitation of GenAI for complex skill assessment (Dede et al., 2021).

Given the affordances and limitations of AI feedback, we see the AI-generated feedback useful for novice learners to develop fluency with strategies and tactics in strategy-focused negotiation. This way, they can quickly receive standardized feedback and adjust their approach to more closely align with the given rules and principles. That said, strategies are not absolute but rather context-dependant. In a real-life negotiation, the strategy appropriate to the interaction might change mid-way. The expert was able to identify context-dependent shifts in the conversation and paid particular attention to the appropriateness of a strategy for the given context. Human expert feedback would be especially helpful for those who are more experienced negotiators looking to advance their adaptability.

A hybrid model to reconcile the black box with the glass box

Application of GenAI in stealth assessment is still at its nascent stage. A fundamental tension discussed earlier in our study was the tension between stealth assessment's "glass box" approach versus the "black box" nature of GenAI. Our findings suggest that this tension could potentially be resolved through a hybrid approach that leverages the strengths of GenAI, human expertise, and the architecture of evidence-based methodology (ECD). Stealth assessment uses ECD to establish connections between

observable behaviours to competencies, which has three components: 1) competency model (knowledge, skills, and attributes to be assessed), 2) evidence model (observable behaviours and link to targeted competency, and 3) task model (tasks and activities that elicit evidence) (Shute et al., 2022). The qualitative analysis of our study elicited a rich set of observable behaviours for judgement-based negotiation that can form part of the evidence model, such as using smart wording to de-escalate conflict, starting on the right footing (for the full list, see Table 6). The four dimensions of negotiation assessment, i.e., persuasiveness, emotion management, creativity, and process management, could form the basis of a competency model.

We suggest a staged approach to GenAI integration as part of stealth assessments rather than using GenAI alone for complex skill assessment. As we found in the study that AI adhered to the negotiation criteria, to start with, it could be leveraged to conduct low-inference work to identify observable behaviour, where complex interpretation is not required. To improve AI's contextual awareness, interpretive capacity, and engagement in metacognitive reflection, the expert think-aloud protocol has delineated sophisticated cognitive processes of a human expert, starting with opening line analysis, line-by-line deep dive, and holistic review, interwoven with dynamic evaluative moves including comparison, anticipation, decoding, and speculation. Future studies should structure AI prompts to mimic this process, using a chain-of-thought prompting strategy, to test whether it could improve AI's contextual awareness and interpretive capabilities. We suggest a gradual progression from low-inference to higher-inference tasks with continuous validation, which positions GenAI as a *component* of stealth assessment rather than a standalone evaluator. For example, Henderson and colleagues (2022) used zero-shot prompting with GenAI to produce synthetic data to enhance the performance of competency

models. This staged and integrated approach leverages the capabilities of AI while maintaining evidence-based methodology central to stealth assessment to keep the glass box transparent.

Limitations and future studies

This case study represents an initial step towards understanding the affordances and limitations of AI-generated assessment and feedback for complex skill development. The mixed-method approach allowed a close and in-depth examination of feedback generated by AI and a human expert. While the small sample size (n=7) generated rich findings to serve strong foundation for future studies, it has limited power for inferential statistics and generalizability. Future studies should expand the sample size to enable more robust statistical tests.

This study also relied on a single human expert, and thus, we were unable to establish inter-rater reliability among experts who might have different perspectives and negotiation approaches. Future studies should aim to include human experts with diverse backgrounds, cultures, and gender identities.

This study tested one Large Language Model (GPT4o) with one structured, zero-shot prompting strategy. Future research should test, improve, and validate prompting strategies (e.g., chain of thought) to enhance AI's interpretative capabilities through emulating the cognitive process of human experts as well as their feedback approach and content identified in this study across LLM models and modes (e.g., thinking mode) with various temperature settings. This would potentially allow AI to gradually take on higher-inference work in stealth assessment. The recent emergence of "thinking LLM" models with intermediate reasoning steps could potentially enhance GenAI's interpretive capability and contextual awareness (Jahrens & Martinetz, 2025; Wu et al., 2024). Future studies should evaluate GenAI's assessment and feedback for

judgement-focused negotiation using the combination of "thinking LLM" with a refined prompting strategy found in this study. Last but not least, the evidence model of ECD requires further development based on our findings, especially in terms of creating weighted scoring rules and statistical models to link observable behaviours with the competency model.

Conclusion

This case study fills a research gap in understanding the affordances and limitations of AI-generated and human feedback in a complex skill like judgement-focused negotiation. We shed light on the similarities and differences between the two types of feedback as they relate to numeric ratings on criteria, feedback approach, and feedback content. With respect to ratings, we found that, in alignment with previous studies (Misiejuk et al., 2024; Shea et al., 2024; Wang et al., 2023), AI tends to produce ratings clustered in the middle of the range, whereas a human expert gives ratings within the full range available, delineating between strong and weak performance. Additionally, in approach and content, AI displays more rigid outputs, focusing primarily on assessment with some observation and sticking strictly to the prompted criteria, while the human expert also heavily draws upon interpretation and metacognitive analysis, paying particular attention to dynamic shifts in context during the negotiation. Through the think-aloud, we extracted the cognitive processes of an expert when reviewing a negotiation and providing feedback to the learner. While this case study is limited in its small scale and focuses on the feedback methodology of an expert who espouses adaptability in his teaching of negotiation, the findings are an important first step towards better understanding when and how to leverage AI versus human versus a combination of AI-human feedback.

The findings caution against using GenAI alone for complex skill assessment,

especially when the goal is to build learners' judgment and adaptability. For this reason, we recommend a staged approach to GenAI integration as part of stealth assessments. Given that the technology strictly adheres to criteria, in the first phase it could be deployed for low-inference work pinpointing specific, formulaic observable behaviours that do not need complex, judgment-oriented interpretation. Before moving to the second phase, we suggest exploring a chain-of-thought prompting strategy, such as the one illustrated in the think-aloud protocol, to test whether AI's contextual awareness and interpretive capacity can be improved. Our findings indicate that further research is necessary to keep the glass box of stealth assessment transparent as integrations with GenAI are considered.

References

Braun, V., & and, V. C. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Chien, C.-C., Chan, H.-Y., & Hou, H.-T. (2024). Learning by playing with generative AI: Design and evaluation of a role-playing educational game with generative AI as scaffolding for instant feedback interaction. *Journal of Research on Technology in Education*, 1–20. https://doi.org/10.1080/15391523.2024.2338085

Deale, D., & Pastore, R. (2014). Evaluation of simSchool: An Instructional Simulation for Pre-Service Teachers. *Computers in the Schools*, *31*(3), 197–219. https://doi.org/10.1080/07380569.2014.932650

Dede, C., Etemadi, A., & Forshaw, T. (2021). *Intelligence Augmentation: Upskilling Humans to Complement AI*. Next Level Lab, Harvard Graduate School of Education.

Dinnar, S. “Mooly,” Dede, C., Johnson, E., Straub, C., & Korjus, K. (2021). Artificial Intelligence and Technology in Teaching Negotiation. *Negotiation Journal*, *37*(1), 65–82. https://doi.org/10.1111/nejo.12351

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Hatano, G., & Oura, Y. (2003). Commentary: Reconceptualizing School Learning Using Insight From Expertise Research. *Educational Researcher*, *32*(8), 26–29. https://doi.org/10.3102/0013189X032008026

Henderson, N., Acosta, H., Min, W., Mott, B., Lord, T., Reichsman, F., Dorsey, C., Wiebe, E., & Lester, J. (2022). *Enhancing Stealth Assessment in Game-Based Learning Environments with Generative Zero-Shot Learning*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED624031

Jahrens, M., & Martinetz, T. (2025). *Why LLMs Cannot Think and How to Fix It* (No. arXiv:2503.09211). arXiv. https://doi.org/10.48550/arXiv.2503.09211

Ma, Zi., Mei, Y., Bruderlein, C., Gajos, K. Z., & Pan, W. (2024). *“ChatGPT, Don’t Tell Me What to Do”: Designing AI for Context Analysis in Humanitarian Frontline Negotiations* (No. arXiv:2410.09139). arXiv. https://doi.org/10.48550/arXiv.2410.09139

Mell, J., & Gratch, J. (2016). IAGO: Interactive Arbitration Guide Online (Demonstration). *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 1510–1512.

Misiejuk, K., Kaliisa, R., & Scianna, J. (2024). Augmenting assessment with AI coding of online student discourse: A question of reliability. *Computers and Education: Artificial Intelligence*, *6*, 100216. https://doi.org/10.1016/j.caeai.2024.100216

Mislevy, R. J., Steinberg ,Linda S., & and Almond, R. G. (2003). Focus Article: On the Structure of Educational Assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Movius, H. (2008). The Effectiveness of Negotiation Training. *Negotiation Journal*, *24*(4), 509–531. https://doi.org/10.1111/j.1571-9979.2008.00201.x

Murawski, A., Ramirez-Zohfeld, V., Mell, J., Tschoe, M., Schierer, A., Olvera, C., Brett, J., Gratch, J., & Lindquist, L. A. (2024). NegotiAge: Development and pilot testing of an artificial intelligence-based family caregiver negotiation program. *Journal of the American Geriatrics Society*, *72*(4), 1112–1121. https://doi.org/10.1111/jgs.18775

Rahimi, S., & Shute, V. J. (2024). Stealth assessment: A theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. *Educational Technology Research and Development*, *72*(5), 2417–2441. https://doi.org/10.1007/s11423-023-10232-1

Rose, C. (2025). *SAB member professor Carolyn Rosé: "As we clarify what the frontier of capability is both for humans and for AI, it will be more clear where the sweet spot is for synergy" | University of Oulu* [Interview]. https://www.oulu.fi/en/news/sab-member-professor-carolyn-rose-we-clarify-what-frontier-capability-both-for-humans-and-for-ai-it

Saldaña, J. (2025). *The Coding Manual for Qualitative Researchers*. SAGE Publications Ltd.

Schneider, J., Haag, S., & Kruse, L. C. (2025). Negotiating with LLMs: Prompt Hacks, Skill Gaps, and Reasoning Deficits. In H. Plácido Da Silva & P. Cipresso (Eds.), *Computer-Human Interaction Research and Applications* (Vol. 2371, pp. 238–

259). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-83845-3_15

Shea, R., Kallala, A., Liu, X. L., Morris, M. W., & Yu, Z. (2024). *ACE: A LLM-based Negotiation Coaching System* (No. arXiv:2410.01555). arXiv. https://doi.org/10.48550/arXiv.2410.01555

Shute, V. J. (2009). Simply Assessment. *International Journal of Learning and Media*, *1*(2), 1–11. https://doi.org/10.1162/ijlm.2009.0014

Shute, V. J., Lu, X., & Rahimi, S. (2022). Stealth Assessment. In *Stealth Assessment*. Routledge. https://doi.org/10.4324/9781138609877-REE58-1

Strauss, A. L. (1990). Systematic Coding in Qualitative Research. *BMS: Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique*, *27*, 52–62.

Wang, R. E., Wirawarn, P., Goodman, N., & Demszky, D. (2023). *SIGHT: A Large Annotated Dataset on Student Insights Gathered from Higher Education Transcripts* (No. arXiv:2306.09343). arXiv. http://arxiv.org/abs/2306.09343

Wheeler, M. (2006). Is Teaching Negotiation Too Easy, Too Hard, or Both? *Negotiation Journal*, *22*(2), 187–197. https://doi.org/10.1111/j.1571-9979.2006.00094.x

Wheeler, M. (2013). *The art of negotiation: How to improvise agreement in a chaotic world* (First Simon&Schuster hardcover edition.). Simon & Schuster.

Wheeler, M. (2021). Introduction to Special Issue: Artificial Intelligence, Technology, and Negotiation. *Negotiation Journal*, *37*(1), 5–12. https://doi.org/10.1111/nejo.12352

Wu, T., Lan, J., Yuan, W., Jiao, J., Weston, J., & Sukhbaatar, S. (2024). *Thinking LLMs: General Instruction Following with Thought Generation* (No. arXiv:2410.10630). arXiv. https://doi.org/10.48550/arXiv.2410.10630

Appendices

Appendix 1: Think-aloud protocol

Think-aloud Protocol with Human Negotiation Expert

Your task is to provide a numerical assessment (between 1 and 5, with 5 being the highest performance) for each of the four dimensions: 1) Persuasiveness, 2) Creativity, 3) Emotional management; 4) Managing the negotiation process. Provide feedback to the learner for each of the four dimensions, as you normally would in your classes.

As you engage in this task, try to say everything that comes to your mind while you are rating the transcript of negotiation.

- Speak all thoughts, even if they are unrelated to the task;
- Not try to plan out what to say;
- Imagine you are alone and speak to yourself;
- Speak continuously. We will also remind you to keep talking

1. *One-minute of silent reflection*

Do you have any additional thoughts about the rating process? /Any additional thoughts came to you?

2. *Immediate follow-up/ retrospective Reporting*

- How did you feel when you were rating this transcript? (affective dimension of human judgement)
- Why stress these specific elements/dimensions? /Why are they important? (rationale of judgement)
- Choose one thing the learner did well, and one thing the learner could work on and give them some feedback. Why did you choose these as the foci? (choice-making)
- Did any of your ratings feel intuitive rather than analytical? Describe that intuition? (intuition vs explicit reasoning can you describe that analytical process?