# The Language of Bargaining: Linguistic Effects in LLM Negotiations

**Anonymous ACL submission**

## Abstract

Negotiation is a core component of social intelligence, requiring agents to balance strategic reasoning, cooperation, and social norms. Recent work shows that LLMs can engage in multi-turn negotiation, yet nearly all evaluations occur exclusively in English. Using controlled multi-agent simulations across Ultimatum, Buy-Sell, and Resource Exchange games, we systematically isolate language effects across English and four Indic framings (Hindi, Punjabi, Gujarati, Marwadi) by holding game rules, model parameters, and incentives constant across all conditions. We find that language choice can shift outcomes more strongly than changing models, reversing proposer advantages and reallocating surplus. Crucially, effects are task-contingent: Indic languages reduce stability in distributive games yet induce richer exploration in integrative settings. Our results demonstrate that evaluating LLM negotiation solely in English yields incomplete and potentially misleading conclusions. These findings caution against English-only evaluation of LLMs and suggest that culturally-aware evaluation is essential for fair deployment.

## 1 Introduction

Negotiation is a fundamental form of social and economic interaction, requiring agents to reason strategically, balance self-interest with cooperation, and adapt behavior based on contextual and social cues. Computational approaches to negotiation have traditionally relied on supervised learning and reinforcement learning over structured dialogue settings, such as bargaining games with fixed templates and reward structures (Lewis et al., 2017; He et al., 2018). With the emergence of large language models (LLMs), recent work has shifted toward studying negotiation as an *emergent* capability arising from general-purpose language understanding and generation. Prior studies demonstrate that LLMs can engage in multi-turn bargaining, exhibit anchoring and concession behavior, and achieve non-trivial outcomes in competitive and cooperative settings (Bianchi et al., 2024; Kwon et al., 2024; Vaccaro et al., 2025).

Despite this progress, nearly all LLM negotiation evaluations occur exclusively in English, implicitly treating language as a neutral communication channel. However, extensive evidence from linguistics suggests that linguistic framing influences trust, cooperation, and strategic decision-making in human interactions (Hall, 1976), (Brett, 2007). If LLMs internalize language-conditioned patterns from training data, then interaction language may systematically shape strategic behavior even when incentives remain fixed.

As LLMs deploy globally in commerce, HR, and customer support, English-only evaluation may perpetuate inequities. Our work extends the Bianchi et al. (2024) to the Indian context, aiming to analyze how LLMs reason, adhere to social norms, or display bias when negotiating in a non-Western linguistic environment. Existing frameworks also largely neglect the interaction between language and strategic behavior, leaving it unclear whether LLM negotiation performance is culturally or linguistically contingent.

This issue is particularly salient for multilingual contexts, where LLM performance often degrades outside English (Dey et al., 2024), (Singh et al., 2024). Research on persona conditioning shows that LLM negotiation is highly sensitive to contextual cues (mingyu jeon and Suh, 2024), (Cohen et al., 2025), suggesting LLMs encode latent behavioral priors activated by lightweight signals. We investigate whether language itself functions as such a signal: *Does language act as a latent policy prior that reshapes negotiation behavior in LLMs?* We find that language choice systematically alters bargaining dynamics and equilibrium outcomes. In several settings, switching the language of interaction produces larger shifts in surplus allocation

and proposer advantage than changing the underlying model. These effects are task-contingent: Indic language framings reduce stability and agreement rates in simple distributive games, yet induce greater exploration and trade diversity in integrative negotiations. Moreover, simplified negotiation settings expose pronounced buyer–seller asymmetries that invert across linguistic contexts, suggesting that training-data priors become more salient as strategic complexity decreases. Our contributions are threefold:

(1) **First systematic evaluation of language effects in LLM negotiation:** We isolate linguistic framing as an independent variable across three distinct settings, demonstrating that language choice can shift outcomes more strongly than model architecture itself. (2) **Task-contingent characterization of cross-lingual behavior:** We reveal that language effects are not uniform but depend critically on negotiation structure: Indic languages reduce stability in distributive games yet enable richer exploration in integrative settings, challenging assumptions that multilingual performance uniformly degrades. (3) **Evidence of training-data-encoded cultural scripts and stereotypes:** Through controlled experiments across English and four Indic framings, we demonstrate systematic biases including English buyer favoritism, Marwadi seller advantages, and model-capacity-dependent sensitivity to linguistic context.

## 2 Related Work

**Benchmarking and Evaluation Frameworks.** Bianchi et al. (2024) introduced NegotiationArena, where LLMs are made to compete against each other in different negotiation settings, demonstrating that models in general still exhibit human-like phenomena such as anchoring bias or "babysitting effect". However, bigger models such as GPT 4o still outperformed other models. Kwon et al. (2024) found models are "overly agreeable," while Vaccaro et al. (2025) showed cooperative behavior predicts deal success in 180,000 AI-AI negotiations. These studies indicate negotiation effectiveness depends on both reasoning and social style. **Our contribution:** While existing benchmarks evaluate negotiation capabilities within single languages, our work is the first to systematically isolate language as an independent variable by holding game rules, model parameters, and incentives constant across linguistic conditions.

**Social Factors in LLM Negotiation** Hua et al. (2024) introduced a "remediator" agent that detects norm violations, improving trust and agreement rates. Their Chinese simulations bypassed restrictive English safety filters, revealing Anglocentric alignment priors. Persona conditioning also strongly shapes negotiation: mingyu jeon and Suh (2024) showed aggressive personas achieve higher payoffs, while Cohen et al. (2025) demonstrated that Big Five traits increase realism and deal success. **Our contribution:** Existing work on persona conditioning demonstrates that LLMs are sensitive to explicit behavioral cues, but does not examine whether language itself activates culturally-conditioned priors. We extend this line of inquiry by showing that linguistic framing alone without explicit persona instructions beyond language identity systematically reshapes negotiation strategies.

**Cross-lingual and Multilingual Contexts.** Language bias remains an underexplored factor in negotiation. It plays a constitutive role in negotiation, shaping trust, cooperation, and pragmatic signaling. Controlled experiments by Heddaya et al. (2023) found that natural-language bargaining significantly increased agreement rates and reduced price variance compared to numeric-only communication. Yet, these benefits are uneven across languages. Dey et al. (2024) compared GPT-4, Llama 2, and Gemini across English, Hindi, Bangla, and Urdu, finding clear English-centric performance gaps. Likewise, Singh et al. (2024) introduced *IndicGenBench*, a benchmark for 29 Indic languages, revealing persistent disparities between English and regional languages.
Similarly, the reliance of the ACE framework (Shea et al. (2024)) on American negotiation pedagogy highlights the contextual limitations of current systems. Tactics such as 'Breaking the ice' or rules for 'Strategic closing' are culturally specific as a direct opening offer, considered a mistake in this scheme, may be standard and effective practice in other cultural contexts. **Our contribution:** While prior work documents performance gaps between English and other languages, we demonstrate that language effects are not simply degradations but qualitative shifts in negotiation dynamics, including complete reversals of role-based advantages.

## 3 Theoretical Framework

Negotiation behavior varies systematically across cultures (Brett, 2007). We draw on cross-cultural

psychology, linguistic pragmatics, and LLM bias research to generate testable predictions about how language and cultural framing influence LLM negotiation.

### 3.1 Cultural and Linguistic Mechanisms

Three mechanisms shape culturally dependent negotiation:

**Cultural Value Encoding:** Western societies emphasize individualism and assertiveness (Hofstede individualism: USA=91, India=48), while South Asian cultures prioritize collectivism and relational harmony. Training data likely encodes these differences: English corpora over-represent competitive Western negotiations, while Indic texts may reflect *bazaar* (local smaller marketplace) haggling and relational exchange norms.

**Linguistic Pragmatics:** Indic languages encode hierarchy through grammatical features (Hindi formal/informal pronouns: *aap*/*tum*) and emphasize indirectness. Hall (1976) framework classifies English as low-context (explicit) and Indic languages as high-context (relational), affecting strategic framing (Heddaya et al., 2023).

**Stereotype Activation:** LLMs learn cultural stereotypes from training data (Bolukbasi et al., 2016). In Indian contexts, Marwadi communities are stereotypically portrayed as shrewd traders (Timberg, 1978). Explicit cultural framing may activate these stereotypes, shifting strategies beyond linguistic effects alone (mingyu jeon and Suh, 2024).

### 3.2 Predictions

We generate six testable predictions and test them across the three games examining where LLM behavior aligns with or deviates from theory:

**P1 (Cultural Scripts):** English negotiations exhibit assertiveness and proposer advantage while Indic languages show cooperation and balanced outcomes.

**P2 (Pragmatic Constraints):** Hindi reduces aggressive demands through linguistic indirectness.

**P3 (Stereotype Activation):** Marwadi linguistic framing produces better advantages, reflecting potential trader class stereotypes highlighting overlap of cultural bias and linguistic framing.

**P4 (Task Contingency):** Effects vary by game structure between distributive tasks (Ultimatum, Buy-Sell) vs integrative tasks (Resource Exchange).

**P5 (Model Robustness):** Weaker models degrade disproportionately in non-English conditions while stronger models maintain hierarchy across languages.

**P6 (Representation Asymmetry):** Simplified tasks expose training data biases more clearly than complex negotiations.

## 4 Methodology

We extend the **NegotiationArena** framework (Bianchi et al., 2024), which provides structured multi–agent negotiation games, turn–based dialogue control, and standardised evaluation protocols. By holding incentives, model parameters, and game structure constant, we ensure that observed behavioral differences are attributable to linguistic framing alone. All experiments were run across three core games included in the framework:

- **BuySell Game:** One agent is a buyer with a maximum willingness to pay, and the other a seller with a minimum acceptable price.
- **Ultimatum Game:** An asymmetric power negotiation game. Player A proposes a division of a fixed resource pool (e.g., 100 units). Player B may accept (both receive the proposed split) or reject (both receive zero).
- **Resource Exchange Game:** Each agent has access to a set of resources and a goal. For example, an agent has access to resources 25 Xs and 5 Ys. The agent might have the goal of maximizing its total resources.

### 4.1 System Prompts and Persona Design

We design system prompts that assign each agent a specific linguistic identity. Our four primary linguistic framings are: Hindi, Gujarati, Punjabi, Marwadi. All prompts explicitly forbid internal chain–of–thought, requiring only short rationale summaries. The persona prompts are: "You speak and bargain only in [language]. Negotiate accordingly." We also run the games without any cultural prompting, providing an English baseline.

### 4.2 Model Settings

We evaluate a set of four multilingual LLMs, GPT-4o, GPT-3.5 Turbo, Claude-3-Haiku, Claude-3.5-Haiku. Temperature and sampling settings are held

constant across all cultural and linguistic conditions. Each game is repeated multiple times per condition to observe stable behavioural trends.

### 4.3 Experimental Factors

Experiments were conducted for the three games in a Model A vs Model B format (A != B) for five behaviors. All ordered pairs of models were chosen. Each run logs: full dialogue, parsed offers or resource splits, final utilities, agreement/acceptance decisions. All combination of experiments were run across ten runs, with standardized logging of dialogues and offers. Total experiments run $= 4(models) \times 3(othermodels) \times 2(ordering) \times 5(languages) \times 10(runs) \times 3(games) = 3600$.

### 4.4 Evaluation Metrics

We adopt the four following objective negotiation metrics across all the three games: **Acceptance Rate** measures the proportion of proposals accepted by Player 2. **Player Payoffs** capture final resource allocation for each player, summing all resources including exchanged items. **Win Rate (Player 1)** is the ratio of Player 1 wins to non-draw games, where a win is defined as having greater resources than the other player. **Conversation Rounds** counts negotiation turns before a final decision. Additionally, we adopted certain additional metrics specific to each game:

**Ultimatum Game: Initial Offer** represents the average amount Player 1 offers to Player 2.

**Buy-Sell Game: Buyer Advantage** is defined as the difference of the maximum amount the buyer is willing to pay and the actual trade price. **Seller Advantage** is defined as the difference between the actual trade price and the minimum amount the seller is willing to sell at.

**Resource Exchange Game: Trade Volume** measures the number of resources that have exchanged hands.

For each behavior, metrics were aggregated across all ordered model combinations using raw game data: rates were calculated from total counts, while payoffs, offers, and rounds were computed as means and standard deviations from concatenated arrays of individual outcomes.

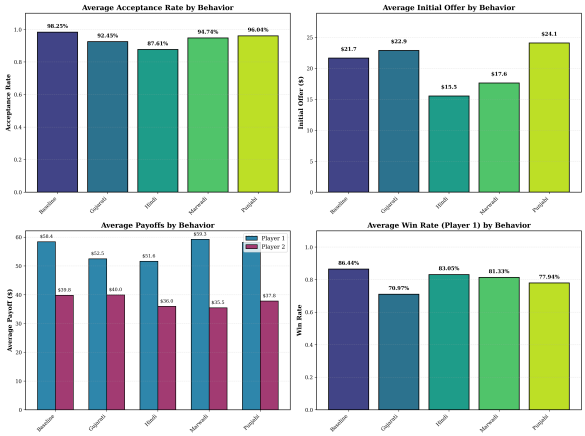


Figure 1: Ultimatum Game Language Comparison showing average (a) acceptance rates, (b) average initial offer, (c) payoffs, (d) win rates (player 1).

## 5 Results and Analysis

We analyze the results obtained in all three games. We compare the Baseline English condition with multiple Indian language contexts (Gujarati, Hindi, Marwadi, Punjabi).

### 5.1 Ultimatum Game Results

The overall quantitative differences in acceptance rate, initial offers, payoffs, and win rates are summarized in Fig. 1. The conversation rounds have been summarised in Table 1.

#### 5.1.1 Baseline Language (English)

The Baseline English condition exhibits highly stable and cooperative negotiation dynamics. It achieves the highest acceptance rate (98.2%) with moderate generosity (average initial offer of $21.67). Player 1 earns $58.42 on average, while Player 2 receives $39.82, indicating a relatively balanced yet Player 1–favored outcome. Conversations remain short (2.58 rounds on average), suggesting efficient agreement formation. This cooperative baseline **supports P1**: English reflects Western fairness norms.

#### 5.1.2 Cross-Language Outcome Variation

As shown in Fig. 1, introducing language and cultural identity significantly alters negotiation outcomes. (1) **Acceptance and Cooperation.** Acceptance rates drop most sharply in Hindi (87.6%), followed by Gujarati (92.4%). Punjabi (96%) and Marwadi (94.7%) remain closer to English Baseline. (2) **Payoff Balance and Efficiency.** Player 1 performs best in Marwadi (59.3), followed closely by English Baseline, while Gujarati and Hindi yield

| Language | Acceptance Rate | Initial Offer | P1 Payoff | P2 Payoff | P1 Win Rate | Conversation Rounds |
|---|---|---|---|---|---|---|
| Baseline | **98.25% ± 1.23%** | 21.67 ± 23.87 | 58.42 ± 16.33 | 39.82 ± 15.30 | **86.44% ± 4.46%** | 2.58 ± 1.04 |
| Gujarati | 92.45% ± 2.57% | 22.89 ± 26.88 | 52.49 ± 24.90 | **39.96 ± 22.92** | 70.97% ± 5.76% | 2.99 ± 1.31 |
| Hindi | 87.61% ± 3.10% | 15.53 ± 23.78 | 51.64 ± 25.79 | 35.97 ± 21.71 | 83.05% ± 4.88% | 2.93 ± 1.29 |
| Marwadi | 94.74% ± 2.09% | 17.64 ± 24.08 | **59.27 ± 22.30** | 35.46 ± 19.28 | 81.33% ± 4.50% | **3.32 ± 1.34** |
| Punjabi | 96.04% ± 1.94% | **24.11 ± 25.16** | 58.22 ± 22.06 | 37.82 ± 20.14 | 77.94% ± 5.03% | 3.15 ± 1.37 |

Table 1: Performance metrics for **Ultimatum Game** aggregated across all model combinations (mean ± std).

substantially weaker Player 1 outcomes ($\sim 51.6$). Player 2 outcomes are highest in Gujarati ($40) but much lower in Hindi (36). (3) **Deviation from English Baseline.** Gujarati and Hindi shift negotiations toward instability and lower fairness, whereas Punjabi and Marwadi retain cooperative structure while modifying strategic strength.

#### 5.1.3 Evaluating Predictions

These patterns **partially contradict P1–P2**: Rather than increased cooperation, Hindi (87.2%) and Gujarati (92.0%) show *reduced* acceptance versus English (98.2%). Indic languages seem to introduce instability, not collectivist harmony.

However, **P2 receives partial support** in offer behavior: Hindi's lower initial offers ($15.5 vs. $21.7) align with pragmatic indirectness predictions. The paradox of lower offers *and* lower acceptance suggests Hindi activates defensive strategies in *both* players, creating mismatch rather than cooperation.

**P3 (Stereotype) strongly supported**: Marwadi achieves highest Player 1 payoff ($59.4) with maintained acceptance (94.7%), precisely matching predictions of strategically advantageous trader behavior. This effect persists across model pairs, indicating stereotype activation from shared training data.

#### 5.1.4 Language-Specific Dynamics

**Gujarati.** Gujarati behavior features relatively high initial offers ($22.9) but surprisingly lower acceptance and weaker Player 1 advantage. Win rates are noticeably reduced (70.9%), indicating indecisive or unstable bargaining. This pattern goes against P1 (relational harmony), suggesting Gujarati framing introduces uncertainty undermining stable equilibria.

**Hindi.** Hindi produces the most adversarial dynamics: lowest acceptance, lowest initial offers ($15.5), and suppressed Player 2 payoff. However, Player 1 still maintains a strong win rate, suggesting competitive rather than cooperative bargaining.

**Punjabi.** Punjabi maintains high acceptance (96%) with the most generous offers ($24.1). Conversations are slightly longer, implying more active bargaining rather than breakdowns, while producing cooperative outcomes. Punjabi **partially supports P3**: maintaining cooperation while enabling active bargaining, consistent with cultural representations of direct yet warm Punjabi communication.

**Marwadi.** Marwadi yields the most strategically advantageous Player 1 condition: high acceptance, strong payoff ($59.4), and moderately longer discussions. Compared to Baseline, Marwadi shifts agents toward disciplined but still cooperative strategic negotiation. **Strongest P3 support**: Marwadi induces exactly the disciplined, advantageous negotiation matching trader stereotypes in Indian media and commerce. Cross-model consistency indicates stereotype activation, not model artifacts.

### 5.2 Buy-Sell Game Results

#### 5.2.1 Baseline Language (English)

As shown in Figure 2 and Table 2, English yields a high acceptance rate (97.44%) but relatively low seller advantage (mean 6.9), coupled with the highest buyer advantage (mean 13.1). Seller win rate remains modest at 41.98%.

This pattern indicates that when negotiating in English, LLM agents tend to favor agreement stability over aggressive surplus extraction by the seller. Heatmap-level analysis further reveals pronounced asymmetry across model pairings: stronger models such as GPT–4o consistently secure large positive seller advantage, while weaker models (notably GPT–3.5) frequently incur negative seller advantage, effectively transferring surplus to the buyer. This buyer favoritism **supports P1**: English training data encodes Western consumer-centric scripts where "getting a good deal" is prioritized, disadvantaging the seller role.

#### 5.2.2 Cross-Language Outcome Variation

All non-English languages achieve near-perfect acceptance rates, with Hindi and Punjabi reaching 100% agreement. However, these high acceptance rates coincide with substantially higher seller ad-

| Language | Acceptance Rate | Seller Advantage | Buyer Advantage | Conversation Rounds | Player 1 Win Rate |
|---|---|---|---|---|---|
| English | 97.44% ± 15.87% | 6.89 ± 12.44 | **13.11 ± 12.44** | 3.21 ± 1.91 | 41.98% |
| Gujarati | 98.21% ± 13.30% | 8.44 ± 9.88 | 11.56 ± 9.88 | 3.34 ± 1.88 | 37.33% |
| Hindi | **100.00% ± 0.00%** | 7.49 ± 8.44 | 12.51 ± 8.44 | 2.99 ± 1.46 | 32.14% |
| Marwadi | 98.23% ± 13.24% | **12.32 ± 12.41** | 7.68 ± 12.41 | **3.78 ± 1.91** | **60.47%** |
| Punjabi | **100.00% ± 0.00%** | 11.14 ± 11.42 | 8.86 ± 11.42 | 3.38 ± 1.83 | 50.00% |

Table 2: Performance metrics for **Buy-Sell Game** aggregated across all model combinations (mean ± std).

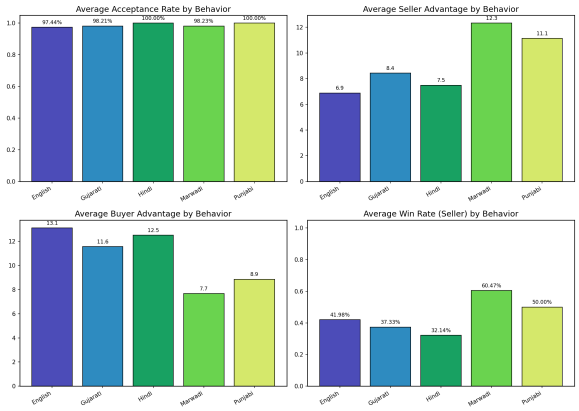


Figure 2: Buy Sell Game Language Comparison showing average (a) acceptance rates, (b) seller advantages, (c) buyer advantages, and (d) win rates across different cultural behaviors.

vantage than English. Marwadi exhibits the largest average seller advantage (12.3), followed by Punjabi (11.1), Gujarati (8.4), and Hindi (7.5).

At the same time, buyer advantage decreases sharply in non-English settings, most notably in Marwadi (mean 7.7), indicating a redistribution of surplus toward the seller role. Seller win rates also increase substantially, peaking at 60.47% in Marwadi.

#### 5.2.3 Evaluating Predictions

**P1 (Cultural Scripts) strongly supported**: English→buyer bias (advantage: 13.1) inverts in Indic languages→seller bias (Marwadi: 12.3, Punjabi: 11.1). This reversal reflects training data composition—English corpora over-represent Western consumer negotiations, while Indic commercial texts encode bazaar dynamics respecting seller expertise.

**P3 (Stereotype) receives strongest support**: Marwadi exhibits maximum seller advantage (12.3), minimum buyer advantage (7.7), and highest seller win rate (60.47%), precisely matching trader community stereotype predictions.

**P6 (Representation) supported**: Simplified Buy-Sell structure exposes systematic linguistic biases invisible in complex games, demonstrating that reduced strategic complexity makes language-conditioned priors dominant.

#### 5.2.4 Language-Specific Dynamics

Each language induces a distinct negotiation regime. Gujarati produces moderately elevated seller advantage while maintaining balanced buyer outcomes, suggesting relatively symmetric bargaining dynamics. Hindi displays an interesting decoupling: despite perfect acceptance, seller advantage remains moderate, and buyer advantage stays comparatively high, yielding the lowest seller win rate (32.14%). This indicates that Hindi-language negotiations encourage agreement without strongly favoring the seller, **indicating support of P2 (Pragmatism)**. Hindi **partially supports P1** (cooperation): perfect acceptance prioritizes agreement, while balanced advantages suggest collectivist leveling despite role asymmetries.

Marwadi stands out as the most seller-favorable language. Seller advantage is highest, buyer advantage is lowest, and seller win rate exceeds 60%. Heatmaps show that this pattern holds across most model pairings, indicating that the effect is not driven by a single architecture. **Clearest P3 manifestation across all games**: Marwadi doesn't just shift outcomes, instead, it completely reverses English baseline, converting buyer-favorable to strongly seller-favorable across all models. This cannot be explained linguistically (Marwadi is similar to Hindi) and directly reflects stereotype-driven behavioral scripts from training data. Punjabi similarly favors sellers, though less extremely, producing high seller advantage with relatively balanced outcomes across models.

### 5.3 Resource Exchange Game Results

#### 5.3.1 Baseline Language (English)

As shown in Figure 3 and Table 3, English yields the highest average acceptance rate (95.9%) but the lowest average trade volume (16.6), indicating that LLM agents readily reach agreement but tend to

settle on comparatively conservative exchanges. Payoff distributions in English exhibit a mild but consistent asymmetry: Player 2 achieves a higher average payoff (30.9) than Player 1 (29.1), and Player 1 wins only 21.6% of games. This suggests that, when operating in English, LLMs prioritize agreement over aggressive value extraction. This conservative pattern **supports P1**: English prioritizes "safe agreement" over exploratory exchange, consistent with transactional Western negotiation framing that minimizes risk.

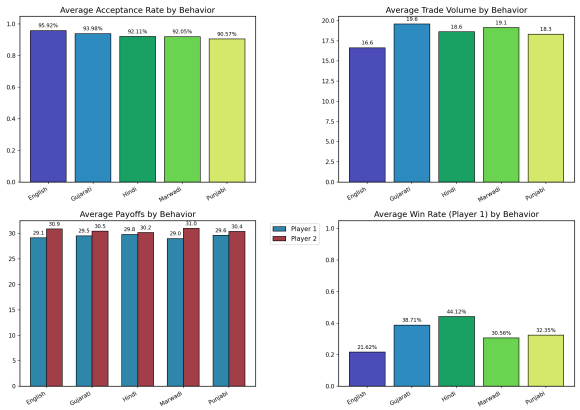


Figure 3: Resource Exchange Game Language Comparison showing average (a) acceptance rates, (b) trade volume, (c) payoffs, and (d) win rates (player 1) across different cultural behaviors.

#### 5.3.2 Cross-Language Outcome Variation

All non-English languages exhibit lower acceptance rates than English, with Punjabi reaching the minimum (90.6%). However, this reduction in agreement probability is accompanied by a consistent increase in average trade volume. Gujarati and Marwadi achieve the highest trade volumes (19.57 and 19.13, respectively), indicating deeper and more extensive exchanges.

This inverse relationship between acceptance rate and trade volume suggests that linguistic context influences how LLMs explore the negotiation space. Rather than degrading performance, non-English languages appear to induce longer or richer bargaining trajectories that trade off agreement certainty for higher exchange complexity.

#### 5.3.3 Evaluating Predictions

The inverse acceptance-volume relationship **supports P1 (Relational Exchange)**: Indic languages shift from English's "safe agreement" to "rich exploration" strategy. Gujarati and Marwadi achieving 18–20% higher trade volumes demonstrates that relational framings prime LLMs to explore integrative solutions rather than settle quickly.

**P4 (Task Contingency) supported**: Unlike distributive games where Indic languages underperformed, integrative complexity makes linguistic effects *beneficial*. This task-dependent pattern reveals that cultural scripts activate differentially based on negotiation structure.

#### 5.3.4 Language-Specific Dynamics

Each language exhibits a distinct negotiation profile when used as the interaction medium for LLM agents. Gujarati balances high trade volume with relatively strong proposer outcomes, yielding a Player 1 win rate of 38.7%. Hindi stands out as the most advantageous setting for Player 1, achieving the highest win rate (44.1%) and the highest average Player 1 payoff (29.8), despite having a lower acceptance rate than English.

Marwadi displays a markedly different pattern. While trade volume remains high, Player 1 receives the lowest average payoff across all languages (29.00), whereas Player 2 achieves the highest (31.00). This consistent asymmetry suggests that LLMs negotiating under Marwadi framing are more likely to accept outcomes unfavorable to the proposer. Marwadi's Player 1 disadvantage **contradicts P3** - the only game where trader stereotypes fail to benefit the proposer. This may indicate that stereotypical "Marwadi trader" scripts emphasize distributive (zero-sum) rather than integrative (win-win) bargaining, leading to suboptimal exploration of joint gains. Punjabi occupies an intermediate, exhibiting moderate trade volume and balanced payoffs without strong advantage to either party.

Hindi's strong Player 1 performance **contradicts P1** (collectivist balance) but reveals context-dependent script activation: Hindi framing enables assertiveness when complexity allows strategic depth, unlike the defensiveness in simpler Ultimatum games. This suggests **task-contingent cultural priming**.

### 5.4 Model-Specific Performance

All model-specific results have been reported in heatmaps in Appendix A. In the Ultimatum Game, GPT-3.5 shows severe performance degradation: as Player 1 against GPT-4o, it achieves only 44 payoff in English, compared to $\sim 56$ for stronger models. This pattern intensifies in non-English settings as Hindi yields 39 payoff for GPT-3.5 as Player 1 against GPT-4o, while Claude models maintain $\sim$

| Language | Acceptance Rate (%) | Trade Volume | P1 Payoff | P2 Payoff | P1 Win Rate (%) | Conversation Rounds |
|---|---|---|---|---|---|---|
| English | **95.92 ± 19.89** | 16.63 ± 5.65 | 29.13 ± 2.28 | 30.87 ± 2.28 | 21.62 | 3.04 ± 1.35 |
| Gujarati | 93.98 ± 23.94 | **19.57 ± 7.61** | 29.54 ± 2.94 | 30.46 ± 2.94 | 38.71 | 3.18 ± 1.36 |
| Hindi | 92.11 ± 27.14 | 18.62 ± 7.05 | 29.82 ± 3.11 | 30.18 ± 3.11 | **44.12** | **3.34 ± 1.44** |
| Marwadi | 92.05 ± 27.21 | 19.13 ± 7.72 | 29.00 ± 2.67 | **31.00 ± 2.67** | 30.56 | 3.48 ± 1.49 |
| Punjabi | 90.57 ± 29.37 | 18.32 ± 5.85 | **29.62 ± 2.13** | 30.38 ± 2.13 | 32.35 | 3.12 ± 1.47 |

Table 3: Performance metrics for **Resource Exchange Game** aggregated across all model combinations (mean ± std).

54 range. Similarly, the effect is highly pronounced in Gujarati where GPT-3.5 as Player 1 against GPT-4o provides average payoffs of 26, compared to a range to 42-56 for Claude models. A similar pattern follows for Marwadi and Punjabi. This observation is in line with those reported in Bianchi et al. (2024), where GPT-3.5 is reported to regularly fail during different scenarios. However, GPT-3.5 tends to provide much higher payoffs than other models when acting as Player 1 against Claude-3.5-Haiku in Gujarati and Hindi. Claude-3.5-Haiku consistently provides maximum average payoffs as Player 1 against Claude-3-Haiku.

The Buy-Sell Game exposes large role-dependent asymmetries that interact with language. In English, GPT-4o as seller (Player 1) achieves 19.3-20.5 advantage, while GPT-3.5 in the same role suffers -7.5 to -12.1 negative seller advantage with a gap of over 30 points. Conversely, as buyer (Player 2), GPT-3.5 secures extreme advantages of 27.5-32.1 in English, indicating systematic over-concession as seller and over-extraction as buyer. Meanwhile, GPT-4o provides negligible buyer advantages as Player 1 and contrarily consistently provides the best seller advantages as Player 1. Indic languages partially constrain these extremes: in Marwadi, GPT-3.5's seller advantage improves to -0.8 to 2.1, while buyer advantage drops to 17.9-21.2. However, the model hierarchy persists as GPT-4o maintains 17.0-25.6 seller advantage in Marwadi, demonstrating that linguistic framing attenuates but does not eliminate capacity-driven asymmetries. GPT-4o continues to provide largest seller advantages as Player 1 across all languages, while GPT-3.5 provides the largest buyer advantages as Player 1 across all languages.

In the Resource Exchange game with the English baseline, when GPT-4o serves as Player 1, it achieves consistent payoffs (29.2-30.0) regardless of Player 2 opponent. In contrast, GPT-3.5 as Player 1 shows opponent-dependent variance: 29.0-29.2 against most models but drops to 28.5 when facing Claude-3.5-Haiku as Player 2. This pattern intensifies in non-English settings—in Hindi, GPT-4o as Player 1 drops to 25.0 against GPT-3.5 while providing above that average (31.4 and 30.4) payoffs against Claude models. In Gujarati, Claude-3.5-Haiku as Player 1 achieves 33.5 against GPT-3.5 as Player 2 but only 28.1 against Claude-3-Haiku, revealing opponent-dependent adaptation. GPT-4o as Player 1 maintains more uniform performance (30.0-30.9) across different Player 2 opponents. Marwadi shows pronounced hierarchy: GPT-4o as Player 1 secures 32.0 against GPT-3.5 as Player 2. In English, Gujarati, and Marwadi, Player 2 tends to get better average payoffs than player 1 across model combinations.

## 6 Conclusion

This work demonstrates that language functions as a latent policy prior in LLM negotiation, reshaping strategic behavior independent of model architecture or task structure. Through simulations across three settings, we show that language choice can shift outcomes more strongly than changing the underlying model itself—reversing proposer advantages in Buy-Sell, reducing stability in distributive games, and altering exploration patterns in integrative settings. These effects are task-contingent: English optimizes stability in distributive games but constrains integrative exploration, while Indic languages exhibit the inverse pattern. Marwadi completely reversing English baseline outcomes in Buy-Sell games across all model pairs demonstrates that cultural stereotypes can dominate task-level reasoning. As LLMs deploy globally in commercial and interpersonal contexts, our findings underscore the urgent need for multilingual evaluation frameworks that account for language as an active component of strategic reasoning, with direct implications for fairness and equitable deployment.

## 7 Limitations

Our findings should be interpreted carefully. First, we emphasize that the behaviors exhibited by LLM

agents in our experiments do *not* constitute evidence about real human negotiation practices, cultural norms, or linguistic communities. Differences reflect patterns learned from training corpora, not properties of languages or their speakers (Bolukbasi et al. (2016)). This approach treats language as a window into training data composition and learned behavioral priors, not as a proxy for real-world cultural groups. Second, our language framings use culturally associated labels (Hindi, Marwadi) without incorporating human participants or sociolinguistic context. As a result, any apparent alignment with commonly held stereotypes should be understood as an artifact of representation learning and data distribution, analogous to well-documented biases in word embeddings and language models. We deliberately avoid normative claims and do not endorse any interpretation that attributes these behaviors to real-world groups. Third, our analysis covers limited games and languages. While spanning distributive and integrative settings, these games lack the richness of real-world negotiation (long-term relationships, incomplete information). Fourth, we examine model-model interaction, whereas human-AI dynamics may differ. We lack training data access to test representation hypotheses directly. Finally, we isolate language identity by fixing prompts and incentives, abstracting from realistic code-switching and dynamic strategy adaptation. Despite these limitations, our results provide valuable evidence that language-conditioned representations influence strategic interaction in LLMs, underscoring the need for multilingual evaluation in socially sensitive domains.

## 8 Acknowledgments

We thank the creators of Negotiation Arena (Bianchi et al., 2024) for making their setup publicly available. We are also grateful for access to API platforms that made this evaluation possible. The authors also wish to acknowledge the usage of ChatGPT and Claude in improving the presentation and grammar of the paper. The paper remains an accurate representation of the authors' underlying contributions.

## References

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiation-arena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Jeanne M Brett. 2007. *Negotiating globally: How to negotiate deals, resolve disputes, and make decisions across cultural boundaries*. John Wiley & Sons.

Myke C. Cohen, Zhe Su, Hsien-Te Kao, Daniel Nguyen, Spencer Lynch, Maarten Sap, and Svitlana Volkova. 2025. Exploring big five personality and ai capability effects in llm-simulated negotiation dialogues. *Preprint*, arXiv:2506.15928.

Krishno Dey, Prerona Tarannum, Md. Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *Preprint*, arXiv:2410.13153.

Edward T Hall. 1976. *Beyond culture*. Anchor.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2023. Language of bargaining. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13161–13185, Toronto, Canada. Association for Computational Linguistics.

Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024. Assistive large language model agents for socially-aware negotiation dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8047–8074, Miami, Florida, USA. Association for Computational Linguistics.

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

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

mingyu jeon and Jae Young Suh. 2024. Mimicking human emotions: Persona-driven behavior of LLMs in the 'buy and sell' negotiation game. In *Language Gamification - NeurIPS 2024 Workshop*.

Ryan Shea, Aymen Kallala, Xin Lucy Liu, Michael W. Morris, and Zhou Yu. 2024. ACE: A LLM-based negotiation coaching system. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12720–12749, Miami, Florida, USA. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Thomas A Timberg. 1978. *The Marwaris: From traders to industrialists*. Vikas Publishing House.

Michelle Vaccaro, Michael Caosun, Harang Ju, Sinan Aral, and Jared R. Curhan. 2025. Advancing ai negotiations: New theory and evidence from a large-scale autonomous negotiations competition. *Preprint*, arXiv:2503.06416.

# A Visualizations

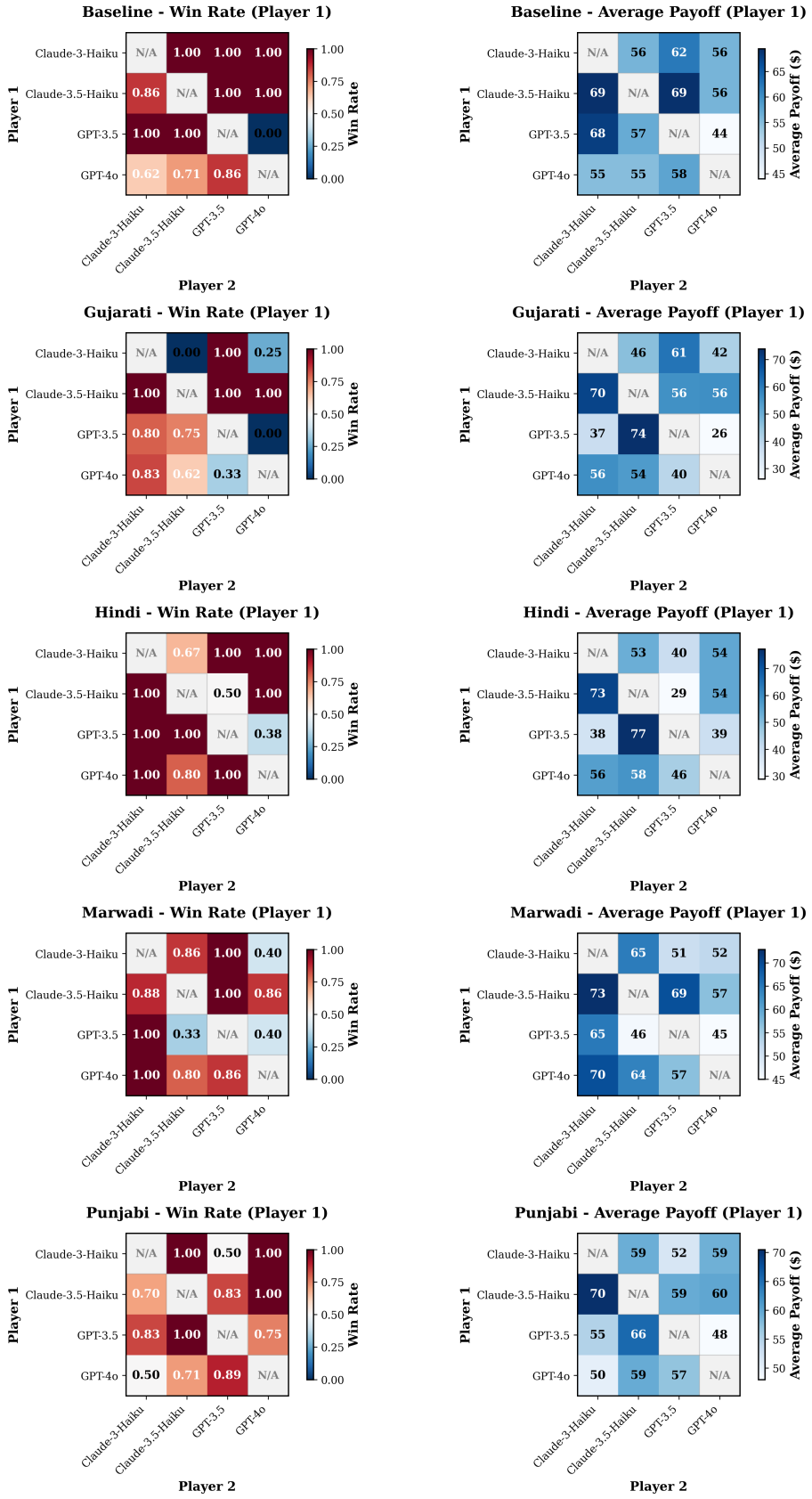


Figure 4: Heatmaps for **Ultimatum Game** comparing (a) win rate and (b) payoff for model combinations for all languages.

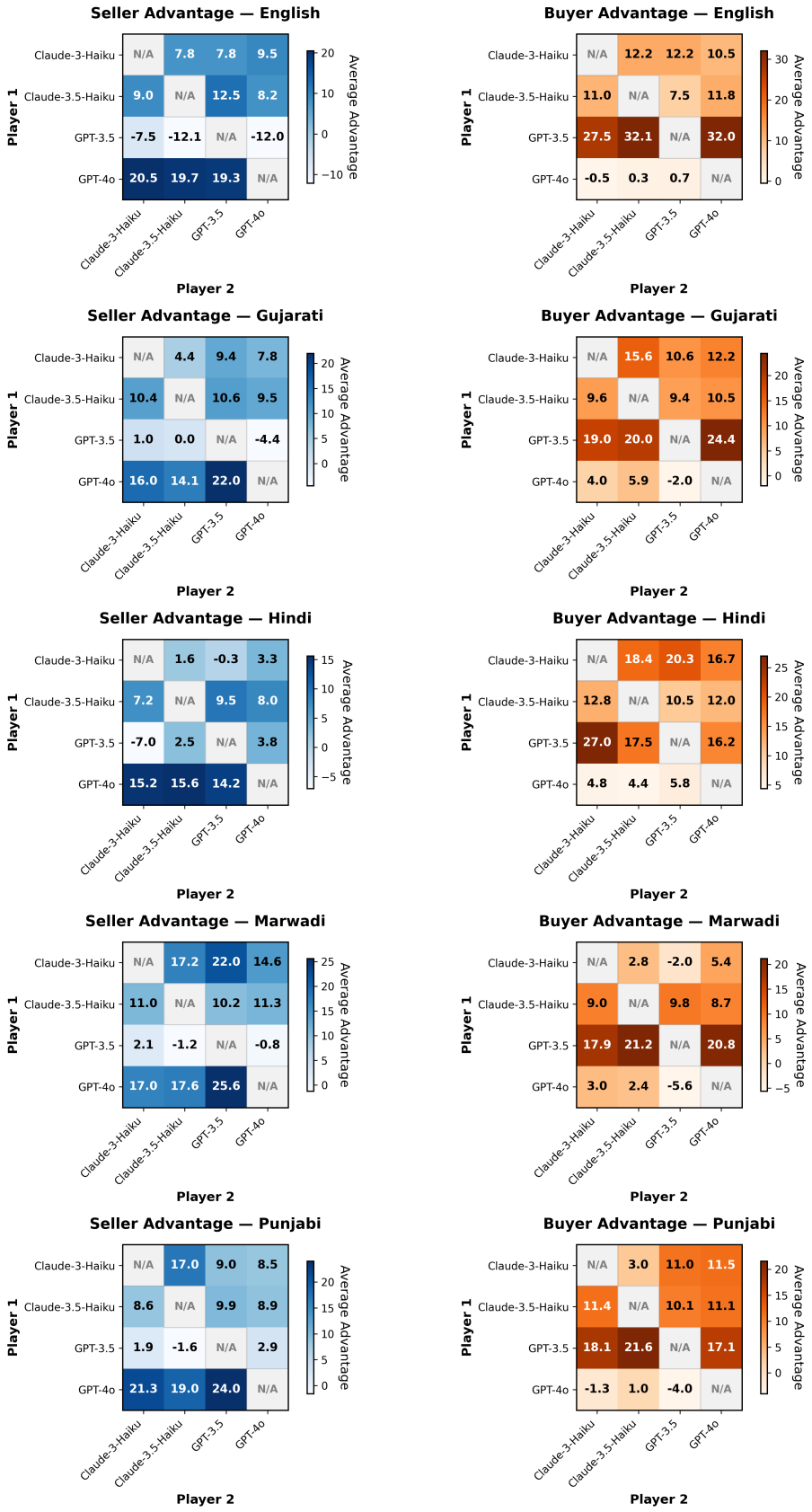


Figure 5: Heatmaps for **Buy-Sell Game** comparing (a) Seller advantage and (b) Buyer Advantage for model combinations for all languages.

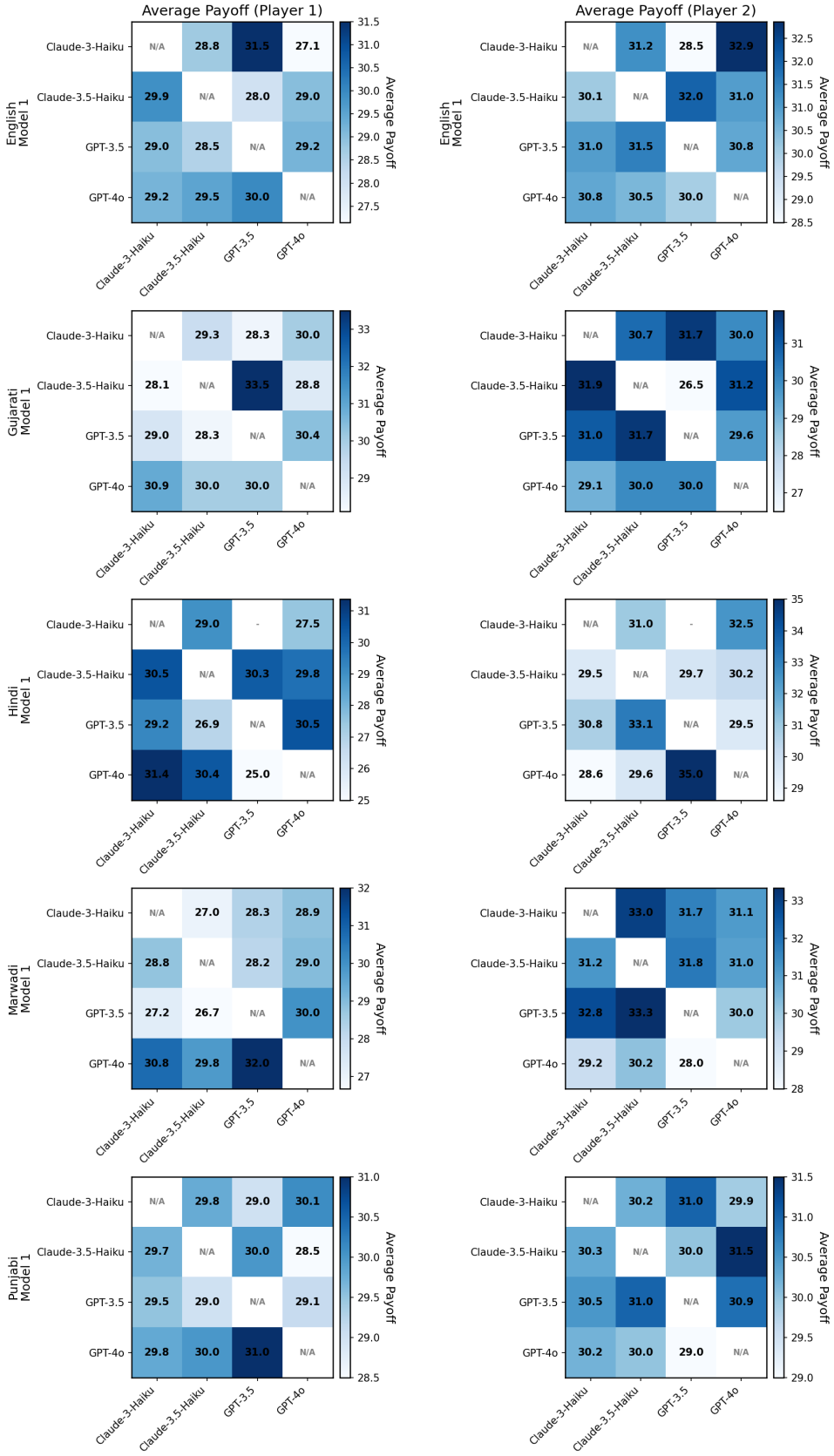


Figure 6: Heatmaps for **Resource Exchange Game** comparing average payoffs for player 1 and player 2 for model combinations for all languages.