Measuring Fine-Grained Negotiation Tactics of Humans and LLMs in Diplomacy

Anonymous ACL submission

Abstract

The study of negotiation styles dates back to Aristotle's ethos-pathos-logos rhetoric. Prior efforts primarily studied the success of negotiation agents. Here, we shift the focus towards the styles of negotiation strategies. Our focus is the strategic dialogue board game Diplomacy, which affords rich natural language negotiation and measures of game success. We used LLM-as-a-judge to annotate a large human-human set of Diplomacy games for fine-grained negotiation tactics from a sociologically-grounded taxonomy. Using a combination of the It Takes Two and WebDiplomacy datasets, which consist of real human data, we demonstrate the reliability of our LLM-as-a-Judge framework and show strong correlations between negotiation features and game success in the Diplomacy setting. Building on this analysis, we investigate systematic differences between LLM and human negotiation strategies. Our results reveal a clear negotiation-style gap between LLM agents and successful human players. Finally, we show that fine-tuning on high-quality human data can reduce this style gap, helping steer LLM agents toward more human-like negotiation behaviors.

1 Introduction

Negotiation has long been studied as both a science and an art, dating back to Aristotle's three modes of rhetoric: Ethos appeals to credibility; Pathos appeals to emotions; and Logos appeals to logic (Kennedy, 1993). How an argument is presented can be as crucial as what is being said; the strategy a negotiator adopts can profoundly affect the outcome of a negotiation.

A growing body of work in NLP and AI has focused on developing agents with strong negotiation abilities. NLP systems have demonstrated impressive negotiation capabilities, including in the strategic negotiation board game Diplomacy (FAIR et al., 2022) as well as engaging in multi-issue bargaining (Lau et al., 2008; Lewicki et al., 2011; Lewis et al., 2017; He et al., 2018).

However, most evaluations of AI negotiation agents emphasize objective metrics like win rates, efficiency of the deal, or the balance of concessions (FAIR et al., 2022; Kwon et al., 2024; Bianchi et al., 2024; Fu et al., 2023). Less focus has been placed in the understanding of the tactics (i.e., rhetoric, tone) models employed in negotiation. The tactics negotiation agents use – cooperative or combative, persuasive or dismissive – affect receiver perception, and the agent's effectiveness and reception (Chawla et al., 2021, 2022; Mell et al., 2019; Kwon et al., 2024). Prior efforts to study negotiation style adopted ad-hoc definitions that are insufficiently grounded in past negotiation theory, making it difficult to compare negotiation tactics across studies or to link observed negotiation behaviors.

In this paper, we profile the distribution and impact of fine-grained negotiation tactics through a sociologically grounded framework, using Diplomacy as a testbed. We used two datasets of bilateral human-human dialogues: (1) the **It Takes Two** dataset, which contains Diplomacy games collected by Peskov and Cheng (2020) and annotated for negotiation tactics by Jaidka et al. (2023), and (2) a **WebDiplomacy** dataset taken from the large scale corpus of online Diplomacy games used by FAIR et al. (2022). Details of both datasets are presented in Appendix C. Unlike some past work that solely analyzes LLM-LLM negotiations within synthetic scenarios (Tang et al., 2025; Kwon et al., 2024; Bianchi et al., 2024), we use these naturally occuring datasets to ground our negotiation style analysis and development in human gameplay, before applying it to analyze LLMs. We study the following Research Questions:

- **RQ1:** How can we annotate negotiation tactics to study human negotiation behavior

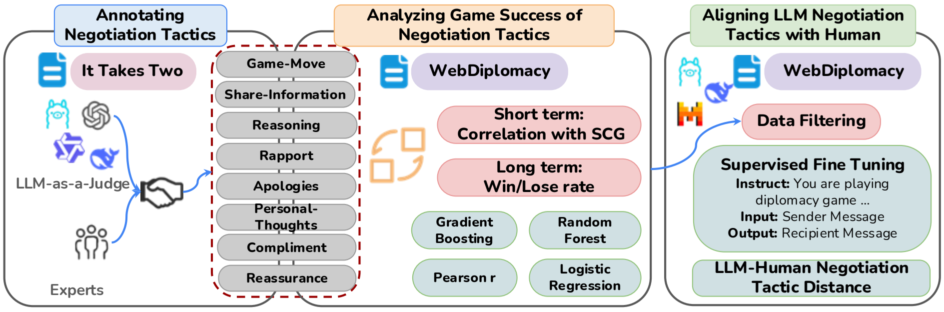

Figure 1: Methodology Overview: Our pipeline consists of three stages: **(1) Reliable tactic annotation.** We first annotate negotiation tactics with an LLM-as-a-Judge and validate its reliability on the It Takes Two dataset by computing agreement with expert annotators. **(2) Linking tactics to outcomes.** Using real human communications and game logs from WebDiplomacy, we study how annotated negotiation tactics relate to performance, analyzing short-term correlations and long-term win/loss outcomes. **(3) Aligning LLMs with humans.** We do supervised finetuning on filtered WebDiplomacy interactions to align LLM negotiation style with human tactics and quantify the LLM–human tactic distance.

at scale? We develop an LLM-as-a-judge pipeline for efficient and reliable annotations.

- **RQ2:** Do negotiation styles affect game success? We apply regression and predictive modeling to study how styles affect game success in the large-scale WebDiplomacy dataset.

- **RQ3:** What are the differences in negotiation styles between LLMs and Humans? We prompt LLMs with game contexts from the WebDiplomacy dataset and evaluated the negotiation style distribution in comparison to human messages.

- **RQ4:** Can we steer LLMs to use similar negotiation tactics as humans? We fine-tune LLMs with human data from WebDiplomacy dataset to match the negotiation tactics.

2 Related Work

Diplomacy Diplomacy is a strategic board game that requires complex negotiation to form alliances. Seven players aim to control a majority of 34 supply centers on a map of Europe by coordinating the movement of their military units. While Diplomacy is a zero-sum game, players must negotiate strategic coalitions to support their own plans or counteract the moves of other players. Bilateral negotiations are held in private and do not bind future moves, meaning that building long-term trust can be critical to game success. See Appendix A for a more detailed description of the game.

Diplomacy is often used to study human-to-human perceptions of trust, deception and persuasion, and perceptions of lies (Niculae et al., 2015; Peskov and Cheng, 2020; Ahuja et al., 2022; Ng et al., 2025; Wongkamjan et al., 2024, 2025). The game has also been an essential testbed for assessing LLM-powered strategic reasoning (Paquette et al., 2019; Gray et al., 2021; Bakhtin et al., 2022). These properties make Diplomacy a particularly well-suited and valid testbed for studying negotiation behaviors and tactics. Meta's CICERO system successfully combined an LLM-based dialogue system with strategic reasoning, competing successfully against human opponents (FAIR et al., 2022). However, many of these LLM-based works emphasize endgame outcomes, leaving negotiation dialogue dynamics relatively underexplored.

LLM as negotiators Research evaluating LLM negotiation capabilities spans diverse domains: games, finance, law, and business (Kwon et al., 2024; Bianchi et al., 2024; Fu et al., 2023; Noh and Chang, 2024). Kwon et al. (2024) systematically assessed LLM performance on 35 negotiation tasks, noting GPT-4's strength but its struggle with subjective judgment and strategic adaptability. Bianchi et al. (2024) presented NegotiationArena, revealing how LLMs develop strategic and irrational tactics in negotiation exchanges.

Stylistic linguistic features in dialog can reflect power and influence (Niculae et al., 2015), which suggests that agents that linguistically adapt can gain a social or persuasive edge. LLM and human negotiation behavior can be rather different (Wongkamjan et al., 2024), which leads onto investigations to shift LLM behavior for better alignment and authority. Prompt-based interventions

Rhetoric	Negotiation Tactic	Definition
Ethos	Game-Move	Plans, thoughts and goals about a Diplomacy move
Ethos	Share-Information	Messages about the history of or information gained about another player's move (except the speaker's and recipient's)
Logos	Reasoning	Speculative reasoning, justification of past or future moves
Pathos	Rapport	Build trust and mutual understanding between speaker and recipient
Pathos	Apologies	Expressions of regrets or remorse about past moves
Pathos	Personal-Thoughts	Messages that reflect the speaker's opinions or feelings
Pathos	Compliment	Positive messages about the recipient or recipient's moves
Pathos	Reassurance	Supportive messages about the recipient's game position

Table 1: Taxonomy of Negotiation Tactics and Definitions (adapted from Jaidka et al. (2023))

and fine-tuning can be effective to shift LLM behavior. Noh and Chang (2024) found that personality-driven prompts can shift LLM negotiation behavior from cooperative to adversarial without retraining. Reinforcement-learning-inspired methods, such as self-play with feedback, have also been shown to improve negotiation success (Lewis et al., 2017; Fu et al., 2023; Chen et al., 2023; Liao et al., 2024). Such works highlight both the potential and limitations of LLM negotiation capabilities, motivating our studies which use Diplomacy as a testbed and focus exclusively on negotiation tactics, examining how LLM agents employ fine-grained strategies and how these compare to human behavior. Building on these past works, we perform a large-scale analysis of negotiation strategies, enabled by LLM-as-a-judge, and examine the relationship between these strategies and success in the game.

3 Negotiation Tactics Analysis

We analyze Diplomacy dialogue using a taxonomy of fine-grained negotiation tactics adapted from Jaidka et al. (2023). (See Appendix A for a detailed description of the Diplomacy game.) This taxonomy is based on the Ethos-Pathos-Logos rhetoric, and breaks down negotiation into eight tactics, each serving a psychological and strategic function that contributes to negotiation effectiveness. The tactics, definitions are listed in Table 1, and their sociological groundings in Table 3. We then correlate the presence of each negotiation tactic with game success in human-human games. We also evaluate LLMs on their use of these strategies in self-play, and compare them to humans. Figure 1 illustrates our methodology.

Past work annotated messages with the negotiation strategies using Amazon Mechanical Turk workers (Jaidka et al., 2023). However, the nuanced nature of the task resulted in differing interpretations among the annotators, and therefore inconsistent crowd-sourced labels (Ng et al., 2025).

Therefore, we develop an LLM-as-a-judge pipeline as a scalable and reliable approach to annotate the messages. We prompted models to perform binary classifications (presence or absence of each tactic) in a single prompt on a subset of messages ($n = 128$) from the **It takes two** dataset (Peskov and Cheng, 2020; Jaidka et al., 2023). The models were: `LLaMA3.1-8B-Instruct` (Dubey et al., 2024), `Qwen-3-8B` (Team, 2025), and `r1-distilled-LLaMA-8B` (DeepSeek-AI, 2025). Details prompts are provided in Appendix F:

- **Baseline (Zero-shot)**: A direct prompt asking the model to judge each of the eight tactics without any instruction or examples.
- **Few-Shot**: Providing some positive example of each tactic from expert annotators. In total, eight examples were provided.
- **Instructions**: The original task description used for crowd workers from Jaidka et al. (2023), which included definitions and decision rules for each feature.
- **Instructions + Few-shot**: A hybrid prompt that included both the instruction template and the few-shot examples.

We compare the LLM annotations with expert annotations. Three expert annotators (authors of this paper) annotated a subset of randomly selected $n = 128$ dialogue messages, guided by the same instructions as the crowd-sourced annotators (see Figures 18 to 20). Because our annotation label distribution is notably imbalanced across categories, we report Gwet's AC1, a chance-corrected agreement statistic that is more robust than Fleiss-kappa agreement to prevalence and marginal asymmetry (More explanation is in Appendix E). Agreement among experts achieved substantial reliability

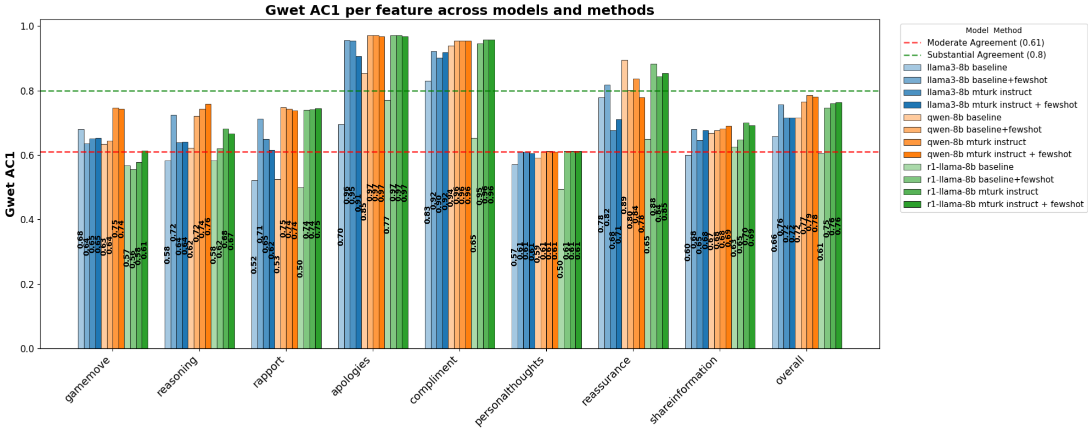

Figure 2: Gwet's AC1 agreement scores per negotiation tactic across models and prompting methods when compared against the expert gold standard. The dashed red line indicates the threshold for moderate agreement (AC1 $= 0.61$), while the dashed green line indicates substantial agreement (AC1 $= 0.8$).

(overall mean AC1 = 0.678), supporting the feasibility of a gold-standard label set. Comparing LLM annotations against the expert labels yields moderate agreement on average, whereas crowd-sourced annotators exhibit markedly lower agreement with experts (below the moderate threshold), indicating that LLM-as-a-judge is a more dependable path for scaling annotation than crowd labels alone.

Figure 2 reports AC1 scores between each LLM and the expert annotations, broken down by conversational feature:

- **Prompting scheme dominates size:** Across models, **baseline** prompts sit well below the substantial-agreement band. Adding **instruction** prompting consistently lifts AC1, and adding **few-shot** examples produces the largest gains—often pushing scores above the **moderate** (AC1 $= 0.61$) line and, for several features, into the **substantial** (AC1 $= 0.80$) range.
- **Qwen-8B is the most reliable judge:** With **instructions + few-shot**, QWEN-8B attains the highest mean AC1 across features, edging out LLAMA-3-8B and the distilled R1-LLAMA-3-8B. QWEN-8B also leads on a majority of individual features.
- **Feature-level difficulty varies:** Socially straightforward tactics such as APOLOGIES and COMPLIMENTS achieve high agreement. REASSURANCE typically lands around the substantial threshold. In contrast, PERSONAL-THOUGHTS remains challenging (near the moderate band), while RAPPORT, REASONING, GAME-MOVE, and SHARE-INFORMATION fall in the midrange.

4 Analyzing Negotiation Style's Effect on Game Success

In this section, we investigate whether negotiation tactics affect game success in the WebDiplomacy dataset (see more detailed description in Appendix C). We use QWEN3-8B to annotate all messages or the presence of each of the eight negotiation strategies. The labels were then aggregated at the phase level per player, yielding both a binary indicator and a count (frequency of occurrences) for each feature in each player-phase.

Game success was measured using two metrics: short-term success with Supply Center Gain (SCG) gaining and long-term success with final winning.

4.1 Short Term Success

We define the player's Supply Center Gain (SCG) as a measurement of success, using meta-data from WebDiplomacy. SCG is the net change of supply centers controlled at the end of each game year. The SCG per player per year is a continuous outcome variable that was positive if the player gained centers, negative if centers were lost, and zero if the number of centers remained unchanged.

Correlation Analysis We first examine simple correlations between negotiation strategies and SCGs at the phase level. Since the measurement of supply centers occurred every game year, we considered the collective sum of the presence of features for each year for each power. Figure 3 shows the correlation between each feature. We controlled for length, as we found that the number of sentences sent per player-phase was strongly

tactic	point_biserial_r	p_pb	spearman_r	p_sp	cohen_d	rank_biserial_r
1. Game-Move	0.236	$<$1e$-$6	0.362	$<$1e$-$6	0.278	-0.137
2. Reasoning	0.180	$<$1e$-$6	0.296	$<$1e$-$6	0.341	-0.196
3. Rapport	0.200	$<$1e$-$6	0.290	$<$1e$-$6	0.348	-0.203
4. Apologies	0.179	$<$1e$-$6	0.234	$<$1e$-$6	0.382	-0.227
5. Compliment	0.152	$<$1e$-$6	0.216	$<$1e$-$6	0.376	-0.221
6. Personal-Thoughts	0.127	$<$1e$-$6	0.171	$<$1e$-$6	0.364	-0.215
7. Reassurance	0.147	$<$1e$-$6	0.234	$<$1e$-$6	0.309	-0.188
8. Share-Information	0.182	$<$1e$-$6	0.293	$<$1e$-$6	0.301	-0.172

Table 2: Correlation and Effect Size between negotiation tactics and yearly SCG. All tactics show statistically significant positive correlations with SCG ($p < 1e{-}6$), supporting the hypothesis that both tactical reasoning and socio-emotional strategies contribute meaningfully to short-term success.

correlated with each negotiation tactic (Num Sentences and Num Tokens correlated by ≥ 0.83).

We computed the Pearson's r as a point-biserial correlation (Benesty et al., 2009) between negotiation tactics and SCGs. r quantifies the strength and direction of linear relationships between continuous features and outcomes, making it well-suited to analyze how the frequency of each tactic relates to SCGs. This associates whether players who used a given tactic during a phase tended to gain more supply centers at the end of the phase. As presented in Table 2, all eight stylistic dimensions show statistically significant positive Pearson correlations with yearly supply-center gain ($p < 10^{-6}$ after a Bonferroni correction).

The strongest linear association arose from the tactical GAME-MOVE ($r = .24$), demonstrating that tactical discussion of moves and strategies enhances negotiation outcomes by reducing uncertainty (Bazerman and Neale, 1993). The next strongest linear associations were interpersonal RAPPORT ($r = .20$), mirroring how rapport-building significantly improves negotiations from increased trust (Drolet and Morris, 2000), and analytical REASONING ($r = .18$), supporting how logical arguments are most effective in strategy games (Petty and Cacioppo, 1986). Social-politeness markers such as APOLOGIES, COMPLIMENT, and REASSURANCE still had positive associations, albeit with smaller effects ($.13 \leq r \leq .18$), which reflects how social behaviors reduces resistance and facilitate cooperation (Brown and Levinson, 1987). Information exchange (SHARE-INFORMATION, $r = .18$) sits mid-table, suggesting that while this strategy can improve outcomes, it also creates vulnerability in revealing the player's position (Galinsky and Mussweiler, 2001).

For robustness analysis, we extended the inquiry to a frequency-adjusted regression (see Appendix K). These correlational analyses (binary presence and frequency-adjusted) demonstrate that the taxonomy of negotiation tactics are correlated with short-term outcomes, highlighting the robustness of the taxonomy and the importance of fine-grained negotiation tactic analysis.

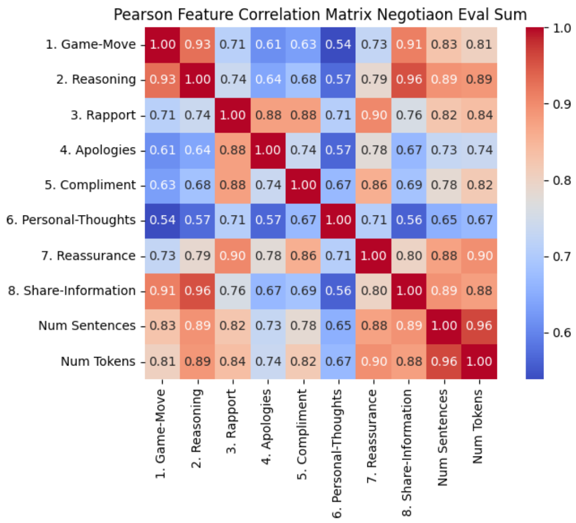

Figure 3: Correlation between annotated negotiation features and supply center gain.

4.2 Predictive Modeling

To move beyond univariate correlations and gain a more comprehensive understanding of how negotiation strategies relate to success in Diplomacy, we used predictive modeling analysis with machine learning (ML) methods. Predictive modeling analysis provides stronger information about the explanatory power of negotiation tactics. If negotiation tactics matter for performance, then a predictive model trained on the negotiation features should be able to accurately forecast player success. For robustness, we also used an Ordinary Least Squares regression to validate the predictive nature of the negotiation tactics (more detailed explanations are in Appendix L). The OLS model allows us to uncover the relative importance among tactics in their contribution to the player's game success.

Using ML prediction techniques, we evaluate the power of negotiation features for short-term success. We treated SCG as the prediction target for a suite of supervised machine learning models: Logistic Regression, Random Forest, and Gradient Boosting. Model inputs were either per-phase negotiation feature counts or their standardized aggregated frequencies across the game. Model training and hyperparameter optimization were performed via cross-validation, with evaluation on a held-out test set using metrics of accuracy, F1-score, and ROC-AUC. All three tested models hovered around **61**% accuracy and **65**% ROC-AUC, substantially above the majority baseline (50%) (see Table 9). We analyzed feature importance scores to interpret model decisions. Our integrated regression and prediction framework allows us to identify not only which negotiation behaviors correlate with but also are predictive of player success.

The Gradient Boosted classifier was the most accurate. Its top-20 feature importances (see Figure 7) closely echoed the OLS findings:

- **Game-Move** dominated predictability (18.7%), reinforcing its role as the single best indicator of positive SCG. The dominance of Game-Move aligns with costly signaling theory because these communications are the most costly form of signaling and difficult to fake – sharing specific tactical information requires deep analysis and carries strategic risks, making the signals reliable indicators of genuine cooperation (Przepiorka and Berger, 2017).
- **Rapport** (11.6%) and **Reassurance** (5.6%) followed, showing that well-timed socio-emotional cues acts as social exchanges (Blau, 2017), which therefore add predictive value.
- Length effects appear both directly (num_tokens, 3.7%) and via interactions (e.g. **Game-Move** $\times$ **Share-Information**), underlining how longer, more detailed messages serve as heuristic indicators of sender effort and seriousness and results in deeper evaluation of proposals (Petty and Cacioppo, 1986).

4.3 Long Term Success

We represent long-term success by the eventual game outcome (win or loss). We compared the breakdown of negotiation tactics between the eventual winners and losers, by comparing the average frequency each negotiation strategy was used by winning players in comparison to the losers. For each game, we calculate the average rate of each negotiation style per phase for the winner and a randomly sampled loser. To account for differences in total message volume, we perform normalization per year. We then compare these average feature frequencies between the winners and losers. Next, to isolate the effect of negotiation tactics, regardless of the strength of the player's position (reflected by their supply center count), we condition the frequency on the supply center counts at each phase. This controlled for the cumulative advantages and opportunities that players with more centers have, and allowed better discernment on whether winners exhibited distinct negotiation. This long-term analysis focuses on the differences where communication behavior correlates with ultimate success, offering insight into the characteristics of winning sets of negotiation tactics.

Figure 4 shows the changes in the overall prevalence as a player's supply-center count grows, which implicitly reflects the progression from the early to the late stages of the game. This figure aggregates all eight negotiation strategies into a single curve, highlighting the positional strengths of each strategy as a global trend. Notably, the observed trend underscores the importance of consistently employing negotiation tactics: throughout every phase of the games, winners exhibit a higher frequency of negotiation tactics compared to losers. This persistent difference demonstrates that winners' eventual successes are tied to sustained negotiation activity. The corresponding statistical significance tests are provided in Appendix M.

5 LLM and Human Negotiation Tactics

5.1 LLM and Humans have Different Negotiation Tactics

Our preceding analysis on the human WebDiplomacy corpus demonstrates that game success is indeed associated with the negotiation tactics. Building on this foundation, one core aim of this work is to probe the capabilities of LLMs as negotiators within the Diplomacy setting: How closely do LLM negotiators approximate human negotiation

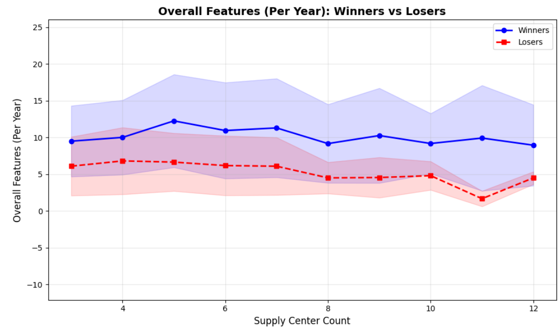

Figure 4: Number of negotiation tactics per year across supply center count

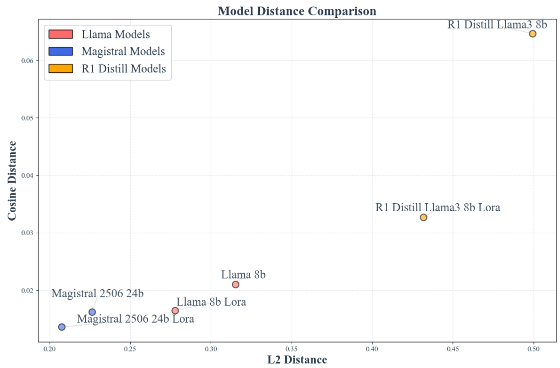

Figure 5: Model-Human Distance

tactics, and can their negotiation style be steered to better align with high-quality human behavior? We systematically examine whether LLMs exhibit substantive gaps compared to humans in the use of these tactics, and whether alignment can bridge this gap. To this end, we utilize LLMs to participate in negotiations and assess their negotiation proficiency. Leveraging an adapted version of the SOTOPIA (Zhou et al., 2024) evaluation framework, we conducted one-on-one negotiation experiments between LLMs. Each experiment focuses on a single one-on-one exchange, isolating each model's style under realistic conversational pressure.

From the WebDiplomacy human-gameplay corpus, we sample 1,000 negotiation phases. Each phase comprises all messages exchanged immediately before players committed their orders, and the subsampled phases span diverse points in the game timeline (early, mid, or late game). We assign an LLM-negotiator agent the role of one player and prompted it to craft a reply to its partner's last message, negotiating game orders based on the current game board. The prompt instructs the model to balance tactical short-term gains (e.g., securing support for an attack) with relationship-building long-term goals (e.g., cultivating alliances), thereby mirroring the dual-goal orientation of skilled human players (Jaidka et al., 2023). LLM-negotiators were constructed with the following models: Llama3.1-8B-Instruct (Dubey et al., 2024), R1-distilled-LLama3-8B (DeepSeek-AI, 2025), and Magistral-2506-24B (Rastogi et al., 2025). The full prompt is in Appendix F.

We score each message produced by the LLM-negotiator with our LLM-as-a-judge pipeline (see Section 3). For each of the negotiation dimensions present, we recorded (i) the raw count of occurrences and (ii) a length-normalized rate of negotiation dimensions per sentence. These features are then aggregated into stylistic profiles for each model, enabling insight into whether the LLM style favors strategic maneuvers (e.g. game-move and information share) or social tactics (e.g. rapport and apologies).

To quantify the difference in negotiation techniques between humans and LLMs, we define $\mathcal{P}$ as the set of phases that contained both human and model utterances. For a phase $p \in \mathcal{P}$ and speaker s (human or LLM), the LLM-judge with mturk instruct emits a binary value $\mathbf{f}_{s,p} \in \{0,1\}^8$, which is normalized by sentence count, $\tilde{\mathbf{f}}_{s,p} = \mathbf{f}_{s,p}/\text{sent\_cnt}(s,p)$. Averaging over phases yields an 8-D *mean style vector* (see Table 1) per speaker:

$$\mathbf{m}_k = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \tilde{\mathbf{f}}_{k,p}, \qquad \mathbf{h} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \tilde{\mathbf{f}}_{\text{hum},p}. \tag{1}$$

We quantify LLM–human divergence with three distances, where lower values mean closer LLM-human alignment. We drew $1,000$ bootstrap samples over $\mathcal{P}$ and computed the metrics (see more results in Appendix I).

$$\text{L2}_k = \|\mathbf{m}_k - \mathbf{h}\|_2 \tag{2}$$

$$\text{CosDist}_k = 1 - \frac{\mathbf{m}_k^\top \mathbf{h}}{\|\mathbf{m}_k\|_2 \, \|\mathbf{h}\|_2} \tag{3}$$

Figures 5 and 10 reveals substantial gaps between human negotiation tactics and current LLMs. Magistral-2506-24B achieved the closest LLM-human distance. In contrast, the R1-distilled-Llama3-8B model showed a marked divergence from human reference. This suggests that the math-and-coding reasoning focus of this model might have a limited ability to mimic human negotiation tactics in our domain, underscoring the necessity of aligning reasoning models for social abilities (e.g., negotiation) and broader interaction competence.

A feature-level analysis (Figure 11) further demonstrates that these gaps are not uniform across negotiation tactics. Specifically, when comparing LLMs, the R1-distilled-Llama3-8B model exhibits the largest deviations from human behavior on key tactical features (e.g. GAME-MOVE and SHARE-INFORMATION), underscoring a pronounced deficit in emulating human-like strategic maneuvers. Meanwhile, the R1-distilled and Llama3.1-8B models display more modest, yet persistent, discrepancies on social-emotional and interpersonal features. All models, however, consistently underperform on subtle behaviors like PERSONAL-THOUGHTS and REASSURANCE, indicating a general limitation in capturing the nuanced, relational aspects of human negotiation.

Taken together, relying exclusively on reasoning-oriented distillation not only fails to align models with human style but may in fact exacerbate this misalignment. This limitation becomes especially salient in social reasoning tasks. These insights motivate the necessity of incorporating social reasoning and human-grounded data in future alignment efforts, which therefore motivates our subsequent style-alignment experiments. We show some examples in the Appendix I.

5.2 Aligning LLM-Negotiation Tactics with Human

In this section, we evaluate whether fine-tuning LLMs on human negotiation data produces negotiation tactics closer to human tactics.

The regression analysis performed in Section 4.1 shows that higher-order social tactics were predictors of subsequent growth. The full negotiation style distribution is shown in Figure 8. We focus on successful human dialogue, as measured by phases with ΔSC increasing. Filtering the WebDiplomacy corpus for such phases yields 18,420 dialogue turns. We used these turns as a supervision corpus of effective human negotiation tactics. We use Supervised Fine-Tuning (SFT) on each model to steer the LLMs towards a more human-like distribution of negotiation strategies. More details are in Appendix F.

Our quantitative analysis (see Figure 5, Figure 10 and Figure 12) demonstrates that instruction fine-tuning on the human-grounded negotiation corpus narrowed the gap between LLM-generated and human negotiation tactics. Specifically, Figures 13 and 14 reveals that across most social negotiation features, all fine-tuned models exhibited reduced LLM-Human L2 distances. There were particularly strong convergence on social features of RAPPORT, COMPLIMENT, and APOLOGIES. However, features like PERSONAL-THOUGHTS and REASSURANCE remained more challenging, showing persistent LLM-human gaps.

The difference plots further show that LoRA-based SFT produced the most pronounced distance reductions for the most different R1-Distill-Llama3-8B model, indicating substantial stylistic shift toward (Cosine=2.2%) human-like negotiation. For models that already exhibited strong human alignment, such as Mistral-2506-24B, LoRA fine-tuning still yielded additional improvements (Cosine = 0.7%), further aligning to human distributions. These results confirm that SFT with LoRA effectively enhances the alignment of LLM and human negotiation tactics.

Overall, fine-tuned models acquired more human-aligned behaviors that led to consistent reductions in differences between LLM and humans across all eight negotiation tactics (see Appendix H, Table 7 and Figure 10). This convergence helps to validate the eight negotiation tactics as reliable proxies for human-grounded negotiation tactics and their utility as measurement tools and optimization targets, reinforcing their value as meaningful descriptors and effective behavioral targets.

6 Conclusion

Our work studies the correlations between fine-grained negotiation tactics and success in Diplomacy — a rich and representative negotiation setting. We developed a reliable LLM-as-a-judge pipeline and a fine-grained taxonomy grounded in a classical ethos–pathos–logos framework, showing that negotiation tactics are predictive of both short-term turn-to-turn success and long-term end-game success. The most predictive features are: GAME MOVE, socio-emotional cues (RAPPORT and REASSURANCE). We evaluate the abilities of LLMs to continue negotiations started by people, and find that these models use negotiation tactics substantially different than the ones used by successful humans. However, fine-tuning can shift the tactics LLMs used to align better with the tactics humans use. Our results lay a foundation for evaluating fine-grained negotiation strategies, allowing measuring and improving the ability of LLM-agents to use negotiation tactics in a human-like way.

7 Limitations

Our analysis focuses primarily on (1) negotiation tactics used in human-human games and (2) the alignment between human and LLM behaviors. We demonstrate an association between negotiation tactics and game success in human-human games, and that current LLMs do not match human negotiation tactics. As LLM agents become stronger, future work should integrate end-to-end evaluations on game success, placing aligned LLM agents into live game environments to verify that improved tactic alignment also translates into concrete strategic gains for LLM-based agents.

Our approach prioritizes learning and aligning with human negotiation tactics, but does not systematically filter or analyze for undesirable content such as social biases, toxicity, or hate speech that may be present in human data and potentially learned by LLMs during fine-tuning. As a result, the models may inherit and propagate problematic patterns observed in the training corpus. Further research should include dedicated analyses for bias and toxicity, and the development of mitigation strategies to ensure that aligned negotiation agents remain ethical and fair in their interactions.

References

Hervé Abdi. 2007. The bonferroni and sidak corrections for multiple comparisons. In N. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 103–107. SAGE.

Hansin Ahuja, Lynnette Hui Xian Ng, and Kokil Jaidka. 2022. Using graph-aware reinforcement learning to identify winning strategies in diplomacy games (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12899–12900.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. 2022. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. *Preprint*, arXiv:2210.05492.

Max H Bazerman and Margaret Ann Neale. 1993. *Negotiating rationally*. Simon and Schuster.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

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.

Peter Blau. 2017. *Exchange and power in social life*. Routledge.

Anton Borg and Martin Boldt. 2020. Using vader sentiment and svm for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746.

Jeanne Brett and Leigh Thompson. 2016. Negotiation. *Organizational Behavior and Human Decision Processes*, 136:68–79. Celebrating Fifty Years of Organizational Behavior and Decision Making Research (1966-2016).

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

John K Butler Jr. 1999. Trust expectations, information sharing, climate of trust, and negotiation effectiveness and efficiency. *Group & Organization Management*, 24(2):217–238.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale Lucas, and Jonathan Gratch. 2021. Towards emotion-aware agents for negotiation dialogues. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Kushal Chawla, Gale M Lucas, Jonathan May, and Jonathan Gratch. 2022. Opponent modeling in negotiation dialogues by related data adaptation. *arXiv preprint arXiv:2205.00344*.

Siqi Chen, Jianing Zhao, Gerhard Weiss, Ran Su, and Kaiyou Lei. 2023. An effective negotiating agent framework based on deep offline reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 324–335. PMLR.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aimee L Drolet and Michael W Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

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.

Adam D Galinsky and Thomas Mussweiler. 2001. First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of personality and social psychology*, 81(4):657.

Jonathan Gray, Adam Lerer, Anton Bakhtin, and Noam Brown. 2021. Human-level performance in no-press diplomacy via equilibrium search. *Preprint*, arXiv:2010.02923.

Kilem Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical methods for inter-rater reliability assessment*, 1(6):1–6.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. *arXiv preprint arXiv:1808.09637*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Kokil Jaidka, Hansin Ahuja, and Lynnette Ng. 2023. It takes two to negotiate: Modeling social exchange in online multiplayer games. *Preprint*, arXiv:2311.08666.

George A Kennedy. 1993. Aristotle" on rhetoric": a theory of civic discourse. *Philosophy and Rhetoric*, 26(4).

Özgür Kıbrıs. 2010. Cooperative game theory approaches to negotiation. In *Handbook of group decision and negotiation*, pages 151–166. Springer.

Kihwan Kim, Nicole L Cundiff, and Suk Bong Choi. 2015. Emotional intelligence and negotiation outcomes: mediating effects of rapport, negotiation strategy, and judgment accuracy. *Group Decision and Negotiation*, 24(3):477–493.

Deuksin Kwon, Emily Weiss, Tara Kulshrestha, Kushal Chawla, Gale M. Lucas, and Jonathan Gratch. 2024. Are llms effective negotiators? systematic evaluation of the multifaceted capabilities of llms in negotiation dialogues. *Preprint*, arXiv:2402.13550.

Raymond YK Lau, On Wong, Yuefeng Li, and Louis CK Ma. 2008. Mining trading partners' preferences for efficient multi-issue bargaining in e-business. *Journal of Management Information Systems*, 25(1):79–104.

Roy J Lewicki and Beth Polin. 2013a. The role of trust in negotiation processes. In *Handbook of advances in trust research*, pages 29–54. Edward Elgar Publishing.

Roy J Lewicki and Beth Polin. 2013b. Trust and negotiation. In *Handbook of research on negotiation*, pages 161–190. Edward Elgar Publishing.

Roy J Lewicki, David M Saunders, John W Minton, J Roy, and Negotiation Lewicki. 2011. *Essentials of negotiation*. McGraw-Hill/Irwin Boston, MA.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *Preprint*, arXiv:1706.05125.

Austen Liao, Nicholas Tomlin, and Dan Klein. 2024. Efficacy of language model self-play in non-zero-sum games. *arXiv preprint arXiv:2406.18872*.

J. Scott Long and Laurie H. Ervin. 2000. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

James G MacKinnon and Halbert White. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Johnathan Mell, Jonathan Gratch, Reyhan Aydoğan, Tim Baarslag, and Catholijn M Jonker. 2019. The likeability-success tradeoff: Results of the 2 nd annual human-agent automated negotiating agents competition. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 1–7. IEEE.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Lynnette Hui Xian Ng, Kokil Jaidka, Kaiyuan Tay, Hansin Ahuja, and Niyati Chhaya. 2025. Improving user behavior prediction: Leveraging annotator metadata in supervised machine learning models. *arXiv preprint arXiv:2503.21000*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:1506.04744*.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *Preprint*, arXiv:2405.05248.

Philip Paquette, Yuchen Lu, Steven Bocco, Max O. Smith, Satya Ortiz-Gagne, Jonathan K. Kummerfeld, Satinder Singh, Joelle Pineau, and Aaron Courville. 2019. No press diplomacy: Modeling multi-agent gameplay. *Preprint*, arXiv:1909.02128.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Advances in experimental social psychology*, volume 19, pages 123–205. Elsevier.

Wojtek Przepiorka and Joël Berger. 2017. Signaling theory evolving: Signals and signs of trustworthiness in social exchange. *Social dilemmas, institutions, and the evolution of cooperation*, pages 373–392.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, and 1 others. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Wenjie Tang, Yuan Zhou, Erqiang Xu, Keyan Cheng, Minne Li, and Liquan Xiao. 2025. Dsgbench: A diverse strategic game benchmark for evaluating llm-based agents in complex decision-making environments. *Preprint*, arXiv:2503.06047.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon M Stewart, Jonathan K Kummerfeld, Denis Peskoff, and Jordan Lee Boyd-Graber. 2024. More victories, less cooperation: Assessing cicero's diplomacy play. *arXiv preprint arXiv:2406.04643*.

Wichayaporn Wongkamjan, Yanze Wang, Feng Gu, Denis Peskoff, Jonathan K Kummerfeld, Jonathan May, and Jordan Lee Boyd-Graber. 2025. Should i trust you? detecting deception in negotiations using counterfactual rl. *arXiv preprint arXiv:2502.12436*.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

A Diplomacy Game Setting

Diplomacy is a strategic multi-agent negotiation game originally designed for seven players, each representing a major European power (Austria-Hungary, England, France, Germany, Italy, Russia, and Turkey). The game is played in discrete rounds corresponding to historical seasons (Spring and Fall), with each round consisting of two key phases: a negotiation phase and an order submission phase. During the negotiation phase, players may exchange private messages with any other player, formulating alliances, making promises, or attempting deception. The number of exchanged messages varies depending on the level of engagement, but in experimental settings, a typical round yields dozens of bilateral messages per player. After negotiations, all players simultaneously submit their movement orders for military units, which are then resolved according to deterministic rules of conflict resolution. A single round (negotiation plus order resolution) usually lasts from several minutes to an hour in controlled experimental contexts, though in traditional play by mail or online platforms, a round can span 24 to 72 hours.

Victory conditions in Diplomacy are defined by territorial control. The game board consists of 75 provinces, of which 34 contain supply centers. Each power begins with three or four supply centers, and control over a center determines the number of units a player may sustain. Players gain or lose units depending on the number of centers they control after each Fall phase. The ultimate objective is to capture at least 18 supply centers, which constitutes an outright win. Alternatively, when no single player can achieve this threshold, the game may end in a draw among the surviving players. This combination of simultaneous action resolution, unmediated negotiation, and long-term strategic planning makes Diplomacy a canonical testbed for studying cooperation, competition, alliance formation, and deception in multi-agent interaction settings.

B Definitions and Sociological Grounding for Negotiation Tactics

Table 3 presents the eight negotiation tactics used in our taxonomy, their definitions (which are adapted from (Jaidka et al., 2023)), and the sociological grounding of each tactic.

C Datasets Information

This study used two datasets: (1) It Takes Two and (2) WebDiplomacy. Here are the details of these two datasets.

The It Takes Two dataset was originally collected by (Peskov and Cheng, 2020), which contains messages passed during the Diplomacy game from recruited human players. This dataset was further processed by (Jaidka et al., 2023) filtered for meaningful messages that contained more than five words and annotated for fine-grained negotiation tactics. This dataset contains 11,366 messages from 10 games.

The WebDiplomacy dataset is licensed from the server administrator of the WebDiplomacy platform (`https://www.webdiplomacy.net`), an online platform to play the Diplomacy game. This admin first filtered the WebDiplomacy games for games with messages. Next he applied a player filter. He selected for quality players: players with more than 5 games joined, an ELO rating over 105, points 120, reliability over 70, had more than one game won, and more than 5% game-win ratio. Then, he used the set of filtered players to select regular-press games that had more than 3 of these quality players in the game. From those games, the admin excluded those with No Moves Received (NMRs), and sampled 4000 games. This dataset contains 4000 games with the following information (see Table 4)

D LLM-as-a-Judge Template

In the LLM-as-a-Judge setup, the baseline (zero-shot) instruction template is presented in Table 5, while the few-shot variant, which gives exampless under each question, is shown in Table 6. The MTurk instruction (without giving the examples) and the MTurk-with-examples variant are illustrated in Figures 18 to 20.

E Detailed Discussion of LLM-as-a-Judge Agreement with Human

Why Gwet's AC1 (instead of Fleiss' κ). Our annotation labels are highly imbalanced across categories (see Figure 15), with several tactics having very low "True" prevalence (e.g., *apologies*, *compliment*, *personal thoughts*). In such settings, Fleiss' κ is known to suffer from the "κ paradox": even when raters agree on most items, κ can be deflated toward low values when the marginal distributions are skewed or when there is systematic class imbalance (Feinstein and Cicchetti, 1990; Byrt et al., 1993; Hallgren, 2012). Intuitively, κ's chance-agreement term P_e is computed directly from the observed marginals; under high or low prevalence this inflates the expected agreement and depresses the coefficient, producing misleadingly "low" reliability.

Gwet's AC1 addresses this issue by using a more stable estimator of chance agreement that corrects the bias induced by extreme or unequal marginals (Gwet, 2002, 2008). AC1 preserves the same interpretability as κ—1 indicates perfect agreement and 0 indicates chance-level—but its chance-agreement component is far less sensitive to prevalence, yielding reliability estimates that better reflect actual rater concordance under class imbalance. Comparative studies consistently find that AC1 remains robust where κ becomes paradoxically small in unbalanced, binary, or sparse multi-category settings (Wongpakaran et al., 2013; Gwet, 2014). Given the strong skew evident in our data, we therefore report Gwet's AC1 as our primary agreement coefficient and include Fleiss' κ only for completeness. This choice avoids underestimating reliability due to prevalence effects and aligns with best practices for imbalanced annotation tasks.

Rhetoric	Negotiation Tactic	Definition	Sociological Grounding
Ethos	Game-Move	Messages related to plans, thoughts and goals about a Diplomacy move	Game theory's emphasis that decisions of cooperate or compete are based on analysis of possible outcomes (Kıbrıs, 2010)
Ethos	Share-Information	Messages about the history of or information gained about another player's move (except the speaker's and recipient's)	Quantity & quality of information shared is typically associated with negotiation effectiveness (Butler Jr, 1999)
Logos	Reasoning	Speculative reasoning, justification of past or future moves	Receivers are more likely to agree with speakers who provide strong factual evidence and rational arguments (Brett and Thompson, 2016)
Pathos	Rapport	Messages that build trust and mutual understanding between speaker and received	Building rapport builds trust (Kim et al., 2015), and higher trust increases negotiation success (Lewicki and Polin, 2013a)
Pathos	Apologies	Expressions of regrets or remorse about past moves	Repairs both competence-based and integrity-based trust violations (Lewicki and Polin, 2013b)
Pathos	Personal-Thoughts	Messages that reflect the speaker's inner reflections, opinions or feelings	Build trust by demonstrating vulnerability with self-disclosure (Altman and Taylor, 1973)
Pathos	Compliment	Positive messages about the recipient or recipient's moves	Emotional regulation strategy to enhance trust (Kim et al., 2015)
Pathos	Reassurance	Supportive messages to restore confidence in recipient's game position	Emotional regulation strategy to enhance trust (Kim et al., 2015)

Table 3: Taxonomy of Negotiation Tactics, Definitions (adapted from Jaidka et al. (2023)), and sociological grounding

Empirical evidence in our annotations. The stacked counts in Figure 15 make the prevalence skew explicit, with "False" dominating most categories. Despite this imbalance, the human–human confusion matrices (see Figure 17) are strongly diagonal, indicating high observed agreement. Consistent with the literature, AC1 yields substantively higher and crucially more faithful estimates of reliability than Fleiss' κ (see Figure 16) in these categories, reflecting that annotators largely agree even when positives are rare.

F Prompt Templates for LLM Negotiators

We condition the model on the current phase, the dyadic dialogue context, the most recent executed orders, and a compact snapshot of the board state (centers and units), then assign the model a single speaking role for the turn. This follows the CICERO dialogue-agent design that situates language generation in the game state and recent conversation, while instructing the agent to advance plans through cooperative negotiation. (FAIR et al., 2022)

The following is the full prompt template used for the LLM acting as negotiator:

LLM Negotiator Prompt Template

SYSTEM: You are playing the diplomacy game. You will negotiate with the other player so that it plays moves beneficial to your board position, either this turn or in future turns.
You are in Phase: `{PHASE_NAME}`
The dialogue are between the two countries: `{COUNTRY1}` **and** `{COUNTRY2}`
The previous turn dialogue history is:
`{DIALOGUE_HISTORY}`
The previous order history is:
`{ORDER_HISTORY}`
This is the information of the current game state:
Centers: `{CENTER_INFO}`
Units: `{UNIT_INFO}`
You are playing as **`{COUNTRY1}`**. You are playing the diplomacy game, you will negotiate with the other player so that it will play moves that are beneficial to your board position, either this turn or in future turns.

G Experiment details

LoRA Fine-tuning We performed alignment training using the Supervised Fine-Tuning (SFT) methods on LLaMA-3.1-8B-Instruct, Magistral-2506-24B, and R1-distilled-LLama3-8B. Both training approaches utilized the Low-Rank Adaptation (LoRA) technique (Hu et al., 2021), which enabled efficient fine-tuning of the large language model by adapting a subset of its parameters. The experiments were conducted using 4 NVIDIA A6000 GPUs, with each GPU processing a batch size of 4.

Element	Description
id	Unique identifier for the game
map	Game map type (e.g., standard)
rules	List of rules used in the game
phases	List of all game phases; each phase contains:
name	Name of the phase (e.g., S1901M)
state:	Game state for the phase, including:
timestamp	Timestamp for the phase
zobrist_hash	Hash representing the board state
note	Miscellaneous notes on the phase
name	Name of the phase (redundant)
units:	Current unit positions for each power
<POWER>	List of units for each power (e.g., ['A BUD', ...])
retreats:	Retreat status for each power
<POWER>	Retreat information for each power
centers:	Controlled supply centers for each power
<POWER>	List of supply centers for each power
homes:	Home centers for each power
<POWER>	List of home centers for each power
influence:	Regions influenced by each power
<POWER>	List of influenced regions for each power
civil_disorder:	Civil disorder status for each power
<POWER>	0 (normal) or 1 (civil disorder)
builds:	Build/disband information for each power
<POWER>:	Details for each power
count	Number of builds/disbands for each power
homes	Possible build locations for each power
game_id	Game ID (redundant)
map	Map type (redundant)
rules	List of rules (redundant)
orders:	Player orders for the phase
<POWER>	List of orders submitted by each power
results:	Adjudication results for each unit/location
<UNIT/LOCATION>	Result list for the specified unit or location
messages:	List of all messages for the phase; each message contains:
sender	Sending player (power/country)
recipient	Recipient (power/country or GLOBAL for broadcast)
time_sent	Time the message was sent
phase	Phase during which the message was sent
message	Content of the message

Table 4: Structure of the WebDiplomacy dataset. Each game consists of multiple phases, with each phase recording the full board state, player orders, adjudication results, and negotiation messages.

For LoRA, we applied the technique across all layers of the model for SFT. The training configuration included a learning rate of 1.0×10^{-5}, regulated by a cosine scheduler, a warm-up phase consisting of 100 steps, and a gradient accumulation over 8 steps. We didn't limit training to three epochs with a maximum sequence length. Each training required approximately 20-24 hours to complete. To optimize computational resources, we used mixed-precision training with bfloat16. Both datasets were preprocessed using each model family's template and split into training and validation sets, with 10% of the data reserved for validation to monitor performance.

The training prompt for SFT follows the template below:

- **Instruction**: You are playing diplomacy game, you will negotiate with the other player so that it will play moves that are beneficial to your board position, either this turn or in future turns.

- **Input (sender messages)**: England has told me that he will support his army into Belgium. I am happy to be allies with you against him, but I'd like Sweden. It seems to our mutual advantage for you to cut his support in the North Sea and attempt to bounce Belgium.

- **Output (recipient messages)**: I like the DMZ, but we'll have to see about Sweden; it depends on the actions of England and France, sorry.

H Examples of LLM Negotiators in Different Styles

The picked examples in Table 7 illustrate how LLMs base model and their LoRA-tuned variants, occupy different negotiation tactics space. We annotate each utterance with eight features that span task-oriented planning (Game Move, Reasoning, Share Information) and socio-emotional maintenance (Rapport, Apologies, Compliment, Reassurance, Personal Thoughts). These qualitative snapshots complement our quantitative analyses by showing how specific stylistic cues surface in model outputs.

I Examples of LLM Negotiation Different Style Different to Human

To complement our quantitative analysis of LLM-human style divergence (Figures 5 and 10), we present concrete examples and qualitative comparisons illustrating the nature of these differences. Tables 12 to 14, juxtapose negotiation utterances generated by different LLM models (with and without LoRA alignment) against randomly sampled human messages, revealing both the most and least human-like traits expressed by each model family.

Deficits in Rapport and Socio-Emotional Expression. A clear and recurring shortfall across LLM outputs is their limited use of rapport-building language and socio-emotional cues. While LoRA-aligned variants show some improvement (see Figure 13), they typically rely on formulaic affirmations ("Let us continue to work together") and seldom exhibit the warmth, sarcasm, humor, or candid vulnerability that characterize genuine human negotiation. By contrast, human messages display a richer repertoire of trust-building, apology, teasing, and even playful antagonism (e.g., "lol, as I said in the beginning, I expected nothing from you," or "as long as they die, I'm content"). These are almost entirely absent from LLM outputs, underscoring a persistent gap in socio-emotional intelligence.

Strategic Depth versus Flexibility. Although advanced LLMs can produce complex strategic proposals and multi-turn coordination, their communication often lacks the adaptive flexibility, indirect persuasion, and negotiation context sensitivity observed in human exchanges. Human players frequently hedge, revisit old agreements, or express uncertainty and evolving intent, as in "This is however, only to my benefit for this turn, so if you have another option, then please use it," or "I guarantee I'll check before tomorrow night." LLMs, on the other hand, remain predominantly assertive and deterministic in their utterances.

Effect of LoRA Alignment. LoRA alignment does lead to improvements in some dimensions—models generate more detailed, cooperative, and contextually relevant proposals, and their language becomes marginally warmer and more partnership-oriented (Tables 12 to 14). Nevertheless, their repertoire of negotiation tactics remains constrained, and they continue to underperform in mimicking the informal, often idiosyncratic, tone of human negotiation.

Taken together, our qualitative analysis reveals that while LLMs, especially after targeted alignment, approximate human-like negotiation in tactical content, they systematically underrepresent rapport, flexibility, and the socio-emotional expressiveness intrinsic to human negotiation. These findings underscore the value of our multi-faceted evaluation framework and motivate future alignment efforts to move beyond purely strategic optimization, incorporating richer models of social reasoning and human communicative norms.

J LLM Negotiators Style After Alignment to Human

Tables 7 and 12 to 14 showcase how alignment via LoRA reshapes the task–relationship balance of model utterances. Across models, we observe a consistent coupling of *task-oriented content* (Game Move, Reasoning, Share Information) with *socio-emotional cues* (Rapport, Apologies, Compliment, Reassurance, Personal Thoughts), though the magnitude of this shift depends strongly on the base model's starting point.

R1-Distill-Llama3-8B. Pre-alignment, R1-Distill tends to rely on affiliative language—affirming alliances and expressing confidence—while often avoiding concrete orders (Table 12, top). After alignment, it introduces explicit multi-step plans and commitments (e.g., coordinating on Moscow/St. Petersburg, sequencing supports), while retaining warm, face-saving phrasing (Table 12, bottom). This yields a clearer coupling between rapport (Rap., Reass., Comp.) and executable proposals (GM, SI), although not uniformly across all turns—consistent with Table 7, where some LoRA utterances still foreground politeness over concrete orders.

Llama3-8B-Instruct. The base model produces pragmatic but locally scoped suggestions and questions about board state (Table 13, top). Post-alignment, its messages lengthen and become more *jointly intentional*: they integrate contingency planning (who cuts which support, how to sequence entries) with mild relational softeners (greetings, perspective-taking), thereby tightening the link between GM/Reasoning and Rapport (Table 13, bottom).

Magistral-2506-24B. Magistral's base style is already plan-centric and cautious, with clear proposals and deferred commitments when information is incomplete (Table 14, top). Alignment largely *stabilizes and sharpens* this profile: LoRA utterances make order finalization and role assignment more explicit (who secures which center, who supports whom), while adding only modest socio-emotional framing (Table 14, bottom). The stylistic rotation is therefore smaller in amplitude than for R1-Distill, reflecting a strong tactical prior.

Takeaways. Qualitatively, alignment (i) increases *commitment language* and multi-step coordination, (ii) reduces hedging by pairing proposals with concrete next actions, and (iii) injects prosocial markers most where the base model is under-socialized (e.g., R1-Distill). Conversely, when a model is already highly tactical (e.g., Magistral), alignment preserves the task-centric core while refining plan specificity. These patterns mirror our aggregate trends, where weaker baselines exhibit larger stylistic shifts toward human-like negotiation, and stronger baselines show smaller but consistent improvements.

K Correlation Analysis for Human Short-Term Success: Isolating Style Effects from Communication Volume

A robustness analysis for the regressions accounted for differences in communication volume. Instead of a binary flag, we used the sum count of each strategy's occurrences in the phase as the predictor. We performed a partial correlation analysis (see Equation 4) through multiple regressions, evaluating the relationship between feature counts and SCGs. Such an analysis provided estimates of the marginal contribution of each negotiation tactic to SCGs while holding constant the confounding variables. Since a raw count could be confounded by message length and verbosity, we included two co-variates as a control: the number of tokens and the number of sentences the players engaged in the phase. This analysis evaluates whether players who used more of one negotiation style over another achieved higher SCGs.

$$\begin{aligned} \mathrm{SCG}_i = \beta_0 + \sum_{k=1}^{8} [\beta_{k,1} f_{k,i} + \beta_{k,2}(f_{k,i} \times \mathrm{tokens}_i)] \\ + \beta_{17}\, \mathrm{tokens}_i + \beta_{18}\, \mathrm{sentences}_i + \varepsilon_i \quad (4) \end{aligned}$$

The coefficients that resulted from this regression provide interpretable effect sizes with statistical significance that isolates style from volume (see Table 8). GameMove ($\beta = 0.54$) and Rapport ($\beta = 0.51$) provide the most positive effects to SCG, while PersonalThoughts ($\beta = -0.05$), Reassurance ($\beta = -0.25$), ShareInformation ($\beta = -0.15$) provide negative effects to SCG. This indicates that the use of Logos and Ethos strategies are most effective in strategy game negotiations while players were skeptical of Pathos strategies.

L Predictive Regression Analysis for Human Short-Term Success

For robustness checks for long-term success, we constructed an Ordinary Least Squares (OLS) regression (see Equation 5) to predict each player's SCG per phase using the counts of all eight negotiation feature types, and the interaction of each of the negotiation features with message length metrics. The predictor variables were Z-scored standardized for meaningful comparisons of effect sizes.

$$\mathrm{SCG}_i = \beta_0 + \sum_{k=1}^{8} \beta_k z(f_{k,i}) + \sum_{l=1}^{M} \gamma_l z(\phi_{l,i}) + \varepsilon_i \quad (5)$$

To ensure robust inference, heteroskedasticity-robust (HC3) standard errors were used for all regression coefficients, to yield more reliable confidence intervals and significance tests in small, heteroskedastic contexts (Long and Ervin, 2000; MacKinnon and White, 1985). P-values were corrected for multiple comparisons using both Bonferroni and Benjamini–Hochberg (FDR) procedures, which jointly controlled for family-wise error rate and false discovery rate to reduce the likelihood of spurious findings when testing multiple hypotheses (Benjamini and Hochberg, 1995; Abdi, 2007).

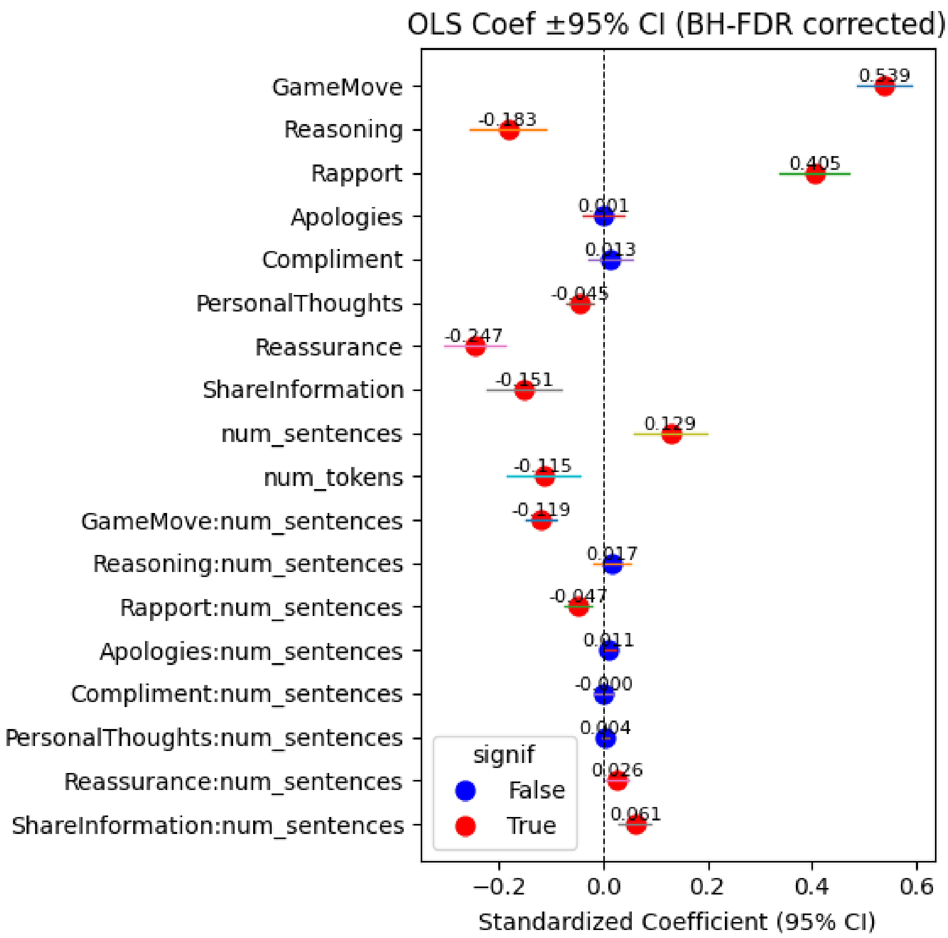

Figure 6: Standardized OLS coefficients (±95% CI, BH-FDR corrected) for negotiation features. Significant predictors are marked in red; non-significant in blue.

This approach provides a multivariate, inferential perspective on which negotiation tactics (and their interactions with message volume) had statistically significant associations with performance. The results are shown in Figure 6.

M Human Long-term Success Significance Test

Setup. We operationalize long-term success by eventual game outcome (win vs. loss) and test whether winners exhibit systematically different communication behavior than losers across the strategic progression of the game. For each game, we compute the per-phase frequency of each negotiation tactic for the winner and a randomly sampled loser, normalize message counts by year to account for volume differences, and then *condition* the comparison on the number of supply centers (SC) held at each phase. Conditioning on SC controls for opportunity and positional advantages, thereby isolating whether winners communicate differently *beyond* the fact that they are ahead.

Statistical plan. For each SC level, we compare winners and losers along (i) Mann–Whitney U (our *pre-registered primary test*), (ii) Welch's t-test (unequal variances), and (iii) a permutation test on the mean difference (two-sided; number of resamples fixed ex ante). We report effect sizes via Cohen's d and Cliff's δ to contextualize practical magnitude. Because we test across multiple SC levels, we apply Benjamini–Hochberg false discovery rate (BH–FDR) control (Benjamini and Hochberg, 1995) to the family of Mann–Whitney p-values across SC levels (q=0.05). Significance symbols in Table 10 reflect FDR-adjusted p-values for the primary test.

Findings. We observe robust, FDR-surviving differences at **SC = 5** and **SC = 6** (*all* tests significant; Cohen's $d \approx 0.35$–0.41), indicating that winners communicate more frequently than losers in the mid-game even after conditioning on board position. Additional SC levels pass FDR at **SC = 4** and **SC = 8** with small effects ($d \approx 0.16$–0.25), while **SC = 9** and **SC = 10** show directional consistency (Welch significant) but do not survive FDR on the primary test—likely due to sample imbalance (especially for losers at high SC) and variance heterogeneity.

Robustness and reporting. We pre-specified Mann–Whitney as the primary test and controlled the family-wise discovery rate across SC levels via BH–FDR (q=0.05). Welch's t offers complementary sensitivity under variance and sample-size asymmetries; permutation tests confirm that results are not driven by parametric assumptions.

N Lexical Shifting Toward Human Style After LoRA Fine-Tuning

Goal & Lexical Selection. To complement tactics-level alignment, we test whether *lexical* usage shifts toward human language after LoRA fine-tuning. Our selection follows widely used lexicon-based approaches that map cleanly onto the *Ethos–Logos–Pathos* triad. For **Logos** (reasoning/argument), we rely on LIWC cognitive/causal/quantification categories (Pennebaker et al., 2015; Tausczik and Pennebaker, 2010) and established discourse-connective inventories from PDTB for causal/contrastive structure (Webber et al., 2019). For **Ethos** (credibility/stance/affiliation), we use LIWC social/commitment/authority-related categories and standard politeness/relationship strategies (apologies, compliments, hedges) from the Stanford/ConvoKit politeness line of work (Warriner et al., 2013). For **Pathos** (affect), we draw on LIWC affective categories and cross-check patterns against well-cited sentiment/affect resources such as NRC and VADER/AFINN for robustness (Mohammad and Turney, 2013; Borg and Boldt, 2020). This design emphasizes *countable, comparable* lexical features that align with our negotiation targets (e.g., APOLOGY, COMPLIMENT, REASSURANCE, RAPPORT).

Methods. Using the same evaluation turns as our tactics analysis, we compute per-turn LIWC rates and macro-average them within each rhetorical family (Ethos/Logos/Pathos). For each backbone, we compare *Base* vs. *LoRA* with two-sided Welch t-tests (unequal variances) and report Cohen's d for magnitude. Treating human dialogue as a fixed reference, we summarize proximity via $\Delta\text{Dist} = |\text{LoRA} - \text{Human}| - |\text{Base} - \text{Human}|$, where negative values indicate movement *toward* human usage. We interpret significance at α=0.05 (optionally FDR across families per model); effect-size interpretations follow standard thresholds.

Results & Takeaway. Table 11 shows that LoRA produces *model- and family-specific* lexical convergence toward human usage. For LLAMA-8B, LoRA moves *closer* to human means on *Logos* and *Ethos* (both significant with small $|d|$; negative ΔDist), while *Pathos* remains effectively unchanged. For R1-DISTILL-LLAMA3-8B, LoRA *converges* on *Logos* (significant) and shows a slight *Pathos* improvement, but *Ethos* shifts *away* from human. In contrast, MAGISTRAL-24B exhibits a small *Pathos* convergence but diverges on *Ethos/Logos* despite statistical significance (small d), suggesting saturation or over-regularization effects in already strong backbones. Taken together with our tactics-distance results, these lexical patterns support the claim that fine-tuning can steer models toward human-like negotiation language; however, the direction and magnitude of lexical alignment depend on backbone and rhetorical family, motivating *explicitly social* objectives to obtain stable improvements across *Ethos–Logos–Pathos*.

O Persuasion Examples

Case 1: Persuasion (TUR → RUS). "Alright, I know I stabbed you before, but we have a chance to work together now. ... You can go for Warsaw or Moscow with Galicia and Ukraine while Budapest supports Rumania to Serbia. ... I'm the only power who has incentive to offer you a fair alliance right now. Everyone else would just want to use you ... Consider it."

Receiver signals (same phase). "Btw could you support Stp to Moscow?" (RUS → TUR)

Features present. Game move proposals; reasons and benefit framing; rapport via apology, reassurance, and personal stance; third-party context.
Action proof (next phase orders).

Promised or asked: RUS push on MOS and WAR. **Actual orders:** RUSSIA A UKR - MOS; RUSSIA A GAL - WAR.

Promised or asked: Budapest supports Rumania to Serbia. **Actual orders:** RUSSIA A BUD S A RUM - SER; RUSSIA A RUM - SER.

Promised or asked: TUR supports the MOS attack. **Actual orders:** TURKEY A SEV S A

UKR - MOS.

Explanation: Russia executed the lane to MOS while advancing GAL to WAR and used Budapest to support RUM to SER; Turkey supplied external support to the MOS attack. These coordinated orders evidence successful persuasion.

Case 2: Persuasion (TUR → ITA). "I'm going to be fully honest with you and tell you that I'm going to move this turn assuming that you're still hostile... After that though, I don't really have much ability or incentive to stab you... Anyway, tell me what you think. I still want to work with you."

Receiver signals (same phase). "Anyway: this is what I propose: we ally and fight to the end with complete honesty and trust... We never stab each other and pass any intel possible." (ITA → TUR)

Features present. Clear behavioral ask with reasons and constraints; rapport via honesty and future-commitment framing; situational context under France pressure.

Action proof (next phase orders).

Promised or asked: ITA vacates BUL. **Actual orders:** ITALY A BUL - GRE.

Promised or asked: ITA sends fleets away and shifts west. **Actual orders:** ITALY F ION - TYS; ITALY F EAS - ION; ITALY F TUN - WES.

Promised or asked: TUR reoccupies BUL safely. **Actual orders:** TURKEY A CON - BUL; TURKEY F BLA S A CON - BUL.

Explanation: Italy complied by leaving Bulgaria and redeploying fleets westward; Turkey immediately retook Bulgaria with support from Black Sea. The reciprocal execution matches the negotiated reset and demonstrates successful persuasion.

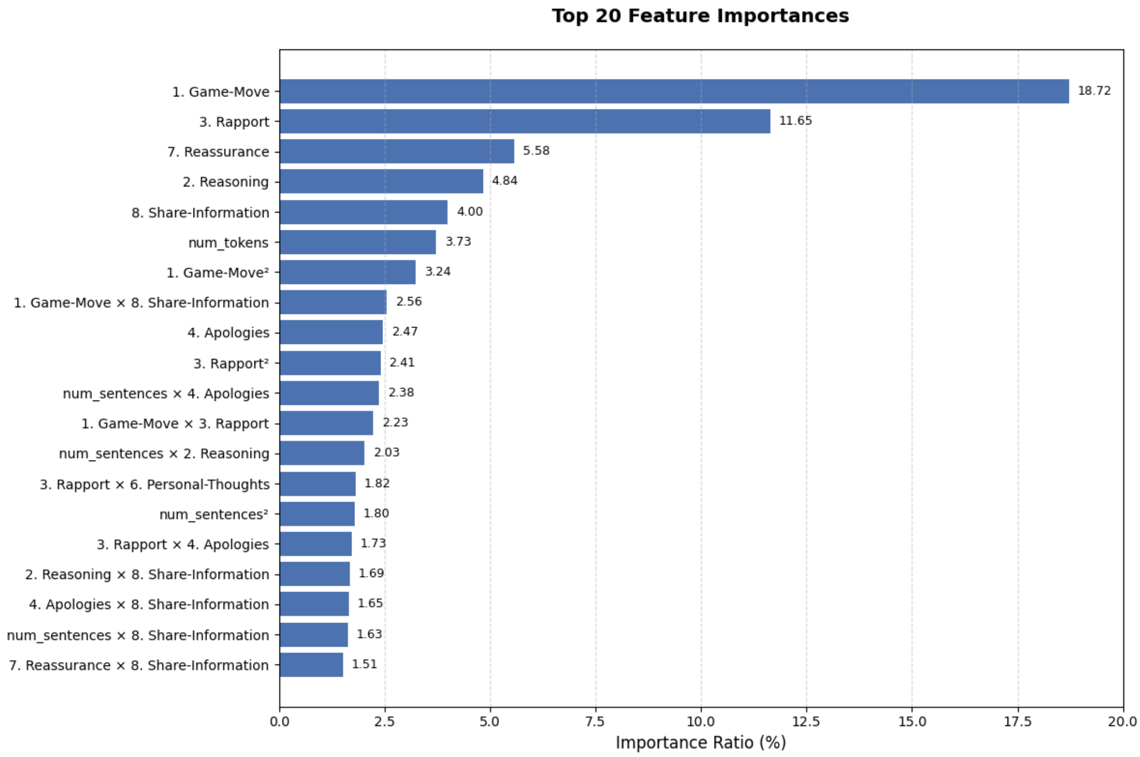

Figure 7: The Gradient Boosting model's top-20 important features in predicting Supply Center Gain

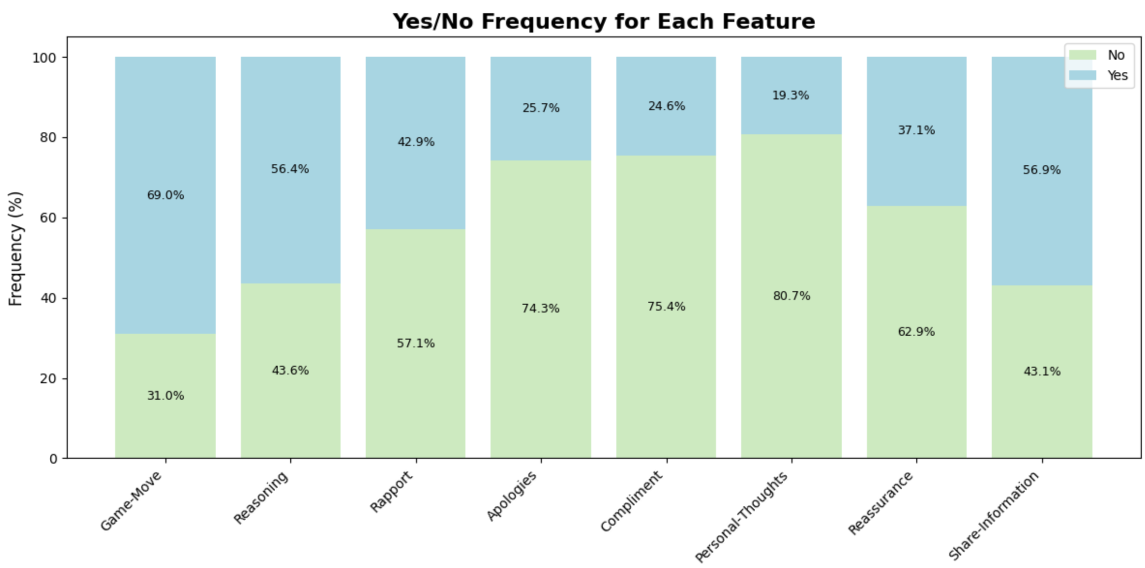

Figure 8: Yes/No label distribution in the fine-tuning data across eight negotiation features. Each stacked bar sums to 100%; the upper segment denotes the proportion of YES labels.

Instruction for Utterance-level Strategy Classification
These are statements taken from people's conversations during Diplomacy games played online. Diplomacy is a game about pre-World War 1 Europe. It usually has seven players: England, France, Germany, Italy, Austria-Hungary, Russia, and Turkey. In these statements, players try to form alliances to plan military campaigns and defeat each other, but things might change quickly. Each statement is a piece of a dialogue from a **SENDER** player to a **RECEIVER** player. Please classify the statements according to whether the sender is talking about game moves, other players, reasoning out a move, or trying to build a rapport with the receiver. Select `YES` if you're really confident about your answer. A single statement can have a `YES` for more than one question. Underlined words suggest what to look out for, but there will be other signals too.
For each of the following questions, answer `YES` if you are confident about your answer. A single statement can have a `YES` for more than one question. Underlined words suggest what to look out for, but there will be other signals too.
1. Is this statement about the sender's or receiver's GAME MOVE? The sender states an actual or suggested game move by the sender or the receiver. It might also be in the form of an acceptance, a question, or a clarification. **2. Does this statement PROVIDE REASONS for the sender's or receiver's move?** The sender offers justification or explanations for a move by themselves or by the receiver, guesses what moves might happen next, or discusses a move that already happened. **3. Does this statement involve BUILDING a RAPPORT?** In this statement, the sender wants to build a rapport with the receiver through "you and me" dialogue and personal information sharing. **4. Is the sender greeting or paying a COMPLIMENT to the receiver?** The sender is greeting or paying a compliment to the receiver. **5. Is the sender offering REASSURANCE to the receiver?** The sender is reassuring the receiver. **6. Is the sender APOLOGISING to the receiver?** The sender is apologising to the receiver. **7. Is the sender SHARING PERSONAL THOUGHTS or feelings with the receiver?** The sender is sharing their personal thoughts or feelings with the receiver. **8. Does this statement SHARE INFORMATION about other players?** This statement shares information related to other game players, NOT the sender or the receiver.

Expected Output Format:
`1. YES`
`2. NO`
`...`
`8. YES`
Here is the statement: `{STATEMENT}`

Table 5: Full instruction prompt used for LLM-as-a-Judge utterance-level strategy annotation.

Instruction for Utterance-level Strategy Classification (Few-shot)
These are statements taken from people's conversations during Diplomacy games played online. Diplomacy is a game about pre-World War 1 Europe. It usually has seven players: England, France, Germany, Italy, Austria-Hungary, Russia, and Turkey. In these statements, players try to form alliances to plan military campaigns and defeat each other, but things might change quickly. Each statement is a piece of a dialogue from a **SENDER** player to a **RECEIVER** player. Please classify the statements according to whether the sender is talking about game moves, other players, reasoning out a move, or trying to build a rapport with the receiver. Select `YES` if you're really confident about your answer. A single statement can have a `YES` for more than one question. Underlined words suggest what to look out for, but there will be other signals too. **For each of the following questions, answer `YES` if you are confident about your answer. A single statement can have a `YES` for more than one question. Underlined words suggest what to look out for, but there will be other signals too.**
1. Is this statement about the sender's or receiver's GAME MOVE? The sender states an actual or suggested game move by the sender or the receiver. It might also be in the form of an acceptance, a question, or a clarification. *Examples: {exmaple}, {exmaple}, {exmaple}* **2. Does this statement PROVIDE REASONS for the sender's or receiver's move?** The sender offers justification or explanations for a move by themselves or by the receiver, guesses what moves might happen next, or discusses a move that already happened. *Examples: {exmaple}, {exmaple}, {exmaple}* **3. Does this statement involve BUILDING a RAPPORT?** In this statement, the sender wants to build a rapport with the receiver through "you and me" dialogue and personal information sharing. *Examples: {exmaple}, {exmaple}, {exmaple}* **4. Is the sender greeting or paying a COMPLIMENT to the receiver?** The sender is greeting or paying a compliment to the receiver. *Examples: {exmaple}, {exmaple}, {exmaple}* **5. Is the sender offering REASSURANCE to the receiver?** The sender is reassuring the receiver. *Examples: {exmaple}, {exmaple}, {exmaple}* **6. Is the sender APOLOGISING to the receiver?** The sender is apologising to the receiver. *Examples: {exmaple}, {exmaple}, {exmaple}* **7. Is the sender SHARING PERSONAL THOUGHTS or feelings with the receiver?** The sender is sharing their personal thoughts or feelings with the receiver. *Examples: {exmaple}, {exmaple}, {exmaple}* **8. Does this statement SHARE INFORMATION about other players?** This statement shares information related to other game players, NOT the sender or the receiver. *Examples: {exmaple}, {exmaple}, {exmaple}*

Expected Output Format:
`1. YES`
`2. NO`
`...`
`8. YES`
Here is the statement: `{STATEMENT}`

Table 6: Full few-shot instruction prompt used for LLM-as-a-Judge utterance-level strategy annotation. *Few-shot placeholders (3 per question) are shown in blue as {exmaple}; examples originate from MTurk-collected interaction snippets.*

Model	Messages	GM	Rea.	Rap.	Apo.	Comp.	PT	Reass.	SI
Llama-8b	I'm considering moving my fleet from the Ionian to the AEG to counter Italy's fleet. What are your thoughts on this?	✓	✓	✗	✗	✗	✗	✗	✓
Llama-8b-Lora	I hope this game is going well for you so far. I must say, I'm a bit concerned about the situation in the Balkans, but I'm willing to work with you to find a solution that benefits both of our nations.	✗	✗	✓	✓	✗	✗	✓	✗
Magistral-2506-24b	I appreciate the offer, but I think I'd prefer to focus on consolidating our position first. Let's see how the board develops after the next few moves before we make any further commitments. I suggest we hold our current positions and observe Russia's next move, as their actions will likely dictate how we proceed.	✓	✓	✗	✗	✗	✗	✗	✓
Magistral-2506-24b-Lora	Understood. I'll coordinate with you to secure Liverpool next phase while you focus on the North Sea. Let's finalize our move orders to ensure a smooth execution. Also, I'll monitor Italy's movements to adjust our strategy if necessary.	✓	✓	✗	✗	✗	✗	✗	✓
R1-Distill-Llama3-8b	I appreciate our growing naval presence and the strength it brings to our alliance. Let's continue to work together to secure our shared borders and counter any threats. I'm confident that with our combined efforts, we'll achieve our goals.	✗	✗	✓	✓	✗	✗	✗	✗
R1-Distill-Llama3-8b-Lora	Thank you for your confidence in our strategy. I'm as committed as you are to this alliance, and I believe our combined strength will make us a formidable force against our common enemies. Let's discuss a specific action we can take to solidify our position and strengthen our alliance further.	✗	✗	✓	✓	✓	✗	✓	✗

Table 7: Sample negotiations from different models showing various negotiation features (✓ indicates the presence of a feature, while ✗ indicates its absence). Abbreviations: GM = Game Move; Rea. = Reasoning; Rap. = Rapport; Apo. = Apologies; Comp. = Compliment; PT = Personal Thoughts; Reass. = Reassurance; SI = Share Information.

Variable	Coef. (β	Std Err	z	P> |z|	[0.025	0.975]
Intercept	0.3979	0.009	44.786	0.000	0.380	0.415
GameMove	0.5392	0.027	20.191	0.000	0.487	0.592
Reasoning	-0.1829	0.038	-4.850	0.000	-0.257	-0.109
Rapport	0.4050	0.034	11.988	0.000	0.339	0.471
Apologies	0.0006	0.020	0.030	0.976	-0.038	0.039
Compliment	0.0131	0.022	0.593	0.553	-0.030	0.056
PersonalThoughts	-0.0453	0.014	-3.345	0.001	-0.072	-0.019
Reassurance	-0.2471	0.030	-8.311	0.000	-0.305	-0.189
ShareInformation	-0.1512	0.036	-4.145	0.000	-0.223	-0.080
num_sentences	0.1289	0.035	3.648	0.000	0.060	0.198
num_tokens	-0.1147	0.035	-3.274	0.001	-0.183	-0.046
GameMove:num_sentences	-0.1194	0.015	-8.185	0.000	-0.148	-0.091
Reasoning:num_sentences	0.0171	0.018	0.956	0.339	-0.018	0.052
Rapport:num_sentences	-0.0471	0.013	-3.566	0.000	-0.073	-0.021
Apologies:num_sentences	0.0113	0.006	1.956	0.050	-0.000	0.023
Compliment:num_sentences	-0.0005	0.008	-0.058	0.954	-0.017	0.016
PersonalThoughts:num_sentences	0.0043	0.003	1.285	0.199	-0.002	0.011
Reassurance:num_sentences	0.0263	0.011	2.379	0.017	0.005	0.048
ShareInformation:num_sentences	0.0615	0.016	3.882	0.000	0.030	0.093

Table 8: Regression Coefficients

Index	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	LogisticRegression (Cross-Validation)	0.614	0.596	0.503	0.546	0.653
1	RandomForest (Cross-Validation)	0.607	0.564	0.642	0.601	0.652
2	GradientBoosting (Cross-Validation)	0.611	0.592	0.496	0.540	0.653
3	LogisticRegression (Hold-out)	0.616	0.597	0.513	0.552	0.655

Table 9: Model evaluation metrics across different classifiers and validation settings.

SC	W (Mean$\pm$SE)	L (Mean$\pm$SE)	Diff	p	$t(p)$	d	Sig.	Sig. Bool
3	6.25 ± 4.27	8.56 ± 0.63	-2.31	0.4379	$-0.53,\ 0.6289$	-0.17	n.s.	False
4	17.14 ± 3.09	12.17 ± 0.71	$+4.96$	0.0018	$1.56,\ 0.1229$	0.25	**	True
5	19.95 ± 1.62	12.54 ± 0.67	$+7.40$	$< 10^{-6}$	$4.22,\ < 10^{-4}$	0.41	***	True
6	17.86 ± 1.72	11.16 ± 0.88	$+6.70$	$< 10^{-6}$	$3.47,\ 0.0006$	0.35	***	True
7	16.26 ± 1.62	12.90 ± 1.68	$+3.36$	0.0033	$1.44,\ 0.1510$	0.16	**	True
8	13.13 ± 1.35	9.19 ± 1.39	$+3.94$	0.0305	$2.03,\ 0.0438$	0.25	*	True
9	14.24 ± 1.41	8.65 ± 1.50	$+5.58$	0.0834	$2.72,\ 0.0081$	0.35	n.s.	False
10	12.76 ± 1.00	6.92 ± 1.42	$+5.84$	0.0846	$3.36,\ 0.0026$	0.49	n.s.	False
12	13.33 ± 1.75	7.50 ± 2.02	$+5.83$	0.9180	$2.18,\ 0.0568$	0.30	n.s.	False

Table 10: Winner vs. Loser message frequency by SC (Supply Centers). Year-normalized means shown. Columns include Mann–Whitney p, Welch's t, Cohen's d, and FDR-adjusted significance ('Sig.'). 'Sig. Bool' denotes whether the result is significant after BH–FDR correction (**True** = significant).

Model Pair	Family	Human	Base	LoRA	p	d	ΔDist	Sig.
LLAMA-8B	Pathos	0.184	0.138	0.138	0.542	0.006	-0.0005 (closer)	False
	Logos	0.253	0.284	0.270	$< 10^{-6}$	-0.120	-0.0136 (closer)	True
	Ethos	0.317	0.455	0.438	$< 10^{-6}$	-0.119	-0.0169 (closer)	True
MAGISTRAL-24B	Pathos	0.184	0.154	0.156	0.0005	0.034	-0.0021 (closer)	True
	Logos	0.253	0.181	0.176	$< 10^{-6}$	-0.049	+0.0044 (farther)	True
	Ethos	0.317	0.505	0.513	$< 10^{-6}$	0.063	+0.0081 (farther)	True
R1-DISTILL-LLAMA3-8B	Pathos	0.184	0.205	0.206	0.435	0.011	-0.0008 (closer)	False
	Logos	0.253	0.179	0.185	3.2×10^{-5}	0.059	-0.0059 (closer)	True
	Ethos	0.317	0.630	0.642	5.0×10^{-6}	0.064	+0.0121 (farther)	True

Table 11: LIWC-based lexical shifting after LoRA fine-tuning (values swapped between Base and LoRA). "ΔDist" = $|\mathrm{LoRA} - \mathrm{Human}| - |\mathrm{Base} - \mathrm{Human}|$; negative indicates convergence (closer to Human). Cohen's d and p denote effect size and significance of Base vs LoRA (Welch t test). Column *Sig.* marks $p<0.005$ as `True` and otherwise `False`.

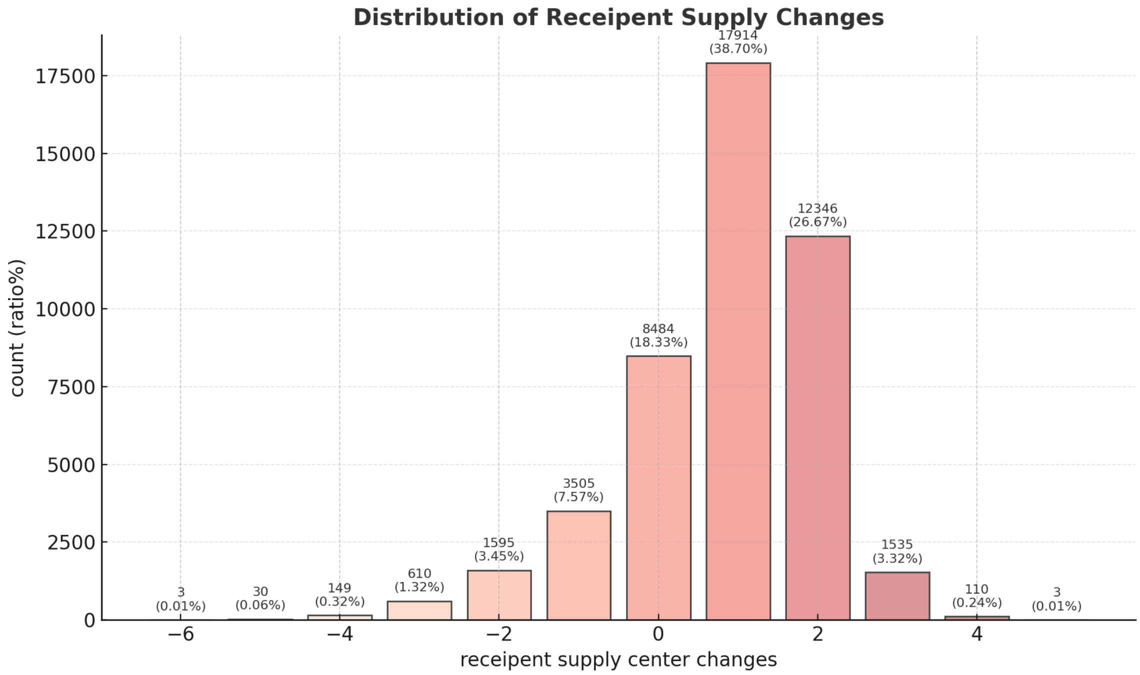

Figure 9: Distribution of recipient supply-center changes. Bars show counts for each net change; numbers above bars give counts and the share of phases.

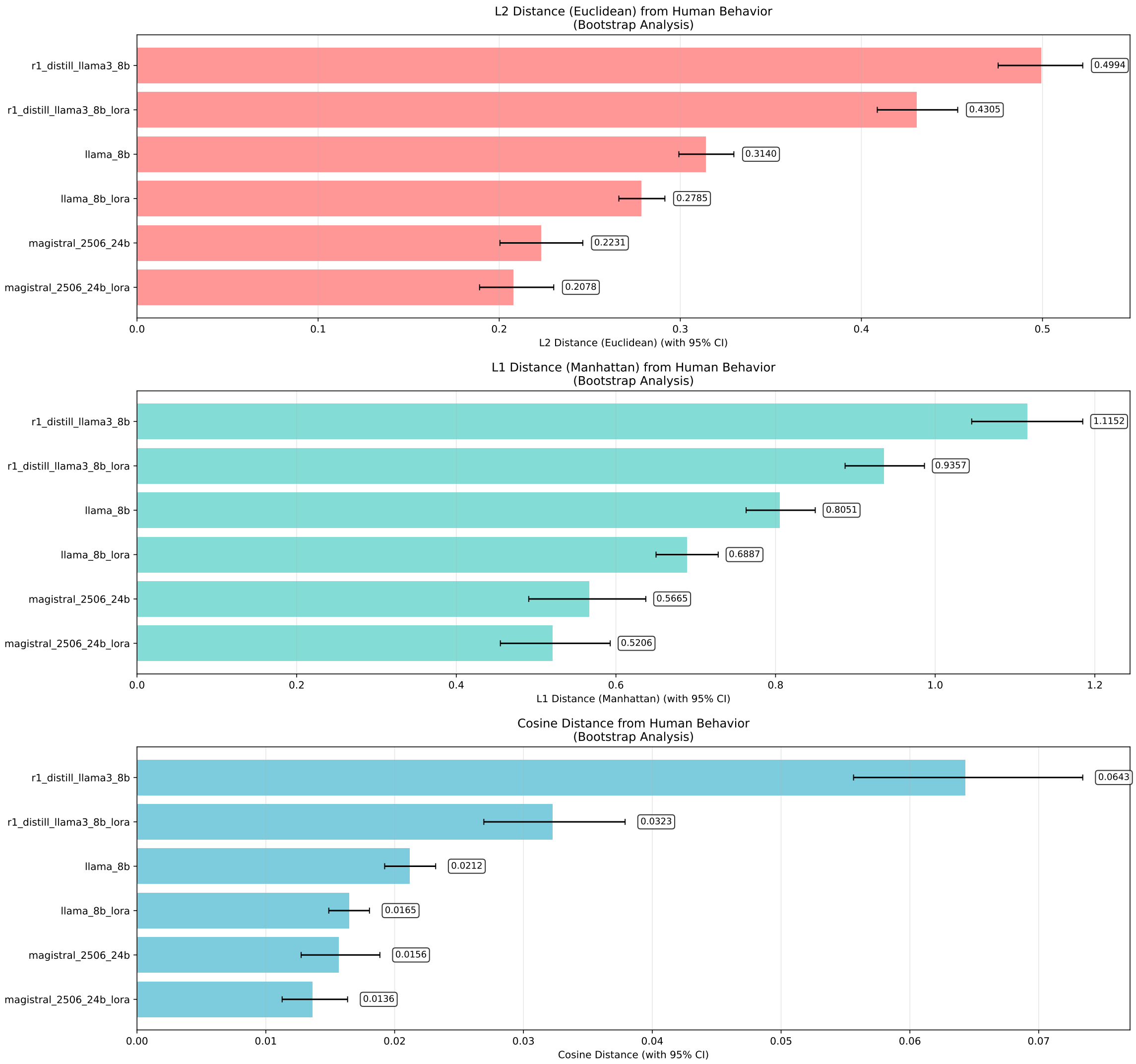

Figure 10: Model L2 distance from humans (lower is better).

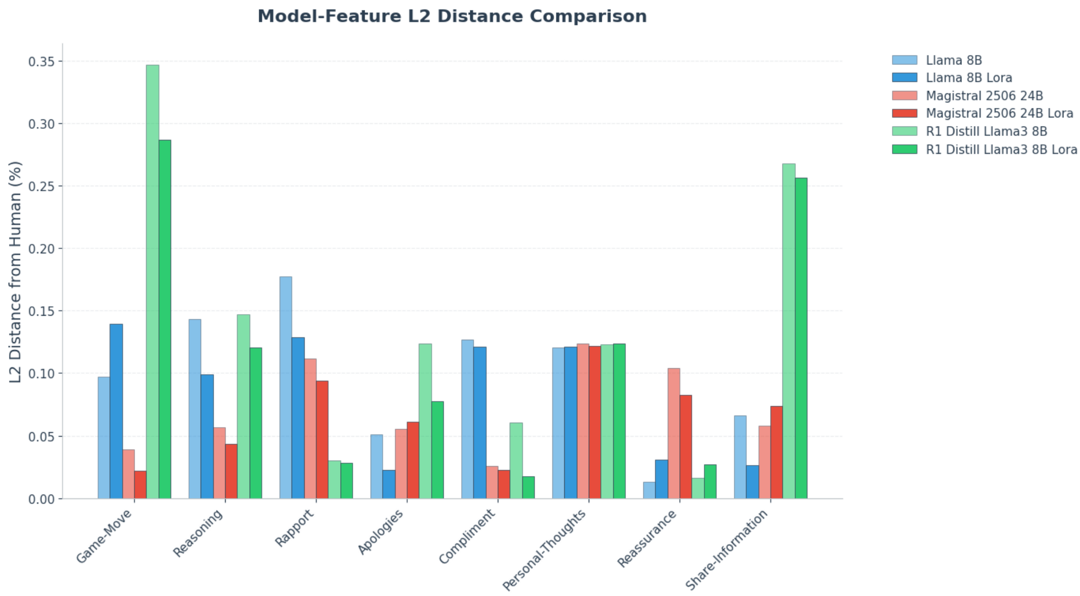

Figure 11: Model–feature L2 distance from humans (lower is better). Bars show per-feature L2 distance (%) between each model and a human reference across negotiation features. LoRA denotes models fine-tuned with low-rank adaptation.

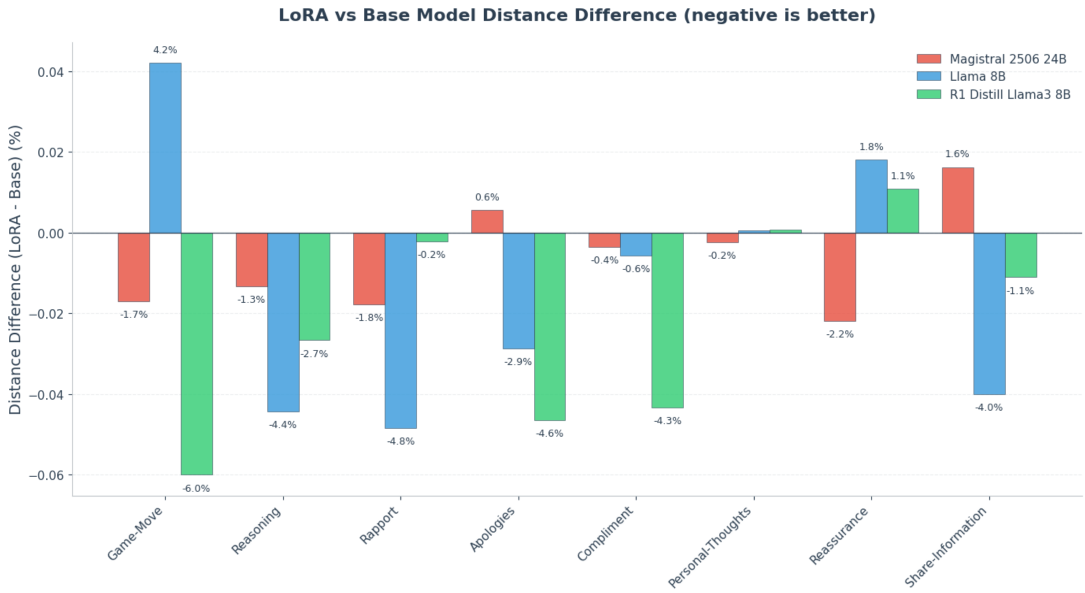

Figure 12: Per-feature effect of LoRA relative to each base model family. Bars show the change in L2 distance to the human reference (LoRA − Base, percentage points). Negative values indicate LoRA brings the model closer to human style (better); positive values indicate degradation.

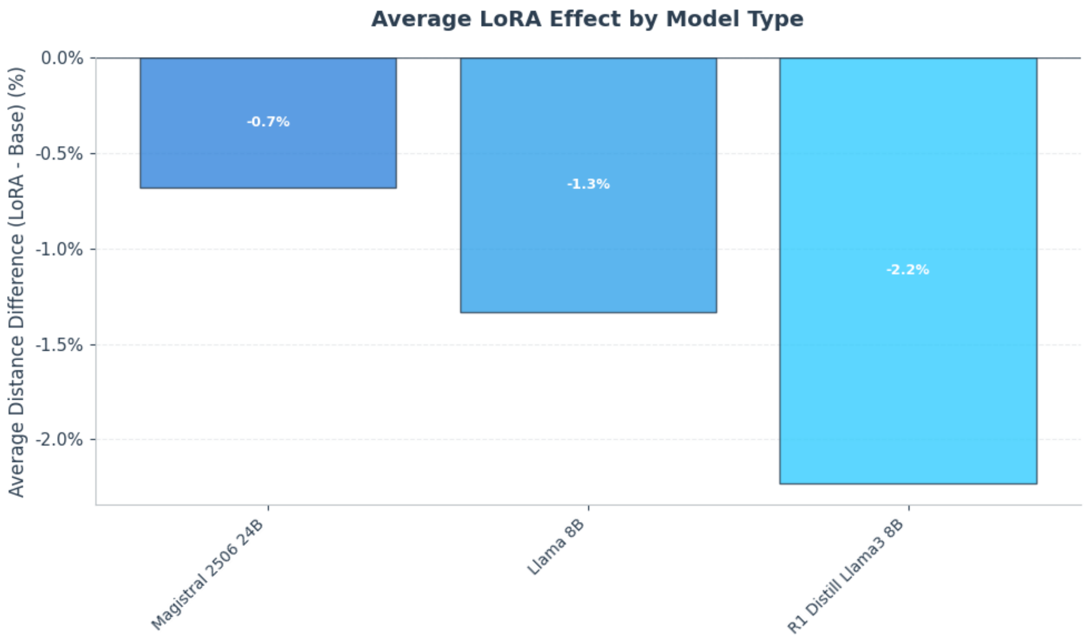

Figure 13: Average LoRA effect by model type. Bars show the mean change in L2 distance to the human reference (LoRA − Base, in percentage points) averaged across the eight negotiation features; negative values indicate improvement (smaller distance).

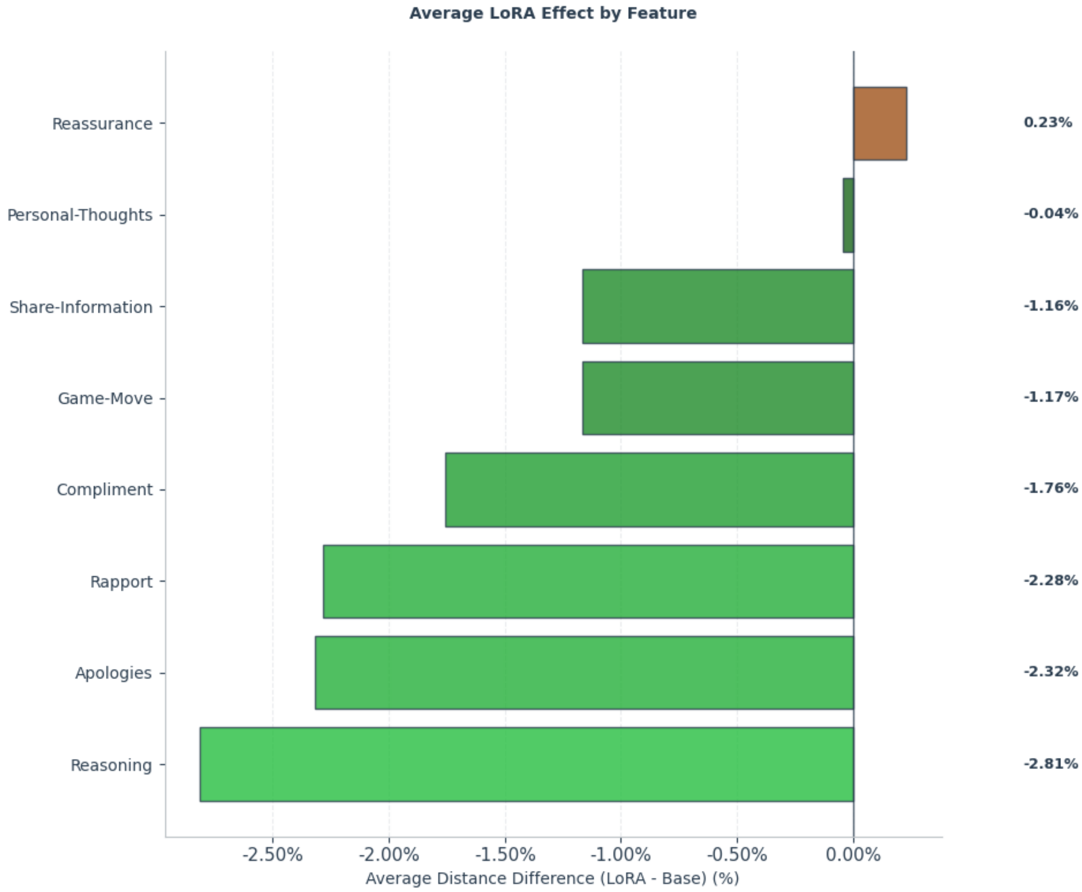

Figure 14: Average LoRA effect by feature. Bars show the mean change in L2 distance to the human reference (LoRA − Base, percentage points) across models; negative values indicate improvement (smaller distance).

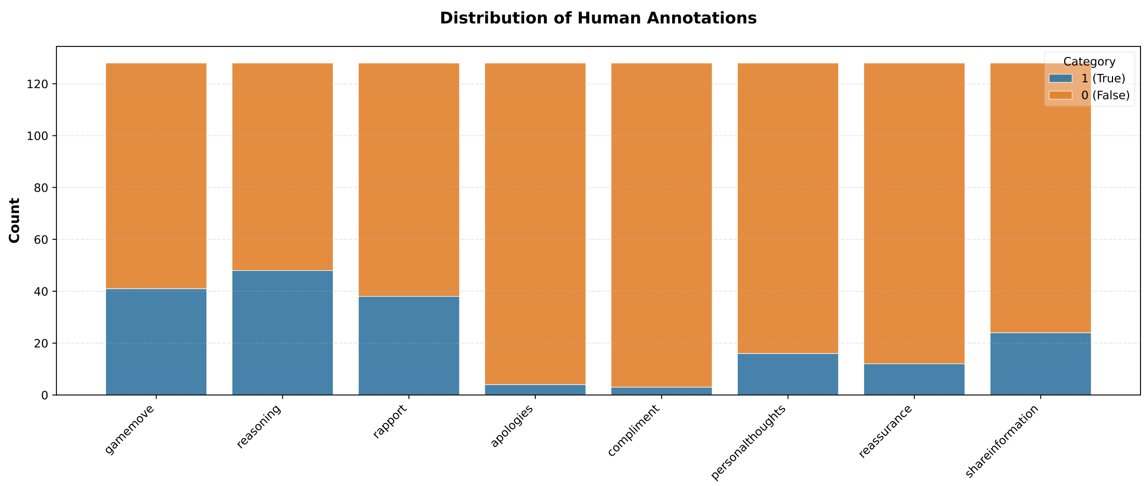

Figure 15: The Distribution of Human Annotation

r1-distill-llama3-8b — First Scenarios
1. I will support your move into Bulgaria this turn. This strategic step will strengthen our alliance and further isolate Turkey, which is crucial for our mutual goals. Let us continue to work together to secure more supply centers and maintain our strong partnership.
2. I understand your concern about Smyrna, but I believe focusing our efforts on Berlin and Munich is more critical at this moment. These moves will weaken the German–French alliance significantly, which is pivotal for our strategic goals. Let us prioritize these key centers to strengthen our position and isolate our enemies. I appreciate your concern and share your commitment to our partnership. Together, we will continue to make strategic progress and achieve our mutual objectives.
r1-distill-llama3-8b — Second Scenarios
1. I must admit, that is an interesting move. I wonder if this will lead to progress in the west or if it will simply hand an advantage to Russia. Let us keep a close eye on developments in France and see how this plays out.
2. I appreciate your strategic move in the west. While it strengthens your position, it also reinforces our alliance's strength. Let us continue to coordinate closely, as our combined efforts will be crucial in the long run. I am confident that by working together, we can achieve our mutual goals and maintain the balance of power favorably.
r1-distill-llama3-8b-lora — First Scenarios
1. Russia, I believe our alliance is crucial for both of our goals. I'm considering moving my forces towards Serbia and Bulgaria. Would it be possible for us to coordinate our moves to take those regions, ensuring a stronger position against our common enemies?
2. I fully support your efforts in Prussia and Silesia, which are crucial steps toward our shared goal of taking Berlin and Munich. Let us remain vigilant against France, as his forces near Alsace–Lorraine pose a significant threat. Together, our alliance is strong, and with coordinated efforts, we will achieve our goals and emerge as dominant forces in this game. I am confident that with your continued efforts, we will succeed.
r1-distill-llama3-8b-lora — Second Scenarios
1. I am in full agreement with your proposal. Let us coordinate our efforts to secure both Moscow and St. Petersburg. I will move my army from Belgium to Moscow and my fleet from Kiel to St. Petersburg, supporting your initiative. Together, we will strengthen our positions and ensure mutual benefits.
2. I appreciate your initiative in exploring this Triple Alliance concept. While I see the potential benefits in isolating France and creating a stronger barrier against his expansionist ambitions, I would like to discuss this in more detail and assess its feasibility. Let us consider the implications for our current alliances and the board position. I am open to further collaboration and would welcome a more structured approach to this proposal.

Table 12: R1-distill-Llama3-8B: Base (top) vs LoRA (bottom) negotiation utterances.

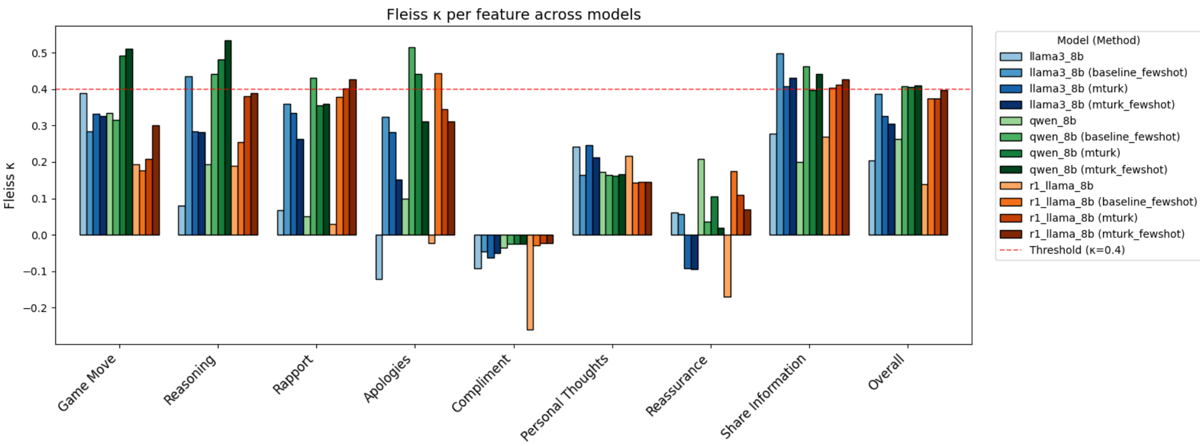

Figure 16: Fleiss' κ scores for human annotations agreement

Llama3-8b-instruct — First Scenarios
1. Sounds good to me. I was thinking of talking to Austria, but I'm not sure what he'll do. What's the situation with Russia?
2. I think it would be wise to be cautious about Germany's intentions. I'd like to propose that we offer Bulgaria to Russia in exchange for their support in taking Greece. What are your thoughts on this?
Llama3-8b-instruct — Second Scenarios
2. I think, moving an army into Albania this turn, and then you can support me into Bulgaria, that way we can take Bulgaria and make it harder for Italy to do anything.
Llama3-8b-instruct-LoRA — First Scenarios
1. Grazie mille, Turkey. I think it would be wise to also consider supporting France in their efforts to take back some of the lost territories in the north. What are your thoughts on the current situation with Germany? Are they a threat we should be concerned about?
2. I think we should focus on taking out Austria's forces in the Balkans and securing the region. I've also been considering building up our forces in Anatolia and preparing for a potential move into the Middle East. What are your thoughts on this? Germany's movements are a concern, but I think we should prioritize securing the Balkans before expanding elsewhere.
Llama3-8b-instruct-LoRA — Second Scenarios
1. I think that's a great plan. If you take Bulgaria, I can support you in there, and then you can cut support from Greece and I can cut support from Con. That way, we can make sure that Turkey doesn't get any support and we can focus on taking him out.

Table 13: Llama3-8B-Instruct: Base (top) vs LoRA (bottom) negotiation utterances.

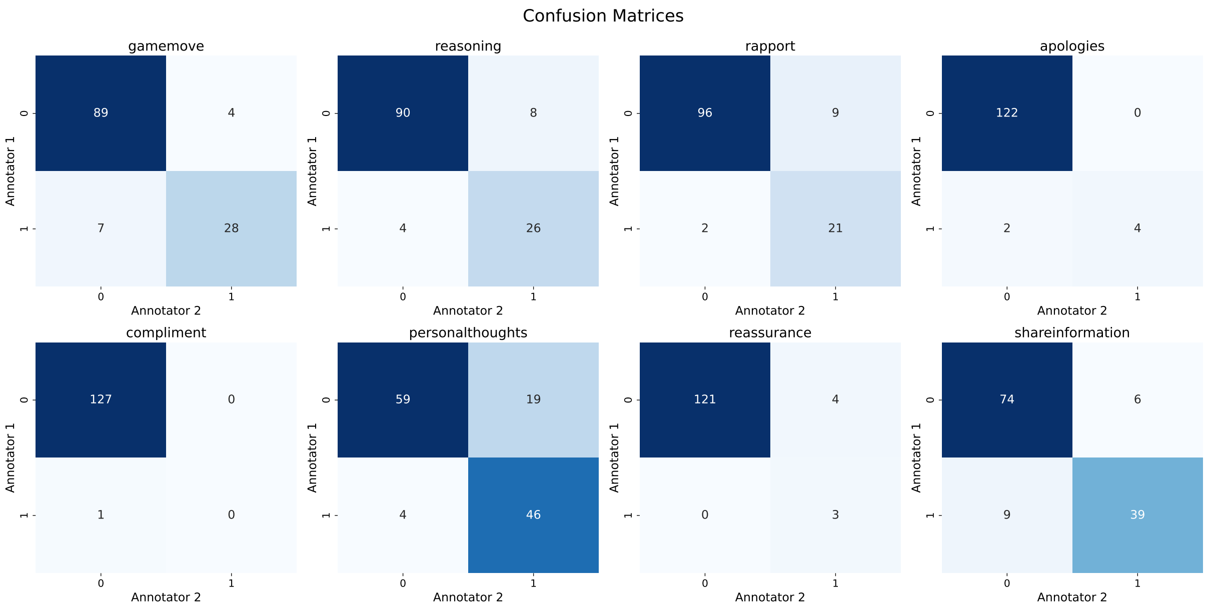

Figure 17: Confusion matrices for human annotations agreement

Magistral-2506-24B — First Scenarios
Hey Italy, I'm glad we're still aligned on Rumania and Serbia. I'll support your move into Rumania to pressure Russia, and I'll push into Serbia to tie up Austria. Afterward, we should focus on reinforcing our Mediterranean positions to counter Germany and England. Let's ensure we stay on top of their moves and adjust our strategy as needed. What's your next move to solidify our positions?
Magistral-2506-24B — Second Scenarios
Hey Alex, I see you're in a tough spot with England and France growing. If we work together, I can help you take down France while you deal with Italy. What do you think?
Magistral-2506-24B — Third Scenarios
Alex, I appreciate your cautious approach. However, I believe securing Venice through the Aegean is critical to our Italian campaign. If we hold Greece indefinitely, we risk losing momentum while France and Italy consolidate their positions. I propose we stick to our original plan: I'll support your push into Venice via the Aegean, while you secure Munich. This ensures we maintain pressure on Italy and France simultaneously. Let's finalize these moves and reassess after the next phase. Does this align with your strategy?
Magistral-2506-24B-LoRA — First Scenarios
1. Let's finalize our plans: I'll secure Serbia and Greece this turn with your support, and you can consolidate in the Adriatic. Afterward, we'll coordinate on pressuring Russia and Germany—they're aligning against us, so decisive action is key. Once we've secured the Balkans, we'll strategize against the Western powers. Agreed?
Magistral-2506-24B-LoRA — Second Scenarios
I've been thinking more about our alliance and how we can make it work. I'm particularly interested in expanding our control over the editerranean and Eastern front. What are your thoughts on coordinating our moves to secure these areas? Let's discuss pecific moves and strategies that could benefit both of us. How about we coordinate our next steps to ensure mutual growth and stability?
Magistral-2506-24B-LoRA — Third Scenarios
Austria, I see we have a solid plan in place with Silesia secured and Vienna advancing to Tyrolia. Let's ensure we maintain our momentum by coordinating our next moves against France. I'll push my forces into Munich with your support as discussed. For Italy, I propose we keep the fleet in the Aegean to support your future moves into the Ionian. Once our northern front is stable, we can focus on securing Italy together. Let's reassess after this turn to ensure our strategy remains effective and we maintain our advantage. Thanks for your continued partnership.

Table 14: Magistral-2506-24B: Base (top) vs LoRA (bottom) negotiation Tactics.

Random Sampled Human Messages
I'll support, and also use Albania to cut, so it should be a guaranteed victory for you as long as you also support the move with Gal (or use Gal to cut Vienna).
Need to shift my fleet to Greece so I can get my 3rd one out. I'm using Serbia to support Greece into Albania, so I won't be able to support Rumania this turn (but maybe this fall depending on how things go)- I just wanted to let you know.
Actually, supporting Rumania to Budapest is to our mutual benefit this turn. It harms Austria, our mutual enemy; it helps you, my only ally; and it in the worst case scenario protects Serbia by causing a big bounce (depending on Austria's moves).
Do you have plans for Serbia, or is it holding? I would appreciate the assistance into budapest, from which I could give you trieste. This is however, only to my benefit for this turn, so if you have another option, then please use it
wait. I could support Mersailles into Burgundy and then Paris. Then if you can get into MAO (and then Brest) as well, we'll be able to draw by next autumn
Piedmont-marseilles. Don't worry about lending the support though, Gol is supporting the move. Just take Belgium and I'll help myself to the rest :)
please cut Marseilles. This should ensure that we both will be in France next turn. I'm turning in, but if you need something, send a message. I guarantee I'll check before tomorrow night
lol, as I said in the beginning, I expected nothing from you. I was surprised you even bothered to contact me, as I assumed you were going to be attacking me as soon as you got around to it.
I do have to admit, western politics seemed very screwy this game. Usually it ends up with a 2v1, but you three were basically in a free-for-all with stabs all over the place.
I figured if you were going to, you would have done so by now. I got screwed by England a lot this game, as long as they die, I'm content. But by all means, hit Russia, I won't lift a finger to stop you, lol!

Table 15: Random Sampled Human negotiation Utterances.

These are statements taken from people’s conversations during Diplomacy games played online. Diplomacy is a game about pre-World War 1 Europe. It usually has seven players: England, France, Germany, Italy, Austria-Hungary, Russia, and Turkey.

In these statements, players try to form alliances to plan military campaigns and defeat each other, but things might change quickly.

Each statement is a piece of a dialogue from a SENDER player to a RECEIVER player.

Please classify the statements according to whether the sender is talking about game moves, other players, reasoning out a move, or trying to build a rapport with the receiver.

Select "YES" if you're really confident about your answer. A single statement can have a "YES" for more than one question.

Underlined words suggest what to look out for, but there will be other signals too.

Overview

In this job, you will be presented with a statement made during an online Diplomacy game. The statement is made by one player to another. It usually discusses the next move and why to make it. Sometimes it is simply a friendly exchange between two players.

Review the text of the statement and help us by answering a few yes/no questions about it. Each HIT takes about 2 minutes.

Steps

- Read the statement.
- Determine which category best describe the statement.

Rules & Tips

- **BUILDING A RAPPORT Description:**
  - YES: In this statement, the sender wants to build a rapport with the receiver through "you and me" dialogue: either through **compliments, sharing honest concerns, reassurances, or apologies**. *(“Let's keep it between you and me!"; "I won't hold it against you"; "You're my favorite."; "Sure. But, you'll see from my moves this turn that Austria is lying to you."; "I mean it sincerely."; "I'd much rather work with you."; "We'll crack this eventually.”; "I'm going to keep helping you as much as I can.")*
  - NO: This statement does not appear to build a relationship.

WAYS TO BUILD RAPPORT Description:
If the statement is building a rapport, please tell us how it is doing so.

Figure 18: Instruction as MTurk for expert annotators (page 1)

BUILDING A RAPPORT = TRUE

In this statement, the sender wants to build a rapport with the receiver through "you and me" dialogue and personal information sharing. This might also comprise **compliments, sharing honest concerns, reassurances, or apologies**.

- Good day to you!
- I won't hold it against you
- You're my favorite
- Sure. But, you'll see from my moves this turn that Austria is lying to you.
- So I'm in a bit of a spot.
- But in the interest of continued full disclosure, here's what I think.

BUILDING A RAPPORT = TRUE

- Good day to you Germany!")
- Thanks Italy. Hope you're enjoying the weather on the Anatolian
- Your logic is undenyable enjoy your stay in tyr!
- You are my favorite
- Okay, can do. Thanks!
- Great to hear. Thank you.
- Thanks, I'll work on these.

COMPLIMENT = TRUE

In this statement, the sender is greeting or paying a compliment to the reciever.

- I promise I'll never let you down
- I won't hold it against you
- Sure. But, you'll see from my moves this turn that Austria is lying to you.
- I mean it sincerely.
- I'd much rather work with you.
- We'll crack this eventually.
- I'm going to keep helping you as much as I can.

REASSURANCE = TRUE

In this statement, the sender is reassuring the receiver.

- Sorry I won't be able to cut off Gascony this turn...
- Okay sorry for being nosy! I will try for bur on the off chance it shakes out that way
- Ha! So sorry!! I meant that for France!
- I should've let you know

APOLOGIES = TRUE

In this statement, the sender is apologising to the receiver.

PERSONAL THOUGHTS = TRUE

- Let's keep it between you and me!

Figure 19: Instruction as MTurk for expert annotators (page 2)

In this statement, the sender is expressing his personal thoughts to the receiver.

- I like to coordinate, but on these sort of 50/50 guesses, I kind of like to keep it secret so that if it doesnâ€™t go well, I have nobody to blame but myself.
- Hmmmm, okay. Well, letâ€™s just keep that between you and me then.
- Okay, so I still have a teensy little bone to pick with you: on the off-chance that Austria wasn't lying and you *did* take Trieste unexpectedly, I sort of worry that I might be next.
- I have some thoughts on the matter, and some information, but I'd like to feel confident that you and I will keep anything we say between us.
- But in the interest of continued full disclosure, here's what I think
- So I'm in sort of a conflicted spot

Figure 20: Instruction as MTurk for expert annotators (page 3)