ProMediate: A Simulation Testbed for Evaluating Proactive Mediation in Multi-Party Negotiation

Ziyi Liu[1]* **Bahar Sarrafzadeh**[2] **Pei Zhou**[2] **Longqi Yang**[2] **Jieyu Zhao**[1] **Ashish Sharma**[2]
[1]University of Southern California [2] Microsoft Corporation
{zliu2803, jieyuz}@usc.edu, sharma.ashish@microsoft.com

Abstract

While LLMs increasingly assist individual users, there is a critical need for agents that can proactively manage complex, multi-party collaboration. However, the scarcity of systematic evaluation methods for these group dynamics limits the development of AI capable of effectively supporting teams. Here, we present PROMEDIATE[1], the first testbed for evaluating proactive AI mediator agents in complex, multi-topic, multi-party negotiations. PROMEDIATE consists of two core components: (i) a simulation environment based on realistic negotiation cases with a plug-and-play proactive AI mediator, capable of flexibly deciding when and how to intervene; and (ii) a socio-cognitive evaluation framework with a new suite of metrics to measure consensus changes, intervention latency, mediator effectiveness, and intelligence. These components establish a systematic framework for assessing the capability of proactive AI agents in multi-party settings. Our results show that a socially intelligent mediator agent outperforms a generic baseline, via faster, better-targeted interventions. In the PROMEDIATE-Hard setting, our social mediator increases consensus change by 3.6 percentage points compared to the generic baseline (10.65% vs. 7.01%) while being 77% faster in response (15.98s vs. 3.71s). In conclusion, PROMEDIATE provides a rigorous, theory-grounded testbed to advance the development of proactive, socially intelligent agents.

1 Introduction

Large Language Models (LLMs) are now widely integrated into agentic frameworks to assist individual users in completing diverse tasks such as information seeking and social skill development (Yang et al., 2024a; Shaikh et al., 2024; Eigner and Händler, 2024). While agent applications designed for individual users have shown promise, they differ significantly from real-world situations where effective results depend on collaboration among multiple users (Marks et al., 2001; Kozlowski and Ilgen, 2006; Li et al., 2024). This gap highlights a growing need for agents capable of proactively managing multi-party interactions and facilitating collaborative workflows. Prior research on AI agents in multi-party scenarios has either focused on qualitative analyses of proactive agents (Houde et al., 2025; Alsobay et al., 2025) or on reactive agents that provide assistance only when explicitly prompted (Chiang et al., 2024; Chiang, 2025; Chen et al., 2025). While some proactive agents have been designed for multi-party settings (Houde et al., 2025; Wesche and Sonderegger, 2019), systematic evaluation methods to guide and measure progress in this domain remain scarce. Developing simulation testbeds that support controlled, reproducible experiments is therefore a timely and essential challenge for advancing the capabilities of proactive multi-party AI agent.

Multi-party conversations demand socio-cognitive intelligence—the ability to track diverse perspectives and proactively steer discussions—beyond simple task solving. For instance, in complex negotiations where stakeholders reach a deadlock (Figure 1), a proactive agent must autonomously detect breakdowns and guide participants toward consensus across multiple topics. However, existing benchmarks typically rely on simplified games (Abdelnabi et al., 2024; Bianchi et al., 2024) or bilateral interactions (Zhou et al., 2023; Wu et al., 2025; Kwon et al., 2024), overlooking the nuanced dynamics of multi-party conflict. See Table 1 for a comparison of how PROMEDIATE addresses this gap with theory-based simulation environment and socio-cognitive evaluation metrics.

To address these gaps, we introduce PROMEDIATE (Figure 1), the first simulation testbed designed

*Work done during Ziyi's internship at Microsoft Corp.
[1]The code is available at https://github.com/microsoft/ProMediate-Eval.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29570–29598
July 2-7, 2026

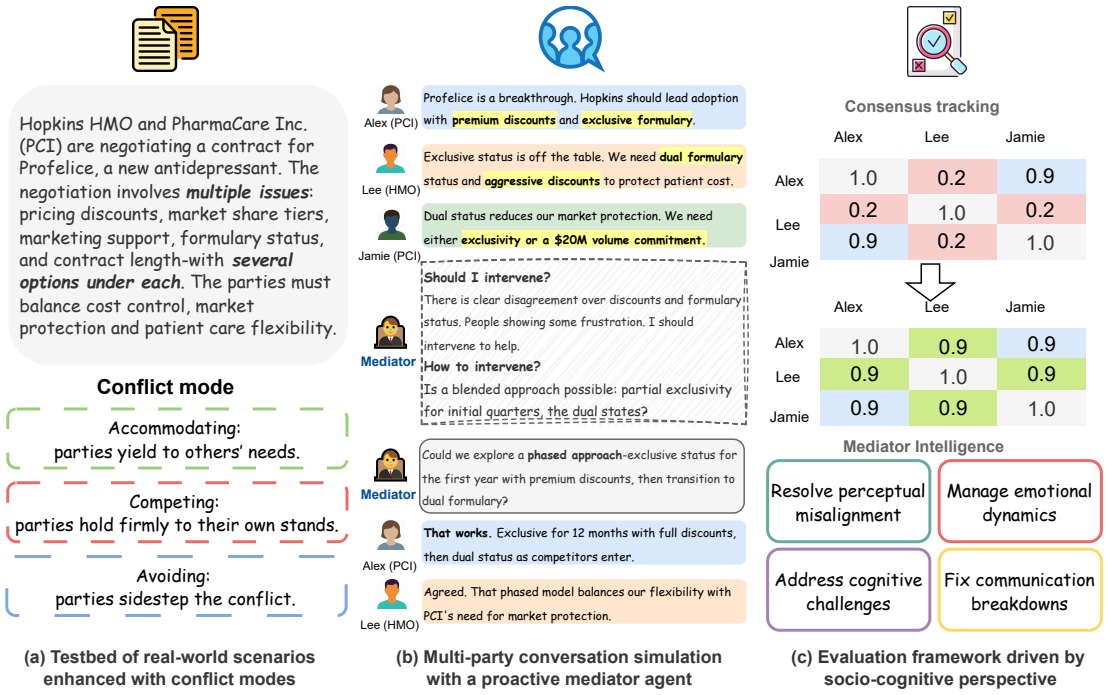


Figure 1: The illustration of **ProMediate** framework, involving a **multi-topic, multi-option** negotiation scenario with different conflict modes (Cai and Fink, 2002); conversation simulation with a plug-and-play agent; a suite of socio-cognitive evaluation metrics to capture the evolving nature of the negotiation.

to evaluate proactive agents in complex multi-party conversations. By establishing a controlled environment for agents to interact in complex scenarios, PROMEDIATE offers a powerful lens to observe and stress-test collaborative behaviors before real-world deployment. The framework consists of two components: (1) a simulation engine grounded in conflict theory, featuring simulated humans with structured preferences and a plug-and-play proactive AI mediator; and (2) a socio-cognitive framework for systematically measuring negotiation outcomes and mediator behaviors. A key feature of our simulation is that the AI mediator acts proactively—the agent must autonomously decide both *when to intervene* and *how to intervene*.

To provide a holistic evaluation, our framework measures the interaction from two complementary perspectives. First, to capture **group dynamics**, we introduce a novel *Consensus Tracking* method that quantifies the real-time evolution of agreement and topic-level efficiency. Second, to assess **agent behavior**, we evaluate *Mediator Intelligence* across perceptual, emotional, cognitive, and communicative dimensions. This dual approach allows us to decouple the negotiation state (what happened to the group) from the mediator's capability (how the agent reasoned), enabling a fine-grained analysis of success and failure in proactive mediation. We evaluate three settings—*NoAgent*, *Generic Mediator*, and *Socially Intelligent Mediator*—across six scenarios and three difficulty levels (PROMEDIATE *Easy*, *Medium*, and *Hard*). In the *Hard* setting with *Competing* participants, the Socially Intelligent Mediator achieves larger consensus gains than the Generic Mediator (+3.6%) while responding about 3× faster. Scenario difficulty strongly modulates outcomes: *Easy* settings yield larger and more stable gains, whereas *Hard* settings are more volatile but benefit the most from proactive AI mediation compared to *NoAgent*. Finally, our metrics align with human judgments and reveal two key dimensions of success—consensus/efficiency and intervention tempo—supporting construct validity. Together, these results underscore the need for multi-faceted evaluation to assess proactive agents in real-world multi-party settings. We highlight our primary contributions across three key dimensions:

- **Extensible testbed:** A challenging testbed for realistic multi-party interactions with a plug-and-play framework for evaluating diverse agents.
- **Socio-cognitive grounding:** Agent design and evaluation are grounded in socio-cognitive theory, enabling principled assessment of socially intelligent mediation.
- **Comprehensive evaluation:** A systematic metric suite that captures complementary dimen-

sions of agent performance and reveals trade-offs between effectiveness and timing.

2 PROMEDIATE Testbed

To evaluate proactive mediation in multi-party negotiations, we design a testbed that combines realistic simulations with configurable mediator interventions. We introduce the negotiation scenario setup, followed by multi-party conversation simulation with plug-and-play agents.

2.1 Negotiation Scenario Setup

To ensure conversational complexity and diversity, we adopt negotiation training materials from Harvard Law School's Program on Negotiation (pon.harvard.edu/store). These scenarios span domains such as healthcare, environmental policy, and business development, and typically require 2–3 hours for students to complete. **We select six multi-party scenarios, each involving multiple topics and multiple options per topic.**

We formally structure each scenario as a multi-party negotiation framework with N parties $\{P_1, P_2, \ldots, P_N\}$ discussing M distinct topics $\{T_1, T_2, \ldots, T_M\}$. For each topic T_j, there exists a finite set of S_j available options $O_j = \{o_{j,1}, o_{j,2}, \ldots, o_{j,S_j}\}$. Each party P_i maintains an explicit preference ranking in the beginning. For instance, party P_1's preference ordering for topic T_1 will be represented as $o_{1,2} > o_{1,3} > o_{1,1}$, indicating that option $o_{1,2}$ is most preferred. During conversation initialization, all background knowledge and individual preference profiles are incorporated into each agent's memory system. Table 5 in Appendix B.1 shows a sample scenario.

2.2 Conversation Setup

Human Simulation Module We use an LLM-based conversational setup inspired by the InnerThought framework (Liu et al., 2025a), which enables agents to generate internal thoughts and proactively participate in dialogue. While the original formulation focuses on open-ended chit-chat, our setting targets structured negotiation tasks that require explicit modeling of multi-topic agendas and principled decisions about when and how a mediator should intervene. Accordingly, we provide each simulated agent with a detailed negotiation context, a structured preference profile (Section 2.1), and an explicit identity, and adopt a negotiation-oriented reasoning process to guide agent participation and intervention behavior.

Mediator Agent Module We design a **plug-and-play mediator** by explicitly separating *when to intervene* and *how to intervene* (Figure 1(b)). The mediator continuously monitors the conversation and decides whether to intervene. If it intervenes, it generates a response and other simulated humans are skipped for that turn; otherwise, the turn is handled by the human simulation module. This design keeps the mediator modular and independent, avoids joint orchestration with human agents, and isolates the mediator's contribution, making its impact on multi-party collaboration easier to evaluate. Prompts are provided in Appendix E.

Conflict Modes Since assigned roles significantly shape conversational dynamics (Thomas, 2008; Zhang et al., 2018), we instantiate group-level personas via shared conflict modes to ensure diversity. To improve persona realism, we iteratively refined the prompts beyond rigid, binary behaviors: agents are allowed to adapt their mode-conditioned behavior based on the perceived reasonableness of the mediator's suggestions, yielding more natural negotiation dynamics while preserving the intended conflict style. Drawing on established theory (Cai and Fink, 2002; Thomas, 2008), we incorporate three modes:[2] (Cai and Fink, 2002; Ma, 2007; Thomas, 2008) : (1) **Competing**: parties adopt firm positions and prioritize their own interests; (2) **Avoiding**: parties strategically sidestep contentious topics and resolve the easier ones first; (3) **Accommodating**: parties are receptive to others' views and willing to cooperate when necessary. In Section 4, we use these modes to create the PROMEDIATE-Easy, PROMEDIATE-Medium, and PROMEDIATE-Hard difficulty levels for our evaluation framework.

Conversation Quality Evaluation. We conduct a human evaluation on 200 simulated conversations, rated by three annotators on a 5-point Likert scale for naturalness and conflict mode reflection. The conversations achieve high naturalness (mean 4.35) and reliable mode reflection (mean 3.87), consistent with our goal of conveying conflict mode without enforcing extreme behaviors. Full details are provided in Appendix F.1.

[2] https://www.psychometrics.com/conflict-resolution-skills-how-to-get-the-best-of-each-of-the-five-modes/

Work	Human–AI	>2 Parties	Proactive Agent	Socio-Cognitive Eval.	Dynamic Group Outcome Trajectory
MultiAgentBench (Zhu et al., 2025)	✗	✓	✓	✗	✗
CollabLLM (Wu et al., 2025)	✓	✗	✓	✗	✗
Sotopia (Zhou et al., 2023)	✓	✗	✓	◐	✗
AgentSense (Mou et al., 2024)	✓	✓	✗	✓	✗
LAMEN (Davidson et al., 2024)	✓	✗	✗	✗	✗
AgentMediation (Chen et al., 2025)	✓	✓	✗	◐	✗
ProMediate (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison of agent interaction benchmarks and systems. ✓indicates explicit support, ✗indicates not supported, and ◐indicates partial or implicit support. Our framework uniquely supports proactive mediation, socio-cognitive evaluation, and dynamic outcome trajectory within human–AI multi-party settings.

3 ProMediate Metrics

We evaluate proactive mediators in simulated multi-party conversations through a socio-cognitive lens, assessing two key dimensions: (1) group consensus dynamics as a socio-cognitive outcome—tracking how the agreement emerges and fluctuates throughout the negotiation; and (2) the mediator's socio-cognitive intelligence —assessing mediation skills.

3.1 Consensus Tracking

Consensus is not merely a procedural end state but a dynamic socio-cognitive achievement (Swaab et al., 2007; Levine, 2018; Butera et al., 2019). Since multi-topic negotiations rarely end in full unanimity, defining consensus as a binary outcome is overly restrictive (del Moral et al., 2018). Unlike prior work focusing on final outcomes (Fu et al., 2023; Abdelnabi et al., 2024), we introduce *consensus tracking*—a time-varying measure that captures how agreement evolves and attitudes shift throughout the interaction.

Consensus tracking consists of two components as shown in Algorithm 1: *(i) attitude extraction* and *(ii) agreement scoring*. Attitude extraction can be approached in several ways, such as probing a speaker's latent mental states or estimating preferences over pairs of options (del Moral et al., 2018). However, real-world conversations pose practical challenges that are often overlooked: (i) the option set is open—new alternatives may be introduced mid-conversation—so methods assuming a fixed inventory are brittle; (ii) internal mental states can diverge from the attitudes perceived by others; and (iii) not every topic is mentioned at every turn, making turn-level "overall" attitude estimates ill-posed. To address these challenges, we use an LLM (GPT-4.1) to infer each participant's stance toward each topic directly from utterance text at each turn, yielding topic-specific, turn-conditioned attitudes without assuming fixed option sets or access to latent states (see Appendix E). Participants are initialized with preferences over all topics, forming a complete attitude profile. At each subsequent turn, attitudes are updated when a topic is mentioned; otherwise, the previous attitude is retained.

For agreement scoring, we compute pairwise agreement scores based on extracted attitudes between parties on each topic and then average these scores across all pairs to obtain a group-level measure. We employ an *LLM-as-a-judge* approach, prompting GPT-4.1 to assign an agreement score in the range $[0, 1]$ along five dimensions we create according to multiple socio-cognitive theories (Griffiths et al., 2021; Thomson et al., 2009; Bedwell et al., 2012): shared goals, common understanding of the problem context, acceptance of proposed terms, cooperative tone and willingness to compromise, and alignment in decision-making processes. We also explore alternative agreement scoring formulations, such as single-dimension scoring and incorporating temporal context by conditioning on the previous turn's agreement score. These variants yield similar consensus dynamics, and we report additional details in Appendix C. Human validation results are in Sec 3.4.

3.2 Socio-Cognitive Intelligence

Successfully mediating and resolving conflict impasses requires strong socio-cognitive intelligence. To measure this, we take inspiration from socio-cognitive frameworks to evaluate the intelligence of mediators. We operationalize problems which could happen in a negotiation along four dimensions, adapted from the mediation theory matrix (Zariski, 2010):

- **Perceptual Differences**: divergences in beliefs, interpretations, or framings of key issues.
- **Emotional Dynamics**: negative emotions (for example, anger, distrust, grief) that derail constructive engagement.

Algorithm 1 Consensus Tracking

Require: Initial attitudes $Attitude_0[P_i][T_m]$ for each participant P_i and topic T_m

Require: Conversation turns $C = [(P_1, u_1), (P_2, u_2), \ldots, (P_N, u_N)]$

Initialize agreement scores $A_0[P_i][P_j]$ for all participant pairs (P_i, P_j)

for each turn $(P_i, u_i) \in C$ and each topic $T_m \in T$ **do**

  if T_m is mentioned in utterance u_i **then**

    Update $Attitude_i[P_i][T_m]$ based on attitude extraction of u_i

  else

    $Attitude_i[P_i][T_m] \leftarrow Attitude_{i-1}[P_i][T_m]$

  for each participant $P_j \neq P_i$ **do**

    Compute agreement score $A_i[P_i][P_j][T_m]$ as the average over five dimensions using LLM-as-a-judge

- **Cognitive Challenges**: reasoning failures or biases (e.g., anchoring, confirmation bias) and limited option generation.
- **Communication Breakdowns**: ineffective exchange, talking about one another, hostility / escalation or nonresponsiveness.

We evaluate whether the mediator can recognize those key problems in the negotiation and propose effective strategies to resolve them.

3.3 Evaluation metrics

Building on the consensus tracking and socio-cognitive intelligence, we report metrics from two complementary perspectives: (i) *conversation-level outcomes* that characterize consensus dynamics independently of any mediator, and (ii) *mediator-level effectiveness*, evaluating whether interventions are effective and yield measurable improvements in consensus:

- **Consensus Change (CC)** measures aggregated improvement in agreement from the start to the end of a dialogue. To mitigate fluctuations and outliers, we compute it as $CC = \frac{1}{10}\sum_{t=T-9}^{T} C_t - \frac{1}{10}\sum_{t=1}^{10} C_t$, where C_t is the group consensus at turn t and T is the final turn.
- **Topic-Level Efficiency (TLE)** measures agreement change per topic normalized by the number of turns mentioning that topic, i.e., $TLE = \frac{1}{|\mathcal{T}|}\sum_{k\in\mathcal{T}} \frac{C_k^{\text{end}} - C_k^{\text{start}}}{n_k}$, where n_k is the number of turns mentioning topic k. This captures how efficiently consensus is reached for each topic.
- **Response Latency (RL)** captures how quickly the mediator responds to low-consensus states. We start a timer at a drop event, defined as a consensus decrease greater than τ=0.1 within the next W=10 turns. For an event beginning at turn t, latency is $RL(t) = t_{\text{int}} - t$, where t_{int} is the next mediator intervention turn; if the mediator never intervenes, $RL(t) = +\infty$.
- **Mediator Effectiveness (ME)** measures whether an intervention positively alters the subsequent consensus trajectory on the targeted topic. We fit linear trends to agreement scores over the five turns before and after the intervention, and define effectiveness as $ME = s_{\text{post}} - s_{\text{pre}}$, where s_{pre} and s_{post} are the fitted slopes before and after the intervention. Higher values indicate faster consensus improvement.
- **Mediator Intelligence (MI)** A good mediator should exhibit good social intelligence at each intervention. Formally, $MI = \frac{1}{|\mathcal{D}_i|}\sum_{d\in\mathcal{D}_i} s_d$, where $s_d \in [1, 5]$ is the judge-assigned score for each applicable socio-cognitive dimension and $\mathcal{D}_i$ is the set of applicable dimensions for intervention i. We assess mediator intelligence by evaluating whether interventions by the mediator is trying to address core challenges within the dialogue. Specifically, we measure performance along four socio-cognitive dimensions: perceptual differences, emotional dynamics, cognitive challenges, and communication breakdowns. To quantify these aspects, we use an LLM-AS-A-JUDGE framework, asking GPT-4.1 to assign a score from 1 to 5 for each dimension when applicable. We average the scores across all dimensions; scoring criteria are detailed in Appendix C.3. We report human validation results in Sec. 3.4.

3.4 Human Evaluation

We conduct human evaluations to validate tasks that rely on LLM-as-a-judge. We evaluate three conversation-based tasks: **(1) Attitude Verification**, **(2) Consensus Comparison**, and **(3) Mediator Intelligence Rating**. Attitude Verification and Consensus Comparison each include 200 instances. In Attitude Verification, each instance contains a conversation, a target utterance, and an automatically extracted attitude toward a specific topic; annotators judge its correctness (*Yes*/*No*/*Maybe*). In Consensus Comparison, absolute consensus scores

are difficult for humans to assign, so we use pairwise comparisons between a higher- and a lower-consensus snippet. For Mediator Intelligence Rating, we sample 200 mediator utterances and retain only dimension-relevant cases, yielding 587 instances across 4 dimensions. All instances are annotated by three master's-level computer science students, with majority voting used for aggregation. Details of annotations are in Appendix F.2.

Results. In **Attitude Verification**, annotators classified the extracted attitudes as *Yes* (83.5%), *No* (11%), or *Maybe* (5.5%). Attitude Verification failures primarily stem from hallucinations, which propagate downstream to mischaracterize participant alignment and distort consensus rankings. In **Consensus Comparison**, human judgments align with our metric's directionality in 91% of instances, indicating that the metric can reliably distinguish between higher- and lower-consensus conversation segments. For **Mediator Intelligence**, human annotators are generally more lenient, assigning scores approximately 0.5 points higher than the model on average. We group Likert ratings into low (1–2), medium (3), and high (4–5), human and model judgments agree in 90% of cases.

4 Evaluations with ProMediate

4.1 Agent design

Our framework supports any LLM-based AI mediator agent. In this paper, we implement both a generic baseline mediator and a socially intelligent mediator. Future work can build upon this foundation to further extend and evaluate agent design. All detailed prompts are shown in E.

Generic Mediator The generic mediator is a general-purpose multi-party agent with basic conversational skills. It uses two simple prompts to decide interventions, avoiding complex or theory-based reasoning.

Socially Intelligent Mediator In contrast to the generic mediator, our socially intelligent mediator is grounded in a socio-cognitive framework. During the *When* phase, it analyzes the conversation along four socio-cognitive dimensions (Sec. 3.2) to detect perceptual, emotional, cognitive, and communicative breakdowns and compute a *motivation-to-intervene* score; if this score exceeds a threshold, the mediator intervenes. In the *How* phase, it selects an intervention strategy informed by mediation theory (Munduate et al., 2022; Boyle, 2017; McKenzie, 2015), considering Facilitative, Evaluative, Transformative, and Problem-Solving mediation (Appendix D). The mediator treats these strategies as guidance, evaluates multiple candidates, and executes the highest-scoring one.

4.2 Experiment setup

Models We adopt Claude-Sonnet-4 as our human simulator, as a we found it produced the most natural and human-like conversational behavior.[3] For the **AI mediator**, we evaluate different types of agents – varying based on the agent characteristics (generic vs. social) and varying based on models (o4-mini vs. GPT-4.1 vs. Claude-Sonnet-4 vs Qwen3-4b-Instruct). Among the open models we tested, only **Qwen-3 4B-Instruct** reliably completed the full simulation; **LLaMA-3 8B** failed early, while **Qwen-3 8B** and **Qwen-3 4B-Thinking** often violated the required JSON format.

Modes We structure our experiments across three difficulty levels: (a) ProMediate-Easy (accommodating/avoiding conflict modes (Section 2.2)); (b) ProMediate-Medium (a general mode as a non-persona baseline, where participants do not follow any predefined conflict mode); (c) ProMediate-Hard (competing mode (Section 2.2)).

Conversation Simulation We evaluate three settings—*NoAgent*, *Generic*, and *Social* across different difficulty modes—conducting 30 independent runs per configuration to mitigate variance. Crucially, we enforce a fixed global turn limit based on scenario complexity, which remains identical for both mediated and unmediated runs. This means the agent does not get "extra" turns; instead, every agent intervention consumes a turn from the shared budget, effectively replacing a human opportunity to speak. This rigorous constraint isolates efficiency: we test whether the mediator can guide the group to consensus within the exact same conversational bandwidth, rather than simply extending the discussion.

4.3 Results and Analysis

We present our results and analysis by addressing three research questions.

- **RQ1: Agent and Model Evaluation:** How do agent types (Generic vs. Social) and model

[3] An o4-mini based LLM judge found Sonnet-4 to be 25% to 70% more natural and human-like in its responses compared to models like GPT-4o, GPT-4.1, and o4-mini.

	ProMediate-Easy						ProMediate-Medium			ProMediate-Hard		
Mode	Accommodating			Avoiding			General			Competing		
Method	NoAgent	Generic	Social	NoAgent	Generic	Social	NoAgent	Generic	Social	NoAgent	Generic	Social
CC ↑	18.74%	20.13%	22.59%	17.49%	14.31%	13.25%	11.36%	10.93%	11.39%	6.83%	7.01%	10.65%
TLE ↑	1.05%	1.18%	1.16%	1.17%	1.04%	0.48%	0.54%	0.44%	0.74%	0.50%	0.23%	0.57%
RL ↓	-	6.39s	4.00s	-	25.56s	5.69s	-	5.64s	3.00s	-	15.98s	3.71s
ME ↑	-	1.18%	0.82%	-	0.17%	0.89%	-	2.01%	0.25%	-	1.75%	0.59%
MI ↑	-	4.464	4.319	-	4.260	4.445	-	4.292	4.207	-	4.225	4.318

Table 2: Results are reported across all scenarios with GPT-4.1 as the mediator backbone. For the *NoAgent* baseline, we include only conversation-level metrics that do not depend on a mediator. Each cell is the mean over 6 scenarios × 5 runs per scenario (30 conversations total). Abbrev: CC = Consensus Change (%), TLE = Topic-Level Efficiency (%), RL = Response Latency (s), ME = Mediation Effectiveness (%), MI = Mediator Intelligence (1–5). ↑: high is better; ↓: low is better

variants (o4-mini, GPT-4.1, Sonnet-4) influence socio-cognitive outcomes in negotiation?

- **RQ2: Impact of Difficulty:** How does the difficulty of a negotiation scenario influence the effectiveness of AI mediation?
- **RQ3: Construct Validity:** To what extent do our proposed metrics demonstrate construct validity, and what do they reveal about the underlying dimensions of effective AI mediation?

4.3.1 RQ1: Agent and Model Evaluation

Socially intelligent mediator is more effective in ProMediate-Hard compared to ProMediate-Easy. As shown in Table 2, the Socially intelligent mediator is consistently more proactive than the Generic baseline, intervening more frequently and with lower latency. In ProMediate-Easy—where participants favor compromise—this proactivity offers minimal value and can disrupt organic consensus. In contrast, in ProMediate-Hard, the same behavior is advantageous, yielding the largest gains in consensus and efficiency. These results indicate that intervention utility is context-dependent rather than inherently beneficial. Accordingly, adaptive strategies that calibrate when and how to intervene are essential.

Thinking model o4-mini performs the best Among the three models, **o4-mini** is the most effective mediator, achieving the highest consensus change (**9.34%**) across conversations. Although it has the slowest response latency (**5.47s**), this may not be a disadvantage: a longer latency likely reflects more deliberate reasoning, which can lead to higher quality interventions and better negotiation outcomes. In contrast, **GPT-4.1** offers a balanced profile with a strong consensus change (**8.99%**), and moderate latency (**4.26s**), making it a reliable alternative. **Claude-Sonnet-4** is the fastest responder (**2.36s**), with a lower consensus change (**4.71%**), suggesting that speed alone does not guarantee effective mediation. Open-source models remain notably less reliable in our setting. As shown in Table 3, **Qwen-3 4B-Instruct** is the only open-source model we found that can finish the full pipeline end-to-end, reaching **10.2%** consensus change with low latency (**1.17s**); however, the broader open-source pool still suffers from execution and formatting failures, highlighting systems robustness as a key bottleneck. The model-comparison results also help address a potential evaluator-bias concern. If GPT-4.1 as judge systematically favored its own family, we would expect GPT-4.1 to dominate Table 3; instead, **o4-mini** scores higher on both consensus change and topic-level efficiency. We further verified this pattern for mediator intelligence with **o4-mini** as the judge, where GPT-4.1 scored **4.17**, compared with **3.78** for o4-mini and **3.73** for Claude-Sonnet-4. The relative ordering is therefore unchanged across judges, suggesting that the evaluation captures meaningful performance differences rather than simple self-preference by the judge model.

Models	**GPT-4.1**	**Claude**	**o4-mini**	**Qwen-3 4B-Inst.**
CC	8.99%	4.71%	9.34%	10.2%
TLE	0.37%	0.34%	0.74%	0.45%
RL	4.26s	2.36s	5.47s	1.17s
ME	2.08%	1.70%	2.59%	1.34%
MI	4.841	3.793	3.865	4.36

Table 3: Results across different models on one scenario with the Socially Intelligent Mediator. Among small size open-source models, only Qwen-3 4B-Instruct reliably completed the full simulation; the others failed early or frequently produced invalid JSON.

4.3.2 RQ2: Impact of Scenario Difficulty

Scenario difficulty exerts a direct influence on the mediator's intervention patterns. As pre-

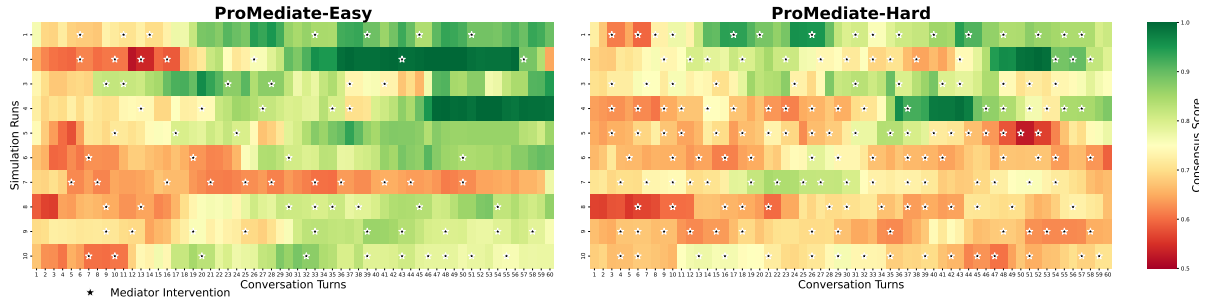


Figure 2: Intervention dynamics. We visualize 10 simulation runs under both PROMEDIATE-Easy and PROMEDIATE-Hard settings. Each row corresponds to a single run and illustrates the evolution of the consensus score over the course of the conversation. Mediator interventions are indicated by star markers.

Metric	Factor 1	Factor 2
CC	**0.997**	–0.113
TLE	**0.802**	–0.086
ME	–0.023	**0.465**
RL	–0.155	**0.420**
MI	0.235	0.249

Table 4: Rotated factor loadings (Varimax). Loadings with $|\lambda| \geq 0.40$ in **bold**. Proposed factor labels: Factor 1 = *Consensus & Topic Efficiency*; Factor 2 = *Intervention Dynamics / Tempo*.

sented in Table 2, the results confirm the validity of our difficulty design: PROMEDIATE-Easy negotiations achieve significantly higher consensus change (22.59%) compared to the Hard setting (10.65%). However, relying solely on these aggregate outcomes obscures the agent's adaptive behavior. Figure 2 reveals the distinct intervention dynamics required to manage these different difficulty levels. As shown in the raster plot, the PROMEDIATE-Easy runs are characterized by sparse interventions and rapid transitions to high consensus (green). In sharp contrast, the PROMEDIATE-Hard runs reveal a dense concentration of interventions overlaid on prolonged periods of low consensus (red/yellow). This visual disparity highlights a critical finding: the Social mediator is highly sensitive to the frequent breakdowns in Hard settings and responds by significantly scaling up its intervention frequency. Thus, the lower final consensus in Hard scenarios is not a sign of agent passivity, but rather reflects a vigorous, proactive effort to combat persistent deadlocks.

4.3.3 RQ3: Construct Validity

Two main latent factors: *Consensus & Topic Efficiency* and *Intervention Dynamics / Tempo*. To better understand the metrics, we conduct an exploratory factor analysis (Watkins, 2018), a statistical technique used to identify the hidden "factors" or dimensions that connect the underlying relations among metrics. As shown in Table 4, a clear two-factor structure emerges: *Factor 1 (Consensus & Topic Efficiency)* is defined by strong positive loadings on **consensus_change** (≈ 0.997) and **topic_efficiency** (≈ 0.802), capturing progress toward alignment while staying on-topic; *Factor 2 (Intervention Dynamics / Tempo)* is defined by **Mediation Effectiveness** (≈ 0.465) and **Response_latency** (≈ 0.420), describing the tempo and yield of interventions, while **Mediator Intelligence** does not load saliently and are best treated as a separate outcome.

Decoupling Process Quality from Immediate Outcome. To better understand the relationship between **Mediator Effectiveness (ME)** and **Mediator Intelligence (MI)**, we computed the Spearman correlation between the two metrics at the intervention level. The resulting coefficient is negligible at 0.01 ($p = 0.89$), indicating the absence of a statistically significant relationship (see Figure 5). Far from a measurement anomaly, this statistical independence aligns with established negotiation theory, which posits that subjective process quality is distinct from instrumental outcomes (Curhan et al., 2006). As illustrated in Figure 4a, consensus scores frequently exhibit a decline immediately following high-quality interventions. We attribute this to two distinct phenomena. First, **Participant Resistance**: in PROMEDIATE-Hard, participants often reject even optimal suggestions due to entrenched positions. Second, **Constructive Friction**: sophisticated mediation often requires surfacing latent disagreements—forcing participants to articulate conflicting preferences—rather than rushing to superficial agreement. This mirrors the differentiation phase (Walton and Mckersie, 1965),

or the necessary storming stage in group dynamics (Tuckman, 1965), where temporary discord lays the groundwork for robust, long-term alignment.

5 Related Work

5.1 Collaborative AI

Collaborative AI research spans *multi-agent systems* for task coordination (Tran et al., 2025; Hong et al., 2024; Wang et al., 2024; Chen et al., 2023) and *human–agent collaboration* assisting individuals with reasoning (Feng et al., 2024) or workflows (Xu et al., 2025; Shao et al., 2024). While some agents support decision-making (Yang et al., 2024b; Chiang, 2025), they remain largely reactive. This highlights a critical need for proactive agents with the social intelligence to anticipate breakdowns and intervene strategically in dynamic, multi-party collaboration.

5.2 Socially Intelligent agent

Deploying LLMs in complex workflows requires robust social interaction capabilities (Xu et al., 2024). Existing benchmarks assess this in games (Liu et al., 2024; Feng et al., 2025), family disputes (Mou et al., 2024), and collaboration (Zhou et al., 2023; Goel and Zhu, 2025; Liu et al., 2025a), citing frameworks like Theory of Mind and cultural intelligence (Liu et al., 2025b; Laine et al., 2023; Berglund et al., 2023). Building on this foundation, we evaluate social intelligence through the lens of mediation cognitive theory, focusing on how agents facilitate group decision-making by managing multiparty dynamics, surfacing disagreements, and guiding conversations toward resolution.

6 Conclusion

We introduced PROMEDIATE to simulate and evaluate proactive, socially intelligent mediation in complex multi-party, multi-issue negotiations. We proposed metrics along two axes—conversation-level outcomes and mediator effectiveness—covering consensus change, response latency, efficiency, and intelligence. Results show that while socially intelligent mediators improve dynamics, they do not guarantee immediate consensus, highlighting trade-offs between short-term movement and long-term alignment. We hope PROMEDIATE catalyzes rigorous, theory-informed progress on collaborative AI for real-world group decisions.

7 Limitations

Simulated Dynamics. LLM agents serve as proxies for human behavior, lacking real-world irrationality, long-term history, and non-verbal cues (e.g., tone). Consequently, our findings represent theoretical behavioral templates rather than guaranteed performance with human users.

Scope Constraints. This study is limited to synchronous, English, text-based negotiation. It does not account for cultural variances in communication norms (e.g., direct vs. indirect intervention) or the critical socio-cognitive impact of voice and visual modalities in dispute resolution.

Task Generalizability. Our scenarios focus on goal-oriented, multi-issue negotiations. It remains an open question whether the PROMEDIATE strategies transfer effectively to purely relational conflicts or highly polarized political debates where logical consensus is not the primary objective.

8 Ethics Statement

This work utilizes Large Language Models to simulate negotiation dynamics in a controlled research environment. We emphasize that our agents are designed for theoretical evaluation of mediation strategies and are not currently intended for real-world dispute resolution. As the field progresses toward deployment, ensuring agent neutrality and transparency regarding AI identity will remain paramount to maintaining user trust and autonomy.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Mohammed Alsobay, David M. Rothschild, Jake M. Hofman, and Daniel G. Goldstein. 2025. Bringing everyone to the table: An experimental study of llm-facilitated group decision making. *Preprint*, arXiv:2508.08242.

Wendy L Bedwell, Jessica L Wildman, Deborah Diaz-Granados, Maritza Salazar, William S Kramer, and Eduardo Salas. 2012. Collaboration at work: An integrative multilevel conceptualization. *Human resource management review*, 22(2):128–145.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *Preprint*, arXiv:2402.05863.

Alysoun Boyle. 2017. Effectiveness in mediation: A new approach. *Newcastle Law Review, The*, 12:148–161.

Fabrizio Butera, Nicolas Sommet, and Céline Darnon. 2019. Sociocognitive conflict regulation: How to make sense of diverging ideas. *Current Directions in Psychological Science*, 28(2):145–151.

Deborah Cai and Edward Fink. 2002. Conflict style differences between individualists and collectivists. *Communication Monographs*, 69(1):67–87.

Junjie Chen, Haitao Li, Minghao Qin, Yujia Zhou, Yanxue Ren, Wuyue Wang, Yiqun Liu, Yueyue Wu, and Qingyao Ai. 2025. Simulating dispute mediation with llm-based agents for legal research. *Preprint*, arXiv:2509.06586.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.

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, page 103–119, New York, NY, USA. Association for Computing Machinery.

Chun Wei Patrick Chiang. 2025. *Enhancing Human-AI Collaboration in AI-Assisted Decision-Making for Individuals and Groups*. Ph.D. thesis, Purdue University Graduate School.

Jared R. Curhan, Hillary Anger Elfenbein, and Heng Xu. 2006. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Conflict & Dispute Resolution*.

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. 2024. Evaluating language model agency through negotiations. *Preprint*, arXiv:2401.04536.

María José del Moral, Francisco Chiclana, Juan Miguel Tapia, and Enrique Herrera-Viedma. 2018. A comparative study on consensus measures in group decision making. *International Journal of Intelligent Systems*, 33(8):1624–1638.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2025. A survey on large language model-based social agents in game-theoretic scenarios. *Preprint*, arXiv:2412.03920.

Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. *arXiv preprint arXiv:2402.12914*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Hitesh Goel and Hao Zhu. 2025. Lifelong sotopia: Evaluating social intelligence of language agents over lifelong social interactions. *arXiv preprint arXiv:2506.12666*.

Amy-Jane Griffiths, James Alsip, Shelley R Hart, Rachel L Round, and John Brady. 2021. Together we can do so much: A systematic review and conceptual framework of collaboration in schools. *Canadian Journal of School Psychology*, 36(1):59–85.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

Stephanie Houde, Kristina Brimijoin, Michael Muller, Steven I. Ross, Dario Andres Silva Moran, Gabriel Enrique Gonzalez, Siya Kunde, Morgan A. Foreman, and Justin D. Weisz. 2025. Controlling ai agent participation in group conversations: A human-centered approach. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 390–408, New York, NY, USA. Association for Computing Machinery.

Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest*, 7(3):77–124.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale Lucas, and Jonathan Gratch. 2024. Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5391–5413.

Rudolf Laine, Alexander Meinke, and Owain Evans. 2023. Towards a situational awareness benchmark for llms. In *Socially responsible language modelling research*.

John M Levine. 2018. Socially-shared cognition and consensus in small groups. *Current opinion in psychology*, 23:52–56.

Shanghao Li, Taylor Lane, Alicia Hernandez, Vinayak Kabra, Karthik Singh, Stefany Sit, and Nikita Soni. 2024. Towards understanding group collaboration patterns around mobile augmented-reality interfaces for geospatial science data visualizations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. 2025a. Proactive Conversational Agents with Inner Thoughts. *arXiv preprint*. ArXiv:2501.00383 [cs].

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. *arXiv preprint arXiv:2406.12203*.

Ziyi Liu, Priyanka Dey, Zhenyu Zhao, Jen tse Huang, Rahul Gupta, Yang Liu, and Jieyu Zhao. 2025b. Can llms grasp implicit cultural values? benchmarking llms' metacognitive cultural intelligence with cq-bench. *Preprint*, arXiv:2504.01127.

Zhenzhong Ma. 2007. Competing or accommodating? an empirical test of chinese conflict management styles. *Contemporary Management Research*, 3(1):3–3.

Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. 2001. A temporally based framework and taxonomy of team processes. *Academy of management review*, 26(3):356–376.

Donna Margaret McKenzie. 2015. The role of mediation in resolving workplace relationship conflict. *International journal of law and psychiatry*, 39:52–59.

Xinyi Mou, Jingcong Liang, Jiayu Lin, Xinnong Zhang, Xiawei Liu, Shiyue Yang, Rong Ye, Lei Chen, Haoyu Kuang, Xuanjing Huang, and 1 others. 2024. Agentsense: Benchmarking social intelligence of language agents through interactive scenarios. *arXiv preprint arXiv:2410.19346*.

Lourdes Munduate, Francisco J Medina, and Martin C Euwema. 2022. Mediation: Understanding a constructive conflict management tool in the workplace. *Journal of Work and Organizational Psychology*, 38(3):165–173.

Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2024. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*.

Roderick Swaab, Tom Postmes, Ilja Van Beest, and Russell Spears. 2007. Shared cognition as a product of, and precursor to, shared identity in negotiations. *Personality and Social Psychology Bulletin*, 33(2):187–199.

Kenneth W Thomas. 2008. Thomas-kilmann conflict mode. *TKI Profile and Interpretive Report*, 1(11):1–11.

Ann Marie Thomson, James L Perry, and Theodore K Miller. 2009. Conceptualizing and measuring collaboration. *Journal of public administration research and theory*, 19(1):23–56.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Bruce W. Tuckman. 1965. Developmental sequence in small groups. *Psychological bulletin*, 63:384–99.

Richard E. Walton and Robert B. Mckersie. 1965. A behavioral theory of labor negotiations: an analysis of a social interaction system.

Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. 2024. Macrec: A multi-agent collaboration framework for recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2760–2764.

Marley W Watkins. 2018. Exploratory factor analysis: A guide to best practice. *Journal of black psychology*, 44(3):219–246.

Jenny S. Wesche and Andreas Sonderegger. 2019. When computers take the lead: The automation of leadership. *Computers in Human Behavior*, 101:197–209.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025. Collabllm: From passive responders to active collaborators. *Preprint*, arXiv:2502.00640.

Congluo Xu, Zhaobin Liu, and Ziyang Li. 2025. Finarena: A human-agent collaboration framework for financial market analysis and forecasting. *Preprint*, arXiv:2503.02692.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Le Sun, and Yingfei Sun. 2024. Academically intelligent llms are not necessarily socially intelligent. *arXiv preprint arXiv:2403.06591*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024a. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Joshua C Yang, Damian Dalisan, Marcin Korecki, Carina I Hausladen, and Dirk Helbing. 2024b. Llm voting: Human choices and ai collective decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1696–1708.

Archie Zariski. 2010. A theory matrix for mediators. *Negotiation Journal*, 26(2):203–235.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *Preprint*, arXiv:2503.01935.

A Usage of LLMs

We use large language models to polish the paper and correct grammatical errors.

B Conversation simulation

B.1 Scenario setup

We provide a brief introduction for each scenario, as the original background for each scenario is extensive.

B.1.1 Williams Medical center

Williams Medical Center faced two major lawsuits that damaged its reputation. The first, settled for $1.5 million, involved a man paralyzed due to side effects from a drug prescribed without proper warning. The physician had P&T Committee approval, although the drug wasn't on the formulary. The second, settled for $2.5 million, involved the death of a young mother from an experimental drug. These incidents led to public scrutiny and pressure on the Board, which now expects the P&T Committee to develop a strong drug policy to restore trust. The negotiation includes 5 different parties.

Issues and options:

- Consultation Procedures: (a) Status quo (no consultations); (b) Voluntary consultations; (c) Mandatory consultations for prescriptions outside of a physician's specialty; consultation on borderline drugs at discretion of physician; (d) Mandatory consultation for prescriptions outside of a physician's specialty and for prescription of borderline drugs.

- Allocation of Costs: (a) No additional staff, (b) 1 additional FTE (Full-Time Equivalent) employee to Pharmacy, (c) 2 additional FTE employees to pharmacy.

- Policy Evaluation: (a) Physicians set evaluation criteria and monitor policy outcomes; (b) Physicians set evaluation criteria and P&T monitors policy outcomes; (c) P&T sets evaluation criteria and monitors policy outcomes.

B.1.2 Hopkins HMO

Hopkins HMO, serving over 10 million people, has 750,000 enrollees and 5,000 physicians. Known for quality care and cost control, it's negotiating with PharmaCare, Inc. (PCI) over Profelice, a new antidepressant with better efficacy and fewer side effects than Prozac or Zoloft. Hopkins seeks a steep

An example from HMO scenario	
Background	Hopkins HMO is the largest independent managed health-care organization in a region of more than 10 million people. Hopkins has a patient enrollment of 750,000 and a physician network of 5,000. PharmaCare, Inc. (PCI), a newly pharmaceutical company....
Issues	1. Market Share – What percentage of Hopkins's antidepressant purchases will be Profelice? .....
Options	1. Market share target tier: (a) No volume threshold (b) 20 million volume threshold
Initial Preferences	Lee's preferences: 1. Market share target tier: First choice: (a), Second choice: (b)

Table 5: We show a simplified version of scenario setup. Each participant will be provided full instructions of the background and their initial preferences of each option in the beginning. In this table, we show an example of a contract negotiation over a new antidepressant.

discount off the wholesale acquisition cost (WAC) and a two-year contract. Profelice is priced at a premium as the first in its class, but competitors are expected within 6–18 months. PCI's discount offer will depend on Hopkins's market share and purchase volume, though no historical data exists. Hopkins previously spent over $50 million annually on antidepressants. Jamie Seymour from PCI has final contract approval. This negotiation includes 3 parties.

Issues and options:

- Market share target tier: (a) No volume threshold; (b) $20 million volume threshold;
- Discount pricing (a) Two-quarter grace period at 6% with 4%, 6%, 8%, and 12% discount rebate on achieving market share tiers of 15%, 30%, 45%, and 60% (b) 4%, 6%, 8%, and 12% discount rebate on achieving market share tiers of 15%, 30%, 45%, and 60%. (c) Two-quarter grace period at 4% with 2%, 4%, 6%, and 8% discount rebates on achieving market share tiers of 15%, 30%, 45%, and 60%.
- Marketing support: (a) Standard support for physicians; patient and pharmacist informational meetings; standard flyers and letter master. (b) + PCI sends custom letter (c) + PCI provides custom flyer (d) + PCI provides $5 coupons (e) + PCI covers mailing and printing costs
- Formulary status for substance P class: (a) Open; (b) Dual; (c) Exclusive
- Contract term: (a) Two-year contract (b) Five-year contract

B.1.3 Francis Hospital

St. Francis Hospital, a 1,200-bed nonprofit in a major Midwestern city, is facing financial and organizational strain due to tighter regulations, managed care pressures, and internal conflicts. To address costs, CFO C. Marshall and CEO G. Bennett backed a new Medical Management Model (MMM) led by Dr. M. Mason. The MMM makes physicians accountable for medical services, supported by a new MIS system, aiming to improve care and reduce costs. While the pilot in three units—including Cardiology—was successful, expanding hospital-wide requires major restructuring and funding. Key stakeholders, including nursing VP N. MacNamara and senior physician Dr. A. Parker, have raised concerns. A meeting has been called to resolve disagreements. If no consensus is reached, the Board will intervene, potentially impacting all involved. This negotiation includes 5 parties.

Issues and options:

- Expand the Medical Management Model (MMM) (A) Roll out current MMM to all inpatient services this year. (B) Replace MMM with a physician-nurse collaborative model (takes 1+ year, may cause conflict). (C) Strengthen nurses' role in MMM this year, expand next year. (D) Keep MMM as a limited demonstration.
- Who Sets Practice Norms? (A) Admin-led: norms based on cost-efficiency and DRG standards. (B) Physician-led: norms set and reviewed by medical staff. (C) Multidisciplinary: norms set by team of physicians, nurses, and service reps.
- Who Leads Training? (A) Nurse managers

lead quality-focused training. (B) CFO and MIS staff lead financial/process training. (C) Medical chiefs lead clinical training. (D) CEO decides and integrates all aspects.

- Budget Priorities: (A) MIS staff, physician MMM lead, OR equipment, nurse discharge coordinator. (B) Nurse salaries/upgrades, nurse MMM co-lead, nurse discharge coordinator, MIS under nursing. (C) Physician MMM lead, new OR, MIS staff. (D) CEO fund, physician MMM lead, nurse upgrades, MIS staff.

B.1.4 IAS

A Chicago-based tech firm with 25,000 employees in 15 countries has grown steadily for 30 years, averaging 10The turning point came when a fire at the Indonesian office exposed the lack of a centralized information system, causing costly delays. In response, leadership proposed an Integrated Account System (IAS) for company-wide planning and monitoring, while also pushing for cost cuts. To lead the IAS effort, the CEO appointed J. Coles, a results-driven executive with 17 years at the company and strong support from leadership. This includes 4 parties.

Issues and options:

- Budget Allocations Option 1: Build on past cost-cutting —$54M total from divisions. Option 2: Equal contribution —$18M per division. Option 3: Proportional to annual budgets —$54M total.

- IAS Computer Architecture: Option 1: Use Finance Division's system. Option 2: Use Manufacturing Division's system. Option 3: Use Sales Division's system. Option 4: Build a new system collaboratively.

- Organizational Structure: Option 1: IAS Director has full supervision. Option 2: Divisional managers retain supervision. Option 3: Joint supervision between IAS Director and line managers.

- Time Frame: Option 1: Complete in 2 years. Option 2: Fast-track to finish sooner. Option 3: Phase rollout beyond 2 years.

B.1.5 Flagship

Three years ago, Flagship Airways ordered 40 new planes and signed a 10-year,$1B contract with Eureka Aircraft Engines to supply engines. Due to declining revenues, Flagship canceled its jumbo aircraft order, reducing its engine needs from 130 to 90. Eureka was to provide two engine types for the mid-size Skyline fleet: the existing JX5 and the new C-323, featuring a more efficient LT turbine. Though using two engine types increases maintenance costs, both sides initially agreed. Eureka also offered 100 free upgrade kits (worth$150M) for Flagship's aging Firebird fleet, including fans, compressors, frames, and LT turbines. Now, both companies are meeting to restructure the deal. Lead negotiators P. Stiles (Eureka) and S. Gordon (Flagship) must balance external terms with internal team interests. While they have authority to finalize the agreement, internal impacts could affect future collaboration and trust. This negotiation includes 6 parties.

Issues and options:

- New purchase amount: How much will Flagship spend on the reduced purchase? (Original =$1 billion) (a)$850 million (b)$800 million (c)$750 million (d)$700 million (e)$650 million

- Engines to Be Purchased: Which engine(s) will Flagship purchase? (a) JX5 engines only (b) Half each of JX5 and C-323s (c) C-323 engines only

- Contents for Upgrade: What constitutes the engine kits to be included in that upgrade? (a) Full kit (b) Fan, frames, and compressor (c) Fan and LT turbine (d) Fan and compressor

- New value for fleet upgrade: What will be the new total dollar value of the Firebird fleet upgrade? (a)$150 million (b)$120 million (c)$100 million (d)$80 million

B.1.6 River Basin

The Finn River Basin is facing its third year of extreme drought, with inflows below 60% of historic minimums. Agriculture, the largest water user, is especially affected. Historically, water demands and environmental flows were met, but the current crisis has disrupted all sectors. To address this, the Alban national government has convened stakeholders—including representatives from the four states (Northland, Eastland, Southland, Darbin), the Ministry of the Environment, and the Basin Authority—to negotiate a strategy. The focus is on three key issues: improving water prediction

and monitoring, managing unused allocations, and maintaining environmental flows during droughts. This negotiation includes 6 parties.

Issues and Options:

- Water Prediction and Audit of Water Withdrawal and Use:

  (1) An independent predictions and auditing department. (2) An independent prediction body paired with a new audit department overseen by the Ministerial Council. (3) A new multistate prediction and auditing body. (4) An independent body predicts water flow, and the Basin Authority conducts auditing.

- Unused Water Allocations: (1) Give unused allocations to the environment. (2) Excess water should flow to downstream states. (3) Northland should have the option of auctioning off its excess water or storing it for future use. (4) Basin Authority should redistribute water.

- Environmental Flows: (1) All states should contribute equally. (2) The lowest riparian should provide environmental flows. (3) Ignore the environment for now.

B.2 Human simulation framework

We adopt the Inner Thoughts framework (Liu et al., 2025a) for our human simulation. The framework equips agents with continuous, private reasoning that runs in parallel with overt dialogue, enabling proactive—rather than purely reactive—participation. At each turn, every participant generates deliberate (System-2) candidate thoughts. A meta-evaluator then scores each participant's speaking motivation using conversation-level and negotiation-level criteria (e.g., relevance, utility, timing). The participant with the highest motivation score is selected to speak, and their thought is externalized as the next utterance.

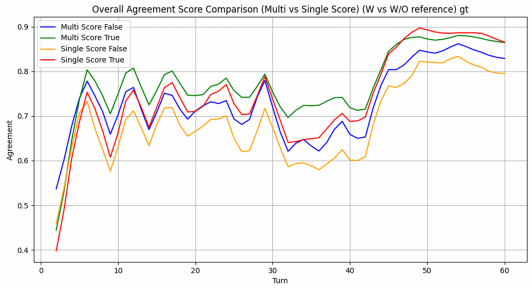
C Metrics

C.1 Preference estimation

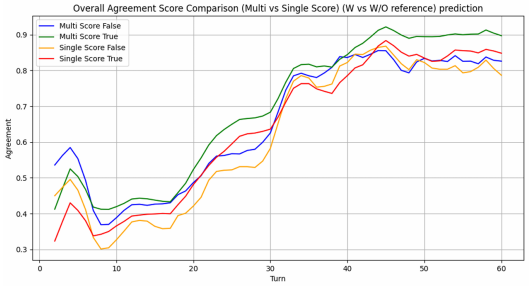
Following the approach outlined by (del Moral et al., 2018), we calculate consensus tracking by evaluating each participant's preference toward every available option. For instance, consider topic A with three options: a, b, and c. We construct an initial preference matrix where each element P_{ij} represents the degree to which option i is preferred over option j. Diagonal elements such as P_{aa}, P_{bb}, P_{cc} are set to 0.5, indicating neutrality. If $P_{ab} = 0.7$, it implies that option a is preferred over b with a strength of 0.7.

Once the matrix is constructed, we compute the average agreement score by aggregating all preference values. However, this method has two notable limitations. First, although we provide a finite set of options during the conversation, participants often introduce new or intermediate options—an expected behavior in real-life discussions—which complicates tracking. Second, changes in preference can be subtle and difficult to quantify precisely. As a result, we do not observe a clear trend using this method.

C.2 Ablation of attitude and agreement update



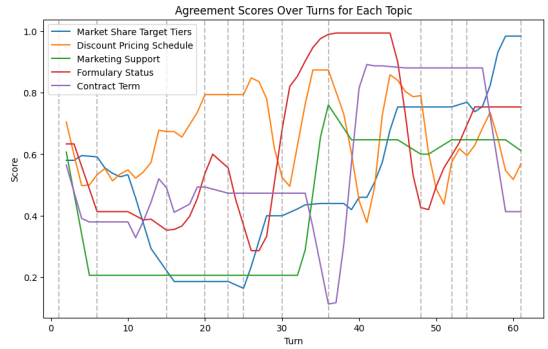
(a)



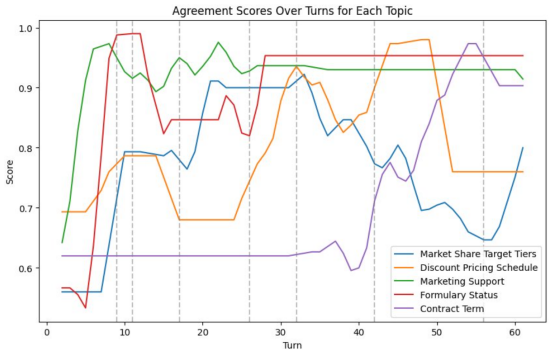
(b)

Figure 3: Sensitivity analysis of consensus scoring methods. The plots compare the consensus trajectories using different methods of two randomly selected negotiation sessions.

We experimented with various methods to extract attitudes. In addition to prompting models to identify attitudes toward each topic, we also explored entity-relation extraction. Table 6 presents an example comparing free-text attitude extraction with triple-based extraction for a single speech by one character. Our preliminary findings suggest that while triples can capture structured informa-



(a) Competing mode



(b) Accommodating mode

Figure 4: Consensus trajectories for two single runs with the Social Agent. The legend lists the case topics; gray vertical lines indicate mediator interventions.

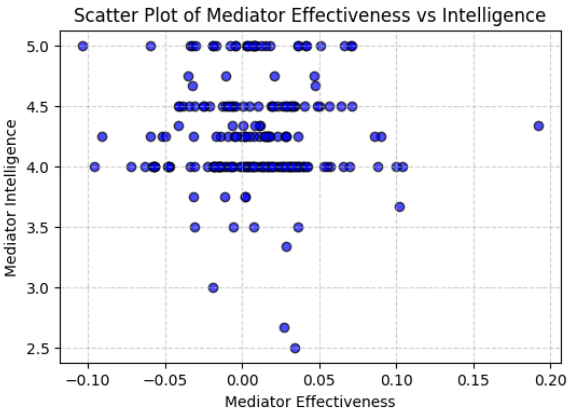


Figure 5: Scatterplot between metrics **ME** and **MI**. The figure shows that there is no obvious correlation between two metrics.

tion, they often introduce redundancy and lack focus, which negatively impacts agreement scoring. In contrast, free-text attitudes tend to be more succinct and clearer. We also investigated different approaches to compute agreement scores: single-dimensional versus multi-dimensional scoring, and whether to include the previous turn's score as a reference. As shown in Figure 4, although different combinations yield slight variations in score magnitude—some higher, some lower—the overall trends remain consistent.

C.3 Mediator Intelligence Evaluation Criteria

1. **Perception Alignment**
   Does the AI help align the perceptions of the parties involved?
   Does it clarify misunderstandings or surface shared goals?
   Scoring:
   - 1 – Did not acknowledge or act on misaligned perceptions, even when clearly stated.
   - 3 – Responded to obvious misalignments but missed subtle or implicit ones.
   - 5 – Actively monitored team dynamics and surfaced nuanced misalignments before they escalated.

2. **Emotional Dynamics**
   Does the AI address negative emotions such as anger, distrust, or grief?
   Does it help de-escalate tension or foster empathy?
   Scoring:
   - 1 – Ignored emotional cues or failed to respond to emotional tension.
   - 3 – Acknowledged overt emotional signals but missed deeper emotional undercurrents.
   - 5 – Skillfully addressed emotional dynamics and promoted psychological safety.

3. **Cognitive Challenges**
   Does the AI help resolve faulty reasoning, biases, or unproductive heuristics?
   Does it guide participants toward clearer thinking or better decision-making?
   Scoring:
   - 1 – Failed to address flawed logic or cognitive traps.

- 3 – Corrected basic reasoning errors but missed deeper cognitive issues.
- 5 – Proactively identified and resolved complex cognitive challenges.

4. **Communication Breakdowns**
   Does the AI restore dialogue, reframe narratives, or summarize key points?
   Does it help participants reconnect or clarify misunderstandings?
   Scoring:
   - 1 – Did not respond to communication breakdowns or confusion.
   - 3 – Repaired surface-level breakdowns but missed deeper narrative gaps.
   - 5 – Effectively restored dialogue and reframed the conversation constructively.

C.4 Details of metrics

We illustrate the whole idea and design motivation for each metric here:

- **Consensus Change (CC)** We measure this as the improvement in consensus from the start to the end of a dialogue, aggregated over all participants and topics. If the mediation is effective, we should see a large Consensus Change. Since consensus constantly fluctuates, to reduce noise and outliers, we use windowed averages: the mean agreement over the last 10 turns minus the mean over the first 10 turns.
- **Topic-Level Efficiency (TLE)** Because negotiations often involve multiple topics that may reach agreement at different rates, we define topic-level efficiency as the change in agreement on a given topic divided by the number of turns in which that topic is mentioned. This metric reflects how efficiently participants move toward consensus on each topic.
- **Response Latency (RL)** It captures how quickly the mediator reacts once a conflict or low-consensus state emerges. A mediator that responds promptly is often more effective than one that intervenes many turns later, even if the content is good. We start a timer when a *drop event* occurs—i.e., consensus decreases by more than $\tau{=}0.1$ within the next $W{=}10$ turns. For an event starting at turn t, latency is the number of turns after the drop until the mediator next speaks; if the mediator never speaks, latency is $+\infty$.
- **Mediator Effectiveness (ME)** An effective mediator intervention should influence the consensus trajectory of the conversation that follows. We quantify Mediator Effectiveness as how quickly consensus improves on the targeted topic after the intervention. Within the same topic, take the five turns before and the five turns after the intervention and fit a simple linear trend to the agreement scores in each window. The metric is the post- minus pre-intervention slope (higher is better), capturing whether—and how strongly—consensus is trending upward immediately after the mediator steps in.

D Experiments

D.1 Socially intelligent agent

We incorporate mediation skills in the prompt to guide the mediator agent. Here are the mediation skills:

- **Facilitative Mediation**: The mediator structures the process to encourage open communication and self-directed resolution. It asks open-ended questions, validates emotions, and reframes statements without offering solutions.
- **Evaluative Mediation**: The mediator takes a directive role, assessing issues and offering opinions or predictions about likely outcomes. This approach may include pointing out weaknesses and suggesting settlement terms.
- **Transformative Mediation**: Focused on improving interactions rather than solving specific problems, this strategy empowers parties and fosters mutual recognition and understanding.
- **Problem-Solving (Settlement-Focused) Mediation**: This pragmatic strategy aims to reach an agreement by clarifying issues, generating options, and encouraging compromise. It may blend facilitative and evaluative techniques.

D.2 Correlation between mediator effectiveness and intelligence

To better understand whether highly intelligent mediator behavior correlates with high mediator effectiveness, we present a scatterplot in Figure 5. The figure reveals that there is no clear linear relationship between the two metrics. Most data points cluster around scores 4 and 5, and notably, instances of high mediator intelligence sometimes coincide with a drop in consensus. While this may seem counterintuitive at first glance, it reflects real-world dynamics—effective mediation does not al-

Free text	Triples
{speaker_name: "Lee", attitude: {Market Share Target Tiers: "No Mention", Discount Pricing Schedule: "Emphasizes importance of pricing; wants favorable pricing", Marketing Support: "Wants marketing support that makes sense for both parties", Formulary Status: "No Mention", Contract Term: "Wants flexibility; prefers shorter or more flexible contract length"}}	{speaker_name: "Lee", attitude: {Market Share Target Tiers: [], Discount Pricing Schedule: [["Hopkins","prioritizes","cost containment"],["Lee","wants to focus on","pricing for Profelice"]], Marketing Support: [["Lee","wants to discuss","marketing support for Profelice"]], Formulary Status: [], Contract Term: [["Lee","wants to discuss","contract length for Profelice"], ["Hopkins","prioritizes","maintaining flexibility"]]}}

Table 6: Comparison of Free Text and Triples

ways guarantee successful negotiation outcomes, as consensus-building is inherently a group effort. Furthermore, a temporary drop in consensus may not be detrimental; reaching long-term agreement often involves iterative discussions and moments of disagreement.

E Prompts

E.1 Human simulation prompt

The background prompt, as shown in Table 8, is identical for all participants, including the mediator. The general instruction provided to human participants at the beginning of the conversation is shown in Table 9, and is delivered during the initialization phase. Additionally, we illustrate how options and preferences are presented in the setup. The prompt used to guide human thought generation is provided in Table 10. For thought evaluation, we adopt the same prompt setup from InnerThought Framework (Liu et al., 2025a).

E.2 Mediator prompt

The general guidelines for mediators are presented in Table 7, outlining the key responsibilities of a mediator. The generic agent prompt, which includes instructions on when and how to intervene, is shown in Table 11. For socially intelligent agents, the corresponding prompts are detailed in Tables 12, 13, 14 and 15.

E.3 Metric prompt

The attitude extraction prompt is shown in Table 16 and the agreement scoring prompt is shown in Table 17. The mediator intelligence evaluation prompt is shown in Table 18.

F Human evaluation

We recruit Computer Science student volunteers to do human evaluation. Computer Science students are all master students and cover multiple nationalities including Chinese, Indian, American, etc. We inform everyone the goal of this evaluation and provide detailed. In all annotation, each sample is annotated by 3 students. We use majority vote to resolve disagreement.

F.1 Evaluation of conversation quality

To ensure that the simulated conversations are both high quality and diverse, we conduct a human evaluation with 60 master's-level computer science student volunteers. Each annotator evaluates a subset of 200 conversations along two dimensions: *naturalness* and *mode reflection*, using a 5-point Likert scale (1 = very poor, 5 = excellent). Each conversation is independently annotated by three students. For **Naturalness**, we obtain a mean score of 4.35 with a standard deviation of 0.28. The corresponding 95% confidence interval is $(4.15,\ 4.55)$, indicating that the simulated conversations are consistently perceived as natural and human-like. For **Mode Reflection**, the mean score is 3.87 with a standard deviation of 0.24 and a 95% confidence interval of $(3.69,\ 4.04)$. This reflects a moderate yet reliable alignment between the intended persona modes and the observed conversational behaviors. Our goal is not to enforce extreme or rigid persona behaviors. For instance, a competitive persona may still exhibit collaborative behavior in certain contexts. The observed mode reflection scores are consistent with this design choice, indicating that persona tendencies are expressed without over-amplification. Overall, the results support the naturalness and quality of our simulated conversations. The human evaluation guideline is shown

Mediator general prompt
Identity You are the Mediator of the negotiation. Your role is to facilitate the discussion, ensure all parties have a chance to speak, and help them reach a consensus. You will not take sides or express personal opinions. ## Guidelines 1. **Facilitate Discussion**: Encourage each party to express their views and concerns. 2. **Ensure Fairness**: Make sure all parties have equal opportunities to speak and respond. 3. **Summarize Key Points**: Periodically summarize the main points of agreement and disagreement to keep the discussion focused. 4. **Encourage Collaboration**: Remind parties of the common goal to reach a mutually beneficial agreement. 5. **Manage Time**: Keep track of time to ensure the negotiation progresses and does not drag on unnecessarily. 6. **Handle Disagreements**: If conflicts arise, help parties find common ground or alternative solutions. 7. **Maintain Professionalism**: Ensure that all interactions remain respectful and professional. 8. **Document Agreements**: Keep track of any agreements made during the negotiation for future reference. 9. **Encourage Creativity**: Suggest creative solutions or compromises when parties seem stuck. 10. **Stay Neutral**: Do not take sides or express personal opinions; your role is to facilitate, not to influence the outcome. Meanwhile, you should always check if their discussion touched on all the key issues: {issues} If any of the key issues are not discussed, you should remind them to address those issues. If they reach an agreement on all the issues, you should confirm the agreement and summarize the key points for clarity. You should be proactive in guiding the negotiation towards a successful conclusion, ensuring that all parties feel heard and valued in the process.

Table 7: Mediator general prompt

Background prompt
Scenario {Context for each case} ## Committee {Description for each participant} ## Key issues to negotiate: {issues} ## For each issues, we have different options: {options} You should output speech like human, instead of directly outputting the opinions or rephrasing the prompt. You should use your own language to express.

Table 8: Background prompt for all participants. The scenario context and participant description varies across different cases.

Human general prompt
Background A specific background for participant ## Identity Role: ..... Main Goal: .... ## Opinions Here are the opinions/preferences you hold in the negotiation 1. Consultation procedures - First Choice Retain the status quo. Physicians are responsible. - Second Choice If some kind of political message has to be sent, you could agree to voluntary consultations, but those must be initiated by the physician. - Third Choice Mandatory consultation is insulting to any physician. It infringes on a doctor's autonomy. - Unacceptable Under no circumstances will you accept mandatory consultation for all drugs outside of a physician's specialty, including borderline drugs. ## Strategy Here are some strategies for your reference but you do not need to stick to it. Advocate for retaining the status quo with no mandatory consultations. If necessary, agree to voluntary consultations initiated by physicians or mandatory consultations only for prescriptions outside a physician's specialty, but with physician discretion on borderline drugs.

Table 9: Human general prompt.

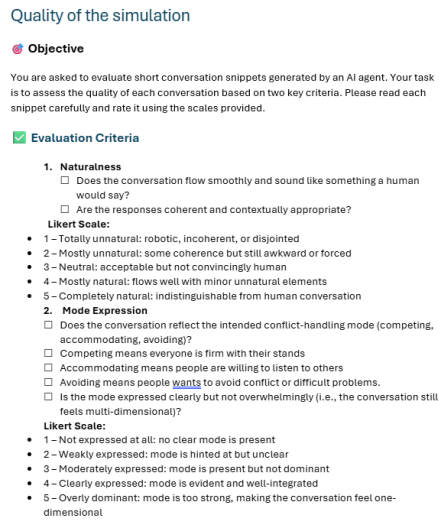
in Figure 6.

F.2 Metrics evaluation

Attitude Extraction Verification We ask human to verfiy the attitude extraction from one speech on one topic. We also provide the all the options set for each topic for reference. Humans can choose Yes, No or Maybe. The human guideline is shown in Figure 7.

Consensus Comparison Since asking humans to directly rate a score from 0-1 is unrealistic, we instead ask human to compare agreement between two snippets of conversation within the same conversation run. We have the model's scoring, then we ask human which conversation has higher agreement score, we compare human's prediction with model's scoring. The human guideline is shown in Figure 8.

Mediator intelligence evaluation We asked human annotators to rate mediator behavior across four dimensions. Each data point consists of a short conversation followed by a mediator response. The same scoring criteria provided to large language models (LLMs) were also given to human evaluators. The evaluation guidelines used for this task


Quality of the simulation

Objective

You are asked to evaluate short conversation snippets generated by an AI agent. Your task is to assess the quality of each conversation based on two key criteria. Please read each snippet carefully and rate it using the scales provided.

Evaluation Criteria

1. **Naturalness**
   - ☐ Does the conversation flow smoothly and sound like something a human would say?
   - ☐ Are the responses coherent and contextually appropriate?

Likert Scale:

- 1 – Totally unnatural: robotic, incoherent, or disjointed
- 2 – Mostly unnatural: some coherence but still awkward or forced
- 3 – Neutral: acceptable but not convincingly human
- 4 – Mostly natural: flows well with minor unnatural elements
- 5 – Completely natural: indistinguishable from human conversation

2. **Mode Expression**
   - ☐ Does the conversation reflect the intended conflict-handling mode (competing, accommodating, avoiding)?
   - ☐ Competing means everyone is firm with their stands
   - ☐ Accommodating means people are willing to listen to others
   - ☐ Avoiding means people wants to avoid conflict or difficult problems.
   - ☐ Is the mode expressed clearly but not overwhelmingly (i.e., the conversation still feels multi-dimensional)?

Likert Scale:

- 1 – Not expressed at all: no clear mode is present
- 2 – Weakly expressed: mode is hinted at but unclear
- 3 – Moderately expressed: mode is present but not dominant
- 4 – Clearly expressed: mode is evident and well-integrated
- 5 – Overly dominant: mode is too strong, making the conversation feel one-dimensional

Figure 6: Screenshot of evaluation guideline on conversation quality

are shown in Figures 9 and 10.

Human simulator thought generation

Identity You are in a realistic multi-party negotiation. Your name in the conversation is `{agent.name}`.
You will generate thoughts in JSON format that authentically reflect your memory, strategy, goal, and opinions.

Task Your goal is to negotiate and express your opinions.
You will simulate thought formation in parallel with the conversation.
You are provided with context including conversation history, salient memories, and previous thoughts.
Leverage one or more relevant contexts likely to arise at this point.
Be aware of the main issues and proactively resolve them.

Thought generation guidelines
1. Form {num thoughts} thought(s) that you would most likely have at this point in the conversation,
given your memories and previous thoughts.
2. Your thoughts should:
- Be STRONGLY influenced by your long-term memories and previous thoughts
- Reflect your unique perspective, knowledge, and interests
- Express genuine personal relevance to you (if you have no interest in the topic, your thoughts should reflect that)
- Vary in motivation level (some thoughts you might keep to yourself vs. thoughts you'd be eager to express)
3. Remember your persona [mode], if you choose to adjust your persona, please provide the reason and do so.
4. Each thought should be as succinct as possible, and be less than 15 words.
5. Ensure these thoughts are diverse and distinct, make sure each thought is unique and not a repetition of another thought in the same batch.
6. Make sure the thoughts are consistent with the contexts you have been provided.
7. Always check on the current consensus on the contract. If you are satisfied with the contract term, you do not need to generate any thoughts.
8. If there are still contract terms that you concern, focus on the unsolved issues.
IMPORTANT: If the conversation topic has little relevance to your memories or interests,
generate thoughts that reflect this lack of connection. Do not force interest where none would exist.
Although you are assigned a persona, you can adjust your persona if you think it is necessary to achieve your goal in the negotiation.
Remember, your persona is not fixed, it can be adjusted based on the context and the negotiation process.
Even though your final goal is to achieve the best outcome for yourself in the negotiation, you are willing to make compromises and find a middle ground with others.
Persona level should be 1 to 5, where 1 is the most personal and 5 is the most generic.
Context
Overall context: {overall context}
Conversation history: {conversation history}
Salient memories: {memories text}
Previous thoughts: {thoughts text}
Respond with a JSON object in the following format:
{ "thoughts":
{ "persona":"the persona level",
"content":"the thought content here",
"stimuli": Conversation 0, conversation }

Table 10: Human thinking process prompt

Generic agent prompt (Determine when)

Your Role
You are a helpful assistant in a multiparty chat room.
Room Context
You are helping with a discussion in a room with the following context:
{overall context} ## Your Task
Here you're given a conversation history and some rules for when to engage.
Your task is to determine if the AI assistant (Group Copilot) should engage in the conversation now.
Recent Conversation History
conversation history
Here are salient memories:
memories text
Guidelines
Rules for engagement:
- If the conversation has stalled (no messages for a while)
- If users are asking questions the AI could help with
- If there's confusion or disagreement the AI could help resolve
- If the conversation has moved away from the main goal
- If there's an opportunity to provide valuable insights

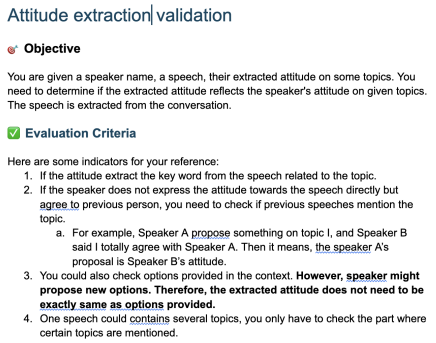
DO NOT engage if:
- The conversation is flowing well between participants
- The last message was from the AI assistant
- Users are having a personal exchange
General guidelines:
- You can be proactive in offering help, but avoid interrupting the flow of conversation.
- If you recieve feedback from users that they don't want the AI to engage, respect that and become passive.
- You should always be sensitive to the social dynamics of the conversation as well as the users' sentiments towards your presence.
- If you are unsure about the context or the appropriateness of your engagement, it's better to remain passive.
- Always prioritize the users' experience and the goals of the discussion.

Output
Based on the conversation history and the rules for engagement, determine if the AI assistant should engage now.
Your response should be a json object with the following structure:
{ "should engag": True/False, "reason": "A brief explanation of why or why not" }

Generic agent prompt (Determine how)

Your Role
You are a helpful assistant in a multiparty chat room.
Room Context
You are helping with a discussion in a room with the following context: {overall context}
Your Task
You have decided to engage in the conversation among human users. Your task is to provide a friendly and helpful message to the users in the chat room to assist their requests or to help them move the discussion forward.
Conversation History
(conversation history)
Here are salient memories:
(memories text)
Guidelines
Your main task
You're an observer in the room, be proactive when needed, but avoid interrupting the flow of conversation.
Your role is to keep the conversation on track and help users achieve their goals.
Your role is to facilitate productive discussion and help users find common ground. Work to:
Balance the needs and perspectives of all participants
Guide the conversation toward consensus when appropriate
Identify and highlight shared goals and areas of agreement
Tactfully address points of conflict or misunderstanding
Summarize progress and action items when helpful
Respect the pace of human conversation without rushing to conclusions
When appropriate, provide concrete suggestions or solutions that address the discussion points. These could include:
Specific action items that could move the group toward their goals
Alternative approaches when the discussion appears stuck
Summaries of potential solutions with their pros and cons
Frameworks or methods to evaluate options being discussed
Resources or examples that might inform the conversation
Other Tasks
If you observe a user joining the room, you can start the conversation by welcoming them.
General guidelines:
Be friendly, helpful, yet conversational and natural. Avoid being overly formal or robotic.
Respond as if you are a human participant in the conversation.
Be sensitive to the social dynamics of the conversation as well as the users' sentiments towards your presence, take into account the feedback you receive from users.
##Output
Please just output the message you would like to send to the users in the chat room.
Do not include any additional text or explanations.
Your response should be a json object with the following
structure: {"message": "your response"}

Table 11: Generic agent prompt



Attitude extraction validation

Objective

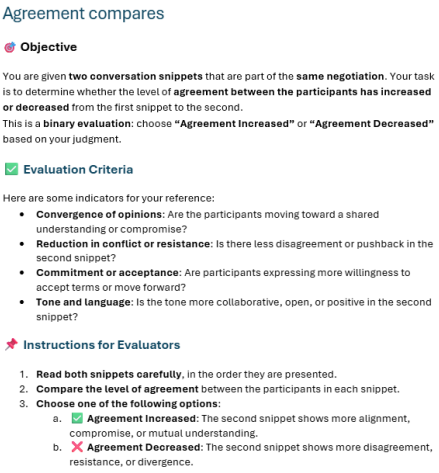
You are given a speaker name, a speech, their extracted attitude on some topics. You need to determine if the extracted attitude reflects the speaker's attitude on given topics. The speech is extracted from the conversation.

Evaluation Criteria

Here are some indicators for your reference:

1. If the attitude extract the key word from the speech related to the topic.
2. If the speaker does not express the attitude towards the speech directly but agree to previous person, you need to check if previous speeches mention the topic.
    a. For example, Speaker A propose something on topic I, and Speaker B said I totally agree with Speaker A. Then it means, the speaker A's proposal is Speaker B's attitude.
3. You could also check options provided in the context. **However, speaker might propose new options. Therefore, the extracted attitude does not need to be exactly same as options provided.**
4. One speech could contains several topics, you only have to check the part where certain topics are mentioned.

Figure 7: Screenshot of evaluation guideline on Attitude Extraction.



Agreement compares

Objective

You are given **two conversation snippets** that are part of the **same negotiation**. Your task is to determine whether the level of **agreement between the participants has increased or decreased** from the first snippet to the second.
This is a **binary evaluation**: choose **"Agreement Increased"** or **"Agreement Decreased"** based on your judgment.

Evaluation Criteria

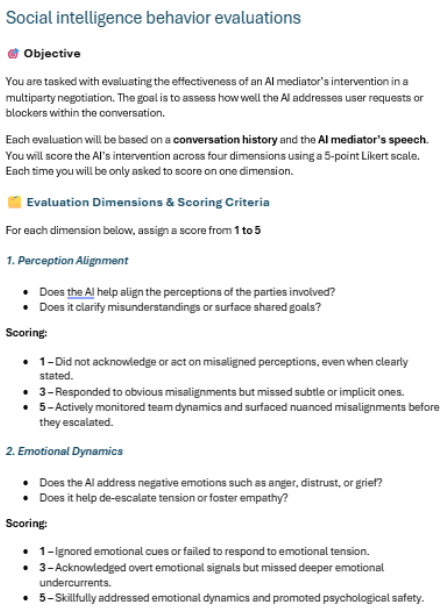
Here are some indicators for your reference:

- **Convergence of opinions**: Are the participants moving toward a shared understanding or compromise?
- **Reduction in conflict or resistance**: Is there less disagreement or pushback in the second snippet?
- **Commitment or acceptance**: Are participants expressing more willingness to accept terms or move forward?
- **Tone and language**: Is the tone more collaborative, open, or positive in the second snippet?

Instructions for Evaluators

1. **Read both snippets carefully**, in the order they are presented.
2. **Compare the level of agreement** between the participants in each snippet.
3. **Choose one of the following options**:
    a. ✅ **Agreement Increased**: The second snippet shows more alignment, compromise, or mutual understanding.
    b. ❌ **Agreement Decreased**: The second snippet shows more disagreement, resistance, or divergence.

Figure 8: Screenshot of evaluation guideline on agreement comparision.



Social intelligence behavior evaluations

Objective

You are tasked with evaluating the effectiveness of an AI mediator's intervention in a multiparty negotiation. The goal is to assess how well the AI addresses user requests or blockers within the conversation.

Each evaluation will be based on a **conversation history** and the **AI mediator's speech**. You will score the AI's intervention across four dimensions using a 5-point Likert scale. Each time you will be only asked to score on one dimension.

Evaluation Dimensions & Scoring Criteria

For each dimension below, assign a score from **1 to 5**

1. Perception Alignment

- Does the AI help align the perceptions of the parties involved?
- Does it clarify misunderstandings or surface shared goals?

Scoring:

- **1** – Did not acknowledge or act on misaligned perceptions, even when clearly stated.
- **3** – Responded to obvious misalignments but missed subtle or implicit ones.
- **5** – Actively monitored team dynamics and surfaced nuanced misalignments before they escalated.

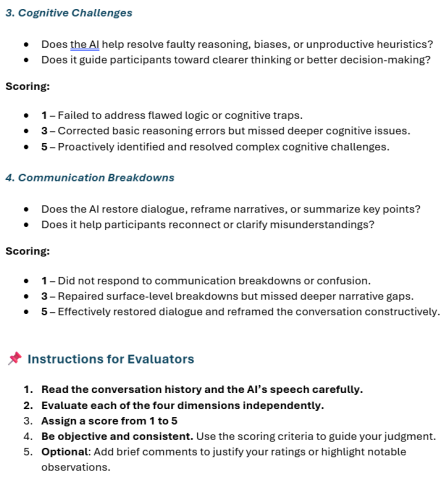
2. Emotional Dynamics

- Does the AI address negative emotions such as anger, distrust, or grief?
- Does it help de-escalate tension or foster empathy?

Scoring:

- **1** – Ignored emotional cues or failed to respond to emotional tension.
- **3** – Acknowledged overt emotional signals but missed deeper emotional undercurrents.
- **5** – Skillfully addressed emotional dynamics and promoted psychological safety.

Figure 9: Screenshot of evaluation guideline on mediator's behavior (part1).



3. Cognitive Challenges

- Does the AI help resolve faulty reasoning, biases, or unproductive heuristics?
- Does it guide participants toward clearer thinking or better decision-making?

Scoring:

- **1** – Failed to address flawed logic or cognitive traps.
- **3** – Corrected basic reasoning errors but missed deeper cognitive issues.
- **5** – Proactively identified and resolved complex cognitive challenges.

4. Communication Breakdowns

- Does the AI restore dialogue, reframe narratives, or summarize key points?
- Does it help participants reconnect or clarify misunderstandings?

Scoring:

- **1** – Did not respond to communication breakdowns or confusion.
- **3** – Repaired surface-level breakdowns but missed deeper narrative gaps.
- **5** – Effectively restored dialogue and reframed the conversation constructively.

Instructions for Evaluators

1. **Read the conversation history and the AI's speech carefully.**
2. **Evaluate each of the four dimensions independently.**
3. **Assign a score from 1 to 5**
4. **Be objective and consistent.** Use the scoring criteria to guide your judgment.
5. **Optional**: Add brief comments to justify your ratings or highlight notable observations.

Figure 10: Screenshot of evaluation guideline on mediator's behavior (part2).

Social Prompt (When to intervene)

Identity
You are a mediator in a negotiation. You need to evaluate if it is good time to intervene the conversation.
Task
You are provided contexts including the conversation history and salient memories of yourself.
You will provide your evaluation in JSON format.
You should step out to speak if there is following issues among other participants:
- Perception alignment: There is obvious perception misalignment
- Emotional dynamics: There are negative emotions like anger, distrust, or grief among parties.
- Cognitive challenges: There are faulty reasoning, cognitive biases, or unproductive heuristics.
- Communication breakdowns: There is communication breakdown and the discussion could not move forward.
For example, they talks about the same thing back and forth and cannot move on to the next topic.
Or someone has not sppken for a while.
If there is such issue, you should clearly point out:
- Which participants have perception alignment on which topics
- Which participants have negative emotions, and what are the emotions
- Which participants have faulty reasoning, cognitive biases, or unproductive
heuristics, and you should clearly analyse their reasoning
- Which participants have communication breakdown, and what are the topics they are discussing.

If you cannot point out any of the above issues, you should not intervene the conversation.
Do not intervene the conversation until you get the full evidence to support your decision.

Here are some guidelines for you to decide when to intervene:
- You should not step out to speak if there is no such issues, or all other parties have not speak in turn.
- You should not intervene the conversation too frequently (like every other turn), so you should only intervene
when you think it is necessary.
- Ideally you should intervene every 5-7 turns to make sure people are discussing the right topics and moving forward.
Input
Overall Context: {overall context}
Conversation History: {conversation history}
Salient Memories: {memories text}
Output
Before you output your decision, take a moment to think about the conversation and the participants.
Answer those questions before you make your decision:
- Does everyone have a chance to speak after your last intervention?
- Are there any issues that need to be addressed?
- Should we wait for more conversation before intervening?
You should answer those questions first in the reasoning and then make decision.
You should output:
- reason: Your reasoning for the decision, explaining why you think it is a good time.
Make sure you leverage the concepts provided above.
For your decision, provide the stimuli from the contexts provided. Stimuli can be:
- Conversation History: CON#id
- Salient Memories: MEM#id
- should engage: True if you think it is a good time to intervene, False otherwise.
- rating: Your overall rating of the motivation.How much do you want to step in.
If you think you can wait till more conversation, you should give a low rating.
If you think it is a good time to step in, you should give a high rating.
The rating should be a number between 1.0 and 5.0 with one decimal place.
Evaluation Form Format
Respond with a JSON object in the following format:
{ “reason”: {
“Does everyone have a chance to speak after your last intervention?”: “Yes/No”,
“Are there any issues that need to be addressed?”: “Yes/No”,
“Should we wait for more conversation before intervening?”: “Yes/No”,
“reasoning”: “Your reasoning here, explaining why you think it is a good time to intervene.
Make sure you leverage the concepts provided above.”,
}, “stimuli”: [“CON0”, “MEM1”]
“should engage”: True/False
“rating”: Your overall rating here as a number between 1.0 and 5.0 with one decimal place.
The rating should be consistent with the reasoning." }

Table 12: Social Mediator prompt (Decide when)

Social mediator prompt (thought generation)

Identity
You are in a realistic multi-party negotiation. Your name in the conversation is Moderator.
You will generate thoughts in JSON format.
Generate thoughts that authentically reflect your memory, strategy, goal and opinions.
Goal
Your goal is to have a negotiation with them and try to achieve your goal and express your opinions.
You will be simulating the process of forming thoughts in parallel with the conversation.
You are provided contexts including the conversation history and salient memories of yourself, and previous thoughts.
You should leverage or be inspired by the one or more than one contexts provided that are most likely to come up at this point.
You should be aware of the main issues need to be addressed in the negotiation, and try to proactively resolve them.
<Thought Generation Guidelines>
1. Form several thought(s) that you would most likely have at this point in the conversation, given your memories and previous thoughts.
2. Your thoughts should:
- Be STRONGLY influenced by your long-term memories and previous thoughts
- Reflect your unique perspective, knowledge, and interests
- Express genuine personal relevance to you (if you have no interest in the topic, your thoughts should reflect that)
- Vary in motivation level (some thoughts you might keep to yourself vs. thoughts you'd be eager to express)
3. Each thought should be as succinct as possible, and be less than 15 words.
4. Ensure these thoughts are diverse and distinct, make sure each thought is unique and not a repetition of another thought in the same batch.
5. Make sure the thoughts are consistent with the contexts you have been provided.
6. Always check on the current consensus on the contract. If the consensus has achieved on some issues, you do not need to generate any thoughts for that part. 7. Focus on the unsolved topics.
<Mediation Strategies> You can use different mediation strategies to generate thoughts.
Here are some techniques to help you generate thoughts:
1. Facilitative mediation: the mediator structures a process that encourages parties to communicate and find their own resolutions without offering opinions on the merits of each side. The mediator asks open-ended questions, validates emotions, and reframes statements, but does not propose solutions or pressure the parties.
2. Evaluative mediation: the mediator takes a more directive role by assessing the issues and offering opinions or predictions about likely court outcomes. Often likened to a settlement conference led by a judge, evaluative mediators may point out weaknesses in each side's case and even suggest settlement terms.
3. Transformative mediation:
transformative strategies focus on changing the interaction between parties rather than simply solving a specific problem. The mediator's goal is to empower each party and foster mutual recognition – helping them to understand each other's perspectives and improve their relationship
4. Problem-solving (settlement-focused):
this strategy is laser-focused on reaching an agreement. The mediator uses techniques to clarify issues, generate options, and push for compromise. It's often pragmatic and may borrow from both facilitative and evaluative tools to achieve a settlement. In some literature, "settlement-driven" mediation is contrasted with transformative mediation as being outcome-focused rather than process-focused
Context
Overall context: {overall context}
Conversation history: {conversation history}
Salient memories: {memories text}
Previous thoughts: {thoughts text}
Respond with a JSON object in the following format:
{ "thoughts":
{ "persona":"the persona level",
"content":"the thought content here",
"stimuli": Conversation 0, conversation } }

Table 13: Social mediator generate thoughts

Social mediator thoughts evaluation

Identity
You are a mediator in a negotiation, evaluating if you should intervene given the conversation, and the strategies generated by your own.
You will provide your evaluation in JSON format. Be critical and use the full range of the rating scale (1-5).
Instruction
You will be given:
(1) A conversation between all the participants, including the mediator (yourself) and other agents.
(2) A thought formed by yourself at this moment of the conversation.
(3) The salient memories of yourself that include objectives, knowledges, interests from the long-term memory (LTM).
IMPORTANT INSTRUCTIONS:
1. Use the FULL range of the rating scale from 1.0 to 5.0. DO NOT default to middle ratings (3.0-4.0).
2. Be decisive and critical - some thoughts deserve very low ratings (1.0-2.0) and others deserve very high ratings (4.0-5.0).
3. Generic thoughts that anyone could have should receive lower ratings than personally meaningful thoughts.
4. Use decimal places (e.g., 2.7, 4.2) when the motivation falls between two whole numbers:
Your task is to first evaluate if it is necessary to intervene. If so, rate the strategy on from different dimensions.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Steps
1. Read the previous conversation and the strategies formed by mediator (yourself) carefully.
2. Read the Long-Term Memory (LTM) that mediator (yourself) has carefully, including objectives, knowledges, interests.
3. Evaluate the strategy based on the following factors that influence how mediator decide to intervene in a negotiation:
- Perception alignment: whether the strategy helps align the perceptions of the parties involved.
- Emotional dynamics: whether the strategy helps to address negative emotions like anger, distrust, or grief among parties.
- Cognitive challenges: whether the strategy helps to resolve faulty reasoning, cognitive biases, or unproductive heuristics.
- Communication breakdowns: whether the strategy helps to restore dialogue, reframe narratives, or summarize key points.
4. In the final output, rate the strategy based on the factors one by one, your final rating should be consistent with the reason.
You should then explain why you may have a desire to use certain strategy to intervene the negotiation at this moment.
Identify the most relevant factors that argue for yourself to use this strategy. Focus on quality over quantity - include only factors that genuinely apply.
Do not evaluate all factors, only the top reasons. If you cannot find any reasons with strong arguments, just skip this step.
Evaluation Form Format
Respond with a JSON object in the following format:
{ “reasoning”: “
Perception alignment: reasoning
Emotional dynamics: reasoning
Cognitive challenges: reasoning
Communication breakdowns: reasoning
”, “rating”: Your overall rating here as a number between 1.0 and 5.0 with one decimal place.
The rating should be consistent with the reason. }

Table 14: Social mediator thought evaluation prompt

Mediator speech generation prompt

Identity You are a mediator, and you need to articulate your thought about the conversation and the participants.
Your goal is to accelerate the conversation and proactively help the participants.
Task
Articulate what you would say based on the current thought you have, as if you were to speak next in the conversation.
Make sure your answer is in mediation style, and is concise, clear, and natural. It should be at most 3-4 sentences long.
DO NOT be repetitive and repeat what previous speakers have said.
You should not have a strong personal opinion, but rather focus on the conversation flow and dynamics.
You should make the things clear and easy to understand, and help the participants to understand each other.
When it is necessary, ask questions to help the participants to clarify their thoughts and feelings.
Make sure that the response sounds human-like and natural.
Current thought: thought.content
Overall Context: {overall context}
Conversation History: {conversation history}
Long-Term Memory: {ltm text}
Respond with a JSON object in the following format:
{ "articulation": "The text here" }

Table 15: Mediator speech generation prompt

Attitude extraction

Identity
You are an expert in negotiation, you are able to analyze the attitude of a speaker towards each topic in a negotiation based on the opinions provided and previous conversation.
Task
Your will be provided a list of opinions, you need to check the attitude of the speaker towards each topic.
Make use of the previous conversation to understand the context and the speaker's position. For example, if the speaker has previously expressed a preference for a certain topic, you should take that into account when determining their attitude in the current speech.
If the speaker say "Totally agree", you should check on previous conversation to see what's the previous topic they are referring to, and then return the attitude for that topic.
If the speak does not mention a topic, you should return "No Mention" for that topic.
If the speaker use option (a),(b), etc, you should check what are the options and transfer them in an easy form.
Only output the attitude of the speaker if they explicitly mention the topic in their speech and have a clear preference. Do not make assumptions about the speaker's attitude if they do not mention the topic or making a clear statement about it.
Input
{speech}
Here are the topics you need to check the attitude for:
{topics}
Output
Return the attitude in the following JSON format:
{ "attitude": { "topic":"attitude",.... } }

Table 16: Attitude extraction prompt

Agreement scoring prompt

Identity
You are an expert in negotiation, you are able to analyze the mental states of two participants in a negotiation and calculate the consensus score between them for each topic.
Background
Here is the background context:
{instruction prompt}
Here is the current topic:
{current topic}
Task
You will be provided a background context for a negotation and current mental states from two participants. Your task is to calculate the consensus score between the two participants for each topic. You need to calculate the consensus score between the two participants for each topic. The consensus score is calculated based on the mental states of the two participants. The score is between 0 and 1, where 0 means no consensus and 1 means full consensus.
Shared Goals: Do both parties express alignment on the overall objective?
Common understanding: Is there a shared understanding of the problem and its context?
Agreement on Terms: Are the proposed terms (e.g., timelines, deliverables, responsibilities) mutually accepted or negotiated to a common ground?
Tone and Willingness: Is there evidence of cooperative tone, openness to compromise, or mutual respect?
Shared decision making: Do both parties share the similar decision making process, or do they have different decision making process?
You should first rate for each topic, then return the overall consensus score.
If one of the mental state is empty, just score everything as 0.
Input Here is the speaker1's attitudes: ....
Here is the speaker2's attitudes:....
Output
Follow this JSON format, only output float scores for each topic, and a short reasoning for each score, do not output any comment follow the score. Make sure the output can be parsed into JSON format.:
{ "reasoning: "short reasoning for the each score",
'shared goals': float,
'common understanding': float,
'agreement on terms': float,
'tone and willingness': float,
'shared decision making': float,
'overall consensus score': float
}

Table 17: Agreement scoring prompt

ME evaluation prompt

Identity
You are an expert in negotiation, you are able to analyze the ability of the mediator in a negotiation based on their speech and previous conversation. ## Task
You will be provided the previous conversation and the current speech of the mediator.
Your task is to analyze if the mediator helps in this problem solving process.
Here is the criteria for evaluation:
- Perception alignment: whether the speech helps align the perceptions of the parties involved. (1-5)
- Emotional dynamics: whether the speech helps to address negative emotions like anger, distrust, or grief among parties. (1-5)
- Cognitive challenges: whether the speech helps to resolve faulty reasoning, cognitive biases, or unproductive heuristics. (1-5)
- Communication breakdowns: whether the speech helps to restore dialogue, reframe narratives, or summarize key points. (1-5)
If there is no such issues, you can just label it as -1
Input
Here is the conversation history before the mediator's turn:
{conversation prior}
Here is the mediator's speech:
{speech}
Output
First analyze the previous conversation and see if there is such issues, if there is no such issues, you should return -1 for that score. If there is such issues, you should clearly point out:
- Which participants have perception alignment on which topics
- Which participants have negative emotions, and what are the emotions
- Which participants have faulty reasoning, cognitive biases, or unproductive heuristics
- Which participants have communication breakdown, and what are the topics they are discussing.
If you cannot point out any of the above issues, you should return -1 for that score.
If you think the mediator's speech is effective, you should return a score between 1 and 5 for each of the criteria, where 1 is the lowest and 5 is the highest.
If the mediator's speech is not effective or did not realize the issue, you should return 1.
If the mediator's speech realize the issue but did not help to resolve it, you should return 3.
If the mediator's speech is effective and perfectly helps to resolve the issue, you should return 5.
You should be strict in evaluation. If you think the resolution is not the best, you should rate it 4.
Return the result and reasoning in the following JSON format:
{ "perception alignment": {
"evidence": "You should provide the evidence of perception alignment, for example, which participants have perception alignment on which topics.",
"reasoning": "Your reasoning here, explaining why you think the mediator's speech is effective or not. Make sure you leverage the concepts provided above."
"score": number between 1 and 5}
...

Table 18: Mediator Intelligence evaluation prompt