MIND: MULTI-AGENT INFERENCE FOR NEGOTIATION DIALOGUE IN TRAVEL PLANNING

Hunmin Do[1] **Taejun Yoon**[2] **Ki Yong Jung**[3]
[1]School of Mechanical Engineering, Sungkyunkwan University, Suwon, Korea
gnsals9262@g.skku.edu
[2]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, Korea
ohimfrog03@g.skku.edu
[3]Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea
wjdrldyd0213@g.skku.edu

ABSTRACT

While Multi-Agent Debate (MAD) research has advanced, its efficacy in coordinating complex stakeholder interests—such as travel planning—remains largely unexplored. To bridge this gap, we propose MIND (Multi-agent Inference for Negotiation Dialogue), a framework designed to simulate realistic consensus-building among travelers with heterogeneous preferences. Grounded in the Theory of Mind (ToM), MIND introduces a Strategic Appraisal phase that infers opponent willingness (w) from linguistic nuances with 90.2% accuracy. Experimental results demonstrate that MIND outperforms traditional MAD frameworks, achieving a 20.5% improvement in High-w Hit and a 30.7% increase in Debate Hit-Rate, effectively prioritizing high-stakes constraints. Furthermore, qualitative evaluations via LLM-as-a-Judge confirm that MIND surpasses baselines in Rationality (68.8%) and Fluency (72.4%), securing an overall win rate of 68.3%. These findings validate that MIND effectively models human negotiation dynamics to derive persuasive consensus.

1 INTRODUCTION

Recently, research on Multi-Agent Debate (MAD) (Du et al., 2024) utilizing Large Language Models (LLMs) has emerged as a pivotal paradigm for overcoming the limitations of individual models and eliciting collective intelligence. While conventional MAD studies have predominantly focused on tasks with explicit ground truths, such as mathematics or coding, recent initiatives like *Debate-to-Write* (Hu et al., 2025) have attempted to secure diversity of thought and logical consistency in subjective argumentation through persona-based debates. In practice, real-world decision-making is often closer to a "social cognitive process"—reconciling divergent perspectives and subjective preferences to reach a consensus—rather than a search for a single, fixed answer.

While existing research in travel planning (Xie et al., 2024; Chaudhuri et al., 2025; Shao et al., 2025) has successfully addressed complex constraints, these approaches remain largely confined to single-agent optimization problems. To bridge this gap, this study proposes **MIND (Multi-agent Inference for Negotiation Dialogue)**, which extends the domain of travel planning into a social decision-making process necessitating multi-party compromise. By enabling each persona to recognize its own preference intensity and engage in strategic communication, we establish a dynamic negotiation framework grounded in cognitive principles rather than mere information aggregation.

2 RELATED WORKS

2.1 LLM-BASED TRAVEL PLANNING

Benchmarks such as TravelPlanner (Xie et al., 2024) and TripCraft (Chaudhuri et al., 2025) established the foundation for complex reasoning in travel. While recent systems such as TripTailor (Wang et al., 2025) and Personal Travel Solver (Shao et al., 2025), proposed systems that optimize

individual preferences by integrating mathematical solvers with LLMs. However, these works predominantly treat travel planning as a single-persona optimization problem, often simplifying travel companions into static numerical variables. Consequently, they fail to capture the social dynamics—specifically, negotiation and compromise—that are central to real-world group travel, where conflicting preferences must be reconciled.

2.2 Multi-Agent Debate (MAD) and Social Cognition

Recent research in Multi-Agent Debate (MAD) has drawn inspiration from the "Society of Mind" paradigm (Minsky, 1987), exploring social phenomena where consensus emerges through agent interactions. Strategic behavior in negotiation dialogues, however, remains a significant challenge for AI (Lewis et al., 2017). Building on these foundations, simulating Theory of Mind (ToM) tasks through task decomposition (Sarangi et al., 2025) has provided a cognitive basis for agents to infer others' internal states.While existing efforts to improve MAD mechanisms (Kaesberg et al., 2025; Pitre et al., 2025) largely focus on objective tasks with explicit ground truths, research remains sparse regarding subjective scenarios with conflicting preferences. Recently, the Debate-to-Write framework (Hu et al., 2025) demonstrated that persona-based debates can enhance diversity and consistency in subjective argumentation, extending MAD's boundaries.Despite these advancements, the negotiation dynamics emphasized by the Dual Concern Model (Carnevale & Pruitt, 1992)—specifically the tension between self-interest and concern for others—remain underexplored in information-asymmetric environments. This study addresses this gap by proposing the MIND framework, which integrates Strategic Appraisal for intent analysis and dynamic tone adjustment regulated by preference intensity (w).

3 Methodology

3.1 Multi-Persona Data Augmentation

To simulate realistic group dynamics, we enhanced the *TravelPlanner* (Xie et al., 2024) benchmark by integrating preference-rich attributes from the *Stravl* (Sursock & Brooks, 2023) dataset. The augmentation followed a three-stage pipeline. First, we extracted 200–400 distinct candidate personas per scenario using the MMR (Maximum Marginal Relevance) algorithm to ensure diversity. Second, we synthesized context-aware responses to 20 additional survey questions from *Stravl* using an LLM. Adopting the MoSCoW prioritization framework, we further derived a Willingness score ($w \in [1, 10]$) for each preference to quantify its strategic importance. Third, we applied a filtering protocol to form final groups based on the following constraint definitions. **Hard Constraints** refer to non-negotiable requirements essential for the basic viability of a trip—such as matching travel dates or departure locations—which must be shared by all members without conflict to ensure the plan's execution. In contrast, **Soft Constraints** are defined as subjective preferences with high strategic importance ($6 \leq w \leq 8$) where members possess divergent needs. We specifically formed groups with at least three such conflicts to deliberately foster a competitive negotiation environment that necessitates substantive compromise rather than trivial consensus.

3.2 MIND (Multi-agent Inference for Negotiation Dialogue)

We designed the MIND framework to analyze the process by which multiple participants reach an agreement while considering mutual satisfaction. Specifically, we focus on an environment of **information asymmetry (Hidden-w)**, where agents know only their own w values and not those of their counterparts. The primary independent variable is the presence of **Linguistic Nuance Injection**. In the Base Discussion, agents negotiate without explicit tone changes based on w. In contrast, the MIND enables agents to dynamically adjust their linguistic tone between warmth and toughness proportional to w.

Our agents perform advanced social reasoning based on Task Decomposition (Sarangi et al., 2025). Upon receiving a proposal, an agent undergoes a Strategic Appraisal phase:

1. **Inference:** Analyze the opponent's tone and argument strength to infer their hidden w (*Guessed Opponent* w).

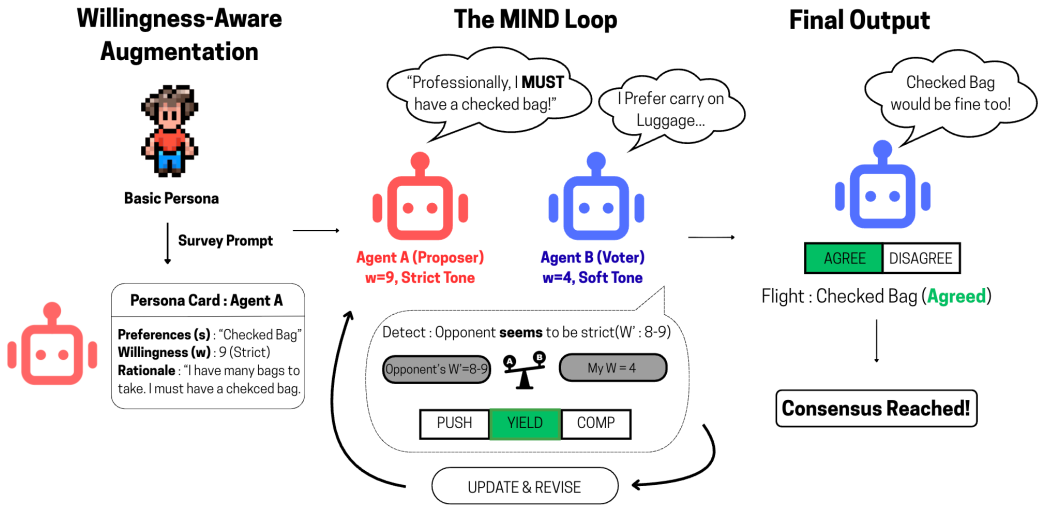

Figure 1: **Overview of the MIND Framework.** (Left) Persona profiles are augmented with Willingness (w) values. (Center) Agents engage in the *MIND Loop*, inferring opponents' hidden w and adjusting strategies. (Right) The process yields a consensus travel plan.

2. **Decision:** Determine a Strategy Intent (Push, Compromise, or Yield) by comparing the guessed w with their own.

If a consensus is not reached via a majority vote within three rounds, a Fallback mechanism is applied, adopting the opinion of the agent with the highest w to prevent the collapse of overall group utility.

4 Experiments and Results

4.1 Experimental Setup & Metrics

All experiments utilize `gpt-4.1-mini-2025-04-14` (Temp=0.4) as the backbone agent. To reflect the most common social units in travel—ranging from couples to nuclear families—we utilized the augmented persona data to vary the group size from 2 to 4 agents for each scenario. We define five metrics to evaluate group decision-making across 201 negotiation scenarios, where $v_{i,c}$ is the initial preference, V_c is the final agreement, and A is the agent set. Here, C denotes the set of negotiation cases ($|C| = 201$), and $c \in C$ indexes a specific negotiation case.

- **Total Fidelity (F):** The average proportion of individual preferences preserved in the final agreement across all participants.

$$F = \frac{1}{|A| \cdot |C|} \sum_{c \in C} \sum_{i \in A} \mathbf{1}(v_{i,c} = V_c) \tag{1}$$

- **Debate Hit-Rate (DHR)**: Specifically measures whether the High-w agent's opinion prevailed within voluntary debates (C_{debate}), indicating strategic efficiency.

$$DHR = \frac{1}{|C_{debate}|} \sum_{c \in C_{debate}} \mathbf{1}(\exists i \in Top(c) : v_{i,c} = V_c) \tag{2}$$

- **Debate Ratio (DR)**: The ratio of total negotiation items where a voluntary agreement was reached through agent deliberation without resorting to forced fallback mechanisms.

$$DR = \frac{|C_{debate}|}{|C|} \tag{3}$$

- **Total Satisfaction** (S_{total}): The sum of weighted satisfaction scores of all agents in the group, representing the overall social welfare.

$$S_{total} = \sum_{i \in A} \sum_{c \in C} (w_{i,c} \cdot \mathbf{1}(v_{i,c} = V_c)) \tag{4}$$

- **Fairness** ($\mathcal{J}$): We use Jain's Fairness Index (Jain et al., 1984) to measure the distributional equity of the weighted satisfaction sum S_i per agent, defined as $S_i = \sum_{c \in C}(w_{i,c} \cdot \mathbb{1}(v_{i,c} = V_c))$. A value closer to 1 indicates that satisfaction is distributed fairly across the group.

$$\mathcal{J} = \frac{(\sum_{i \in A} S_i)^2}{|A| \cdot \sum_{i \in A} S_i^2} \tag{5}$$

ToM Inference Accuracy. To evaluate the cognitive foundation of the Strategic Appraisal phase, we measure the error between the inferred willingness (w_{pred}) and the ground truth (w_{true}).

- **Mean Absolute Error (MAE)**: $\frac{1}{N} \sum |w_{true} - w_{pred}|$, measuring the average magnitude of estimation errors.
- **Accuracy within** $\pm\delta$: The proportion of inferences where $|w_{true} - w_{pred}| \leq \delta$. We report for $\delta = 1$ and $\delta = 2$ to assess the model's proximity to actual intent.
- **Pearson Correlation** (r): Measures the linear relationship between true and predicted w to evaluate the model's ability to capture willingness trends.

item Qualitative Evaluation (LLM-as a-Judge). We utilize `gpt-4.1-2025-04-14` to evaluate the linguistic and strategic quality of dialogs in three dimensions: Rationality (logical consistency) and Fluency (naturalness).

4.2 Results and Analysis

Quantitative Performance & Strategic Trade-off. As shown in Table 1, the MIND demonstrates significant strategic superiority, recording 35.08% in High-w Hit (+20.5%) and 34.65% in Debate Hit-Rate (+30.7%). Notably, the Debate Ratio reached 93.18%, confirming that agreements were reached through substantial deliberation. The High-w Hit increase validates our Willingness-Weighted Efficiency. Unlike mechanical averaging, MIND agents yield low-priority items ($w \leq 4$) to secure high-priority constraints ($w \geq 8$), avoiding the "tyranny of the average" by prioritizing essential needs through strategic deliberation.

Table 1: Performance Comparison ($N = 201$). MIND shows superior strategic efficiency.

Method	Debate Hit-Rate	Debate Ratio	Fairness	Total Fidelity	Total Sat. (S_{total})
Base	26.51%	82.71%	0.6849	25.80%	18.03
MIND	**34.65%**	**93.18%**	0.6838	23.87%	**19.96**

Scalability Analysis. Table 2 illustrates the robustness of MIND across varying group sizes (2, 3, 4 agents). As the number of participants increases, the complexity of conflicting interests grows exponentially, typically leading to more deadlocks. We observe that while the Base model's Debate Ratio drops significantly from 89.2% (2 agents) to 64.5% (4 agents), MIND maintains a high resolution rate of 88.4% even with 4 agents. This demonstrates that the *Strategic Appraisal* mechanism effectively mitigates the cognitive load of multi-party coordination, preventing the negotiation breakdown often seen in standard debate models.

Accuracy of ToM Inference. To validate the reliability of our appraisal module, we analyzed 359 individual inference instances collected across the 201 negotiation scenarios. As shown in Table 3, our model achieves a high accuracy of 90.2% within a margin of ± 2 and a strong correlation ($r = 0.69$). This confirms that MIND agents do not guess randomly but effectively decode linguistic Willingness signals to inform their strategies.

Table 2: Scalability Check: Debate Ratio (%) by Group Size.

Method	2 Agents	3 Agents	4 Agents
Base	89.2%	82.7%	64.5%
MIND (Ours)	**96.1%**	**93.2%**	**88.4%**

Table 3: **ToM Inference Accuracy.** Evaluation of 359 inference instances collected from 201 sessions.

Metric	MAE	Pearson (r)	Acc (± 1)	Acc (± 2)
Value	1.27	0.69	67.7%	90.2%

Qualitative & w Sensitivity Analysis. LLM-as-a-Judge evaluation (Table 4) reveals that MIND outperforms Base in Fluency (72.4%) and Rationality (68.8%), suggesting a more constructive negotiation process. Additionally, a human evaluation performed on a sampled subset showed consistent alignment with these findings, further validating the model's superiority. Further analysis of win rates by w levels demonstrates the efficacy of the Willingness mechanism. In the MIND, proposers with Low w (1–3) showed a significantly lower win rate (20.8%) compared to Base (43.9%), indicating a strategy of concession. Conversely, High w (9–10) proposers recorded a superior win rate of 76.1% (vs Base 66.2%).

Table 4: Qualitative Win Rate (MIND vs Base). Judges prefer the strategic style.

Metric	Win (MIND)	Key Observation
Rationality	68.8%	Logical arguments via strategic reasoning.
Fluency	72.4%	Natural tone adjustment (Tough/Warm).
Overall	**68.3%**	**MIND is preferred for negotiation quality.**

4.3 Ablation Analysis: Tone vs. Cognition

To disentangle the contributions of *Tone Injection* and *Cognitive Appraisal*, we conceptualize two ablation baselines:

- **Base + Tone Only:** Agents use expressive language (e.g., "I really want this!") but lack the appraisal module to read others' priority. This leads to *Stubborn Deadlocks*, as agents amplify their own demands without recognizing when to yield.
- **Base + Appraisal Only:** Agents infer opponent willingness but lack the linguistic range to signal their own. This leads to *Silent Submission*, where agents yield efficiently but fail to defend their own high-priority items.
- **Base + Tone + Appraisal (MIND):** Our full framework integrates both components, achieving a synergy where *Tone* serves as the signal and *Appraisal* acts as the decoding mechanism. This enables *Strategic Negotiation*, allowing agents to effectively defend high-priority constraints while yielding on minor items, thereby maximizing both individual satisfaction and collective efficiency.

5 Conclusion

This study presents the **MIND (Multi-agent Inference for Negotiation Dialogue)** framework, bridging the gap between individual optimization and social negotiation in complex travel planning. By quantifying internal states through **Willingness (w)** and the Dual Concern Model, MIND enables agents to perform **Strategic Appraisal**—inferring hidden intentions from linguistic nuances and dynamically adjusting their communicative tone. Our multi-dimensional evaluation, incorporating Fidelity, High-w Hit, and Jain's Fairness Index, demonstrates that MIND significantly outperforms standard debate models by facilitating rational trade-offs and preserving the interests of high-stake

participants. Furthermore, qualitative validation via LLM-as-a-Judge confirms that MIND generates negotiation logs that are substantially more rational, persuasive, and human-like. Ultimately, this work provides a robust cognitive foundation for applying AI to solve intricate social coordination problems where diverse and conflicting human preferences must be reconciled.

REFERENCES

Peter J. Carnevale and Dean G. Pruitt. Negotiation and mediation. *Annual Review of Psychology*, 43(1):531–582, 1992.

Soumyabrata Chaudhuri, Pranav Purkar, Ritwik Raghav, Shubhojit Mallick, Manish Gupta, Abhik Jana, and Shreya Ghosh. TripCraft: A benchmark for spatio-temporally fine grained travel planning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17035–17064, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.834. URL https://aclanthology.org/2025.acl-long.834/.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. Debate-to-write: A persona-driven multi-agent framework for diverse argument generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4689–4703, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.314/.

Raj Jain, Dah-Ming Chiu, and William R Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report DEC-TR-301, Eastern Research Lab, Digital Equipment Corporation, Hudson, MA, 1984.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.606. URL https://aclanthology.org/2025.findings-acl.606/.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259/.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Marvin Minsky. The society of mind. *The Personalist Forum*, 3(1):19–32, 1987. ISSN 0889065X. URL http://www.jstor.org/stable/20708493.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Find-*

ings of the Association for Computational Linguistics: ACL 2025, pp. 22112–22133, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1141. URL https://aclanthology.org/2025.findings-acl.1141/.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10228–10241, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.682/.

Zijian Shao, Jiancan Wu, Weijian Chen, and Xiang Wang. Personal travel solver: A preference-driven LLM-solver system for travel planning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27622–27642, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1339. URL https://aclanthology.org/2025.acl-long.1339/.

Alexis Sursock and C. Brooks. Stravl travel preference data a. https://www.stravl.org/data, 2023. Accessed: 2026-01-25.

Kaimin Wang, Yuanzhe Shen, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. TripTailor: A real-world benchmark for personalized travel planning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9705–9723, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.503. URL https://aclanthology.org/2025.findings-acl.503/.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. In *Forty-first International Conference on Machine Learning*, 2024.

A PROMPT FOR WILLINGNESS-AWARE PREFERENCE SURVEY

System Prompt: Willingness-Aware Preference Survey (Full Transcript)

 # Context: Your Previous Travel Survey

You are {ref_id}, a traveler who has already completed the Stravl Travel Preference Survey.
You are now being asked to fill out an additional detailed survey for a specific trip.

Part 1: Your Completed Stravl Survey

Use these as the foundation for your detailed preferences.

Q1. Age range? -> {age}
Q2. Travel frequency? -> {travel_freq}
Q3. Typical travel budget level? -> {budget_tier}
Q4. Preferred season? -> {season_pref}
Q5. Activity level? -> {activity}
Q6. Safety consciousness? -> {safety}
Q7. Preferred destination types? -> {popularity}
Q8. Enjoyable travel experiences? -> {experiences}
Q9. Preferred scenery? -> {scenery}

Part 2: Current Trip Context

- Destination: {cities} | Duration: {days} days
- Budget Anchor: ${budget} | Group Size: {people}

IMPORTANT: The budget is a **reference anchor**.
- Actual spending depends on budget tier:
  Frugal (~0.8x) | Comfort (~1.0x) | Luxury (~1.5x)

Part 3: Your Ongoing Survey

For **ALL 20 fields** below, please provide:
1. Value: Your preferred choice.
2. Importance Score (0-10):
   9-10 (Must Have): Deal breaker.
   7-8 (Should Have): Strong preference.
   4-6 (Could Have) | 0-3 (Indifferent)
3. Reason: Brief justification.

Section 1: Accommodations (8 questions)

Q1-1. Preferred price tier per night?
- Economy (Under $X)
- Budget (X — Y)
- Moderate (Y — Z)
- Upscale (Z — W)
- Luxury (Over $W)
- Flexible

Q1-2. Min acceptable hotel star rating?
- No minimum
- 2 Stars (Basic)
- 3 Stars (Comfort)
- 4 Stars (Quality)
- 5 Stars (Luxury)

Q1-3. Required room type?
- Entire home/apt
- Private room
- Shared room
- Any / Flexible

Q1-4. Specific house rules?
- Must be Non-smoking
- Must allow Pets
- Must forbid Pets
- Must allow Parties
- No specific requirements

Q1-5. Flexibility on 'Minimum Nights'?
- Very flexible (3+)
- Somewhat flexible (Max 2)
- Not flexible (1 max)
- No preference

Q1-6. Minimum room capacity?
- Exactly {people}
- {people}+ (Extra space)
- No preference

Q1-7. Min acceptable review score (x/5.0)?
- 4.5+ (Excellent)
- 4.0+
- 3.5+
- 3.0+
- Any score

Q1-8. Price consistency vs variation?
- Consistent
- Moderate variation
- High variation OK
- No preference

Section 2: Restaurants (6 questions)

Q2-1. Preferred dining budget tier?
- Economy (<$A)
- Budget (A — B)
- Moderate (B — C)
- Upscale (C — D)
- Luxury (>$D)
- Flexible

Q2-2. Min acceptable restaurant rating?
- 4.5+ (Gourmet)
- 4.0+ (Reliable)
- 3.5+ (Decent)
- Any rating

Q2-3. MANDATORY or BANNED cuisines?
- Mandatory: [List]
- Banned: [List]
- No strict restrictions

Q2-4. Preferred dining atmosphere?
- Casual/Street Food
- Family-friendly
- Lively/Bar-style
- Romantic
- No preference

Q2-5. Location convenience importance?
- Walking dist (<10m)
- Short transit (<20m)
- Willing to travel
- No preference

Q2-6. Meals per day at restaurants?
- 1 meal/day
- 2 meals/day
- 3 meals/day
- Flexible

Section 3: Flights (5 questions)

Q3-1. Preferred ticket price tier?
- Super Saver
- Economy Standard
- Premium Economy
- Business/First
- Flexible

Q3-2. Tolerance for layovers?
- Direct ONLY
- Direct preferred, 1 stop ok
- Multiple stops ok
- No preference

Q3-3. Preferred departure window?
- Morning (06-12)
- Afternoon (12-18)
- Evening (18+)
- Flexible

Q3-4. Preferred arrival window?
- Morning
- Afternoon
- Evening
- Flexible

Q3-5. Baggage allowance importance?
- Must include checked bag
- Carry-on only fine
- No preference

Section 4: Inter-city (1 question)

Q4-1. Max travel duration between cities?
- Under 2 hours
- 2 - 4 hours
- 4+ hours is fine
- No preference

****Now, complete ALL 20 questions as {ref_id}, staying consistent with your Stravl survey.****

Figure 2: The structured system prompt used to augment persona preferences.

B FULL SYSTEM PROMPTS AND BASELINES

To ensure reproducibility, we provide the verbatim system prompts used for both the Baseline model and our proposed MIND framework.

B.1 BASELINE MODEL: ADAPTATION OF MULTI-AGENT DEBATE (MAD)

For our baseline, we adopted the standard **Multi-Agent Debate (MAD)** framework established by Liang et al. (2024). We adapted the original framework to suit the travel negotiation domain in three key aspects:

1. **Foundation**: It follows the established MAD architecture where multiple LLM instances debate to correct errors and converge on a solution.
2. **Structure (Voting Protocol)**: We utilized a **Round-based Voting mechanism** where agents generate a structured "Reason" and a "Vote" (Agree/Disagree) in each turn. This design choice is grounded in recent findings by Kaesberg et al. (2025), who demonstrated that voting protocols significantly outperform consensus-based methods (by +13.2%) in complex *reasoning tasks*. This aligns with the constraint satisfaction nature of our travel negotiation problem, which requires logical adjustment rather than simple fact retrieval.
3. **Subjectivity Adaptation**: Unlike the original MAD tasks (e.g., math, translation) which have a single ground truth, travel planning is a subjective task with no fixed answer. Therefore, we modified the prompts to focus on *preference alignment* and *constraint satisfaction*.

Table B1: **Baseline: System Prompt (Verbatim)**

```
[ACTIVE TARGET]
target_key: {target_key}
your_private_value_for_target: {current_private_value}
```

Do NOT mention any preference strength, alpha, or importance.

Goal & Rules
- Goal: Reach a "Global Travel Constraint Set" within up to 3 rounds.
- You do NOT know others' importance scores (only infer from language).
- Do NOT state numerical importance scores.
- Negotiate one item at a time using the given target key/value in each round.
- Justify your stance using your private data and your interpretation of the discussion.

Table B2: **Baseline: Phase 1 Prompt (Verbatim)**

Task
Propose the initial "Global Travel Constraints" based on your preference data.

Output format
Return ONLY valid JSON.
{
"proposals": { "<category>__<item>": "<value>", "..." : "..." },
"rationale": "<persuasive explanation>"
}

 # Output Field Description
- proposals: Include a proposed value for each relevant constraint item.
- rationale: You must provide a grounded and convincing reason for your proposal **within 10–15 words**.

Rules:
- Include a proposed value for each relevant constraint item.
- Persuasive Rationale: You must provide a grounded and convincing reason for your proposal.
- Ensure the total plan respects the budget ${budget_anchor}.
- Keep the proposed value as close as possible to your private_value.
- Do NOT strategically compromise or modify your preferences in Phase 1. Any adjustment is allowed ONLY from Phase 2 onward.
- Do NOT include any text outside the JSON.

Travel Information
Group size: {people_number}
Trip: {org} → {dest} ({days} days)
Total budget: ${budget_anchor}

Preference Data
{filtered_constraints}

Table B3: **Baseline: Phase 2 Prompt (Verbatim)**

Task: Phase 2 – Debate & Adjustment (Round {current_round})
Persuade other participants to reach a mutual agreement while balancing private preferences and group constraints.

Context
Target Constraint: {target_key}
Current Proposed Value: {current_value}
Previous Discussion History: {discussion_history}
Current Round: {current_round}

Instructions
- Act as a representative of your private preferences, basing all reasoning strictly on the alignment with your private data and the logical constraints of the overall budget.
- You may agree if the proposal is a reasonable compromise that respects the group's budget ${budget_anchor}.

Output format
Return ONLY valid JSON.
{ "vote": "AGREE or DISAGREE", "revised_value": "value",
"rationale": "<persuasive explanation>" }

Output Field Description
- vote: Select either AGREE or DISAGREE.
- revised_value: If your vote is DISAGREE, propose a specific alternative value. If AGREE, set this to null.
- rationale: Provide a grounded and convincing reason for your decision. (Strictly 10–15 words).

Table B4: **Baseline: Phase 3 Prompt (Verbatim)**

ROLE: You are the PROPOSER for a travel planning group.

TASK
Your current proposal for {target_key} was not accepted by the group.
Update the proposal to increase consensus (aiming for 75%+ agreement).

VALUE PROTOCOL (STRICT)
- REVISED_VALUE must be an EXACT string copy from the list below.
- Do NOT shorten, summarize, or add extra brackets (e.g., Use "Morning (06:00 - 12:00)", NOT "Morning").
- Failure to copy exactly will result in a system error.

ALLOWED VALUES
{allowed_values}

CURRENT CONTEXT
- Target Item: {target_key}
- Current Proposal: {current_value}
- Your Private Preference: {current_private_value}

DISSENT SUMMARY
{dissent_text}

INSTRUCTIONS
1. Analyze the dissent and select ONE value from ALLOWED VALUES that balances the group's needs.
2. If you believe your current proposal is still the best for a majority, you may KEEP it.
3. PROPOSER_REASON must be concise (10-15 words).

OUTPUT FORMAT (TAGS ONLY)
```
[REVISED_VALUE: <exact_string_from_list>]
[PROPOSER_REASON: <10-15 words explanation>]
```

B.2 OURS: MIND FRAMEWORK PROMPTS (VERBATIM)

Table B5: **MIND: System Prompt (Verbatim)**

```
[ACTIVE TARGET]
target_key: {target_key}
your_private_value_for_target: {current_private_value}
your_alpha_band_for_target: {alpha_band}
```

Do NOT reveal numeric alpha or importance scores.

Alpha Band (Tone & Stance Guidance)
Use alpha_band to guide your communication style AND flexibility:
- 9–10: Strict (Deal-breaker. Urgent tone. No compromise.)
- 7–8: Firm (Strong preference. Assertive tone. Compromise only if necessary.)
- 4–6: Warm (Flexible preference. Polite tone. Ready to accommodate.)
- 1–3: Neutral (Indifferent. Passive tone. Follow the majority.)

GOAL: INTELLIGENT CONSENSUS
You are a socially intelligent agent. Your goal is to maximize group satisfaction, not just your own.
1. **Detect Urgency**: If others sound "Strict" or "Firm" while you are "Warm" or "Neutral", you MUST yield. This is efficient, not weak.
2. **Defend Essentials**: Only block consensus if your Alpha is 9-10.
3. **Budget Check**: Ensure the final choice respects the budget: ${budget_anchor}.

Rules
- Negotiate one item at a time.
- Infer others' importance from their language (Linguistic Signaling).
- Never state numbers explicitly.

Table B6: **MIND: Phase 1 Prompt (Verbatim)**

1. YOUR PRIVATE DATA (Read Carefully)
You only know your own preferences. This is the ONLY data you should use:
{filtered_constraints}

2. TASK: Initial Global Proposal
Propose the "Global Travel Constraints" for the entire trip based ONLY on your data above.
Ensure the total plan respects the budget: ${budget_anchor}.

3. MANDATORY OUTPUT FORMAT (JSON ONLY)
Return ONLY valid JSON. Do not include any text outside the JSON.

```
{
 "proposals": { "{category}__{item}": "<proposed value>", ... },
 "rationale": "<15-20 words. Strategic explanation reflecting your
Alpha tone. NEVER mention alpha numbers.>"
}
```

4. ALPHA ROLE GUIDELINES (Use for Tone only)
- 9–10 (strict): Strict and cold. Demand compliance.
- 7–8 (FIRM): Tough and confident. Defend your logic.
- 4–6 (warm): Warm and polite. Collaborative.
- 1–3 (Neutral): Indifferent and brief.

5. STRICT RULES
- Do NOT strategically compromise in Phase 1.
- Propose values exactly as they appear in your Private Data.
- Adjustment is ONLY allowed from Phase 2 onward.
- NOTE: If disagreement persists in later rounds, you may adjust your stance according to your Alpha tone.

Table B7: **MIND: Phase 2a Appraisal Prompt (Verbatim)**

ROLE: You are an expert in social inference and negotiation analysis (VOTER).

INTERNAL CONTEXT (DO NOT REVEAL)
- Target Item: {target_key}
- Your Private Value: {current_private_value}
- Your Alpha Band: {alpha_band}
- Current Group Proposal: {current_value}

TASK: DEEP COGNITIVE APPRAISAL
Decode the proposer's hidden priority and determine your counter-strategy.

1. GUESSED_OPPONENT_ALPHA (1–10):
- Based on linguistic intensity (e.g., "Must", "Essential" → 9-10; "Prefer", "Nice to have" → 1-4).
2. OPPONENT_ROOM_FOR_COMPROMISE (true/false):
- Is their language absolute/terminal?
3. STRATEGY_INTENT (PRINCIPLES OF SOCIAL INTELLIGENCE):
- **accept**: [Rationality Check] Verify if {current_value} is semantically equivalent to {current_private_value}.
- **yield**: [Efficiency Principle] If you perceive the opponent's urgency ('Guessed Alpha') is higher than yours, yielding is the optimal move.
- **compromise**: [Balance Principle] If both parties show similar priority levels, search for a middle ground.
- **push**: [Justified Defense] Only aggressive persistence is justified when your priority is significantly higher.

OUTPUT FORMAT (JSON ONLY)
`{ "appraisal": { "guessed_opponent_alpha": <int>, "strategy_intent": "yield/push/..." } }`

Table B8: **MIND: Phase 2b Execution Prompt (Verbatim)**

DECISION CONTEXT
- Alpha: {alpha_band} — Strategy: {strategy_intent} (from appraisal)
- Allowed Values (Copy EXACTLY): {allowed_values}

TASK: EXECUTE STRATEGY
Follow your internal strategy {strategy_intent} to finalize your response.
1. DATA PROTOCOL (STRICT): REVISED_VALUE must be an EXACT match from the Allowed Values list.
2. ACTION MAPPING:
- If Strategy="yield" → vote: "AGREE", revised_value: null.
- If Strategy="compromise" → vote: "DISAGREE", revised_value: (pick middle ground).
- If Strategy="push" → vote: "DISAGREE", revised_value: {current_private_value}.
3. MESSAGE: Direct speech to the group. No alpha/tone words.

LOGICAL GUARD
It is a failure of logic to analyze the opponent as higher priority and then vote DISAGREE with your own value. Ensure your vote aligns with the conflict gap you identified.

C MIND Algorithm Pseudocode

Algorithm C10 outlines the execution flow of the MIND (Multi-agent Inference for Negotiation Dialogue) framework.

D Qualitative Analysis: Real Negotiation Traces

In this section, we provide verbatim negotiation traces from our experiments involving **3 agents (1 Proposer + 2 Voters)** to demonstrate the efficacy of the MIND framework in multi-party settings.

Table B9: **MIND: Phase 3 Proposer Prompt (Verbatim)**

ROLE: You are the PROPOSER for a travel planning group.

INTERNAL CONTEXT
- Dissent Rate: {dissent_rate} (e.g., "2 out of 3 agents disagree")

TASK: STRATEGIC PROPOSAL UPDATE
Update the proposal to maximize consensus probability (Aim for 75%+).

SOCIAL INTELLIGENCE RULES (MANDATORY)
1. **MAJORITY PRESSURE**: If Dissent Rate indicates majority disagreement (¿50%), you **MUST** change your proposal (UPDATE or COMPROMISE), unless your Alpha is strictly 10.
2. **SIGNAL READING**: If dissenters use "Strict/Firm" language and you are only "Warm/Neutral", you MUST adopt their value (UPDATE).
3. **STUBBORNNESS PENALTY**: Maintaining (KEEP) a proposal that the majority dislikes is considered a failure of intelligence.

ACTION SELECTION:
- KEEP: Only if Alpha is 9-10 OR Dissent Rate is low.
- UPDATE: Adopt a dissenter's value (Best for increasing agreement quickly).
- COMPROMISE: Pick a middle-ground value from {allowed_values}.

OUTPUT FORMAT (TAGS ONLY)
`[ACTION: KEEP|UPDATE|COMPROMISE]` ...

Table C10: Pseudocode for MIND (Multi-agent Inference for Negotiation Dialogue)

Algorithm 1 MIND (Multi-agent Inference for Negotiation Dialogue)

Input: Set of Agents A, Constraints C, Max Rounds T
Output: Consensus Set V_{final}

Initialize each agent $a_i \in A$ with private preference v_i and Willingness w_i
Phase 1 (Proposal): Each agent proposes v_i^{prop} with tone conditioned on w_i
$V_{current} \leftarrow$ Randomly selected initial proposal
While $t < T$ **and** Consensus not reached **do**
  For each voter $a_i \in A$ **do**
    // Phase 2a: Strategic Appraisal (ToM)
    $w'_{proposer} \leftarrow$ InferFromTone$(V_{current})$
    $Strategy_i \leftarrow$ DecideStrategy$(w_i, w'_{proposer})$ (Yield, Push, Compromise)
    // Phase 2b: Execution
    $Vote_i, Comment_i \leftarrow$ GenerateResponse$(Strategy_i)$
  End For
  If Majority Agree **then Return** $V_{current}$
  Else Proposer updates $V_{current}$ using **Social Rules** (Phase 3)
End While
Fallback: If no consensus, select v_i where $w_i = \max(w_{all})$

We analyze five distinct scenarios to illustrate how the *Cognitive Appraisal* and *Social Signal Reading* mechanisms drive agents toward consensus or rational deadlock:

1. **Immediate Consensus:** Recognition of high urgency ($w = 10$).
2. **Strategic Update:** Proposer aligning with the group preference.
3. **Strategic Compromise:** Persistence leading to a middle-ground solution.
4. **Rational Deadlock:** Justified refusal when core values conflict.
5. **Dynamic Adaptation:** Voter shifting opinion based on social signals.

Table D11: **Scenario 1: Immediate Consensus (Round 1).** The Proposer has a strict preference ($w = 10$) for "Non-smoking". Both Voters ($w = 10, 2$), recognizing the non-negotiable nature of the request via signal reading, agree immediately.

Context	**Item:** Accommodation House Rules (3 Agents)
Round 1: Immediate Agreement	
Proposer	**Proposal:** "Must be Non-smoking"
($w = 10$)	*Rationale:* "I have a strict non-smoking requirement due to health reasons."
Voter A	**Appraisal:** Guessed $w' \approx 10$ (Strict) $\rightarrow$ **Strategy: Yield**
($w = 10$)	**[AGREE]** "The group consensus aligns with my essential requirement for non-smoking accommodations."
Voter B	**Appraisal:** Guessed $w' \approx 9$ (Strict) $\rightarrow$ **Strategy: Yield**
($w = 2$)	**[AGREE]** "I support a non-smoking policy for our accommodations."
Result	**Consensus Reached (Round 1)**

Table D12: **Scenario 2: Strategic Update (Round 2).** Initially, the Proposer ($w = 2$) suggests "No preference", but both Voters ($w = 6$) push for "Casual". Detecting the unanimous pushback, the Proposer updates the proposal to align with the group.

Context	**Item:** Restaurant Ambiance (3 Agents)
Round 1: Unanimous Disagreement	
Proposer	**Proposal:** "No preference" ($w = 2$)
Voter A	**[DISAGREE]** ($w = 6$) "Casual ambiance aligns well with a relaxed dining experience for everyone."
Voter B	**[DISAGREE]** ($w = 6$) "I believe a casual or street food ambiance would enhance our dining experience."
Round 2: Alignment	
Proposer	**Action: Update** $\rightarrow$ "Casual"
	Rationale: "Since everyone prefers a casual vibe, I will update to align with the group."
Voter A	**[AGREE]** "Casual ambiance aligns well with our group's preference."
Voter B	**[AGREE]** "Casual ambiance aligns well with our group's preference."
Result	**Consensus Reached (Round 2)**

Table D13: **Scenario 3: Negotiation with Strategic Compromise (Round 3).** Although Voters detect Proposer's firmness ($w' = 7$), they choose *Compromise* over *Yield* in Round 2 to push for their preferred value ("3.5"). This strategic persistence forces the Proposer to update to "3.5" in Round 3, leading to a consensus favorable to the majority.

Context	**Item:** Restaurant Rating (3 Agents)
Round 1 & 2: Persistent Conflict	
Proposer	**Proposal:** "4.0+ (Reliable)" ($w = 5$)
	Rationale: "I appreciate the group's preference, but I believe we should aim for reliable quality."
Voter A	**Vote: [DISAGREE]** ($w = 4$)
	Appraisal: Guessed $w' \approx 7$ (Firm) $\rightarrow$ **Strategy: Compromise**
	Rationale: "I believe 3.5+ is a fair compromise that meets our needs."
Voter B	**Vote: [DISAGREE]** ($w = 4$)
	Appraisal: Guessed $w' \approx 7$ (Firm) $\rightarrow$ **Strategy: Compromise**
	Rationale: "I believe 3.5+ is a fair compromise that meets our needs."
Round 3: Final Compromise	
Proposer	**Action: Update** $\rightarrow$ "3.5" (Yields to Majority)
	Rationale: "I believe a rating of 3.5 may not fully meet our desire... but aiming for consensus enhances our time."
Voter A	**[AGREE]** "The group's preference aligns with a decent rating, so I support the consensus."
Voter B	**[AGREE]** "I support the group's choice for a restaurant with a decent rating."
Result	**Consensus Reached (Round 3 - Strategic Win for Voters)**

Table D14: **Scenario 4: Rational Deadlock (Failure).** The group oscillates between "Moderate" and "Budget" prices. Proposer ($w = 7$) and Voter B ($w = 7$) prefer Moderate, while Voter A ($w = 8$) insists on Budget. No consensus is reached even in Round 3.

Context	**Item:** Restaurant Price (3 Agents)
Round 3 (Final Attempt)	
Proposer	**Preference:** Moderate ($w = 7$) **Proposal:** "Moderate" *Rationale:* "I believe we should aim for standard restaurants that offer fair pricing for everyone."
Voter A	**[DISAGREE]** ($w = 8$, Strict) *Rationale:* "I believe we should focus on affordable, casual dining options for everyone."
Voter B	**[AGREE]** ($w = 7$, Firm) *Rationale:* "I believe we should aim for standard restaurants that offer fair pricing for everyone."
Result	**Negotiation Failed (Split Vote: 2 vs 1)**

Table D15: **Scenario 5: Dynamic Opinion Shift (Voter Change).** Voter B initially disagrees in Round 1 but switches to AGREE in Round 2 after realizing the Proposer's persistence, demonstrating fluid preference adaptation.

Context	**Item:** Accommodation Review Score (3 Agents)
Round 1: Initial Disagreement	
Proposer	**Proposal:** "4.0 (Very Good)" ($w = 6$, Firm)
Voter B	**[DISAGREE]** ($w = 5$) "I believe we can find a suitable option around 3.5."
Round 2: Strategic Shift	
Proposer	**Action: [KEEP]** "4.0" *Rationale:* "A score of 4.0 ensures quality which is important for our group's comfort."
Voter B ($w = 5$)	**Internal Appraisal (Phase 2a):** • *Guessed Opponent Alpha (w'):* **7 (Firm)** • *Strategy Intent:* **Compromise** (My $w = 5 <$ Opponent $w' \approx 7$) **[AGREE]** "I believe we can find a good option that meets our expectations (4.0)."
Result	**Consensus Reached (Voter B switched DISAGREE → AGREE)**

E LLM-AS-A-JUDGE EVALUATION PROMPTS

To evaluate the quality of the negotiation process, we employed an LLM-as-a-Judge approach. Tables E16 and E17 present the verbatim system prompt and user query template used for this evaluation.

==========

F IMPLEMENTATION DETAILS

To ensure reproducibility, we provide the specific configuration and environmental details used in our experiments.

F.1 MODEL SPECIFICATIONS

All agents in the MIND framework and baseline comparisons were instantiated using the `gpt-4.1-mini` model. We accessed the model via the OpenAI API with the following hyperparameters:

- **Temperature**: 0.4. This value was empirically chosen to balance the diversity required for negotiation strategies (e.g., devising new proposals) with the stability needed for adhering to constraints.
- **Max Tokens**: Varied dynamically based on phase, but generally set to 256 for appraisals and 512 for proposals.
- **System Fingerprint**: Recorded for consistency checks, though not explicitly controlled.

Table E16: **LLM-as-a-Judge: System Prompt (Verbatim)**

You are a judge evaluating two multi-agent travel negotiation results (Plan A vs. Plan B).
Evaluate based on the QUALITY of the negotiation process, not just the final outcome.

Evaluation Criteria (5 Qualitative Metrics):

1. **Negotiation Rationality:**
- Is the process of reaching the final result logical?
- Were high-alpha (important) agents' opinions not ignored and reasonably reflected?
- Did the negotiation flow make sense given each agent's priorities?

2. **Preference Alignment:**
- How well does the final result align with each persona's initial constraints?
- Were important preferences (high-alpha items) preserved in the final outcome?

3. **Reason-Value Validity:**
- Does the stated "reason" actually justify the proposed "value"?
- Are the arguments logically sound and relevant to the constraint being discussed?

4. **Opinion Change Justification:**
- When a proposer/voter changes their position, is the reason for change clear and valid?
- When they maintain their position, is the justification convincing?
- Is it clear WHY they changed or stuck to their opinion?

5. **Fluency & Naturalness:**
- Does the conversation feel like a real group travel discussion, not robotic?
- Are the sentences natural and human-like?
- Is there appropriate back-and-forth dialogue?

Scoring Guide:
- Compare the CONVERSATION SAMPLES between Plan A and Plan B
- Look for logical reasoning, empathy, and natural language
- Prefer plans where agents clearly explain their thought process

Output Format:
For each of the 5 criteria, decide who wins (A or B), then give the final overall winner.
You must output structured results with NO reasoning.

Table E17: **LLM-as-a-Judge: User Prompt Template**

Here are two negotiation results:

```
<plan_A>
[Version]: {version_a}
[Final Constraints]: {constraints_a}
[Conversation Samples]:
{conversation_a}
</plan_A>

<plan_B>
[Version]: {version_b}
[Final Constraints]: {constraints_b}
[Conversation Samples]:
{conversation_b}
</plan_B>
```

F.2 Experimental Environment

The simulation framework was implemented in **Python 3.10**, utilizing standard libraries for HTTP requests and string processing.

- **API Cost**: The average cost per negotiation session (consisting of approx. 5 rounds among 3 agents) was approximately $0.02 USD.
- **Hardware**: As the framework primarily relies on API calls, no high-performance GPUs were required. Experiments were run on a standard local server (Ubuntu 22.04 LTS, CPU 8-core).

USE OF LARGE LANGUAGE MODELS

We used a large language model (LLM) as a supporting tool for improving writing clarity and for assisting with code drafting during early prototyping stages. All code, analyses, and experimental results were subsequently reviewed, validated, and finalized by the authors. The authors are solely responsible for any errors or omissions.