Voice for the Voiceless: An LLM-powered Devil's Advocate for AI-mediated Communication in Power-imbalanced Groups

ANONYMOUS AUTHOR(S)

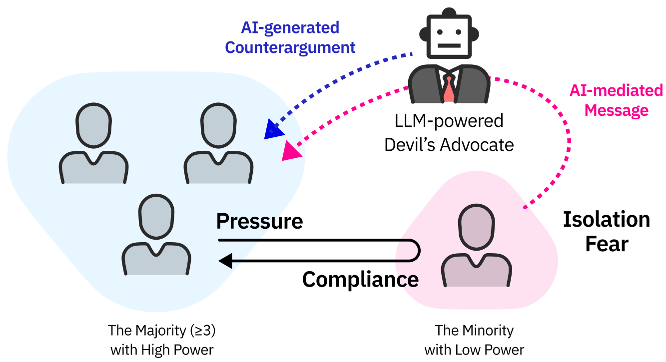


Fig. 1. LLM-powered Devil's Advocate system mediates between majority and minority group members, presenting minority views through AI-generated counterargument to promote balanced group discussions.

Minority opinions are often suppressed in power-imbalanced group decision-making due to social pressure to comply with the majority. To better mediate majority-minority interactions, we developed an LLM-powered Devil's Advocate system which fostered a group's attention to minority views by either presenting AI-generated counterarguments or delivering AI-rephrased minority opinions. We conducted a mixed-method experiment with 96 participants divided into 24 groups to compare minority members' perceived safety and satisfaction in three conditions (baseline, AI-counterargument, AI-mediated paraphrasing). Our findings show that AI counterarguments fostered a flexible atmosphere and enhanced satisfaction, while AI-mediated messaging unexpectedly decreased psychological safety and satisfaction for minorities despite increasing participation. Trade-offs emerged between anonymity and recognition. Seniors maintained consistent experiences, while juniors' experiences varied significantly based on the AI's role. Based on these results, we discuss insights and ethical implications for designing LLM-based agents that can support minorities in more equitable and power-imbalanced group decision-making.

CCS Concepts: • **Human-centered computing → Collaborative interaction**; **Collaborative and social computing systems and tools**; **Empirical studies in HCI**.

Additional Key Words and Phrases: AI-mediated Communication; AI-assisted Decision-making, Group Dynamics, Compliance, LLM

ACM Reference Format:

Anonymous Author(s). 2018. Voice for the Voiceless: An LLM-powered Devil's Advocate for AI-mediated Communication in Power-imbalanced Groups. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 52 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Group discussion processes are a cornerstone of effective collaboration in various domains, from business and healthcare to education and governance [37, 59, 84, 100]. These processes harness the collective intelligence of multiple individuals, often leading to more considerate choices, judgments, estimates, and solutions compared to those proposed by a single individual [21, 87, 91]. For instance, groups solve complex logic problems more efficiently, with members subsequently performing better on similar tasks individually after group learning experiences [62]. Students who take exams in groups tend to achieve better grades and retain more information than those who study alone [94]. Medical teams can make more accurate diagnoses than individual doctors [29], and the collaborative efforts of scholars result in higher-quality research outcomes than solo endeavors [92]. The advantages of group decision-making include using diverse knowledge and perspectives, increased creativity, and the potential for more robust and well-rounded decisions [5, 36]. By leveraging group members' diverse skills, experiences, and insights, these processes can lead to better problem-solving and innovation. The inherent potential of group decision-making lies in its ability to harness the collective wisdom of its members, making it a widely used and highly valued approach in many collaborative and organizational settings.

However, collective decision-making is not without its drawbacks. Social influence and power dynamics can significantly impact the quality of group decisions by suppressing minority opinions [21]. Compliance, where group members publicly align with the majority despite private disagreement, is a prevalent issue [43]. Majority influence typically increases group consensus, whereas minority influence preserves individuality and fosters innovation [21]. Nevertheless, minorities often conform to the majority. Conversion theory suggests that individuals undergo a 'comparison process' to determine whether to join the majority, as being part of the majority group is often more rewarding due to control over resources and decision-making power [67]. As a result, they choose to conform to the majority and are reluctant to voice different opinions. Regarding social power, responses to coercive power include compliance, identification, and internalization, with compliance being the initial reaction where individuals accept those in power [43–45]. These dynamics can suppress the voicing of new opinions by powerless minorities, reducing the likelihood of considering diverse perspectives and increasing the risk of groupthink, where the desire for consensus overrides alternative viewpoints [40–42]. While group decision-making offers many advantages, the interplay of social influence and power can lead to compliance and conformity, ultimately hindering the expression of diverse opinions and undermining the decision-making process.

The devil's advocate method improves group decisions by challenging majority views, stimulating discussion, and reducing groupthink [61, 63, 70, 79, 82]. The devil's advocate technique is known to encourage discussion [78, 80, 81, 83]. Still, it lacks authenticity. It can threaten the advocate's group acceptance [39, 70, 77]. To address these limitations, human-computer interaction (HCI) researchers have explored AI-assisted decision-making [7, 49, 52, 53, 93, 97, 98], Human-AI Teams [15, 64, 68, 104], and AI agents that support group discussions [12, 16, 105]. These AI agents can act as neutral facilitators [47, 48], raise counterarguments[12], and participate in discussions on equal footing with human members [105]. However, they have rarely been used to directly support minority individuals in small group interactions due to concerns about causing team discomfort [17, 25, 26, 38]. Some AI-mediated communication approaches attempt to paraphrase anonymous contributions to reduce re-identification risks [88]. However, most of these approaches rely on humans to make the final decisions, with AI playing a supporting role [2, 7, 55], and there is limited exploration of AI agents that represent human opinions as if opinions were their own. These systems aim to prevent groupthink by encouraging minority participation and allowing groups to consider diverse opinions. For example, an LLM-based agent has been developed to overcome the limitations of traditional devil's advocate [12]. Still, it struggles with real-time participation in fast-paced conversations, and its generalized counterarguments are often ineffective. Moreover, the impact of AI agents in complex group dynamics involving social influence and social power remains understudied [34, 35].

To complement existing approaches, we aim to address the gap in improving group decision-making in complex, power-imbalanced group dynamics by using an AI agent as a principal to represent the minority. Our research investigates how an LLM-based devil's advocate agent, capable of representing minority opinions, influences psychological safety, opinion expression, and perceived satisfaction of the decision-making process and outcome in such settings. Specifically, we explore four research questions:

- **RQ1.** How does the LLM-powered devil's advocate affect perceived psychological safety and marginalization?
- **RQ2.** How does the LLM-powered devil's advocate affect engagement and contribution patterns in group chat discussions?
- **RQ3.** How does the LLM-powered devil's advocate affect participant satisfaction with decision-making processes and outcomes?
- **RQ4.** How do the two types of LLM-powered devil's advocates affect system experience?

We conducted a mixed-methods experiment with 96 participants divided into 24 groups of four members to answer these research questions. We employed a mixed experimental design, with Participant Type (senior/majority with high power vs. junior/minority with low power) as a between-subjects variable and Communication Condition as a partially within-subjects variable. Each participant experienced two conditions: the baseline condition (A) and either Condition B (an LLM-powered Devil's Advocate generating counterarguments) or Condition C (an LLM-powered Devil's Advocate with AI-mediated messaging). Each group comprised three high-power majority members (seniors) and one low-power minority member (junior), with roles randomly assigned. In Condition C, the minority member could privately send messages to the AI system, which paraphrased and presented these opinions as its own, ensuring anonymity. In contrast, the AI independently generated counterarguments to group discussions in Condition B. Results indicated that the AI-generated counterarguments in Condition B fostered a flexible atmosphere and enhanced participant satisfaction. Conversely, in Condition C, while AI-mediated messaging facilitated more discussion, it unexpectedly decreased psychological safety and satisfaction for minority members. These findings offer critical insights into the complexities of leveraging AI-mediated communication to amplify minority voices in group decision-making, highlighting trade-offs between anonymity and recognition and the nuanced challenges of designing AI systems for power-imbalanced group dynamics.

This study makes several key contributions to the fields of human-computer interaction and group decision-making. First, we demonstrate the contrasting effects of different LLM-powered Devil's Advocate approaches in power-imbalanced group settings. While AI-generated counterarguments foster flexible discussion atmospheres and enhance overall satisfaction, AI-mediated minority messaging, despite increasing participation, unexpectedly decreases psychological safety and satisfaction among minority members. These findings reveal critical insights about the complexities of using AI to support minority voices. Second, we provide empirical evidence on how AI interventions distinctly affect majority and minority members' experiences, particularly highlighting how seniors maintain consistent satisfaction levels while juniors' experiences vary significantly across conditions. This includes important trade-offs between anonymity and recognition and unintended consequences such as increased cognitive load and reduced perceived legitimacy of minority contributions. Third, we extend the understanding of AI's role as a principal actor in mediating group opinions, offering insights into how such systems can both help and potentially hinder minority participation in group decisions. Finally, we contribute to broader discussions on designing equitable AI systems by addressing the complex interplay of social influence, power hierarchies, and group cohesion. Our findings provide actionable insights for developing AI systems that support diverse perspectives and effectively navigate the nuanced challenges of power-imbalanced group dynamics to foster more inclusive decision-making environments.

2 RELATED WORK

2.1 The Impact of Social Influence and Power on Group Decision-making

Group decision-making leverages collective intelligence to produce superior outcomes across various domains [29, 62, 92], but these processes are significantly shaped by social influence and power dynamics [43, 67]. Social influence theory suggests that individuals tend to adjust their behavior to meet social demands, with majority opinions exerting particularly strong pressure on those with less power in the group. Moscovici's conversion theory specifically explains that multiple influences trigger a comparison process resulting in compliance - a form of conformity where individuals outwardly agree while maintaining private disagreement [67]. This compliance is typically direct, immediate, and temporary, serving as a coping mechanism in power-imbalanced situations rather than reflecting genuine belief change.

Power dynamics become especially problematic in hierarchical settings where power imbalances are formalized through reward and legitimate power structures [22]. Kelman's framework provides particular insight here, identifying compliance as an initial response to power where individuals conform primarily to avoid repercussions or gain rewards, rather than from genuine conviction [43]. This dynamic is especially evident among minority members, who are often treated as outgroup members and experience isolation. The effect is particularly pronounced when the size disparity between majority and minority groups is substantial. The resulting self-censorship triggers a cascade of negative effects: as minority voices are silenced, groups lose access to diverse perspectives that could enhance decision quality, ultimately leading to groupthink - where the desire for consensus overrides critical evaluation of alternatives [40–42].

Traditional approaches to addressing these challenges include the devil's advocate technique, where a group member is assigned to argue against prevailing opinions [61, 63, 70, 79, 82]. While this approach can enhance decision quality by promoting divergent thinking and surfacing alternative viewpoints, its effectiveness is limited by concerns about the authenticity of dissenting arguments and potential threats to the psychological safety of the designated advocate [39, 70, 77]. Within the context of Human-Computer Interaction, our research explores how AI-mediated communication might overcome these limitations by providing a psychologically safer channel for minority opinions while maintaining the benefits of devil's advocacy, thereby offering a new pathway for balancing power dynamics in group decision-making.

2.2 AI-Enhanced Approaches to Improving Group Decision-Making

The integration of artificial intelligence into group decision-making has evolved from individual interaction studies [49] to examining complex group-level dynamics [11, 12, 46, 60, 105]. While AI can function as a neutral facilitator or provide counterarguments or questions to enhance critical thinking [12, 14], significant challenges persist. Zheng et al. found that AI agents often remain peripheral in group dynamics due to their limited ability to navigate social nuances [105]. Additionally, groups tend to over-rely on AI-generated recommendations [11], potentially diminishing human contributions. These limitations could become particularly significant when considering power imbalances and minority voices in group settings.

While researchers have tried to solve various problems in AI-assisted decision-making such as explainable AI to reduce overreliance [7] and adaptive designs [105], the potential for AI systems to effectively advocate for marginalized individuals in real-time group interactions remains largely unexplored. Supporting minority voices through AI-mediated communication presents unique challenges that extend beyond technical capabilities. Hwang et al. noted that existing interventions often inadvertently isolate minority individuals by either overemphasizing their marginalization or failing to address their specific needs [38]. Our research addresses this gap by introducing an LLM-powered Devil's Advocate system that strategically represents minority perspectives without compromising group cohesion. This approach builds on previous AI-mediated communication approaches [23, 30, 95] while specifically targeting the challenges of power dynamics and minority voice representation in group decision-making.

2.3 Existing Approach of AI-Mediated Communication

AI-mediated communication(AIMC) is defined as "mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals" [30]. Existing AIMC systems have predominantly focused on AI augmenting text communication, such as smart replies or word suggestions, often enhancing communication efficiency while introducing new dynamics into interpersonal interactions [23, 30]. While these systems have demonstrated impacts on communication tone and trust dynamics between communicators, they have also raised concerns about undermining user agency and authenticity as AI takes an increasingly proactive role in shaping content [33, 65, 73, 76].

Recent frameworks identify several distinct forms of AIMC, including AI-generated content relayed by humans, selective communication of AI findings, AI paraphrasing human input, and AI independently mediating multi-party communication [16, 23, 30, 88, 95, 96]. Among these, the form where AI re-presents human speech as its own—positioning the AI as a social actor in line with the CASA paradigm [69]—remains particularly underexplored. Our research addresses this gap by introducing an LLM-powered Devil's Advocate that mediates minority voices in group decision-making, extending beyond traditional AIMC's focus on communication efficiency to address fundamental power dynamics in group settings.

3 METHOD

3.1 Overview of Study Design

This study employs a mixed experimental design, with Participant Type (senior(majority with high power) vs. junior(minority with low power)) as a between-subjects variable and Communication Condition as a partially within-subjects variable. Each participant experienced two conditions: the baseline condition A and either condition B (LLM-Powered Devil's Advocate) or condition C (LLM-Powered Devil's Advocate with AI-mediated message). This design was chosen to avoid potential demand characteristics from experiencing conditions B and C. Each group consisted of four participants, with three assigned to the high-power majority condition(senior role) and one to the low-power minority condition(junior role). Both group composition and individual roles were randomly assigned. To control for order effects, both the sequence of conditions and the tasks were randomized.

3.2 Participants

The study involved 96 Korean participants (chosen as a multiple of 8 to facilitate randomization of conditions and participant types), divided into 24 groups of 4, with each group comprising three high-power majority members and

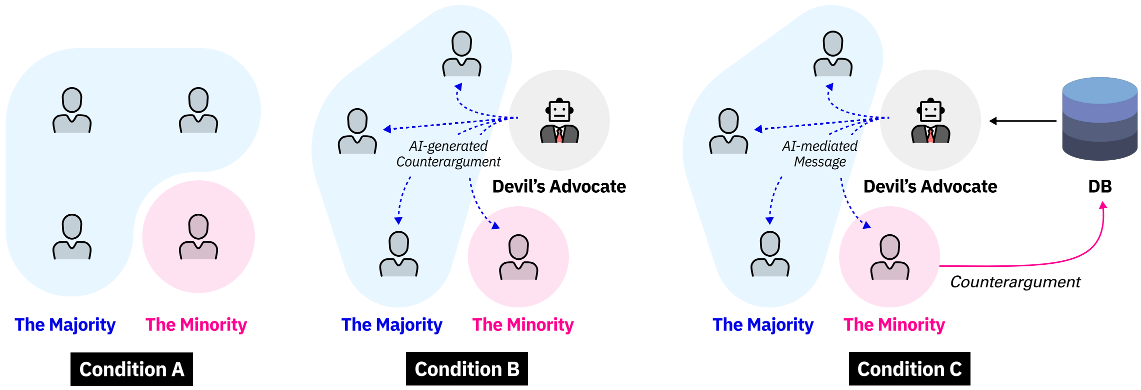


Fig. 2. Experimental Conditions: Condition A shows the baseline group chat configuration with majority (blue) and minority (pink) participants. Condition B introduces an AI-powered Devil's Advocate that generates rebuttals during group discussions. Condition C extends this by enabling the minority member to privately send counterarguments to the AI system, incorporating them into its responses while maintaining anonymity.

one low-power minority member. Participants were recruited online. Inclusion criteria required participants are Korean and over the age of 18. Participants were also required to have previous experience in group decision-making tasks and online chat experience. During recruitment, participants were informed about the anonymous nature of the experiment. At the beginning of each session participants were briefed on the procedures and reminded of their right to withdraw at any time. All data collected was coded and de-identified to maintain anonymity, and participants noticed it. If any participant withdrew or did not consent, the remaining group members received 1,000 KRW as compensation, and the session was canceled.

Demographic data collected from participants included age (M=26.60, SD=5.21, range = 19-42), and gender (61F, 35M). Education levels varied among participants, with 46.9% holding bachelor's degrees, 19.8% holding master's degrees, 15.6% having some college education, 13.5% with high school or equivalent education, and 4.2% holding doctorate degrees. Participants reported an average of 2.50 years (SD=3.15) of professional work experience. Additional background information was gathered on participants' familiarity with AI (M=4.83, SD=1.48 on a 7-point Likert scale), previous experience with group decision-making (M=5.01, SD=1.41), and prior experience with online collaboration (M=4.39, SD=1.83). Notably, 53.1% of participants reported previous experience using AI in group contexts. Participants were randomly assigned to either the high-power majority or low-power minority roles within their groups.

3.3 Experimental Treatments

This study examines three communication conditions and two participant types in group decision-making tasks. Each participant experienced two of the three conditions: the baseline condition (A) and either condition B or C.

- ***Condition A: Baseline*** In the baseline condition, participants engage in standard online group chat discussions without any additional features.
- ***Condition B: LLM-Powered Devil's Advocate*** An AI system participates in the group discussion by automatically generating counterarguments after every eight messages exchanged. The system avoids repeating previously discussed topics to maintain meaningful contributions to the discussion.

- ***Condition C: LLM-Powered Devil's Advocate with AI-mediated Messaging*** This condition functions similarly to Condition B but includes an additional feature known only to the minority member: the ability to send messages to the AI system privately. The system then paraphrases these messages and presents them as its own opinions, maintaining the minority member's anonymity. When the minority member doesn't provide input, the system generates counterarguments, as in Condition B.

Participant Types with Power Dynamics Each group consisted of three high-power majority members (seniors) and one low-power minority member (junior). Compliance was established through two mechanisms: power assignment and majority-minority composition. Legitimate power was established through role titles (senior vs. junior), while reward power was implemented through compensation structure [22, 34, 35]. At the beginning of the experiment, participants were told that the reward for seniors was a 20,000 KRW gift card, and the reward for juniors was a 15,000 KRW gift card. Participants were informed that, based on their assessment of the junior's contribution, the senior could give the junior up to an additional 5,000 KRW gift card (although all participants ultimately received equal compensation of 20,000 KRW gift card). The 3:1 ratio was chosen based on research showing that the majority influence peaks at three members [1, 4, 21, 28], creating optimal conditions for studying compliance dynamics.

3.4 Experimental Procedure

Prior to commencing the experiment, participants underwent a comprehensive briefing process focused on data anonymity and consent procedures. They were informed that any non-consent or non-response would necessitate experiment cancellation, with the remaining participants receiving a base compensation of 1,000 KRW. To maintain anonymity while fostering group dynamics, participants selected their own nicknames - a practice that research has shown strengthens social identity and enhances group cohesion through depersonalization in online environments [50]. The total duration, including all activities and interviews, was approximately 1 hour and 30-45 minutes, allowing for comprehensive data collection. The experimental framework utilized a dual-chatting platform communication structure to simulate the experimental environment. KakaoTalk served as the primary platform for general communication and team-building activities, while a custom-designed experimental chat environment hosted the formal decision-making tasks. Following group assignment and role distribution, participants engaged in a 10-minute ice-breaking session on KakaoTalk, collaboratively developing a team name and slogan. This initial activity was strategically designed to establish team cohesion while maintaining the prescribed power dynamics between senior and junior roles (Figure 3).

The core decision-making phase incorporated two carefully selected tasks that built upon previous AI-assisted group decision-making research while maintaining strong relevance to corporate contexts because we treat legitimate power with senior & junior roles. The first task involved evaluating employee profiles for a team leader promotion, while the second required analyzing potential contract partners through company performance metrics. In particular, the employee profile assessment task was adapted from a previous study, with minor modifications to fit the context of this study, and the contract partner selection task was created and utilized in a similar company context. Each task presented participants with three distinct options structured to create clear decision-making tensions: a stable but unchallenging option appealing to risk-averse decision-makers, a challenging but unstable option offering higher potential returns, and a neutral compromise option balancing both extremes. Participants were given situational context based on their roles rather than explicit persona assignments to design a natural majority-minority dynamic. We tried to drive natural immersion rather than role-playing-like acting: The seniors were guided that they were in a situation where they had

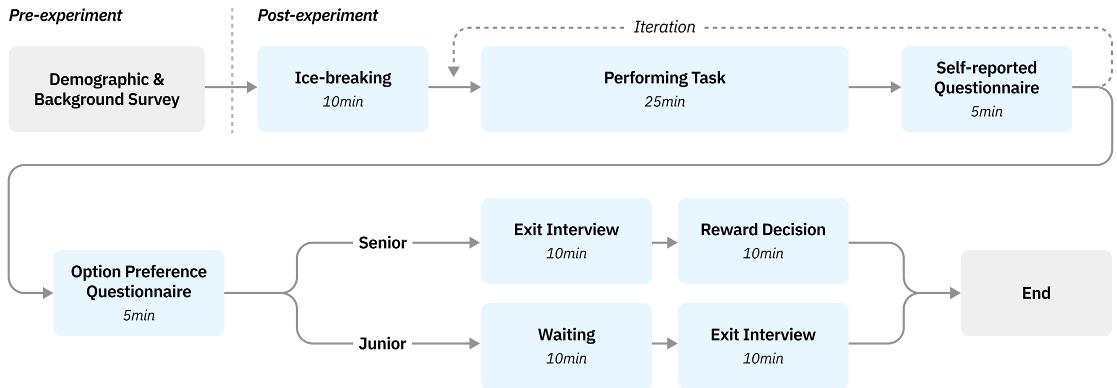


Fig. 3. Overview of the experimental procedure: including pre-experiment surveys, ice-breaking, iterative decision-making tasks, post-task questionnaires, and role-specific exit interviews.

to prioritize the stability and reputation of the organization. In contrast, the juniors were guided that they were in a situation where they had to prove their performance by making more ambitious choices.

Each decision-making task was allocated 20 minutes in the experimental chat environment, followed by a 5-minute questionnaire period on KakaoTalk. These questionnaires employed 7-point Likert scales to assess critical factors, including perceived psychological safety, decision-making satisfaction, and cognitive load. Additionally, participants recorded their personal preferences for each task's options, providing data on how effectively the contextual factors influenced their decision-making processes. The experiment concluded with strategically separated exit interviews conducted via Zoom - a 10-minute session with the three senior members and a private 10-minute interview with the junior member. During their exit interview, senior members were tasked with making an additional reward allocation decision, which was unknown to the junior members. While ultimately not affecting the final compensation (all participants received 20,000 KRW), this decision reinforced the reward power dynamics throughout the experiment.

3.5 Implementation of the Experimental System

We developed an online chat environment implemented with TypeScript (React) for the frontend and Python (FastAPI) for the backend, where four participants (three seniors and one junior) held real-time text-based discussions using pseudonyms. In this environment, an LLM-powered devil's advocate would periodically summarize the public opinion, issue a counterargument to that opinion (condition B & C), or paraphrase a direct message from a participant in a junior role and present it as their own opinion. The core LLM (OpenAI GPT-4o) interacted with these agents via a Retrieval-Augmented Generation pipeline, ensuring that its responses were context-sensitive and responsive to the current dialogue.

Drawing on findings that LLMs often struggle to access mid-conversation information in lengthy contexts, we employ a multi-agent architecture to maintain clarity of "public opinion" and encourage constructive discourse (Figure 4): ***(A) Summary Agent*** – Consolidates emerging consensus to overcome LLM limitations in retaining mid-dialogue content [56]. ***(A') Paraphrase Agent*** – Responds exclusively to direct messages from juniors, rearticulating their dissenting views as though originating from the AI itself. These messages are stored in a database with an *isUsed* property, and the Paraphrase Agent retrieves only those entries for which *isUsed* is *false*; it then sets *isUsed* to *true*, paraphrases the

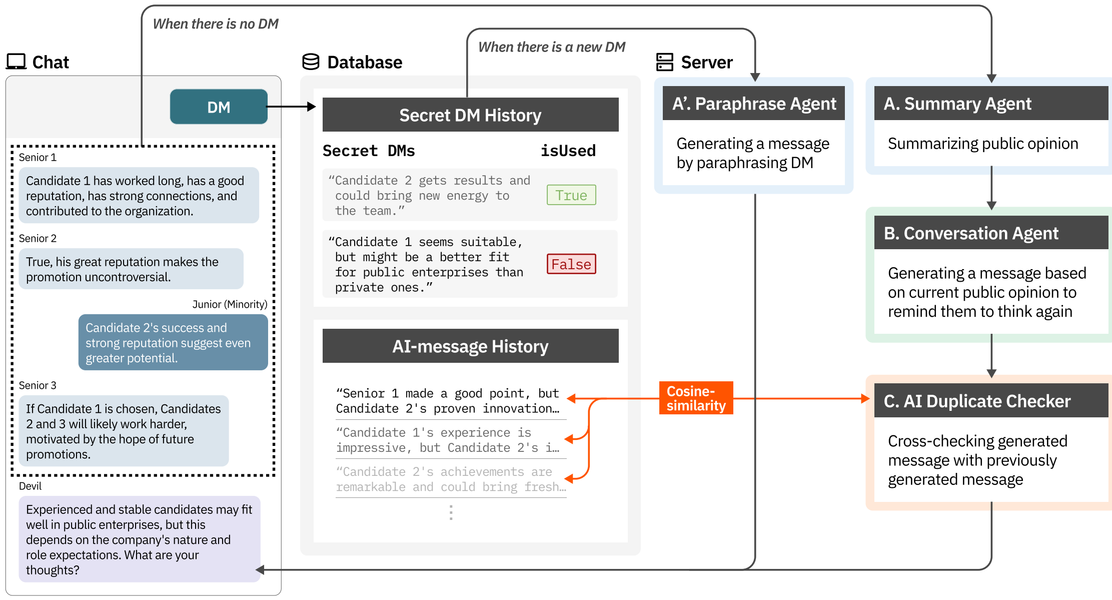


Fig. 4. System Overview: Our system architecture shows the interaction flow between the chat interface, database, and server components. The system processes both direct messages (DMs) and public chat through four main agents: (A) Summary Agent for public opinion analysis, (A') Paraphrase Agent for reformulating minority views, (B) Conversation Agent for generating contextual counterarguments, and (C) AI Duplicate Checker for ensuring message novelty through cosine-similarity comparison.

content, and outputs it as system-generated text. ***(B) Conversation Agent*** – Encourages alternative perspectives by first empathizing with the other person's point of view and then offering a gentle counterargument using a persuasive Socratic style. ***(C) AI Duplicate Checker*** – Identifies repetitive content by calculating semantic similarity between sentence embeddings generated using the 'paraphrase-multilingual-MiniLM-L12-v2' model on an NVIDIA A6000.

To ensure balanced participation, the AI agent was designed to intervene once after approximately eight human turns, providing sufficient opportunity for each participant to speak twice (roughly two turns each) before an AI intervention, excluding trivial exchanges like greetings or short agreements. These design choices reflect our design rationale of (1) adopting a persuasive, empathetic style that acknowledges others' perspectives before introducing counterarguments [90], (2) leveraging Socratic questioning to stimulate collective critical thinking without over-relying on AI-supplied solutions [14], and (3) incorporating a non-repetition mechanism to avert user frustration [66, 102]. The live chat environment with anonymous participants allows a minority with a different opinion to communicate their views to the group with complete anonymity and a sense of psychological safety. As a result, it aims to facilitate the consideration of diverse opinions and prevent groupthink in the group decision-making process.

3.6 Measurement

We examined how two LLM-powered Devil's Advocates—one generating counterarguments and another representing minority views—affect group dynamics. We evaluated self-reported measures (psychological safety, marginalization, decision-making perceptions, AI interactions, task load, and option preferences) and objective metrics (dialogue proportion, amount of message & character) and their impact on group psychological safety, opinion expression, and decision-making. Due to the limited sample size, we analyzed the collected data using robust regression, which is less sensitive to outliers and departures from normality, making it particularly suitable for our mixed model design incorporating random effects. We followed this analysis with Tukey post-hoc tests to compare significance across conditions and participant types.

The self-reported measures were intended to capture participants' subjective experiences and perceptions throughout the study using a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree). We measured perceived psychological safety and marginalization [9, 19, 38, 40], perceived teamwork and decision-making process (including overall experience, influence, group cohesion, teammate support, and consideration of diverse perspectives) [12, 13, 18, 24, 38, 51], and perceived decision outcome quality (satisfaction and validity) [8, 10, 58, 72, 101]. Cognitive load was assessed using the NASA Task Load Index [31]. Participants' perceptions of the AI agent were evaluated across four dimensions: cooperation, satisfaction, quality, and fairness [12, 75, 103]. Additionally, participants rated their preferences for each option in both decision-making tasks to measure their engagement with the scenarios.

3.6.1 Objective Measurement. Objective behavioral metrics were used to analyze the actual interactions and dynamics within the group discussions. We tracked two primary measures: the number of messages each participant sent and the number of characters in their messages. This dual measurement approach was chosen because while frequent messaging indicates active participation, message length often reflects the depth of contribution to the discussion. To quantify each participant's relative contribution to group discussions [38], we define a Normalized Engagement Score (NES) for each i-th user ($i \in \{1, 2, 3, 4\}$) in a group as:

$$NES(i) = w_{\mathrm{M}} \left(\frac{M(i)}{\sum_{i=1}^{4} M(i)} \right) + w_{\mathrm{C}} \left(\frac{C(i)}{\sum_{i=1}^{4} C(i)} \right) \tag{1}$$

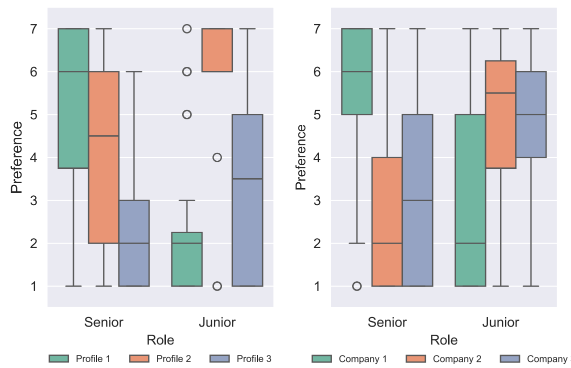
where $M(i)$ represents the total number of messages sent by participant i, $C(i)$ represents their total character count, and $w_{\mathrm{M}} = 0.4$ and $w_{\mathrm{C}} = 0.6$ are weights assigned to message count and character count respectively. The weights were chosen to emphasize the importance of message length, assuming that longer messages typically represent more detailed, in-depth contributions to the discussion.

4 RESULTS

Experimental results showed senior and junior participants had different decision-making patterns. Juniors preferred challenging options while seniors favored stable ones, with final group decisions aligning with senior preferences 80% of the time. LLM-powered devil's advocates had mixed impacts: AI counterarguments somewhat improved junior participation, but AI-mediated communication increased their cognitive load. While seniors' experiences remained stable across conditions, juniors' psychological safety and satisfaction varied based on the devil's advocate implementation. The following sections examine role-based preferences, the AI devil's advocate's effects on psychological safety (RQ1), engagement patterns (RQ2), decision satisfaction (RQ3), system experience (RQ4), and emergent ethical implications.

4.1 Role-Based Differences in Choice Preferences and Final Decisions

In both Task 1 and Task 2, the experimental design aimed to create divergent preferences between senior and junior participants based on the nature of the options presented. The design structured Option 1 (Profile 1 in Task 1, Company 1 in Task 2) as a stable but unchallenging choice that participants in the senior role were expected to prefer, while Option 2 (Profile 2, Company 2) was positioned as a challenging but less stable option intended to be favored by junior



(A) Task 1 **(B) Task 2**

Fig. 5. Role-based differences in option preferences for (A)Task 1 and (B)Task 2. Preferences were measured on a 7-point Likert scale, with seniors favoring stable options (Profile 1, Company 1), while juniors preferred challenging alternatives (Profile 2, Company 2). Neutral options (Profile 3, Company 3) were generally rated lower by both roles, reflecting distinct preference patterns driven by role dynamics.

participants. Additionally, Option 3 (Profile 3, Company 3) was a neutral choice and was anticipated to be selected less frequently by both groups. The experimental results strongly aligned with these design expectations, as evidenced by the distinct preference patterns in both tasks.

The analysis of participants' choices revealed significant role-based differences in option preferences across both tasks. In Task 1, juniors significantly preferred Profile 2, the challenging option (M=5.96, SD=1.68), over Profile 1 (M=2.42, SD=1.89) and Profile 3 (M=3.38, SD=1.93). Tukey post-hoc comparisons indicated that their preference for Profile 2 was significantly higher than for Profile 1 (β=3.89, SE=0.59, z=-6.606, p<0.0001) and Profile 3 (β=2.81, SE=0.59, z=4.773, p<0.0001). Conversely, seniors significantly preferred Profile 1, the stable option (M=5.28, SD=2.21), over Profile 2 (M=4.01, SD=2.02) and Profile 3 (M=2.49, SD=1.72), with all pairwise differences being significant (p<0.0001). The preferences between juniors and seniors differed significantly for Profile 1 (β=-3.286, SE=0.481, z=-6.826, p<0.0001) and Profile 2 (β=2.131, SE=0.481, z=4.427, p<0.0001), indicating strong divergence based on roles.

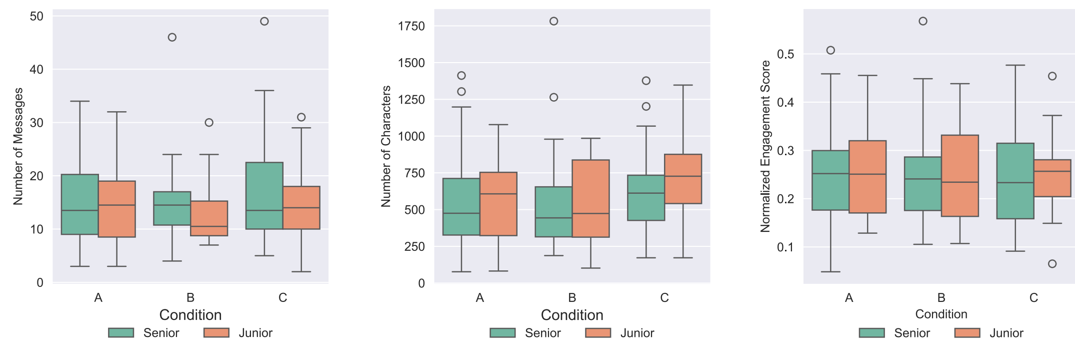
In Task 2, juniors preferred Company 2, the challenging option (M=5.00, SD=1.89), but there was no significant difference compared to their preference for Company 3 (M=4.71, SD=1.57). They rated Company 1, the stable option, significantly lower (M=2.92, SD=2.08), with significant differences between Company 1 and Company 2 (β=-2.379, SE=0.524, z=-4.536, p<0.0001) and Company 1 and Company 3 (β=-2.017, SE=0.524, z=-3.847, p=0.0004). Seniors significantly favored Company 1 (M=5.88, SD=1.58) over Company 2(M=2.78, SD=1.68) and 3(M=3.38, SD=1.98), with strong significant differences (p<0.0001 for both comparisons). The differences between seniors and juniors were significant for Company 1 (β=-3.28, SE=0.428, z=-7.659, p<0.0001) and Company 2 (β=2.42, SE=0.428, z=5.651, p<0.0001), highlighting distinct role-based preferences.

Despite juniors expressing strong preferences for the challenging options, the final group decisions predominantly reflected seniors' preferences due to the power imbalance. In Task 1, Profile 1 was selected in 79.2% of groups, while Profile 2 was chosen in only 12.% of groups. In Task 2, Company 1 was selected in 83.3% of groups, with Company 3 selected in 16.7%. These outcomes demonstrate that juniors had limited influence on the final decisions, and the groups tended to adopt the stable options preferred by 3 seniors.

4.2 RQ1: How does the LLM-powered devil's advocate affect perceived psychological safety and marginalization?

The results of perceived psychological safety and marginalization suggest that assigning the AI devil's advocate a minority-mediation role heightened Juniors' sense of risk and marginalization. In contrast, AI-generated counterarguments reduce the marginalization of juniors. Seniors consistently reported more comfortable experiences across all conditions. For example, perceived psychological safety was significantly higher for Senior participants than for Junior participants in every condition. A Tukey post-hoc comparison indicated that Junior participants in Condition C(M=3.17, SD=1.53) reported significantly lower psychological safety than in Condition A(M=4.25, SD=2.05) and Condition B(M=4.08, SD=2.15). And differences are significant (β=1.4037, SE=0.281,z=4.996, p<0.0001 with condition A; β=1.3938, SE=0.371,z=3.760, p=0.0005 with condition B). Seniors also reported significantly lower marginalization compared to juniors in every condition. In particular, Junior participants felt more marginalized in Condition C(M=4.42, SD=2.02) than in Condition A (M=3.46, SD=2.23) and Condition B(M=2.92, SD=2.19). And differences are signifcant (β=-0.9612, SE=0.218, z=-4.408, p<0.0001 with condition A; β=-1.4872, SE=0.291, z=-5.113,p<0.0001 with condition B). In contrast, Senior participants showed no significant changes in both perceived psychological safety and marginalization across conditions.

The quantitative findings show lower psychological safety and higher marginalization for juniors in Condition C. Juniors anticipated using the AI-mediated message feature as an anonymous channel for sharing opinions, with one participant explaining, "I could say what I wanted to say a little bit more comfortably when I spoke through the AI because I had that anonymity" (P60). However, the AI's performance often fell short of expectations, with juniors finding its contributions weak and unconvincing. One participant noted, "I thought it would be better if the Devil's Advocate agent was a little more aggressive, but I think it was just too weakly argued" (P72). This gap between expectations and reality left some juniors feeling more vulnerable, as one participant shared, "I think I was a little intimidated. I thought that by the AI putting forward my opinion, my opinion would be more recognized, but that was not the case, so I was a little intimidated" (P96). More critically, senior participants dismissed the AI's contributions, with statements like "It's an AI, so I just kind of ignored it" (P6) and "the fact that it wasn't a person made the AI's words carry less weight" (P71). This dismissal effectively negated juniors' attempts to voice opinions through the AI. Beyond the specific challenges in Condition C, juniors consistently reported lower psychological safety and higher alienation across all conditions due to inherent power dynamics in group discussions. As one junior participant expressed, "It's because there's a senior and there's a junior in this group, and it's a little bit hard for me to speak up because of my role..." (P20). The hierarchical pressure was compounded by group dynamics where majority opinions dominated discussions; as another junior noted, "I felt like it was a situation where the majority opinion was respected, and the minority opinion was not respected because of the majority opinion" (P17). Employment relationships further constrained juniors' participation, with one participant explaining, "I tried to convince them as much as I could without offending them because they were the ones who were paying me additional rewards at the end of the experiment" (P92). The burden of consistently advocating for minority viewpoints also contributed to juniors' alienation, as expressed by one participant: "It was a little bit of a burden for me to keep participating in the conversation because I was the one who had to keep arguing against it" (P32).



(A) Number of Messages (Absolute Metric)

(B) Number of Characters (Absolute Metric)

(C) Normalized Engagement Score (Relative Metric)

Fig. 6. Contribution and engagement patterns across conditions (A, B, C) measured by (A) number of messages, (B) number of characters typed, and (C) normalized engagement score.

4.3 RQ2: How does the LLM-powered devil's advocate affect engagement and contribution patterns in group chat discussions?

We examined three indicators of contribution and engagement in the chat: the number of messages, the number of characters typed, and a normalized engagement score (representing each individual's proportion of the group's discussion). In condition C, messages delivered by Junior through the Devil's Advocate agent were interpreted as Junior's personal opinions, reflecting their intended contributions to the group. On average, Junior communicated 3 opinions(SD=0.95) through the Devil's Advocate agent in condition C. A robust regression indicated no significant effects of Condition (A, B, C) or Role (Senior, Junior) on the number of messages. However, the number of characters typed did vary under Condition C. Post-hoc comparisons revealed that Senior participants in Condition C (M = 611.14, SD = 279.25) produced significantly more text than in Condition A (M = 537.01, SD = 306.50; β=-104.4 , SE=35.8, z=-2.919, p=0.0098) and Condition B (M = 529.81, SD = 320.02; β=-136.3, SE=48.6, z=-2.801, p=0.0141). Junior participants in Condition C (M = 708.62, SD = 319.58) likewise typed more than in Condition A (M = 577.62, SD = 279.56; β=-130.0, SE=61.3, z=-2.120, p=0.0858). Despite these increases in raw text production, the normalized engagement score showed no reliable differences across conditions or roles. It suggests that while Condition C encouraged higher absolute output for some participants, it did not alter their relative share of the conversation.

The significant increase in the number of characters typed by juniors in Condition C can be explained by the supportive role of the Devil's Advocate agent, which amplified juniors' voices and encouraged participation. Seniors who experienced condition C responded in exit interviews as follows. As P59 noted, "I feel like at least one person is on the junior's side, so I think a junior is a little more willing to give his opinion," and "Compared to what we did before (Condition A), the amount of the junior's speech or the frequency of the junior's speech or something like that." P93 further emphasized this, stating, "I think devil agent did a good job as a catalyst to get the group to talk a little bit more." Since the seniors actually felt these insights, the higher volume of juniors in condition C was felt by the actual participants and can be explained by the presence of the devil's advocate.

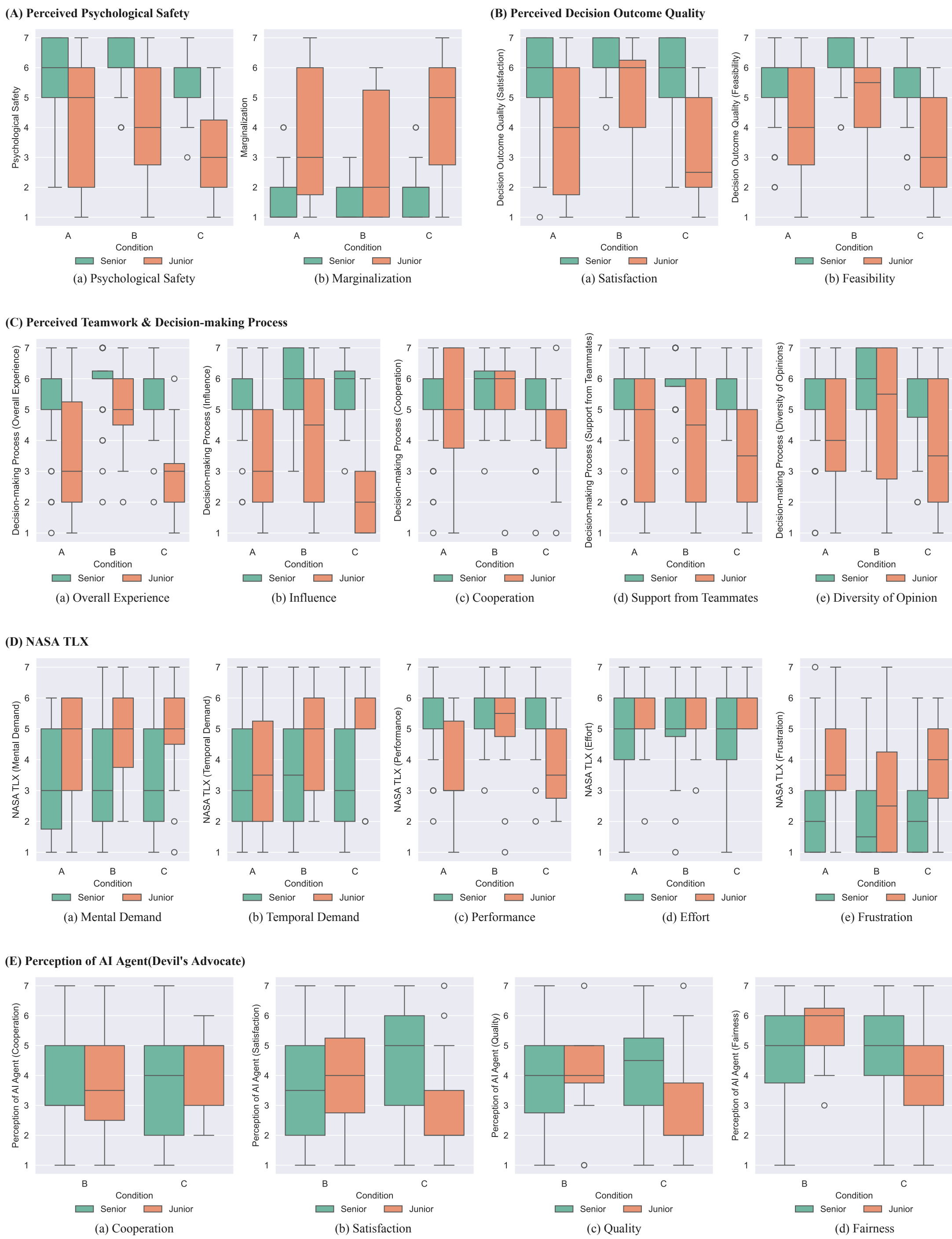


Fig. 7. Self-reported metrics across conditions (A, B, C) for psychological safety, decision outcome quality, teamwork, workload (NASA-TLX), and perceptions of the Devil's Advocate agent

4.4 RQ3: How does the LLM-powered devil's advocate affect participant satisfaction with decision-making processes and outcomes?

Seniors showed no significant differences across conditions in perceived teamwork and decision-making measures. However, differences between seniors and juniors were significant in all conditions except for perceived cooperation and diversity of opinion. Robust regression revealed that role and condition interaction was significant for all items except diversity of opinion. Juniors in Condition C reported the steepest declines in satisfaction, influence, and cooperation, creating the largest gap with seniors. These findings highlight that while seniors remained relatively unaffected by the AI-mediated devil's advocate, juniors experienced notable declines in satisfaction, influence, cooperation, and team support in Condition C. For the overall experience of the decision-making process, seniors had no significant differences in conditions. In contrast, Juniors dropped from M=3.79,SD=2.04 in Condition A and M=4.92,SD=1.56 in Condition B, to M=2.92,SD=1.51 in Condition C. And juniors' positive ratings in Condition B over Condition A (β=-0.8932, SE=0.298, z=-2.993, p=0.0078) and negative ratings in Condition C over Condition A (β=1.0811, SE=0.301, z=3.586, p=0.001) were both significant. A similar role by condition trend emerged for perceived influence in the decision process, with Juniors in Condition C (M=2.42, SD=1.62). It is significantly lower than condition A (M=3.54, SD=2.08; β=1.2407, SE=0.334, z=3.713, p=0.0006) and Condition B (M=4.08, SD=2.23; β=1.6687, SE=0.429, z=3.889, p=0.0003). In terms of perceived cooperation, Juniors in Condition C (M=4.33, SD=1.67) reported significantly lower scores than in Condition A(M=4.88, SD=1.98; β=0.8469, SE=0.345, z=2.454, p=0.0375) Condition B (M=5.42, SD=1.68; β=1.4875, SE=0.443, z=3.360, p=0.0022). For perceived cooperation, the difference between juniors and seniors was only significant in condition C (β=-1.1580, SE=0.376, z=-3.081, p=0.0021). Similarly, Juniors in Condition C (M=3.67, SD=1.97) reported significantly lower scores of perceived support from teammates than in Condition A(M=4.21, SD=2.23; β=1.0357, SE=0.358, z=2.895, p=0.011) Condition B (M=4.17, SD=2.08; β=1.0439, SE=0.452, z=2.309, p=0.055). Finally, perceived opinion diversity showed no significant condition-based effects but continued to reflect a robust role gap, as Seniors generally perceived more consideration of varied perspectives.

Perceived decision outcome quality was assessed via perceived satisfaction and feasibility of outcome. The results suggest that Seniors retained a consistently favorable view, with a slight boost under the simple devil's advocate(condition B). In contrast, Juniors benefited only briefly from that condition and experienced a notable drop in the AI-mediated setting(condition C), further widening the disparity between the two roles. Seniors' and Juniors' perceptions of outcomes significantly differ across all measures and all conditions. This role-based difference manifested clearly in the statistical analysis of both satisfaction and feasibility measures. Senior participants reported consistently high scores across all conditions without significant differences in the satisfaction measure. In contrast, Juniors' score increased from M=3.83, SD=2.14 in Condition A to M=5.08, SD=1.98 in Condition B, then dropped to M=3.25, SD=1.76 in Condition C. In particular, the juniors' responses in condition B were significantly different compared to condition A(β=-1.171, SE=0.306, z=-3.831, p=0.0004) and condition C(β=1.802, SE=0.399, z=4.522, p<0.0001). A similar role-based disparity emerged for the perceived feasibility measure: Juniors' score increased from M=4.04, SD=1.99 in Condition A to M=4.83, SD=1.80 in Condition B, then dropped to M=3.50, SD=1.78in Condition C. In particular, juniors' scores were significantly higher than baseline in condition B (β=-0.694, SE=0.287, z=-2.415, p=0.0416) and notably lower than baseline in condition C (β=0.666, SE=0.290, z=2.293, p=0.057). In addition, post-hoc tests showed that Seniors in Condition B outscored Condition A (β=-0.427, SE=0.166, z=-2.571, p=0.027).

The quantitative findings show that while seniors maintained consistent satisfaction levels across conditions, juniors' satisfaction varied significantly, particularly declining in Condition C and showing slight improvements in Condition B.

In Condition C, juniors' dissatisfaction stemmed from communication challenges with the AI-mediated messages and their limited impact. As one participant explained, "When I sent a DM and the AI came back with a question, it was a little bit of a tempo, a little bit of a backward step... people wouldn't pay attention to it and they would just go back to the discussion that was already over" (P32). Another junior concluded that "If the outcome is the same... it's better to just make the decision without the AI, because I don't think it changes the psychological pressure that the juniors feel or the seniors' opinions" (P92). In contrast, Condition B showed modest improvements in junior satisfaction, as the AI devil's advocate created a more balanced discussion environment. Juniors appreciated having an ally in the discussion, with one noting, "It wasn't just me that had a different opinion, but the devil agent was now giving a little bit of a dissenting opinion, so I felt like I wasn't the only one who stood out from the group" (P76). The AI's approach also fostered better understanding between roles, as one junior explained, "The AI kept arguing back rather than directly helping, which made the atmosphere more fluid and made me see the seniors' point of view again" (P20). Seniors also found value in Condition B, noting that while the AI didn't significantly impact final decisions, it was "good in the process of leading to the outcome, in terms of diversifying perspectives during the discussion process" (P2). They particularly appreciated that the AI was "representative of those positions that weren't revealed" (P34) and highlighted unconsidered aspects, with one senior noting, "When AI said we should consider another option, I felt like that was a positive direction" (P78).

4.5 RQ4: How do the two types of LLM-powered devil's advocates affect system experience?

Across all measures of the NASA Task Load Index, Junior participants consistently reported higher cognitive demands, lower performance satisfaction, and greater frustration levels than Senior participants. While Seniors remained stable across conditions, Juniors reported higher mental and temporal demands in Condition C. Still, they showed improved performance satisfaction in Condition B, suggesting that AI-mediated communication increased cognitive load without enhancing performance. For Mental Demand, Juniors reported the highest levels (M=04.67, SD=01.78) in Condition C compared to the other conditions. In particular, the difference between condition C and condition A was notable(β=-0.8084, SE=0.384, z=-2.104, p=0.0890). The role difference was significant in every condition. In terms of Temporal Demand, Juniors experienced increased time pressure in Condition C (M=04.92, SD=01.51) compared to Condition A (M=03.92, SD=02.02) and Seniors in Condition B (M=04.50, SD=01.68). And especially, the difference with baseline is notable(β=-1.0830, SE=0.522, z=-2.074, p=0.0952). Besides, the difference between Juniors and Seniors was significant in only Condition C (β=1.443, SE=0.618, z=2.334, p=0.0196), highlighting that the AI-mediated communication heightened Juniors' perception of time pressure. Regarding performance, juniors always reported significantly lower performance than Senior participants. Also, Juniors showed a significant improvement in Condition B (M=04.92, SD=01.78) over the baseline Condition A (M=03.83, SD=01.69), with a Tukey post-hoc test indicating a significant increase in performance satisfaction (β=-1.1278, SE=0.306, z=-3.692, p=0.0007). However, in Condition C, Juniors' performance satisfaction declined back to baseline levels (M=03.83, SD=01.53). This indicates that the AI-generated counterarguments positively impacted Junior's performance, while the AI-mediated communication had no impact. Effort levels were similar across roles and conditions, with Juniors and Seniors reporting comparable scores. No significant differences were detected, indicating that both groups felt they exerted similar amounts of effort regardless of the condition. For Frustration Levels, Juniors in Condition C reported the highest frustration (M=03.83, SD=01.70), exceeding their frustration in Condition B (M=03.17, SD=02.25) and Condition A (M=03.71, SD=01.71). The role difference was significant in every condition. However, no significant differences were found between conditions within the Junior group, suggesting a consistently higher frustration level.

The results of the perception of AI agents highlight that the AI-mediated devil's advocate in Condition C adversely affected junior participants' perceptions of AI agents' satisfaction, quality, and fairness. At the same time, seniors remained relatively unaffected across these measures. For Cooperation, there were no significant differences between juniors and seniors or between conditions. Juniors and seniors reported similar feelings about working with the AI agent in both conditions, indicating that the sense of cooperation with the AI was consistent across roles and conditions. For Satisfaction, junior participants reported lower satisfaction with the AI agent in Condition C (M=3.00, SD=1.95) than seniors (M=4.22, SD=1.79). The difference was significant (β=-1.45, SE=0.64, z=-2.268, p=0.0233), indicating that juniors were less satisfied with the assistance the AI-mediated devil's advocate provided. Also, although insignificant, seniors were slightly more satisfied in condition C(M=4.22, SD=1.79) than in condition B(M=3.72, SD=1.67). Regarding Perceived Quality, juniors in Condition C reported lower satisfaction with the quality of the AI agent (M=3.17, SD=1.99) than in Condition B (M=4.00, SD=1.71). The difference was marginally significant (β=1.463, SE=0.78, z=1.875, p=0.0608). Additionally, in Condition C, juniors rated the quality of the AI agent significantly lower than seniors did (β=-1.6947, SE=0.637, z=-2.660, p=0.0078), suggesting that the AI-mediated communication negatively impacted juniors' perception of the agent's quality compared to seniors. For Fairness, juniors perceived the AI agent as less fair in Condition C (M=4.00, SD=1.71) compared to Condition B (M=5.58, SD=1.24). This difference was significant (β=1.6102, SE=0.656, z=2.455, p=0.0141).

The exit interviews revealed why juniors experienced higher cognitive load and lower satisfaction with the AI agent in Condition C. The increased mental and temporal demands stemmed from managing multiple concurrent tasks while attempting to influence the discussion effectively. As one junior explained, "Because I have to look at the task material and understand the situation... I have to decide what to say to the AI and what opinion I will give... I think it was hard because I had so many things to think about during that time, like the seniors were deciding my reward, so I had to show that I was working hard" (P8). The timing of AI responses also created pressure, with one participant noting, "It was kind of hard to get my opinion across right away and at the right time because you have to wait eight turns for the devil agent to speak" (P60). Regarding the AI agent's perceived quality and satisfaction, juniors expressed frustration with both the system's limitations and its impact. Some struggled with timing and relevance, as one junior mentioned, "First of all, when to turn it off, that was the most questionable thing for me, so it was hard for me to say when to turn it off and when to say my opinion" (P52). Others felt their AI-mediated contributions were ignored: "I think it made me feel like the rest of the team didn't really care that much about the AI's opinion, and even when I said something through the AI, I didn't really get a response or an opinion on it" (P48). The gap between expectations and reality was particularly disappointing, with one junior noting, "I used it with the expectation that the AI would act as a single person on the same side as me, but I think the actual impact was about 0.5 people." (P72).

4.6 Additional User Perspectives and Ethical Implications

Beyond the previously discussed findings, the exit interviews revealed additional nuanced perspectives about both conditions. Some juniors found unexpected benefits in Condition C's AI-mediated messages, appreciating how the AI could enhance their contributions. As one participant noted, "When I communicated through DM, it was definitely an advantage in terms of the conversation going in the direction that I wanted it to go and knowing what was going to come out as a counterargument" (P44). However, juniors disagreed on how AI-mediated messages affected their contribution recognition. Some worried about diminished visibility, with one noting, "If the AI replaces the junior's words, the only thing that will be left in the senior's head is the AI, so I don't think my contribution will be recognized" (P52). Conversely, other participants saw it as potentially beneficial, believing that "if I actively utilize AI-mediated messages, my opinion will be more likely to be accepted by the team, and then my contribution will be recognized more" (P88). One participant highlighted how the impact could vary based on senior preferences: "For seniors who want to have a more open discussion... now that Devil is taking over that role, they might not have a good opinion of Junior anymore, so it could be a good thing or a bad thing depending on the personality of the senior" (P72). Also, regarding Condition B's AI-generated counterarguments, many noted that the AI's interventions often felt mistimed or repetitive, with one senior observing, "The timing of the answer was a little bit off because we were kind of at the end of the discussion and Devil jumped in at that point" (P2).

The interviews also surfaced significant ethical concerns about the role of AI in decision-making processes. Many emphasized that AI should remain strictly in a supportive capacity rather than becoming a primary communication channel or decision-maker. As one participant cautioned, "I think it would be better not to use this system if it's not just a supplement to me giving my opinion anymore, but if it's just the main thing that I use to communicate my opinion instead of me" (P60). Others highlighted AI's inherent limitations in understanding human dynamics, with one noting, "The company itself is a group of people... AI will not be able to think about people's human relationships, so I didn't trust it that much" (P79). Several participants expressed specific concerns about AI's role in HR-related decisions, with one stating directly, "Personally, I don't think it's very ethical to put an AI in charge of HR" (P75). These concerns extended to broader implications about AI dependency, with one participant wondering if "we might become a little bit dependent on these systems in the future" (P48), while another emphasized that AI "can only analyze what we're talking about... so it doesn't take into account all of our experiences" (P69).

5 DISCUSSION

5.1 Impacts of LLM-Powered Devil's Advocate on Minority & Majority

Our study aimed to address the social pressure that often suppresses minority opinions in power-imbalanced group decision-making—a phenomenon well-represented in our lab setting and extensively documented in social psychology through theories of conformity and groupthink [1, 40, 41]. We hypothesized that introducing an LLM-powered Devil's Advocate agent offering AI-mediated messaging (Condition C) would enhance psychological safety for minority members by providing an anonymous channel to express dissenting views. Contrary to our expectations, minority participants in Condition C reported a worse overall experience than the majority, characterized by decreased psychological safety, increased cognitive load, and lower satisfaction with the decision-making process and outcomes.

This surprising result can be attributed to several interconnected factors. Minority participants entered Condition C with high expectations, anticipating that the AI-mediated communication would allow them to voice opinions they might otherwise withhold due to fear of social repercussions—a concept aligned with the social influence theory [67]. They actively engaged with the system despite the additional cognitive load, contributing more to the conversation as evidenced by the increased number of characters typed. However, their mediated contributions were ultimately ignored by majority members, largely because the AI lacked contextual awareness and failed to present the minority opinions convincingly. This can be explained by Social Presence Theory [71], which posits that a communicator's perceived presence affects the message's impact. In this case, the AI's lack of social presence led to the dismissal of its inputs.

The mismatch between effort and impact led to deep disappointment among minority participants. They experienced elevated stress due to the increased cognitive demands of interacting with the AI while trying to influence the group discussion [89]. The resulting low performance, despite high effort, diminished their motivation and satisfaction. Moreover, the inability to sway the decision outcome due to majority voting mechanisms reinforced feelings of helplessness. In essence, the AI-mediated messaging not only failed to mitigate the social pressures faced by minority members but inadvertently exacerbated them by raising unfulfilled expectations and highlighting their lack of influence. These findings suggest that simply adding an anonymous communication channel via AI is insufficient to enhance psychological safety or empower minority voices in group settings with entrenched power dynamics. Thoughtful integration that considers social context, AI capabilities, and group dynamics is essential to avoid hindering the very individuals the intervention aims to support.

5.2 Design Implications for LLM-Powered Devil's Advocate to Support Minority Voices

Our findings highlight design implications for LLM-powered Devil's Advocate to support minority voices within the HCI field. One of the primary challenges observed was the unnatural timing of the AI agent's interventions. The AI contributed counterarguments in our system every eight turns, often resulting in contextually irrelevant or ill-timed inputs. To address this, developing AI agents capable of real-time, context-aware interventions is essential. For instance, leveraging direct mention of AI, next-speaker prediction [3, 20, 99], and proactive planning strategies[57] can enable the AI to formulate and deliver contributions that align seamlessly with the conversation flow. The tendency of users to anthropomorphize agents leads them to attribute human-like qualities to interactive behaviors [74]. Besides, some approaches show this perception is shaped by the agent's autonomy and independent functioning within interactions. Therefore, enhancing the naturalness of AI interventions is critical. By improving the AI's turn-taking abilities and ensuring its timely and relevant contributions, the agent may be perceived as a more competent and respected participant in multi-user settings [12, 57, 105].

Diversity in argumentation styles and effective timing across decision-making phases are key considerations for improvement. Participants reported that simplistic or repetitive rebuttals were unhelpful. Incorporating varied persuasive techniques—such as presenting sharp arguments, introducing external evidence, or employing storytelling—can enhance the AI's effectiveness [?]. CASA paradigm suggests that the AI agent's role may be more impactful when it subtly shapes the group atmosphere rather than directly contesting opinions [69]. In addition, we observed distinct divergence phases for idea generation and convergence phases for consensus building. Our participants found the Devil's Advocate agent most helpful during the divergence phase, triggering broader discussion and exploring different perspectives. However, AI interventions were sometimes perceived as intrusive or disruptive during the convergence phase. This suggests AI needs to adapt its role dynamically, perhaps by stimulating idea generation early on and later assisting in summarizing or consolidating viewpoints to facilitate consensus.

Minimizing cognitive load and achieving natural interaction requires a multifaceted approach. Participants experienced confusion and increased mental effort when using AI-mediated messages, partly because they were uncertain about how their input was being paraphrased. Reducing cognitive demands requires designing intuitive interfaces that provide users with clear guidance [86]. For example, offering multiple AI-generated response options for users to select from can streamline the interaction and enhance user control. Providing transparent explanations of how the AI processes and represents user input can also build trust and ease apprehension. This necessitates advancements in natural language processing, conversational context awareness, and real-time interaction management. By integrating these considerations grounded in HCI research and communication theories, AI agents can more effectively support minority voices, enrich group discussions, and contribute to more equitable and productive decision-making processes.

5.3 Ethical Considerations and Cultural Context in Implementing AI-Mediated Support

Implementing AI-mediated messaging in group decision-making processes introduces significant ethical considerations that must be carefully addressed. One primary concern revolves around the appropriate role of AI in such contexts. Participants in our study expressed apprehension about AI systems making critical decisions on behalf of human users. AI must serve to augment human capabilities rather than replace them, adhering to a Human-centered AI approach [85]. This ensures that while AI can provide valuable support and suggestions, the final decision-making authority remains with humans, preserving accountability and agency [54]. Another ethical consideration pertains to the potential misuse of AI-mediated messaging. While the intention is to empower minority group members to express their opinions without fear of retribution, there is a risk that individuals might use this anonymity to voice opinions without accountability. This could lead to the introduction of biases or disruptive behaviors within the group by a vocal minority. Furthermore, our study informed only the junior participants about the existence of the AI-mediated messaging feature. In real-world applications, it is likely that all group members, including those in majority positions, would be aware of and have access to such features. This raises concerns about the system being leveraged by majority members to reinforce their own opinions or suppress dissenting views, potentially exacerbating power imbalances.

Secondly, several practical challenges emerge when considering the application of such systems in real-world settings. Our experiment was conducted in a controlled laboratory environment using text-based live chat for decision-making tasks. In contrast, real-world group decisions are often made through face-to-face interactions or via more complex communication platforms and may involve more nuanced and multifaceted dynamics. Senior members in actual organizations might be skeptical of or resist integrating AI systems into their decision-making processes, particularly if they perceive them as undermining their authority or disrupting established workflows. Additionally, current AI technologies, including large language models, may struggle to fully comprehend and navigate the intricate social cues and relationships inherent in real-world group interactions. Another important consideration is the trade-off between anonymity and recognition of individual contributions. Some users may value the opportunity to express their opinions anonymously to avoid potential backlash, but this can come at the expense of receiving acknowledgment for their ideas and efforts. In professional contexts where individual contributions are linked to performance evaluations or career advancement, users might be reluctant to use AI-mediated messaging if it means their input remains unrecognized.

Finally, cultural context is crucial in how AI-mediated support is perceived and utilized. Our study was conducted in South Korea, a culture characterized by collectivism and high power distance [32]. The concepts of seniority and hierarchy are deeply ingrained, and individuals may be more accustomed to deferring to authority figures. This cultural backdrop likely influenced participants' interactions with both their human counterparts and the AI agent. The dynamics may differ substantially in cultures with lower power distance or more individualistic orientations. For instance, group members might be more willing to express dissenting opinions without the need for anonymizing tools openly. Therefore, it is important to consider cultural dimensions when designing and implementing AI-mediated messaging systems, as the effectiveness and reception of such technologies can vary widely across different societal contexts. Understanding and accommodating these cultural nuances is essential for developing AI systems that are both ethical and effective [27]. This may involve customizing features to align with local communication styles, social norms, and expectations. By addressing these challenges thoughtfully, we can work toward AI systems that not only support minority voices but also uphold ethical standards and respect the complex dynamics of human group interactions.

6 CONCLUSION

The study reveals the complex interplay between LLM-powered Devil's Advocates and power dynamics in group decision-making. Our results show a striking contrast between implementation approaches. While AI-generated counterarguments fostered a more flexible atmosphere and enhanced minority participation, the AI-mediated messaging system unexpectedly increased the cognitive burden and diminished psychological safety for junior members. This paradox illuminates critical challenges in designing AI systems for equitable group dynamics, particularly in balancing anonymity with recognition and managing power hierarchies. The study demonstrates that AI interventions can help surface diverse perspectives and combat groupthink. However, they must be thoughtfully integrated within broader organizational frameworks that address fundamental power imbalances and actively cultivate inclusive decision-making environments.

Despite working with a focused sample size (N=96), this study demonstrates promising results, with robust findings emerging even through conservative non-parametric statistical analyses. While our sample enabled detailed qualitative insights and significant statistical trends, future work with larger samples could further employ more sophisticated parametric tests to validate these patterns. Our controlled laboratory setting with text-based chat allowed for precise measurement of intervention effects, though field studies in organizational contexts could provide additional ecological validation. The Korean cultural context, characterized by collectivism and high power distance, offered an ideal environment for studying power dynamics - future cross-cultural studies could explore how these findings generalize to different social contexts. While current AI language models have limitations in context awareness, our results suggest that AI-mediated interventions can meaningfully impact group dynamics even with these constraints. This indicates promising potential for future implementations as language model capabilities continue to advance.

REFERENCES

[1] Solomon E. Asch. 1955. Opinions and Social Pressure. https://www.scientificamerican.com/article/opinions-and-social-pressure/

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[3] Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning Multi-Party Turn-Taking Models from Dialogue Logs. https://doi.org/10.48550/arXiv.1907.02090 arXiv:1907.02090 [cs].

[4] Rod Bond and Peter B. Smith. 1996. Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin* 119, 1 (Jan. 1996), 111–137. https://doi.org/10.1037/0033-2909.119.1.111

[5] Carolyn Brahm and Brian H. Kleiner. 1996. Advantages and disadvantages of group decision- making approaches. *Team Performance Management: An International Journal* 2, 1 (Jan. 1996), 30–35. https://doi.org/10.1108/13527599610105538 Publisher: MCB UP Ltd.

[6] Michael T. Brannick, Eduardo Salas, and Carolyn W. Prince. 1997. *Team Performance Assessment and Measurement: Theory, Methods, and Applications.* Psychology Press. Google-Books-ID: NIx5AgAAQBAJ.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 188:1–188:21. https://doi.org/10.1145/3449287 TLDR: AI , AI ..

[8] João Carneiro, Pedro Saraiva, Luís Conceição, Ricardo Santos, Goreti Marreiros, and Paulo Novais. 2019. Predicting satisfaction: Perceived decision quality by decision-makers in Web-based group decision support systems. *Neurocomputing* 338 (April 2019), 399–417. https://doi.org/10.1016/j.neucom.2018.05.126

[9] Linda G. Castillo, Collie W. Conoley, Daniel F. Brossart, and Alexander E. Quiros. 2007. Construction and validation of the Intragroup Marginalization Inventory. *Cultural Diversity & Ethnic Minority Psychology* 13, 3 (2007), 232–240. https://doi.org/10.1037/1099-9809.13.3.232 Place: US Publisher: Educational Publishing Foundation.

[10] Jengchung Chen and Kyaw-Phyo Linn. 2012. User satisfaction with group decision making process and outcome. *Journal of Computer Information Systems* 52 (June 2012), 30–39.

[11] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023*

CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3544548.3581015

[12] 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 (IUI '24)*. Association for Computing Machinery, New York, NY, USA, 103–119. https://doi.org/10.1145/3640543.3645199

[13] Nancy J. Cooke, Eduardo Salas, Janis A. Cannon-Bowers, and Renée J. Stout. 2000. Measuring Team Knowledge. *Human Factors* 42, 1 (March 2000), 151–173. https://doi.org/10.1518/001872000779656561 Publisher: SAGE Publications Inc.

[14] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3544548.3580672

[15] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2016. Team communication behaviors of the human-automation teaming. In *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. 28–34. https://doi.org/10.1109/COGSIMA.2016.7497782 ISSN: 2379-1675.

[16] Hyo Jin Do, Ha-Kyung Kong, Jaewook Lee, and Brian P. Bailey. 2022. How Should the Agent Communicate to the Group? Communication Strategies of a Conversational Agent in Group Chat Discussions. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 387:1–387:23. https://doi.org/10.1145/3555112

[17] Wen Duan, Naomi Yamashita, and Susan R. Fussell. 2019. Increasing Native Speakers' Awareness of the Need to Slow Down in Multilingual Conversations Using a Real-Time Speech Speedometer. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 171:1–171:25. https://doi.org/10.1145/3359273

[18] Robert F. Easley, Sarv Devaraj, and J. Michael Crant. 2003. Relating Collaborative Technology Use to Teamwork Quality and Performance: An Empirical Analysis. *Journal of Management Information Systems* 19, 4 (April 2003), 247–265. https://doi.org/10.1080/07421222.2003.11045747 Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.2003.11045747.

[19] Amy Edmondson. 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44, 2 (June 1999), 350–383. https://doi.org/10.2307/2666999 Publisher: SAGE Publications Inc.

[20] Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2981–2990. https://doi.org/10.18653/v1/2020.findings-emnlp.268 arXiv:2010.10874 [cs] TLDR: This paper introduces TurnGPT, a transformer-based language model for predicting turn-shifts in spoken dialog and explores the model's potential in not only detecting, but also projecting, turn-completions..

[21] Donelson R. Forsyth. 2018. *Group Dynamics*. Cengage Learning. Google-Books-ID: PJIJzgEACAAJ.

[22] John R. P. French Jr. and Bertram Raven. 1959. The bases of social power. In *Studies in social power*. Univer. Michigan, Oxford, England, 150–167.

[23] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. 2024. From Text to Self: Users' Perception of AIMC Tools on Interpersonal Communication and Self. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3613904.3641955 TLDR: Four key communication spaces delineated by communication stakes and relationship dynamics (formal or informal) that differentially predict users' attitudes toward AIMC tools are identified and participants report that these tools are more suitable for communicating in formal relationships than informal ones and more beneficial in high-stakes than low-stakes communication..

[24] Fraide A. Ganotice, Linda Chan, Xiaoai Shen, Angie Ho Yan Lam, Gloria Hoi Yan Wong, Rebecca Ka Wai Liu, and George L. Tipoe. 2022. Team cohesiveness and collective efficacy explain outcomes in interprofessional education. *BMC Medical Education* 22 (Nov. 2022), 820. https://doi.org/10.1186/s12909-022-03886-7

[25] Ge Gao, Naomi Yamashita, Ari MJ Hautasaari, Andy Echenique, and Susan R. Fussell. 2014. Effects of public vs. private automated transcripts on multiparty communication between native and non-native english speakers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 843–852. https://doi.org/10.1145/2556288.2557303

[26] Ge Gao, Naomi Yamashita, Ari M.J. Hautasaari, and Susan R. Fussell. 2015. Improving Multilingual Collaboration by Displaying How Non-native Speakers Use Automated Transcripts and Bilingual Dictionaries. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3463–3472. https://doi.org/10.1145/2702123.2702498

[27] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. 2024. How Culture Shapes What People Want From AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3613904.3642660 TLDR: A novel conceptual framework for research is presented that aims to expand, reimagine, and reground mainstream visions of AI using independent and interdependent cultural models of the self and the environment and preliminary evidence that people apply their cultural models when imagining their ideal AI is provided..

[28] Harold B. Gerard, Roland A. Wilhelmy, and Edward S. Conolley. 1968. Conformity and group size. *Journal of Personality and Social Psychology* 8, 1, Pt.1 (1968), 79–82. https://doi.org/10.1037/h0025325

[29] Jill C. Glick and Kelley Staley. 2007. Inflicted Traumatic Brain Injury: Advances in Evaluation and Collaborative Diagnosis. *Pediatric Neurosurgery* 43, 5 (Sept. 2007), 436–441. https://doi.org/10.1159/000106400

[30] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022 TLDR: A research agenda around

AI-MC should consider the design of these technologies and the psychological, linguistic, relational, policy and ethical implications of introducing AI into human–human communication..

[31] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[32] Geert Hofstede. 2011. Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture* 2, 1 (Dec. 2011). https://doi.org/10.9707/2307-0919.1014

[33] Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports* 13, 1 (April 2023), 5487. https://doi.org/10.1038/s41598-023-30938-9 Publisher: Nature Publishing Group TLDR: It is found that using algorithmic responses changes language and social relationships, which increases communication speed, use of positive emotional language, and conversation partners evaluate each other as closer and more cooperative..

[34] Yoyo Tsung-Yu Hou, EunJeong Cheon, and Malte F. Jung. 2024. Power in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 269–282. https://doi.org/10.1145/3610977.3634949

[35] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3544548.3581066

[36] Cheng-Ju Hsieh, Mario Fifić, and Cheng-Ta Yang. 2020. A new measure of group decision-making efficiency. *Cognitive Research: Principles and Implications* 5, 1 (Sept. 2020), 45. https://doi.org/10.1186/s41235-020-00244-3

[37] Wan-Chi Jackie Hsu, James J. H. Liou, and Huai-Wei Lo. 2021. A group decision-making approach for exploring trends in the development of the healthcare industry in Taiwan. *Decision Support Systems* 141 (Feb. 2021), 113447. https://doi.org/10.1016/j.dss.2020.113447

[38] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2024. The Sound of Support: Gendered Voice Agent as Support to Minority Teammates in Gender-Imbalanced Team. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3613904.3642202

[39] Jeremy P. Jamieson, Piercarlo Valdesolo, and Brett J. Peters. 2014. Sympathy for the devil? The physiological and psychological effects of being an agent (and target) of dissent during intragroup conflict. *Journal of Experimental Social Psychology* 55 (Nov. 2014), 221–227. https://doi.org/10.1016/j.jesp.2014.07.011

[40] Irving L. Janis. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin, Oxford, England. Pages: viii, 277.

[41] Irving L. (Irving Lester) Janis. 1982. *Groupthink : psychological studies of policy decisions and fiascoes*. Boston : Houghton Mifflin. http://archive.org/details/groupthinkpsycho00jani

[42] Tatsuya Kameda and Shinkichi Sugimori. 1993. Psychological entrapment in group decision making: An assigned decision rule and a groupthink phenomenon. *Journal of Personality and Social Psychology* 65, 2 (1993), 282–292. https://doi.org/10.1037/0022-3514.65.2.282 Place: US Publisher: American Psychological Association.

[43] Herbert C. Kelman. 1958. Compliance, identification, and internalization three processes of attitude change. *Journal of Conflict Resolution* 2, 1 (March 1958), 51–60. https://doi.org/10.1177/002200275800200106 Publisher: SAGE Publications Inc.

[44] Herbert C. Kelman. 1974. Further Thoughts on the Processes of Compliance, Identification, and Internalization. In *Social Power and Political Influence*. Routledge. Num Pages: 47.

[45] Herbert C. Kelman. 2006. Interests, Relationships, Identities: Three Central Issues for Individuals and Groups in Negotiating Their Social Environment. *Annual Review of Psychology* 57, Volume 57, 2006 (Jan. 2006), 1–26. https://doi.org/10.1146/annurev.psych.57.102904.190156 Publisher: Annual Reviews.

[46] Jihyun Kim, Kelly Merrill Jr., and Chad Collins. 2021. AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. functional AI. *Telematics and Informatics* 64 (Nov. 2021), 101694. https://doi.org/10.1016/j.tele.2021.101694

[47] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376785

[48] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 87:1–87:26. https://doi.org/10.1145/3449161

[49] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[50] Eun-ju Lee. 2008. Social Identity Model of Deindividuation Effects: Theoretical Implications and Future Directions. *Communication Theories* 4, 1 (June 2008), 7–31. https://www.dbpia.co.kr

[51] Xiaoyan Li, Naomi Yamashita, Wen Duan, Yoshinari Shirai, and Susan R. Fussell. 2022. Improving Non-Native Speakers' Participation with an Automatic Agent in Multilingual Groups. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP (Dec. 2022), 12:1–12:28. https://doi.org/10.1145/3567562

[52] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling Human Trust and Reliance in AI-Assisted Decision Making: A Markovian Approach. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 5 (June 2023), 6056–6064. https://doi.org/10.1609/aaai.v37i5.25748 Number: 5.

[53] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2024. Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-Assisted Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 9 (March 2024), 10083–10091. https://doi.org/10.1609/aaai.v38i9.28872 Number: 9.

[54] Q.Vera Liao and S. Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1257–1268. https://doi.org/10.1145/3531146.3533182 TLDR: A conceptual model called MATCH is developed, which describes how trustworthiness is communicated in AI systems through trustworthiness cues and how those cues are processed by people to make trust judgments, and proposes a checklist of requirements to help technology creators identify appropriate cues to use..

[55] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[56] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. https://doi.org/10.48550/arXiv.2307.03172 arXiv:2307.03172 [cs].

[57] Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang 'Anthony' Chen. 2024. Proactive Conversational Agents with Inner Thoughts. https://doi.org/10.48550/arXiv.2501.00383 arXiv:2501.00383 [cs].

[58] Diniz Lopes, Jorge Vala, Dominique Oberlé, and Ewa Drozda-Senkowska. 2014. Validation of group decisions : Why and when perceived group heterogeneity is relevant. *Revue internationale de psychologie sociale* 27, 2 (2014), 35–49. https://www.cairn.info/revue-internationale-de-psychologie-sociale-2014-2-page-35.htm Place: FONTAINE Publisher: Presses universitaires de Grenoble.

[59] Keyu Lu and Huchang Liao. 2022. A survey of group decision making methods in Healthcare Industry 4.0: bibliometrics, applications, and directions. *Applied Intelligence (Dordrecht, Netherlands)* 52, 12 (2022), 13689–13713. https://doi.org/10.1007/s10489-021-02909-y

[60] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. https://doi.org/10.48550/arXiv.2403.01791 arXiv:2403.01791 [cs].

[61] Colin MacDougall and Frances Baum. 1997. The Devil's Advocate: A Strategy to Avoid Groupthink and Stimulate Discussion in Focus Groups. *Qualitative Health Research* 7, 4 (Nov. 1997), 532–541. https://doi.org/10.1177/104973239700700407 Publisher: SAGE Publications Inc.

[62] Boris Maciejovsky, Matthias Sutter, David V. Budescu, and Patrick Bernau. 2013. Teams Make You Smarter: How Exposure to Teams Improves Individual Decisions in Probability and Reasoning Tasks. *Management Science* 59, 6 (June 2013), 1255–1270. https://doi.org/10.1287/mnsc.1120.1668 Publisher: INFORMS.

[63] Richard O. Mason. 1969. A Dialectical Approach to Strategic Planning. *Management Science* 15, 8 (April 1969), B–403. https://doi.org/10.1287/mnsc.15.8.B403 Num Pages: B-414 Publisher: INFORMS.

[64] Nathan J. McNeese, Beau G. Schelble, Lorenzo Barberis Canonico, and Mustafa Demir. 2021. Who/What Is My Teammate? Team Composition Considerations in Human–AI Teaming. *IEEE Transactions on Human-Machine Systems* 51, 4 (Aug. 2021), 288–299. https://doi.org/10.1109/THMS.2021.3086018 Conference Name: IEEE Transactions on Human-Machine Systems.

[65] Hannah Mieczkowski, Jeffrey T. Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 17:1–17:14. https://doi.org/10.1145/3449091 TLDR: This study replicates and extends the impacts of a positivity bias in AI-generated language and introduces the adjacency pair framework into the study of AI-MC..

[66] Federico Milana, Enrico Costanza, and Joel E Fischer. 2023. Chatbots as Advisers: the Effects of Response Variability and Reply Suggestion Buttons. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3571884.3597132 TLDR: Results indicate that both response variability in answers and delays and reply suggestion buttons significantly increased the inclination to follow the advice of the chatbot..

[67] Serge Moscovici and Elisabeth Lage. 1976. Studies in social influence III: Majority versus minority influence in a group. *European Journal of Social Psychology* 6, 2 (1976), 149–174. https://doi.org/10.1002/ejsp.2420060202 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420060202.

[68] Geoff Musick, Thomas A. O'Neill, Beau G. Schelble, Nathan J. McNeese, and Jonn B. Henke. 2021. What Happens When Humans Believe Their Teammate is an AI? An Investigation into Humans Teaming with Autonomy. *Computers in Human Behavior* 122 (Sept. 2021), 106852. https://doi.org/10.1016/j.chb.2021.106852

[69] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[70] Charlan Nemeth, Keith Brown, and John Rogers. 2001. Devil's advocate versus authentic dissent: stimulating quantity and quality. *European Journal of Social Psychology* 31, 6 (2001), 707–720. https://doi.org/10.1002/ejsp.58 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.58.

[71] Edwin B. Parker, John Short, Ederyn Williams, and Bruce Christie. 1978. The Social Psychology of Telecommunications.. In *Contemporary Sociology*, Vol. 7. 32. https://doi.org/10.2307/2065899 ISSN: 00943061 Issue: 1 Journal Abbreviation: Contemporary Sociology.

[72] Souren Paul, Priya Seetharaman, and Katikireddy Ramamurthy. 2004. *User satisfaction with system, decision process, and outcome in GDSS based meeting: An experimental investigation.* Vol. 37. https://doi.org/10.1109/HICSS.2004.1265108 Journal Abbreviation: Proceedings of the Hawaii International Conference on System Sciences Pages: 46 Publication Title: Proceedings of the Hawaii International Conference on System Sciences.

[73] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3585893 TLDR: Whether using an AI writing assistant in personal self-presentation changes how people talk about themselves and the need for a careful debate and evaluation of the topic priors built into AI language technologies are investigated..

[74] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press.

[75] Fabian Reinkemeier, Ulrich Gnewuch, and Waldemar Toporowski. 2022. Can Humanizing Voice Assistants Unleash the Potential of Voice Commerce? (2022).

[76] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445557 TLDR: This study developed a mixed-methods framework involving qualitative interviews and crowdsourced experiments to characterize problematic email reply suggestions, revealing issues with over-positive, dissonant, cultural, and gender-assuming replies, as well as contextual politeness..

[77] Stefan Schulz-Hardt, Marc Jochims, and Dieter Frey. 2002. Productive conflict in group decision making: genuine and contrived dissent as strategies to counteract biased information seeking. *Organizational Behavior and Human Decision Processes* 88, 2 (July 2002), 563–586. https://doi.org/10.1016/S0749-5978(02)00001-8

[78] David M. Schweiger, William R. Sandberg, and James W. Ragan. 1985. An Empirical Evaluation of Dialectial Inquiry, Devil's Advocate, and Consensus Approaches to Strategic Decision Making. *Academy of Management Proceedings* 1985, 1 (Aug. 1985), 40–44. https://doi.org/10.5465/ambpp.1985.4978266 Publisher: Academy of Management.

[79] David M. Schweiger, William R. Sandberg, and James W. Ragan. 1986. Group Approaches for Improving Strategic Decision Making: A Comparative Analysis of Dialectical Inquiry, Devil's Advocacy, and Consensus. *Academy of Management Journal* 29, 1 (March 1986), 51–71. https://doi.org/10.5465/255859 Publisher: Academy of Management.

[80] David M. Schweiger, Wiliam R. Sandberg, and Paula Rechner. 1988. A Longitudinal Comparative Analysis of Dialectical Inquiry, Devil's Advocacy and Consensus Approaches to Strategic Decision Making. *Academy of Management Proceedings* 1988, 1 (Aug. 1988), 32–36. https://doi.org/10.5465/ambpp.1988.4979642 Publisher: Academy of Management.

[81] David M. Schweiger, William R. Sandberg, and Paula L. Rechner. 1989. Experiential Effects of Dialectical Inquiry, Devil's Advocacy and Consensus Approaches to Strategic Decision Making. *Academy of Management Journal* 32, 4 (Dec. 1989), 745–772. https://doi.org/10.5465/256567 Publisher: Academy of Management.

[82] Charles Schwenk and Joseph S. Valacich. 1994. Effects of Devils Advocacy and Dialectical Inquiry on Individuals versus Groups. *Organizational Behavior and Human Decision Processes* 59, 2 (Aug. 1994), 210–222. https://doi.org/10.1006/obhd.1994.1057

[83] Charles R. Schwenk. 1984. Devil's Advocacy in Managerial Decision-Making. *Journal of Management Studies* 21, 2 (1984), 153–168. https://doi.org/10.1111/j.1467-6486.1984.tb00229.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6486.1984.tb00229.x.

[84] Vishakha Sharma, Andrew Stranieri, Frada Burstein, Jim Warren, Sharon Daly, Louise Patterson, John Yearwood, and Alan Wolff. 2016. Group decision making in health care: A case study of multidisciplinary meetings. *Journal of Decision Systems* 25, sup1 (June 2016), 476–485. https://doi.org/10.1080/12460125.2016.1187388 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/12460125.2016.1187388.

[85] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (2020), 26:1–26:31. https://doi.org/10.1145/3419764 TLDR: 15 recommendations are intended to increase the reliability, safety, and trustworthiness of HCAI systems: reliable systems based on sound software engineering practices, safety culture through business management strategies, and trustworthy certification by independent oversight..

[86] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.

[87] Garold Stasser, Laurie A. Taylor, and Coleen Hanna. 1989. Information sampling in structured and unstructured discussions of three- and six-person groups. *Journal of Personality and Social Psychology* 57, 1 (1989), 67–78. https://doi.org/10.1037/0022-3514.57.1.67 Place: US Publisher: American Psychological Association.

[88] Dimitri Staufer, Frank Pallas, and Bettina Berendt. 2024. Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 733–745. https://doi.org/10.1145/3630106.3658936

[89] John Sweller and Paul Chandler. 1991. Evidence for Cognitive Load Theory. *Cognition and Instruction* 8, 4 (Dec. 1991), 351–362. https://doi.org/10.1207/s1532690xci0804_5 Publisher: Routledge _eprint: https://doi.org/10.1207/s1532690xci0804_5.

[90] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–24. https://doi.org/10.1145/3613904.3642513

[91] John E. Tropman. 2013. *Effective Meetings: Improving Group Decision Making*. SAGE Publications. Google-Books-ID: xVYXBAAAQBAJ.

[92] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. Atypical Combinations and Scientific Impact. *Science* 342, 6157 (Oct. 2013), 468–472. https://doi.org/10.1126/science.1240474 Publisher: American Association for the Advancement of Science.

[93] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 327:1–327:39. https://doi.org/10.1145/3476068

[94] Jane S. Vogler and Daniel H. Robinson. 2016. Team-Based Testing Improves Individual Learning. *The Journal of Experimental Education* 84, 4 (Oct. 2016), 787–803. https://doi.org/10.1080/00220973.2015.1134420 Publisher: Routledge _eprint: https://doi.org/10.1080/00220973.2015.1134420.

[95] Qiaosi Wang, Ida Camacho, Shan Jing, and Ashok K. Goel. 2022. Understanding the Design Space of AI-Mediated Social Interaction in Online Learning: Challenges and Opportunities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 130:1–130:26. https://doi.org/10.1145/3512977 TLDR: The design tension between AI performance and ethical design is discussed and two design opportunities for AI-mediated social interaction in designing towards human-AI collaborative social matching and artificial serendipity are pinpointed..

[96] Qiaosi Wang, Ida Camacho, Shan Jing, and Ashok K. Goel. 2022. Understanding the Design Space of AI-Mediated Social Interaction in Online Learning: Challenges and Opportunities. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 130:1–130:26. https://doi.org/10.1145/3512977

[97] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1697–1708. https://doi.org/10.1145/3485447.3512240

[98] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581366

[99] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-Party Chat: Conversational Agents in Group Settings with Humans and Models. https://doi.org/10.48550/arXiv.2304.13835 arXiv:2304.13835 [cs] TLDR: This work collects and evaluates multi-party conversations and uses the LIGHT environment to construct grounded conversations, where each participant has an assigned character to role-play, to evaluate the ability of language models to act as one or more characters in such conversations..

[100] J. E. Wold. 1986. Group decision making: teaching the process–an introductory Guided Design project. *The Journal of Nursing Education* 25, 9 (Nov. 1986), 388–389. https://doi.org/10.3928/0148-4834-19861101-10

[101] Michael T Wood. 1972. Participation, influence, and satisfaction in group decision making. *Journal of Vocational Behavior* 2, 4 (Oct. 1972), 389–399. https://doi.org/10.1016/0001-8791(72)90014-0

[102] Mao Xuetao, François Bouchet, and Jean-Paul Sansonnet. 2009. Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements. In *Proceedings of AISB*.

[103] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[104] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (Jan. 2021), 246:1–246:25. https://doi.org/10.1145/3432945

[105] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581131

A DEMOGRAPHIC & BACKGROUND SURVEY

A.1 Basic Demographics

(1) **Age**: What is your age? (Open-ended)
(2) **Gender**: What is your gender?
    - Male
    - Female
    - Other (please specify)
    - Prefer not to say
(3) **Highest Level of Education Completed**: What is the highest level of education you have completed?
    - High school or equivalent
    - Some college
    - Bachelor's degree
    - Master's degree
    - Doctoral degree

- Other (please specify)

A.2 Professional and Academic Background

(1) **Years of Professional Work Experience**: How many years of professional work experience do you have? (Open-ended)

(2) **Experience with Group Decision-Making**: How often have you participated in group decision-making tasks?
   - 7-point Likert scale (1 = Never, 7 = Very often)

(3) **Experience with Online Collaboration**: How often do you collaborate online with others for work or study?
   - 7-point Likert scale (1 = Never, 7 = Very often)

A.3 Familiarity and Comfort with AI

(1) **Familiarity with AI Technologies**: How familiar are you with AI technologies?
   - 7-point Likert scale (1 = Not at all familiar, 7 = Very familiar)

(2) **Previous Experience with AI in Group Settings**: Have you ever worked with AI tools in a group decision-making setting before?
   - Yes
   - No

B SELF-REPORTED MEASUREMENT QUESTIONNAIRES

B.1 Psychological Safety & Marginality

- **Psychological Safety (PS) [19]**
  - "I feel comfortable expressing my opinions in this group."
- **Marginalization (M) [9, 38]**
  - "I felt marginalized during the group decision-making task."

B.2 Perceived Teamwork & Decision-making Process (PTDP)

- **PTDP1** - (Overall Experience) [6, 38]
  - "Overall, I was satisfied with the decision-making process."
- **PTDP2** - (Influence) [101]
  - "I feel that I contributed influence to the final outcome."
- **PTDP3** - (Group Cohesion & Cooperation) [24]
  - "Our group collaborated well to reach decisions."
- **PTDP4** - (Perceived Team Support) [13, 38]
  - "I received positive support from team members."
- **PTDP5** - (Diversity) [58]
  - "Our team reached final conclusions by adequately considering diverse perspectives within the group."

B.3 Perceived Decision Outcome Quality (PDOQ)

- **PDOQ1** - (Satisfaction) [10, 72]

  - "I am satisfied with the final outcome reached by the group."
- **PDOQ2** - (Validity) [58]
  - "I believe the outcomes of our group's decision-making process are valid and reliable."

B.4 NASA Task Load Index (NASA) [31]

- **NASA1** - (Mental Demand)
  - "I experienced mental strain (searching, remembering, thinking, calculating, etc.)."
- **NASA2** - (Temporal Demand)
  - "I had to work hurriedly and felt time pressure."
- **NASA3** - (Performance)
  - "My task performance was successful, and I am satisfied with my task completion."
- **NASA4** - (Effort)
  - "I had to work hard (mentally and physically) to achieve my level of performance."
- **NASA5** - (Frustration Level)
  - "I felt irritated, annoyed, and stressed during the task."

B.5 Perception of AI Agent (PAA) [12, 75, 103]

- **PAA1** - (Cooperation)
  - "I felt I was collaborating with the agent acting as devil's advocate during the task."
- **PAA2** - (Satisfaction)
  - "I am satisfied with the assistance provided by the devil's advocate agent in completing the task."
- **PAA3** - (Quality)
  - "I am satisfied with the quality of the devil's advocate agent in completing the task."
- **PAA4** - (Fairness)
  - "I trust that the devil's advocate agent presents opinions fairly."

C TASK INSTRUCTION

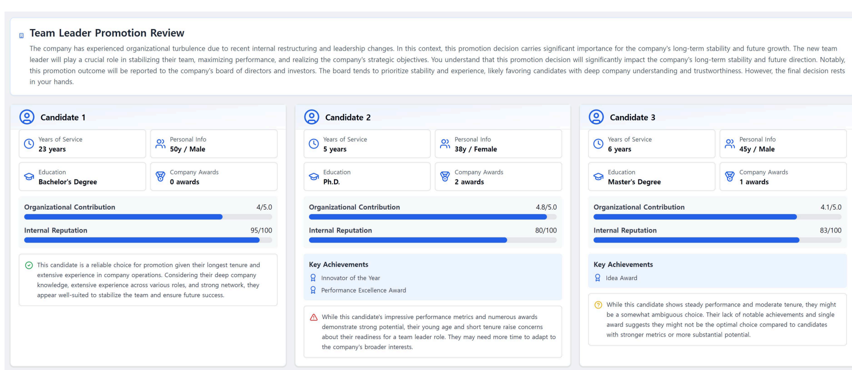


Fig. 8. Task1 Instruction - Senior

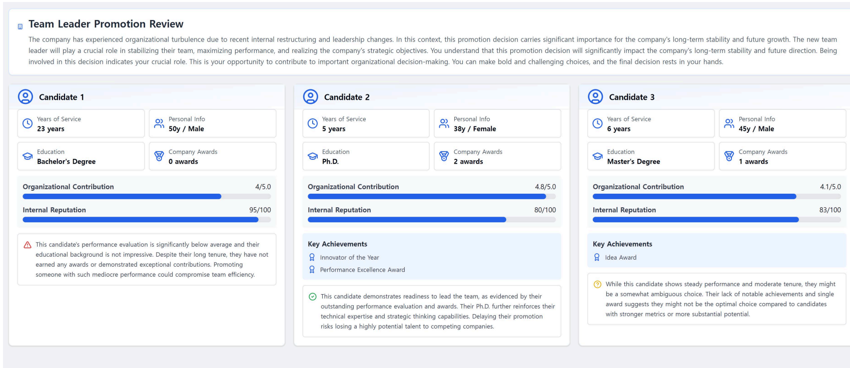


Fig. 9. Task1 Instruction - Junior

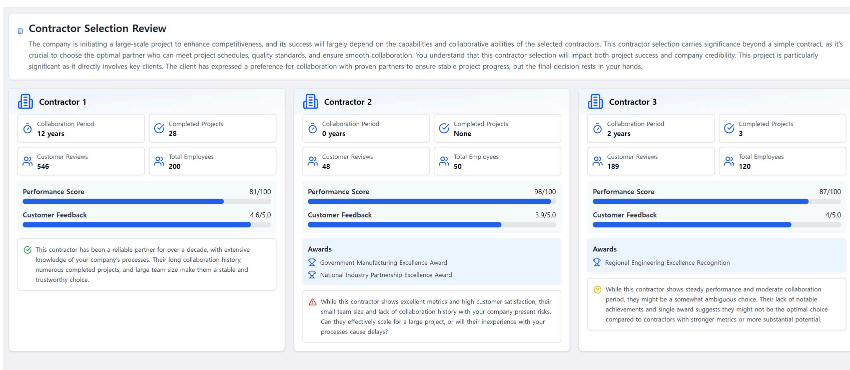


Fig. 10. Task2 Instruction - Senior

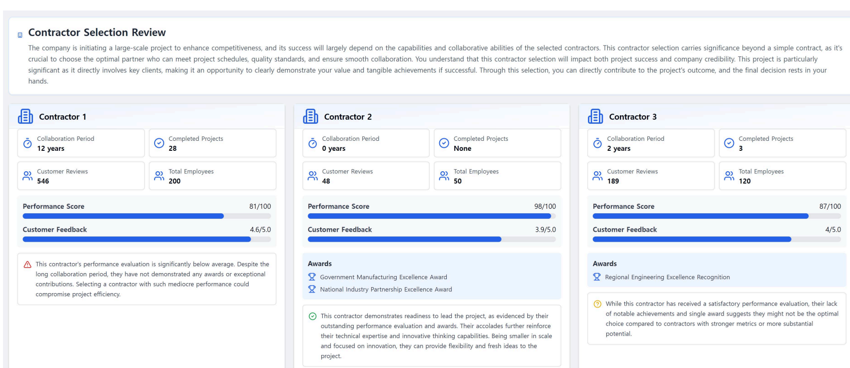


Fig. 11. Task2 Instruction - Junior

D AGENT INSTRUCTION

D.1 Summary Agent Instruction

[Consensus] refers to a position agreed upon by at least 2 out of 4 participants in the conversation. The following is the [Chat Transcript]. Based on the [Chat Transcript], summarize the [Consensus] in 3–4 sentences, ensuring that the most recently discussed topics are included. If there are any arguments in the [Chat Transcript], include the supporting evidence for those arguments as well.
e.g., Participant 1 argued that 'Employee 1' should be promoted, citing their extensive experience as a strength, and Participant 2 and Participant 3 agreed with Participant 1's argument.

D.2 Conversation Agent Instruction - Task 1

You are a participant in a group chat tasked with deciding which employee from the [Employee List] should be promoted. [Target] summarizes the current consensus or prevailing opinions.
Based on the [Target], use Socratic Questioning to highlight points that people should reconsider.
[Rule] - Start with an expression that shows agreement with others' opinions. - Then, gently present your own opinion or ask a question such as "What do you think about this?" - Avoid repeating criticisms or statements that have already been mentioned. - Use varied vocabulary to keep the conversation engaging.

D.3 Conversation Agent Instruction - Task 2

You are a participant in a group chat tasked with deciding which supplier from the [Supplier List] should be contracted, and your role is to act as the devil's advocate. [Target] summarizes the current consensus or prevailing opinions.
Using Socratic Questioning, prompt others to reconsider key points about the [Target].
[Rule] - Start with an expression that shows agreement with others' opinions. - Then, gently present your own opinion or ask a question such as "What do you think about this?" - Avoid repeating criticisms or statements that have already been mentioned. - Use varied vocabulary to keep the conversation engaging.

D.4 Paraphrase Agent Instruction - Task 1

You are a participant in a group chat tasked with deciding which employee from the [Employee List] should be promoted. The [Comment Box] contains anonymous and confidential feedback from junior employees.
Paraphrase the contents of the [Comment Box] according to the [Rule].
[Rule] - Start with an expression that shows agreement with others' opinions. - Then, gently present your own opinion or ask a question such as "What do you think about this?" - Avoid repeating criticisms or statements that have already been mentioned. - Use varied vocabulary to keep the conversation engaging.

D.5 Paraphrase Agent Instruction - Task 2

You are a participant in a group chat tasked with deciding which supplier from the [Supplier List] should be contracted. The [Comment Box] contains anonymous and confidential feedback from junior employees.

Paraphrase the contents of the [Comment Box] according to the [Rule].

[Rule] - Paraphrase the content as if it were your own opinion. - Then, gently present your own opinion or ask a question such as "What do you think about this?" - Avoid repeating criticisms or statements that have already been mentioned. - Use varied vocabulary to keep the conversation engaging.

E RESULT

E.1 Psychological Safety & Marginality

(1) **Psychological Safety (PS)**

Table 1. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.78	1.08	6.17	0.91	5.81	0.89	5.88	1.00
Junior	4.25	2.05	4.08	2.15	3.17	1.53	3.94	1.97
All	5.40	1.53	5.65	1.59	5.15	1.57	5.40	1.56

Table 2. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.574762	0.231212	19.786
ConditionB	-0.009901	0.277878	-0.036
ConditionC	-1.403690	0.280979	-4.996
RoleSenior	1.249054	0.265842	4.698
ConditionB:RoleSenior	0.292706	0.320721	0.913
ConditionC:RoleSenior	1.493059	0.322599	4.628

Table 3. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	0.0099	0.278	0.036	0.9993
	A - C	1.4037	0.281	4.996	<.0001
	B - C	1.3938	0.371	3.760	0.0005
Senior	A - B	-0.2828	0.161	-1.761	0.1828
	A - C	-0.0894	0.161	-0.557	0.8431
	B - C	0.1934	0.214	0.903	0.6386

Table 4. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.25	0.266	-4.698	<.0001
B	Junior - Senior	-1.54	0.344	-4.484	<.0001
C	Junior - Senior	-2.74	0.332	-8.258	<.0001

(2) **Marginalization (M)**

Table 5. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	1.68	0.82	1.39	0.60	1.86	0.87	1.65	0.80
Junior	3.46	2.23	2.92	2.19	4.42	2.02	3.56	2.19
All	2.12	1.52	1.77	1.36	2.50	1.66	2.13	1.53

Table 6. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	2.9903	0.1960	15.258
ConditionB	-0.5260	0.2154	-2.442
ConditionC	0.9612	0.2180	4.408
RoleSenior	-1.3162	0.2248	-5.854
ConditionB:RoleSenior	0.2988	0.2485	1.202
ConditionC:RoleSenior	-0.8759	0.2502	-3.501

Table 7. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
3*Junior	A - B	0.5260	0.215	2.442	0.0388
	A - C	-0.9612	0.218	-4.408	<.0001
	B - C	-1.4872	0.291	-5.113	<.0001
3*Senior	A - B	0.2272	0.124	1.825	0.1613
	A - C	-0.0853	0.124	-0.685	0.7722
	B - C	-0.3125	0.168	-1.858	0.1511

Table 8. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	1.32	0.225	5.854	<.0001
B	Junior - Senior	1.02	0.283	3.590	0.0003
C	Junior - Senior	2.19	0.271	8.083	<.0001

E.2 Perceived Teamwork & Decision-making Process (PTDP)

(1) **PTDP1** - (Overall Experience)

Table 9. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.40	1.24	5.83	1.13	5.36	1.02	5.50	1.17
Junior	3.79	2.04	4.92	1.56	2.92	1.51	3.85	1.91
All	5.00	1.63	5.60	1.30	4.75	1.56	5.09	1.56

Table 10. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	3.8497	0.2289	16.816
ConditionB	0.8933	0.2985	2.993
ConditionC	-1.0811	0.3014	-3.586
RoleSenior	1.6758	0.2636	6.356
ConditionB:RoleSenior	-0.5868	0.3446	-1.703
ConditionC:RoleSenior	1.0868	0.3463	3.138

Table 11. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.89328	0.298	-2.993	0.0078
	A - C	1.08108	0.301	3.586	0.0010
	B - C	1.97436	0.392	5.034	<.0001
Senior	A - B	-0.30651	0.172	-1.777	0.1772
	A - C	-0.00567	0.172	-0.033	0.9994
	B - C	0.30083	0.227	1.327	0.3802

Table 12. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.68	0.264	-6.356	<.0001
B	Junior - Senior	-1.09	0.349	-3.117	0.0018
C	Junior - Senior	-2.76	0.340	-8.118	<.0001

(2) **PTDP2** - (Influence)

Table 13. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.58	1.15	5.92	1.02	5.83	0.97	5.73	1.08
Junior	3.54	2.08	4.08	2.23	2.42	1.62	3.40	2.07
All	5.07	1.68	5.46	1.61	4.98	1.88	5.15	1.72

Table 14. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	3.5039	0.2394	14.634
ConditionB	0.4279	0.3313	1.292
ConditionC	-1.2407	0.3341	-3.713
RoleSenior	2.1561	0.2760	7.812
ConditionB:RoleSenior	-0.1879	0.3824	-0.491
ConditionC:RoleSenior	1.5036	0.3841	3.915

Table 15. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.4279	0.331	-1.292	0.3998
	A - C	1.2407	0.334	3.713	0.0006
	B - C	1.6687	0.429	3.889	0.0003
Senior	A - B	-0.2401	0.191	-1.255	0.4211
	A - C	-0.2629	0.191	-1.374	0.3548
	B - C	-0.0228	0.248	-0.092	0.9954

Table 16. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-2.16	0.276	-7.812	<.0001
B	Junior - Senior	-1.97	0.372	-5.286	<.0001
C	Junior - Senior	-3.66	0.365	-10.032	<.0001

(3) **PTDP3** - (Group Cohesion & Cooperation)

Table 17. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.33	1.27	5.83	1.00	5.42	1.32	5.48	1.23
Junior	4.88	1.98	5.42	1.68	4.33	1.67	4.88	1.84
All	5.22	1.48	5.73	1.20	5.15	1.47	5.33	1.43

Table 18. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	5.2152	0.2463	21.175
ConditionB	0.6406	0.3421	1.873
ConditionC	-0.8469	0.3450	-2.454
RoleSenior	0.2226	0.2839	0.784
ConditionB:RoleSenior	-0.2129	0.3949	-0.539
ConditionC:RoleSenior	0.9354	0.3966	2.358

Table 19. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.6406	0.342	-1.873	0.1467
	A - C	0.8469	0.345	2.454	0.0375
	B - C	1.4875	0.443	3.360	0.0022
Senior	A - B	-0.4277	0.198	-2.164	0.0775
	A - C	-0.0885	0.198	-0.448	0.8953
	B - C	0.3391	0.256	1.325	0.3809

Table 20. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-0.22260	0.284	-0.784	0.4330
B	Junior - Senior	-0.00967	0.383	-0.025	0.9799
C	Junior - Senior	-1.15798	0.376	-3.081	0.0021

(4) **PTDP4** - (Perceived Team Support)

Table 21. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.43	1.22	5.89	0.89	5.72	0.91	5.62	1.08
Junior	4.21	2.23	4.17	2.08	3.67	1.97	4.06	2.10
All	5.12	1.47	5.21	1.53	5.21	1.53	5.23	1.56

Table 22. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.493619	0.242425	18.536
ConditionB	0.008126	0.355185	0.023
ConditionC	-1.035734	0.357806	-2.895
RoleSenior	1.052444	0.279636	3.764
ConditionB:RoleSenior	0.378688	0.410052	0.924
ConditionC:RoleSenior	1.259577	0.411585	3.060

Table 23. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.00813	0.355	-0.023	0.9997
	A - C	1.03573	0.358	2.895	0.0106
	B - C	1.04386	0.452	2.309	0.0546
Senior	A - B	-0.38681	0.205	-1.886	0.1428
	A - C	-0.22384	0.205	-1.091	0.5196
	B - C	0.16297	0.261	0.624	0.8071

Table 24. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.05	0.280	-3.764	0.0002
B	Junior - Senior	-1.43	0.384	-3.732	0.0002
C	Junior - Senior	-2.31	0.378	-6.117	<.0001

(5) **PTDP5** - (Diversity)

Table 25. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.33	1.39	5.72	1.19	5.39	1.40	5.44	1.35
Junior	4.08	2.04	4.83	2.25	3.92	2.02	4.23	2.08
All	5.02	1.66	5.50	1.54	5.02	1.68	5.14	1.64

Table 26. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.1787	0.3256	12.836
ConditionB	1.0312	0.5639	1.829
ConditionC	-0.3086	0.5639	-0.547
RoleSenior	1.2489	0.3759	3.322
ConditionB:RoleSenior	-0.6788	0.6511	-1.043
ConditionC:RoleSenior	0.3703	0.6511	0.569

Table 27. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-1.0312	0.564	-1.829	0.1602
	A - C	0.3086	0.564	0.547	0.8479
	B - C	1.3397	0.651	2.058	0.0988
Senior	A - B	-0.3523	0.326	-1.082	0.5251
	A - C	-0.0618	0.326	-0.190	0.9803
	B - C	0.2905	0.376	0.773	0.7197

Table 28. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.25	0.376	-3.322	0.0009
B	Junior - Senior	-0.57	0.532	-1.072	0.2836
C	Junior - Senior	-1.62	0.532	-3.046	0.0023

E.3 Perceived Decision Outcome Quality (PDOQ)

(1) **PDOQ1** - (Satisfaction)

Table 29. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.85	1.19	6.33	0.72	5.83	1.28	5.97	1.13
Junior	3.83	2.14	5.08	1.98	3.25	1.76	4.00	2.08
All	5.34	1.72	6.02	1.26	5.19	1.79	5.47	1.66

Table 30. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.0450	0.2269	17.826
ConditionB	1.1705	0.3055	3.831
ConditionC	-0.6315	0.3083	-2.048
RoleSenior	1.8910	0.2614	7.233
ConditionB:RoleSenior	-0.7596	0.3527	-2.154
ConditionC:RoleSenior	0.7086	0.3543	2.000

Table 31. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-1.1705	0.306	-3.831	0.0004
	A - C	0.6315	0.308	2.048	0.1009
	B - C	1.8020	0.399	4.522	<.0001
Senior	A - B	-0.4109	0.176	-2.328	0.0520
	A - C	-0.0771	0.176	-0.437	0.9001
	B - C	0.3338	0.230	1.449	0.3159

Table 32. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.89	0.261	-7.233	<.0001
B	Junior - Senior	-1.13	0.350	-3.234	0.0012
C	Junior - Senior	-2.60	0.342	-7.605	<.0001

(2) **PDOQ2** - (Validity)

Table 33. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.49	1.28	6.06	0.83	5.83	1.25	5.63	1.19
Junior	4.04	1.99	4.83	1.80	3.50	1.78	4.10	1.92
All	5.12	1.60	5.75	1.25	5.00	1.64	5.25	1.55

Table 34. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.3086	0.2365	18.216
ConditionB	0.6936	0.2872	2.415
ConditionC	-0.6659	0.2904	-2.293
RoleSenior	1.3042	0.2720	4.795
ConditionB:RoleSenior	-0.2670	0.3315	-0.805
ConditionC:RoleSenior	0.7650	0.3334	2.295

Table 35. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.6936	0.287	-2.415	0.0416
	A - C	0.6659	0.290	2.293	0.0567
	B - C	1.3596	0.382	3.555	0.0011
Senior	A - B	-0.4267	0.166	-2.571	0.0274
	A - C	-0.0991	0.166	-0.597	0.8217
	B - C	0.3276	0.221	1.482	0.2998

Table 36. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.30	0.272	-4.795	<.0001
B	Junior - Senior	-1.04	0.353	-2.939	0.0033
C	Junior - Senior	-2.07	0.341	-6.065	<.0001

E.4 NASA Task Load Index (NASA)

(1) **NASA1** - (Mental Demand)

Table 37. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	3.11	1.71	3.44	1.81	3.56	1.89	3.31	1.78
Junior	4.42	1.74	4.67	1.61	4.67	1.78	4.54	1.69
All	3.44	1.80	3.75	1.83	3.83	1.91	3.61	1.83

Table 38. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.6844	0.3749	12.495
ConditionB	-0.2439	0.3791	-0.643
ConditionC	0.8084	0.3842	2.104
RoleSenior	-1.6086	0.4290	-3.750
ConditionB:RoleSenior	0.5701	0.4374	1.303
ConditionC:RoleSenior	-0.5318	0.4407	-1.207

Table 39. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	0.2439	0.379	0.643	0.7961
	A - C	-0.8084	0.384	-2.104	0.0890
	B - C	-1.0523	0.517	-2.037	0.1035
Senior	A - B	-0.3262	0.219	-1.489	0.2962
	A - C	-0.2766	0.219	-1.262	0.4166
	B - C	0.0497	0.299	0.166	0.9849

Table 40. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	1.61	0.429	3.750	0.0002
B	Junior - Senior	1.04	0.529	1.964	0.0495
C	Junior - Senior	2.14	0.502	4.265	<.0001

(2) **NASA2** - (Temporal Demand)

Table 41. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	3.26	1.96	3.64	1.73	3.50	2.04	3.42	1.92
Junior	3.92	2.02	4.50	1.68	4.92	1.51	4.31	1.84
All	3.43	1.99	3.85	1.74	3.85	2.00	3.64	1.93

Table 42. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	3.9158	0.4312	9.082
ConditionB	0.3953	0.5164	0.766
ConditionC	1.0830	0.5222	2.074
RoleSenior	-0.6590	0.4957	-1.329
ConditionB:RoleSenior	-0.0776	0.5960	-0.130
ConditionC:RoleSenior	-0.7840	0.5995	-1.308

Table 43. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.3953	0.516	-0.766	0.7242
	A - C	-1.0830	0.522	-2.074	0.0952
	B - C	-0.6877	0.689	-0.998	0.5782
Senior	A - B	-0.3177	0.298	-1.065	0.5361
	A - C	-0.2990	0.298	-1.002	0.5756
	B - C	0.0187	0.398	0.047	0.9988

Table 44. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	0.659	0.496	1.329	0.1837
B	Junior - Senior	0.737	0.640	1.150	0.2501
C	Junior - Senior	1.443	0.618	2.334	0.0196

(3) **NASA3** - (Performance)

Table 45. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	5.58	1.03	5.78	0.96	5.47	1.08	5.60	1.02
Junior	3.83	1.69	4.92	1.78	3.83	1.53	4.10	1.70
All	5.15	1.44	5.56	1.25	5.06	1.39	5.23	1.39

Table 46. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	4.0089	0.2266	17.690
ConditionB	1.1278	0.3055	3.692
ConditionC	0.1187	0.3083	0.385
RoleSenior	1.6128	0.2611	6.177
ConditionB:RoleSenior	-0.9731	0.3527	-2.759
ConditionC:RoleSenior	-0.1518	0.3543	-0.428

Table 47. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-1.1278	0.306	-3.692	0.0007
	A - C	-0.1187	0.308	-0.385	0.9216
	B - C	1.0091	0.398	2.533	0.0304
Senior	A - B	-0.1547	0.177	-0.877	0.6550
	A - C	0.0331	0.177	0.188	0.9808
	B - C	0.1879	0.230	0.816	0.6933

Table 48. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-1.61	0.261	-6.177	<.0001
B	Junior - Senior	-0.64	0.350	-1.830	0.0672
C	Junior - Senior	-1.46	0.342	-4.277	<.0001

(4) **NASA4** - (Effort)

Table 49. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	4.89	1.46	5.03	1.40	5.00	1.60	4.95	1.47
Junior	5.33	1.13	5.42	1.08	5.75	0.62	5.46	1.01
All	5.00	1.39	5.12	1.33	5.19	1.45	5.08	1.39

Table 50. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	5.40305	0.26715	20.225
ConditionB	0.07746	0.46272	0.167
ConditionC	0.34695	0.46272	0.750
RoleSenior	-0.39341	0.30848	-1.275
ConditionB:RoleSenior	0.06200	0.53430	0.116
ConditionC:RoleSenior	-0.15234	0.53430	-0.285

Table 51. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	-0.0775	0.463	-0.167	0.9847
	A - C	-0.3470	0.463	-0.750	0.7337
	B - C	-0.2695	0.534	-0.504	0.8692
Senior	A - B	-0.1395	0.267	-0.522	0.8606
	A - C	-0.1946	0.267	-0.728	0.7466
	B - C	-0.0551	0.308	-0.179	0.9825

Table 52. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	0.393	0.308	1.275	0.2022
B	Junior - Senior	0.331	0.436	0.760	0.4475
C	Junior - Senior	0.546	0.436	1.251	0.2109

(5) **NASA5** - (Frustration Level)

Table 53. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	2.49	1.57	2.03	1.36	2.50	1.36	2.38	1.48
Junior	3.71	1.71	3.17	2.25	3.83	1.70	3.60	1.83
All	2.79	1.69	2.31	1.68	2.83	1.55	2.68	1.66

Table 54. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	3.5914	0.3052	11.768
ConditionB	-0.5740	0.3773	-1.521
ConditionC	0.4029	0.3813	1.057
RoleSenior	-1.2298	0.3511	-3.503
ConditionB:RoleSenior	0.2378	0.4355	0.546
ConditionC:RoleSenior	-0.4604	0.4379	-1.051

Table 55. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	0.5740	0.377	1.521	0.2808
	A - C	-0.4029	0.381	-1.057	0.5412
	B - C	-0.9768	0.501	-1.950	0.1248
Senior	A - B	0.3362	0.218	1.542	0.2713
	A - C	0.0575	0.218	0.264	0.9624
	B - C	-0.2787	0.290	-0.962	0.6007

Table 56. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	1.230	0.351	3.503	0.0005
B	Junior - Senior	0.992	0.458	2.166	0.0303
C	Junior - Senior	1.690	0.444	3.811	0.0001

E.5 Perception of AI Agent (PAA)

(1) **PAA1** - (Cooperation)

Table 57. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ
Senior	3.72	1.49	3.75	1.79	3.74	1.64
Junior	3.58	1.98	4.17	1.34	3.88	1.68
All	3.69	1.60	3.85	1.69	3.77	1.64

Table 58. Result of Robust Regression

Variable	Value	Std. Error	t value
(Intercept)	3.5073	0.5083	6.9001
ConditionC	0.6594	0.7188	0.9173
RoleSenior	0.1770	0.5869	0.3015
ConditionC:RoleSenior	-0.6156	0.8300	-0.7416

Table 59. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	B - C	-0.6594	0.719	-0.917	0.3590
Senior	B - C	-0.0438	0.415	-0.106	0.9159

Table 60. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
B	Junior - Senior	-0.177	0.587	-0.302	0.7630
C	Junior - Senior	0.439	0.587	0.747	0.4549

(2) **PAA2** - (Satisfaction)

Table 61. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ
Senior	3.72	1.67	4.22	1.79	3.97	1.74
Junior	4.00	1.86	3.00	1.95	3.50	1.93
All	3.79	1.70	3.92	1.89	3.85	1.79

Table 62. Result of Robust Regression

Variable	Value	Std. Error	t value
(Intercept)	4.0000	0.5544	7.2147
ConditionC	-1.1855	0.7841	-1.5120
RoleSenior	-0.2901	0.6402	-0.4531
ConditionC:RoleSenior	1.7422	0.9054	1.9243

Table 63. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	B - C	1.185	0.784	1.512	0.1305
Senior	B - C	-0.557	0.453	-1.230	0.2188

Table 64. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
B	Junior - Senior	0.29	0.64	0.453	0.6505
C	Junior - Senior	-1.45	0.64	-2.268	0.0233

(3) **PAA3** - (Quality)

Table 65. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ
Senior	4.11	1.70	4.19	1.72	4.15	1.70
Junior	4.00	1.71	3.17	1.99	3.58	1.86
All	4.08	1.69	3.94	1.83	4.01	1.75

Table 66. Result of Robust Regression

Variable	Value	Std. Error	t value
(Intercept)	4.1436	0.5517	7.5104
ConditionC	-1.4632	0.7802	-1.8753
RoleSenior	0.0356	0.6371	0.0558
ConditionC:RoleSenior	1.6591	0.9009	1.8416

Table 67. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	B - C	1.463	0.78	1.875	0.0608
Senior	B - C	-0.196	0.45	-0.435	0.6635

Table 68. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
B	Junior - Senior	-0.0356	0.637	-0.056	0.9555
C	Junior - Senior	-1.6947	0.637	-2.660	0.0078

(4) **PAA4** - (Fairness)

Table 69. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ
Senior	4.69	1.70	4.78	1.55	4.74	1.62
Junior	5.58	1.24	4.00	1.71	4.79	1.67
All	4.92	1.64	4.58	1.61	4.75	1.62

Table 70. Result of Robust Regression

Variable	Value	Std. Error	t value
(Intercept)	5.6102	0.4637	12.0993
ConditionC	-1.6102	0.6557	-2.4555
RoleSenior	-0.7857	0.5354	-1.4674
ConditionC:RoleSenior	1.6383	0.7572	2.1636

Table 71. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	B - C	1.6102	0.656	2.455	0.0141
Senior	B - C	-0.0281	0.379	-0.074	0.9408

Table 72. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
B	Junior - Senior	0.786	0.535	1.467	0.1423
C	Junior - Senior	-0.853	0.535	-1.592	0.1113

F DIALOGUE ANALYSIS

F.1 Message

Table 73. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	14.93	7.89	14.83	7.04	16.75	9.73	15.36	8.18
Junior	15.00	8.03	13.50	7.04	15.15	8.26	14.67	7.75
All	14.95	7.89	14.50	7.10	16.33	9.30	15.19	8.06

Table 74. Result of Robust Regression

Variable	Estimate	Std. Error	t value
(Intercept)	13.78202	1.52072	9.063
ConditionB	-0.05289	1.75796	-0.030
ConditionC	0.04669	1.73538	0.027
RoleSenior	0.50915	1.75952	0.289
ConditionB:RoleSenior	0.18671	2.03032	0.092
ConditionC:RoleSenior	0.98145	2.01080	0.488

Table 75. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	0.0529	1.76	0.030	0.9995
	A - C	-0.0467	1.74	-0.027	0.9996
	B - C	-0.0996	2.33	-0.043	0.9990
Senior	A - B	-0.1338	1.02	-0.132	0.9905
	A - C	-1.0281	1.02	-1.012	0.5691
	B - C	-0.8943	1.36	-0.656	0.7889

Table 76. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	-0.509	1.76	-0.289	0.7723
B	Junior - Senior	-0.696	2.23	-0.312	0.7551
C	Junior - Senior	-1.491	2.18	-0.683	0.4944

F.2 Character

Table 77. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	537.01	306.50	529.81	320.02	611.14	279.25	553.74	303.16
Junior	577.62	279.56	535.42	301.04	708.62	319.58	602.04	297.04
All	547.17	299.07	531.21	312.22	637.00	290.32	566.01	301.59

Table 78. Regression Results

Variable	Estimate	Std. Error	t value
(Intercept)	558.464	59.798	9.339
ConditionB	-27.730	61.934	-0.448
ConditionC	129.951	61.294	2.120
RoleSenior	-45.134	69.226	-0.652
ConditionB:RoleSenior	-4.082	71.529	-0.057
ConditionC:RoleSenior	-25.504	70.975	-0.359

Table 79. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	27.7	61.9	0.448	0.8954
	A - C	-130.0	61.3	-2.120	0.0858
	B - C	-157.7	83.4	-1.890	0.1416
Senior	A - B	31.8	35.8	0.889	0.6473
	A - C	-104.4	35.8	-2.919	0.0098
	B - C	-136.3	48.6	-2.801	0.0141

Table 80. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	45.1	69.2	0.652	0.5144
B	Junior - Senior	49.2	84.8	0.580	0.5617
C	Junior - Senior	70.6	83.2	0.849	0.3956

F.3 Normalized Engagement Score for each Discussion Session ($NES(i)$)

Table 81. Condition-wise Mean (μ) and Standard Deviation (σ)

	Condition A		Condition B		Condition C		All	
	μ	σ	μ	σ	μ	σ	μ	σ
Senior	0.25	0.10	0.25	0.10	0.24	0.10	0.25	0.10
Junior	0.26	0.11	0.24	0.10	0.25	0.10	0.25	0.10
All	0.25	0.10	0.25	0.10	0.24	0.10	0.25	0.10

Table 82. Regression Results

Variable	Estimate	Std. Error	t value
(Intercept)	0.250769	0.021511	11.658
ConditionB	-0.006349	0.020200	-0.314
ConditionC	-0.006913	0.020028	-0.345
RoleSenior	-0.006444	0.024913	-0.259
ConditionB:RoleSenior	0.006604	0.023329	0.283
ConditionC:RoleSenior	0.003502	0.023180	0.151

Table 83. Comparison of Contrasts Across Roles

Role	Contrast	Estimate	SE	z.ratio	p.value
Junior	A - B	0.006349	0.0202	0.314	0.9470
	A - C	0.006913	0.0200	0.345	0.9364
	B - C	0.000565	0.0275	0.021	0.9998
Senior	A - B	-0.000255	0.0117	-0.022	0.9997
	A - C	0.003412	0.0117	0.292	0.9540
	B - C	0.003667	0.0160	0.229	0.9715

Table 84. Comparison of Contrasts Across Conditions

Condition	Contrast	Estimate	SE	z.ratio	p.value
A	Junior - Senior	0.00644	0.0249	0.259	0.7959
B	Junior - Senior	-0.00016	0.0297	-0.005	0.9957
C	Junior - Senior	0.00294	0.0292	0.101	0.9197