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. We systematically isolate language effects across English and three Indic framings (Hindi, Punjabi, Gujarati) by holding game rules, model parameters, and incentives constant for multi-agent simulations for Ultimatum, Buy-Sell, and Resource Exchange games. Our results indicate that language choice is correlated to negotiation outcomes and that this correlation persists across model architectures. Crucially, effects are task-contingent: Indic languages reduce stability in distributive games (where agents compete over a fixed resource) yet are associated with richer exploration in integrative (cooperative) settings. Further, they suggest that findings obtained under English-only conditions may not generalize to other linguistic settings. These findings caution against English-only evaluation of LLMs and suggest that linguistically-grounded 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 (Lewis et al., 2017; He et al., 2018). With the emergence of Large Language Models (LLMs), prior studies demonstrate that LLMs can engage in multi-turn bargaining and achieve non-trivial outcomes in competitive settings (Kwon et al., 2024; Vaccaro et al., 2025). Frameworks such as NegotiationArena (Bianchi et al., 2024) demonstrate that LLMs exhibit human-like negotiation behaviors, including anchoring and concession patterns. Nearly all studies conduct their evaluation exclusively in English, implicitly treating language as a neutral communication channel.

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 the interaction language may systematically shape strategic behavior even when incentives remain fixed. This issue is particularly salient in multilingual contexts, where LLM performance is often framed as a matter of simple degradation (Dey et al., 2024; Singh et al., 2024). However, recent inquiries suggest that language choice may fundamentally alter the structure of interaction (Tam et al., 2025; Wang et al., 2024). We investigate whether negotiation behavior is invariant to language in LLMs. (Hakimov et al., 2025).

To our knowledge, this study is the first to hold game rules, model parameters, and incentives constant while systematically varying linguistic framing across multi-agent negotiation games. We demonstrate that the interaction language acts as a non-trivial strategic prior, introducing variances in outcomes that persist across different model architectures. We demonstrate that a model's strategic intelligence is not a fixed attribute but is qualitatively altered by the language of interaction. Across 4,320 games spanning three canonical settings, we find statistically significant effects. In the **Ultimatum Game**, Gujarati and Punjabi framing significantly reduces acceptance rates relative to English, Punjabi produces significantly lower initial offers, and all Indic conditions yield significantly longer negotiations. In the **Buy-Sell Game**, Punjabi significantly extends negotiation length relative to English. In the **Resource Exchange Game**, all Indic languages significantly increase trade volume over English, while payoffs remain balanced.

These findings show that language is an active component of strategic reasoning rather than a passive

medium. This work establishes that multilingual evaluation is a fundamental requirement for fair and robust AI deployment. We move beyond simple performance degradation narratives to show that language is associated with entirely different behavioral regimes. For developers and policymakers, these findings serve as a necessary caution: an LLM that is cooperative in English may become adversarial or sub-optimal in another language, even when provided with identical instructions.

2 Related Work

Recent work has increasingly challenged the assumption that language is a neutral variable for large language models (LLMs). Beyond basic performance metrics, recent research has highlighted how the choice of language fundamentally alters a model's internal processing and alignment. (Tam et al., 2025) investigate the "English-as-a-hub" phenomenon, demonstrating that large reasoning models often default to English for internal chain-of-thought steps even when the input is in another language, which can lead to a significant performance-alignment trade-off. This linguistic contingency extends to the ethical domain; Agarwal et al. (Agarwal et al., 2024) show that moral value alignment and ethical reasoning in frontier models like GPT-4 are not universal but vary significantly depending on the prompt language. Similarly, the Babel Effect (Wang et al., 2024) demonstrates systematic performance disparities across languages, highlighting the English-centric nature of current LLM training and evaluation. These works establish that language influences reasoning accuracy and consistency, but primarily focus on static tasks.

In parallel, emerging work has begun to explore multilingual effects in interactive settings. (Hakimov et al., 2025) examines how language impacts negotiation efficiency, reasoning cost, and outcome quality. This line of work suggests that multilingual reasoning introduces trade-offs between computational cost and performance. However, prior approaches largely treat language as a factor affecting efficiency or correctness, rather than as a variable that can fundamentally alter interaction.

A separate body of literature studies LLMs as negotiators. Subsequent work shows that cooperation, agreeableness, and persona conditioning significantly influence outcomes (Kwon et al., 2024; Vaccaro et al., 2025). Persona-based studies further highlight that lightweight contextual signals can strongly modulate negotiation strategies (Jeon and Suh, 2024; Cohen et al., 2025). However, these studies are predominantly conducted in English, implicitly treating language as invariant.

Finally, multilingual evaluation studies document persistent performance gaps across languages, particularly in low-resource settings (Dey et al., 2024; Singh et al., 2024). While this literature frames multilingual differences primarily as bias or degradation, it does not examine whether language induces qualitatively different behaviors in interactive tasks.

Unlike prior work that focuses on reasoning accuracy or efficiency, we show that linguistic framing is a non-trivial factor influencing negotiation dynamics in LLMs, with effects that vary across task structure. Specifically, by holding models, incentives, and game structure constant, we present a controlled experimental framework to study the effect of interaction language in multi-agent LLM negotiation across three canonical games. This positions our work at the intersection of multilingual reasoning and multi-agent interaction, highlighting the need for evaluation frameworks that treat language as an active component of strategic reasoning rather than a passive medium.

3 Hypotheses

We generate two testable hypotheses and test them examining where LLM behavior aligns with or deviates from theory:

H1 (Language-Mediated Strategy): LLM negotiations are not language agnostic; instead, the choice of language systematically reshapes strategic behavior, social norm adherence, and equilibrium outcomes.

H2 (Task Contingency): Effects vary by game structure between distributive tasks (Ultimatum, Buy-Sell) vs integrative tasks (Resource Exchange).

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: P1 is the seller with a minimum acceptable price, and P2 is the buyer with a maximum willingness to pay.

Ultimatum Game: An asymmetric power negotiation game. P1 proposes a division of a fixed resource pool (e.g., 100 units). P2 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. P1 initiates the first offer.

4.1 System Prompts and Persona Design

We design system prompts that assign each agent a specific linguistic identity. Our three primary linguistic framings are: Hindi, Gujarati, Punjabi. All prompts explicitly forbid internal chain-of-thought, requiring only short rationale summaries, following the experimental setup of (Bianchi et al., 2024). This constraint is necessary to isolate the model's immediate behavioral outputs and prevent the introduction of an uncontrolled variable: reasoning traces which could mask the direct influence of linguistic framing on decision-making. The persona prompts are: "You speak and bargain only in [language]. Negotiate accordingly." The choice to write system prompts in English was made to maintain consistency with the NegotiationArena framework (Bianchi et al., 2024), which uses English-language instructions. We also run the games without any lingual 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=0.7 and sampling settings are held constant across all linguistic conditions. Each game is repeated thirty 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 ($A \neq B$) for four languages. Model A and Model B simply refer to assigning a Model A to Player 1(P1) and Model B to Player 2(P2). All ordered pairs of models were chosen. Each run logs: full dialogue, parsed offers or resource splits, final utilities, agreement/acceptance decisions. All combinations of experiments were run across thirty runs, with standardized logging of dialogues and offers. Total experiments run $= 4(models) \times 3(othermodels) \times 4(languages) \times 30(runs) \times 3(games) = 4320$.

4.4 Evaluation Principles and Metrics

Following prior work on computational negotiation and multi-agent bargaining evaluation that emphasize that evaluating bargaining systems requires measuring not only whether agreements occur, but also how value is allocated and how negotiation unfolds over time, our evaluation metrics are grouped into three categories: **(i) Outcome Stability Metrics**, which measure whether negotiations successfully reach agreements *(Acceptance Rate)*. **(ii) Value Distribution Metrics**, which capture how negotiated resources are allocated between agents *(Player Payoffs, Win Rate, Buyer/Seller Advantage)*. **(iii) Interaction Dynamics Metrics**, which characterize the negotiation process and strategic exploration *(Conversation Rounds, Trade Volume)*.

We adopt the four following objective negotiation metrics across all the three games: **Acceptance Rate** measures the proportion of proposals accepted by P2. **Player Payoffs** capture final resource allocation for each player, summing all resources including exchanged items. **Win Rate (P1)** is the ratio of P1 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 adopt certain additional metrics specific to each game:

Ultimatum Game: Initial Offer represents the average amount P1 offers to P2.

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.

4.5 Statistical Testing

To assess whether language behavior significantly affects negotiation outcomes, we apply non-parametric statistical tests across all language conditions. For continuous metrics (trade volume, payoffs, and negotiation rounds), we use the Kruskal-Wallis H-test for overall differences, followed by pairwise Mann-Whitney U tests. For binary outcomes (acceptance rate and win rate), we use chi-square tests with pairwise two-proportion z-tests. All pairwise p-values are corrected using the Benjamini-Hochberg false discovery rate procedure. We report significance at three levels: $^{*}(p_{\text{corr}} < 0.05)$, $^{**}(p_{\text{corr}} < 0.01)$, $^{***}(p_{\text{corr}} < 0.001)$, with non-significant results denoted *ns*.

4.6 Language Compliance Analysis

To verify linguistic compliance, we applied the AI4Bharat IndicNLP language identification model (Madhani et al., 2023), which uses FastText embeddings pretrained on a large-scale monolingual Indic corpus covering 22 Indian languages, to all conversation turns across all experimental conditions. We report adherence rates and average confidence scores in Appendix C. Overall, adherence rates are high across all verified languages and games (89.65%–97.85%), with near-perfect confidence scores ($\geq$0.966), indicating that models reliably produced output in the instructed language.

5 Results and Analysis

We compare the Baseline English condition with multiple Indian language contexts (Gujarati, Hindi, Punjabi) in all three games. All results are supported by non-parametric statistical testing (Kruskal-Wallis, Mann-Whitney U, and chi-square tests with Benjamini-Hochberg correction); full test statistics and p-values are reported in Appendix B.

5.1 Ultimatum Game Results

Results are shown in Figure 1 and Table 1.

5.1.1 Baseline Language (English)

The English condition exhibits stable and efficient negotiation dynamics. It achieves a high acceptance rate (93.1%), with an average initial offer of 43.31 ZUP. P1 earns 62.09 ZUP on average, while P2 receives 36.25 ZUP, indicating a moderately P1-favored outcome. Conversations remain short (2.55 rounds on average).

5.1.2 Multilingual Outcome Variation

(1) **Acceptance and Cooperation.** Acceptance rates decrease relative to the English baseline in all other languages: Gujarati (84.4%), Hindi (88.3%), and Punjabi (86.7%). The reductions for Gujarati ($p_{\text{corr}} = 0.0015^{**}$) and Punjabi ($p_{\text{corr}} = 0.0135^{*}$) are statistically significant, while the Hindi difference does not remain significant after correction.
(2) **Payoff Balance and Efficiency.** P1 payoff remains broadly comparable across languages, with no statistically significant differences relative to the baseline after correction. A notable exception is a significant difference between Gujarati and Punjabi ($p_{\text{corr}} = 0.0406^{*}$), where Punjabi yields higher P1 payoff. P2 payoff does not exhibit statistically significant variation across conditions.
(3) **Deviation from English Baseline.** The most pronounced deviations from the baseline arise in initial offer behavior and interaction length. Punjabi produces substantially lower initial offers (mean 37.96), significantly below English, Gujarati, and Hindi. Additionally, all non-English conditions result in longer negotiations, with significantly more turns than the baseline, indicating increased interaction before agreement.

5.1.3 Evaluating Hypotheses

The results provide clear evidence in support of H1 (Language-Mediated Strategy). Several outcome dimensions vary systematically with language. Acceptance rates are significantly lower in Gujarati and Punjabi relative to the English baseline ($p_{\text{corr}} = 0.0015^{**}$ and 0.0135^{*}, respectively), while Punjabi exhibits substantially lower initial offers compared to all other conditions (all pairwise $p_{\text{corr}} < 0.001^{***}$). In addition, all non-English conditions show significantly longer negotiations (all $p_{\text{corr}} < 0.001^{***}$).

5.1.4 Evaluating Prompt Sensitivity

To verify that our findings are not an artifact of the specific persona prompt wording, we replicate the Ultimatum Game under three variants of the prompts: two semantically equivalent phrasings of the original prompt (A1–A2) and a native-script prompt where the persona directive is written directly in the target language (A3), keeping all other conditions constant. While absolute values vary only modestly across variants, the relative behavior replicates consistently: acceptance rates remained high and stable under English (91.9–93.3%), reduced acceptance rates in Gujarati and Punjabi,

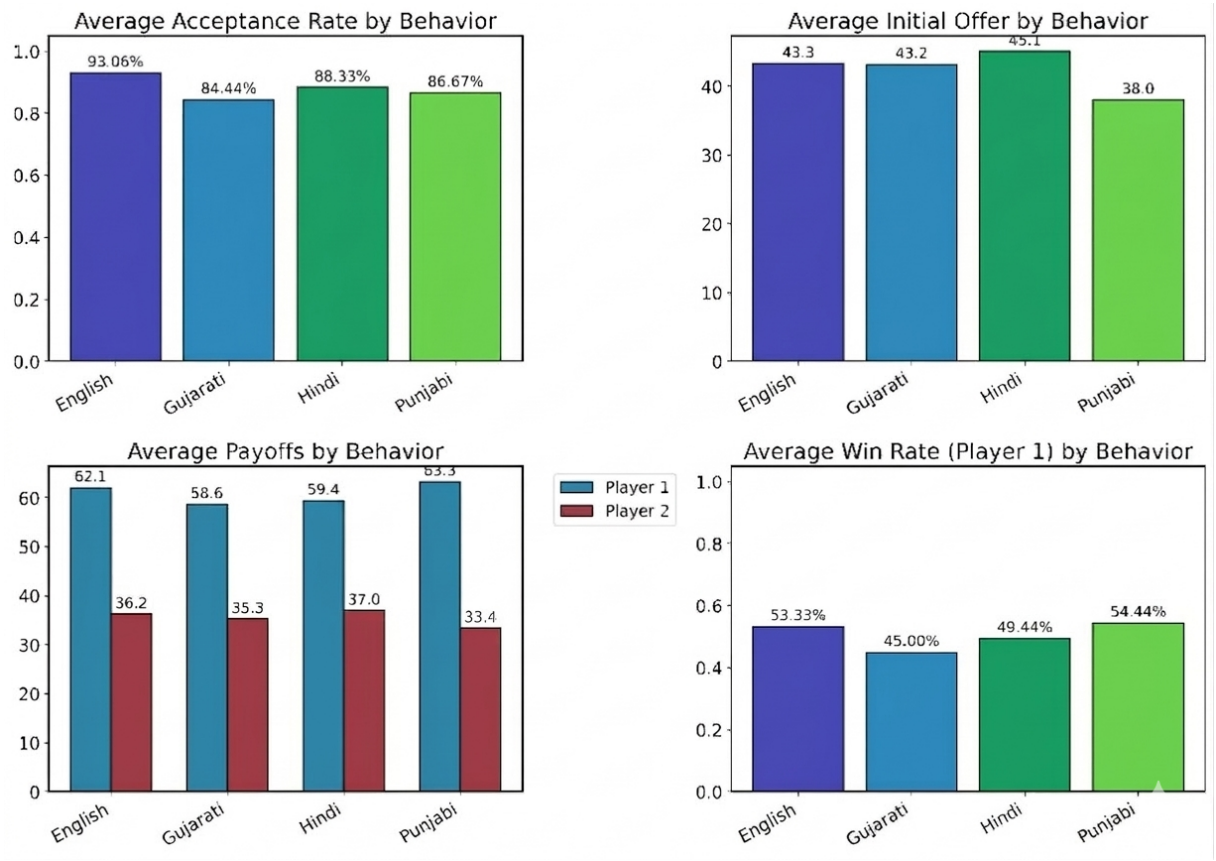

Figure 1: Ultimatum Game Language Comparison showing average (a) acceptance rates, (b) average initial offer, (c) payoffs, (d) win rates (P1).

Language	Acceptance Rate	Initial Offer	P1 Payoff	P2 Payoff	P1 Win Rate	Conversation Rounds
English	**93.06% ± 25.42%**	43.31 ± 11.26	62.09 ± 19.57	36.25 ± 18.44	53.33% ± 49.89%	**2.55 ± 1.03**
Gujarati	84.44% ± 36.24%	43.20 ± 13.59	58.64 ± 26.48	35.25 ± 23.63	45.00% ± 49.75%	3.10 ± 1.41
Hindi	88.33% ± 32.10%	**45.08 ± 13.51**	59.42 ± 22.36	**36.97 ± 20.47**	49.44% ± 50.00%	3.04 ± 1.37
Punjabi	86.67% ± 33.99%	37.96 ± 15.36	**63.29 ± 24.45**	33.38 ± 22.32	**54.44% ± 49.80%**	3.14 ± 1.48

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

Punjabi produces the lowest initial offers across all runs (37.4–41.1 vs. 41.4–46.4 for other languages), and the P1-favored payoff asymmetry persists throughout. Notably, A3, the native-script condition, does not produce systematically different trends from the English-phrased variants, suggesting that the observed effects are driven by the interaction language itself rather than the script or the phrasing of the prompt. The one notable source of variance across prompts is win rate under Punjabi, where A1 yields an elevated P1 win rate of 70.8% compared to 52.2–56.7% in other variants, suggesting this metric carries more noise. Comprehensive per prompt variant results are reported in Appendix D.

5.2 Buy-Sell Game Results

Results are shown in Figure 2 and Table 2.

5.2.1 Baseline Language (English)

English yields a high acceptance rate (97.14%) with seller advantage (mean 7.12, std 12.39) lower than buyer advantage (mean 12.88, std 12.39). The seller win rate remains modest at 41.23%, and negotiations conclude in an average of 3.09 rounds (std 1.90).

5.2.2 Multilingual Outcome Variation

All non-English languages achieve near-perfect acceptance rates, with Hindi reaching 100% agreement and Gujarati (98.77%) and Punjabi (99.35%) closely following.

While average seller advantage increases slightly in Gujarati (7.93), Hindi (8.22), and Punjabi (8.17)

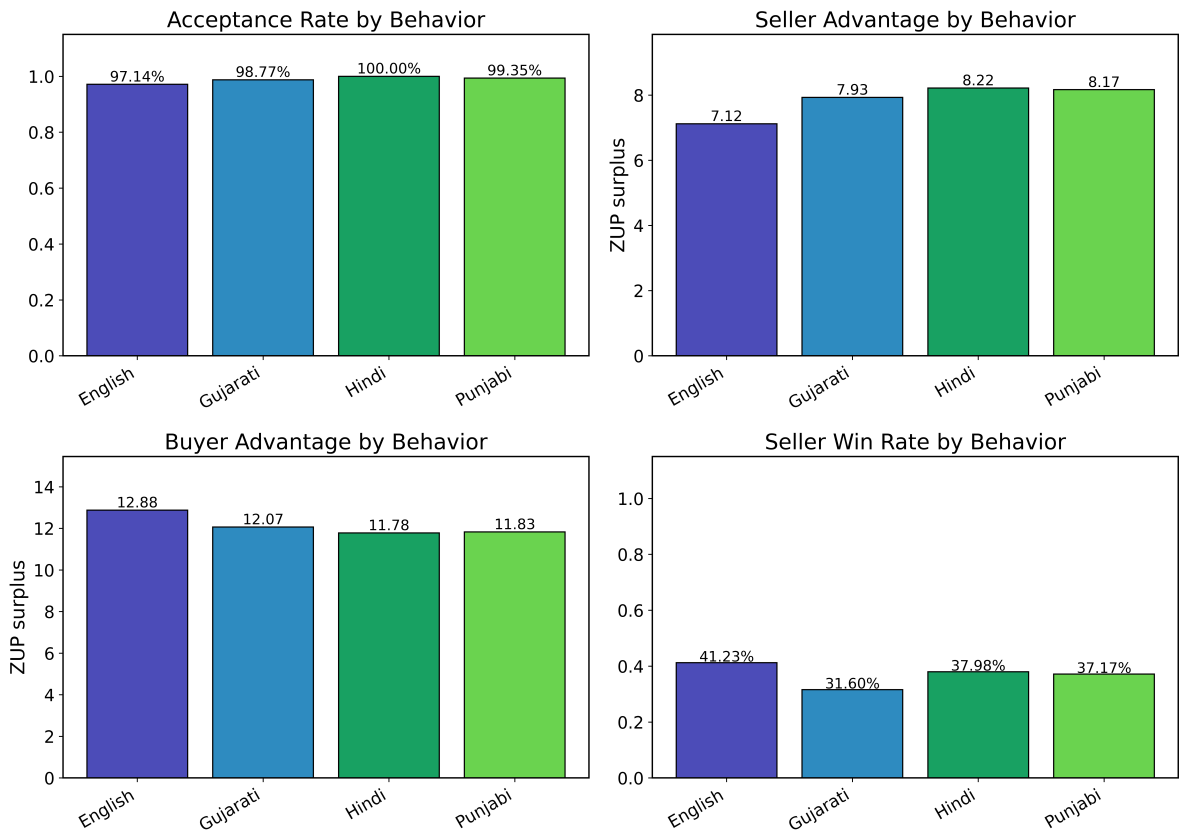

Figure 2: Buy Sell Game Language Comparison showing average (a) acceptance rates, (b) seller advantages, (c) buyer advantages, and (d) win rates across different linguistic behaviors.

Language	Acceptance Rate	Seller Advantage	Buyer Advantage	Conversation Rounds	P1 Win Rate
English	97.14% ± 16.68%	7.12 ± 12.39	**12.88 ± 12.39**	3.09 ± 1.90	**41.23%**
Gujarati	98.77% ± 11.04%	7.93 ± 9.82	12.07 ± 9.82	**3.31 ± 1.92**	31.60%
Hindi	**100.00% ± 0.00%**	**8.22 ± 11.57**	11.78 ± 11.57	3.04 ± 1.46	37.98%
Punjabi	99.35% ± 8.05%	8.17 ± 10.86	11.83 ± 10.86	3.42 ± 1.87	37.17%

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

relative to English, and buyer advantage correspondingly decreases, these differences remain small in magnitude.

More pronounced variation emerges in negotiation dynamics. Punjabi produces the highest average number of negotiation rounds (3.42, std 1.87), followed by Gujarati (3.31) and English (3.09), with Hindi lowest (3.04). The increase in negotiation rounds for Punjabi relative to English is statistically significant ($p_\text{corr} = 0.0012^{**}$), and Punjabi also exceeds Hindi ($p_\text{corr} = 0.0468^{*}$).

5.2.3 Evaluating Hypotheses

The primary statistically supported effect concerns negotiation length. Punjabi is consistently associated with longer negotiations compared to English ($p_{\text{corr}} = 0.0012^{**}$), providing clear evidence that language choice can influence interaction dynamics.

In contrast, differences in seller and buyer advantage across languages are modest and do not yield statistically significant pairwise contrasts after correction. Accordingly, the data do not provide direct evidence for systematic shifts in surplus allocation corresponding to H1 or H2.

5.3 Resource Exchange Game Results

Results are shown in Figure 3 and Table 3.

5.3.1 Baseline Language (English)

English yields a 90.84% acceptance rate, indicating that LLM agents consistently reach agreement under this condition. However, English produces the lowest average trade volume (16.05).

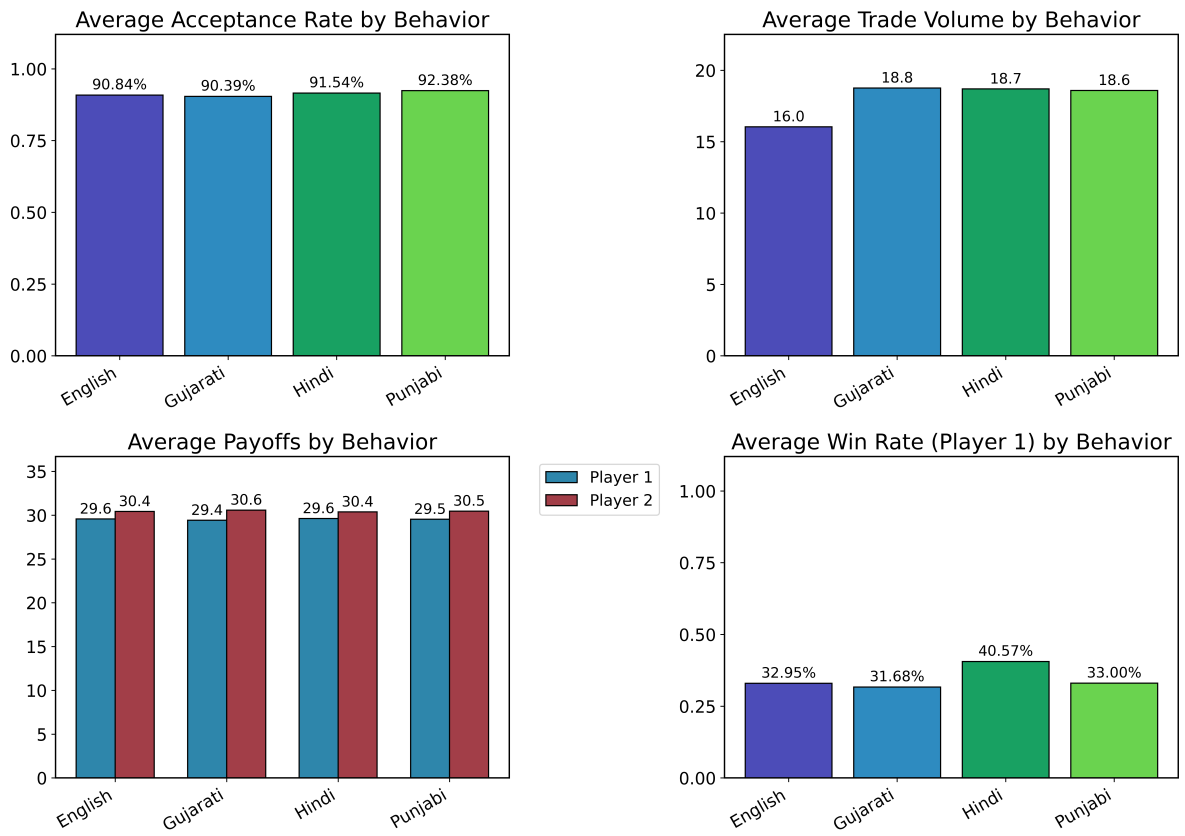

Figure 3: Resource Exchange Game Language Comparison showing average (a) acceptance rates, (b) trade volume, (c) payoffs, and (d) win rates (P1) across different linguistic behaviors.

Language	Acceptance Rate (%)	Trade Volume	P1 Payoff	P2 Payoff	P1 Win Rate (%)	Conversation Rounds
English	90.84 ± 28.90	16.05 ± 6.77	29.57 ± 2.57	30.43 ± 2.57	32.95	3.10 ± 1.42
Gujarati	90.39 ± 29.52	$\mathbf{18.77} \pm \mathbf{7.99}$	29.42 ± 2.76	$\mathbf{30.58} \pm \mathbf{2.76}$	31.68	$\mathbf{3.28} \pm \mathbf{1.41}$
Hindi	91.54 ± 27.88	18.70 ± 6.87	$\mathbf{29.63} \pm \mathbf{2.79}$	30.37 ± 2.79	**40.57**	3.27 ± 1.41
Punjabi	$\mathbf{92.38} \pm \mathbf{26.57}$	18.59 ± 6.70	29.53 ± 2.64	30.47 ± 2.64	33.00	3.15 ± 1.40

Table 3: Metrics for **Resource Exchange** aggregated across all model combinations (mean $\pm$ std).

Payoff distributions in English are closely balanced: P1 achieves an average payoff of 29.53 and P2 30.47, with no statistically significant differences relative to other languages after correction. Similarly, P1 wins 33.0% of non-tied games, a rate that does not differ significantly across languages.

5.3.2 Multilingual Outcome Variation

Trade volume exhibits clear and statistically significant variation. Gujarati (18.77), Hindi (18.70), and Punjabi (18.59) all yield substantially higher average trade volumes than English (16.05). Pairwise comparisons confirm that all non-English languages differ significantly from English ($p_{\text{corr}} = 0.0001^{***}$), while differences among Gujarati, Hindi, and Punjabi are not statistically significant. Other metrics remain largely stable across languages. P1 and P2 payoffs show no significant pairwise differences after correction, with means tightly clustered around 30. Negotiation rounds exhibit minor variation (English: 3.10; Gujarati: 3.28; Hindi: 3.27; Punjabi: 3.15), but none of the comparisons reach statistical significance. Similarly, P1 win rates vary modestly (ranging from 31.7% to 40.6%).

5.3.3 Evaluating Hypotheses

The results provide partial support for **H1 (Language-Mediated Strategy)**. Language choice systematically affects trade volume, with all non-English conditions inducing significantly higher exchange levels than English. However, other key outcomes - including payoffs, negotiation length, and win rates remain invariant across languages af-

ter correction, indicating that the effect of language may be selective rather than global.

5.4 Model-Specific Performance

As the primary focus of this work is on language effects in negotiation, model-specific results are discussed briefly here; detailed heatmaps are provided in Appendix A.

Ultimatum Game. GPT-4o exhibits asymmetric performance across roles: as P2 it is comparatively difficult to win against, whereas as P1 it achieves a lower win rate than GPT-3.5 under comparable conditions. Notably, GPT-4o consistently provides its negotiation rationale in the interaction language, while the other models evaluated often default to English regardless of the instructed language. This difference in response behaviour coincides with the observed performance variation across conditions.

Buy-Sell Game. The Buy-Sell Game exposes large role-dependent asymmetries that interact with language choice. In the English baseline, GPT-4o as seller (P1) achieves a seller advantage of 19.1–20.0, while GPT-3.5 in the same role records only 0.2–0.9—a gap of nearly 20 points. Conversely, as buyer (P2), GPT-3.5 secures extreme advantages of 26.5–28.1, indicating systematic over-concession when selling and over-extraction when buying. GPT-4o as buyer (P1) yields negligible buyer advantages and, conversely, consistently provides the largest seller advantages as P1 across all languages.

Under Indic language conditions, these asymmetries are attenuated but not eliminated. In Gujarati, GPT-3.5's seller advantage as P1 improves to -2.0 to -0.4, while its buyer advantage falls to 20.4–22.0. GPT-4o maintains a seller advantage of 13.3–16.5 in Gujarati, preserving its dominant position. This pattern holds across all tested languages: GPT-4o consistently achieves the largest seller advantages as P1, and GPT-3.5 consistently achieves the largest buyer advantages as P1, across all languages. Taken together, these results suggest that linguistic framing may attenuate but does not eliminate capacity-driven asymmetries between models.

6 Conclusion

This study explores how linguistic framing influences the dynamics of negotiation in Large Language Models across three game-theoretic environments. Our observations suggest that language choice may not be a neutral factor in strategic interactions; rather, it correlates with shifts in negotiation outcomes and surplus allocation, particularly within the Indic language contexts studied. While we find that certain linguistic framings are associated with different patterns of stability or exploration compared to English, these effects appear to be highly task-contingent. The contrast between the Ultimatum and Resource Exchange findings provides support for H2: language effects are task-contingent, reducing stability in distributive games while increasing exploration in integrative ones. These results highlight the potential limitations of evaluating strategic reasoning solely through an English-centric lens. Our findings suggest that incorporating a broader range of languages into the evaluation of LLM agents is an important step toward a more comprehensive understanding of their capabilities in diverse linguistic settings as LLMs deploy globally in commercial and interpersonal contexts, with direct implications for fairness and equitable deployment. Any benchmark or evaluation framework for LLM negotiation or strategic reasoning that is English-only should be considered incomplete, and multilingual evaluation should be treated as a methodological requirement rather than an optional extension.

7 Limitations

Our findings should be interpreted carefully. The behaviors exhibited by LLM agents do not constitute evidence about real human negotiation practices or cultural norms - differences reflect patterns learned from training corpora, not properties of languages or their speakers. Our language framings use culturally associated labels without incorporating human participants or sociolinguistic context; any apparent alignment with stereotypes should be understood as an artifact of representation learning, analogous to well-documented biases in word embeddings. We deliberately avoid normative claims and do not endorse any interpretation that attributes these behaviors to real-world groups. Our system prompts are written in English to maintain consistency with NegotiationArena, which may partially attenuate the linguistic signal being measured. 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). We also examine only

model-model interaction across three games, abstracting away from human-AI dynamics, richer incomplete-information settings, and realistic code-switching such as Hinglish. In addition, a more specific lingual system prompt could be curated; in our results we observed some models would provide their reason for the trade in the respective language while others would stick to English, it would be valuable to evaluate this variable as well. While our results provide a proof-of-concept that linguistic framing functions as a significant behavioral prior, whether these effects generalize to typologically distant languages such as Mandarin or Arabic remains an open question for future work.

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

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.

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.

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.

Sherzod Hakimov, Roland Bernard, Tim Leiber, Karl Osswald, Kristina Richert, Ruilin Yang, Raffaella Bernardi, and David Schlangen. 2025. The price of thought: A multilingual analysis of reasoning, performance, and cost of negotiation in large language models.

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.

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*.

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.

Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Bhasa-abhijnaanam: Native-script and romanized language identification for 22 Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. 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.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *arXiv preprint arXiv:2505.17407*.

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.

Chenxi Wang, Yixuan Zhang, Lang Gao, and Zixiang Xu. 2024. The Babel effect: Analyzing multilingual performance discrepancies in large language models.

A Visualizations

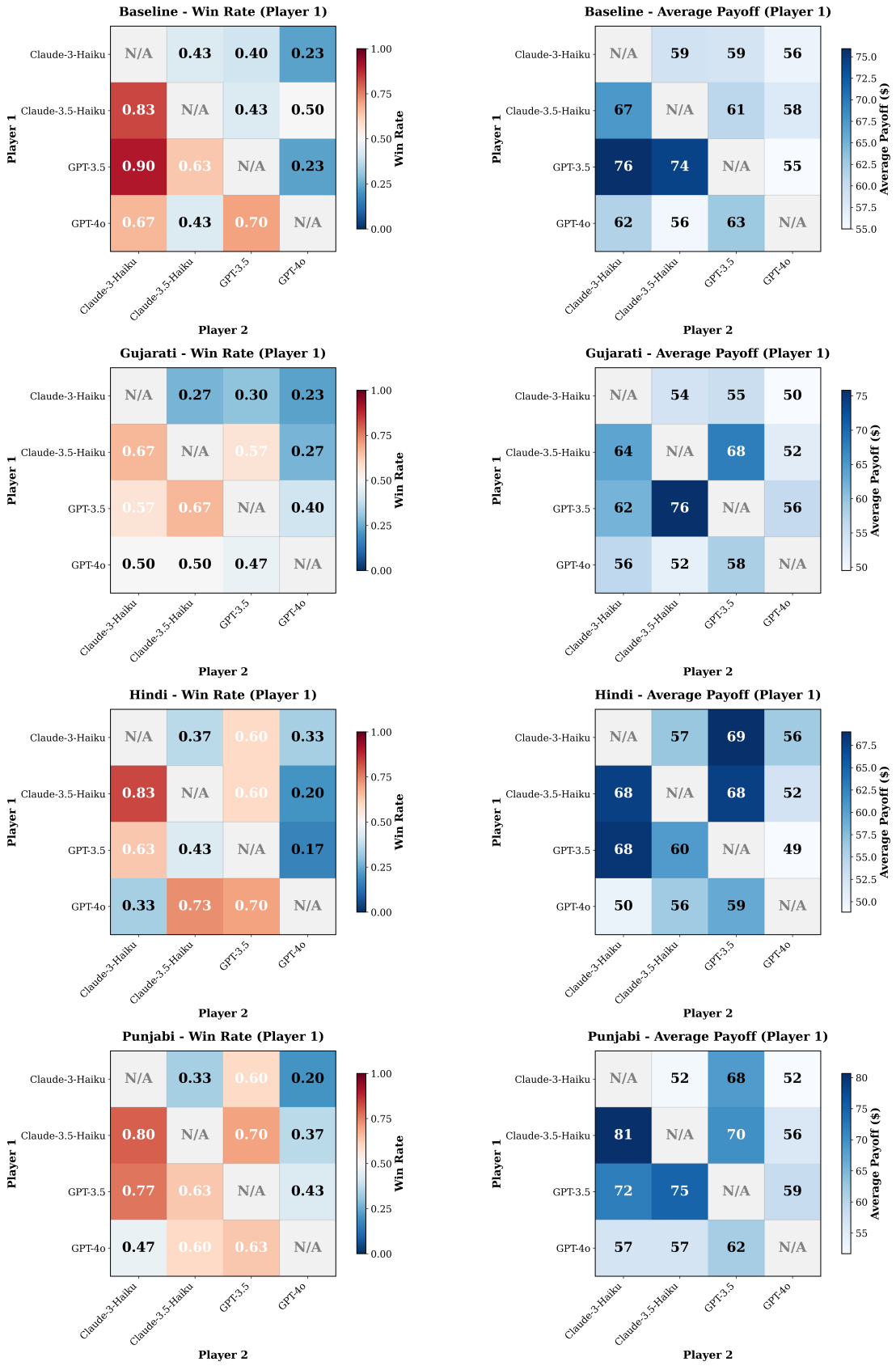

Figure 4: Heatmaps for **Ultimatum Game** comparing (a) win rate and (b) payoff for model combinations for all languages.

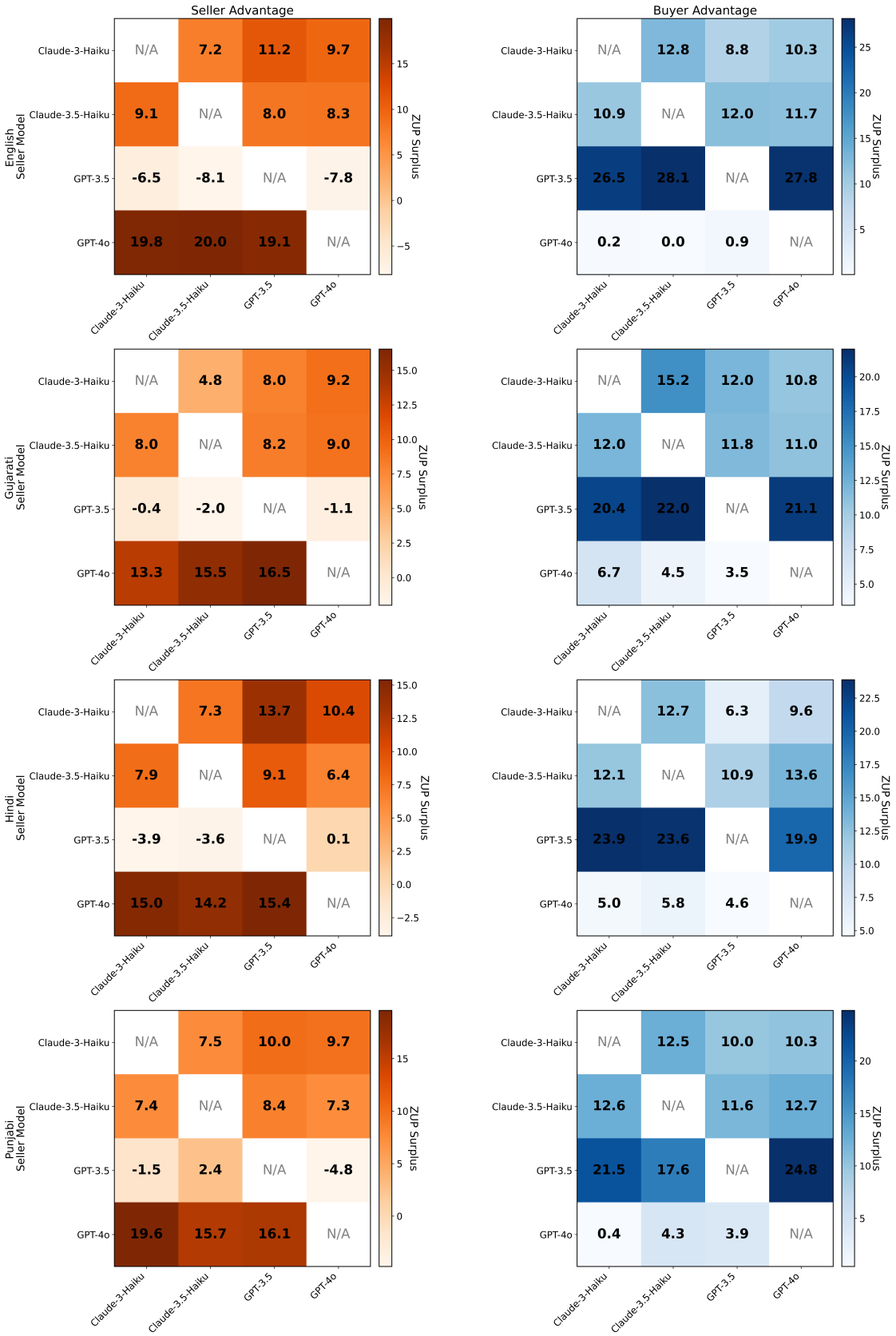

Figure 5: Heatmaps for **Buy-Sell Game** comparing (a) Seller advantage and (b) Buyer Advantage for model combinations for all languages.

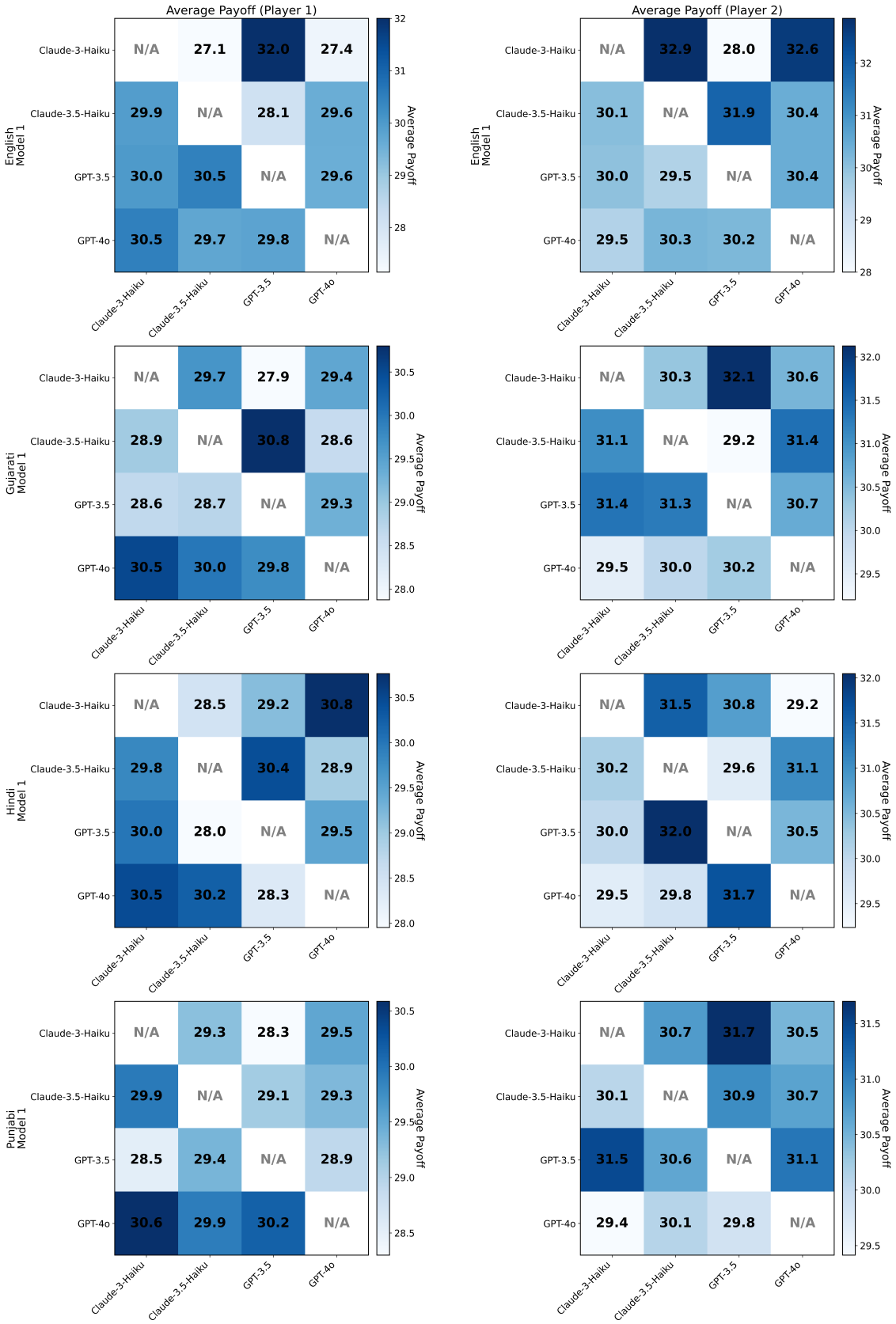

Figure 6: Heatmaps for **Resource Exchange Game** comparing average payoffs for P1 and P2 for model combinations for all languages.

B Statistical Analysis

Metric	Comparison	Mean Diff	U / z	p	p_corr	Sig
P1 Payoff (H=10.69, p=0.0135, *)						
	Baseline vs Gujarati	3.45	70828.5	0.026187	0.052375	ns
	Baseline vs Hindi	2.66	68838.0	0.135475	0.203213	ns
	Baseline vs Punjabi	-1.21	61928.5	0.290632	0.348759	ns
	Gujarati vs Hindi	-0.79	62568.5	0.411617	0.411617	ns
	Gujarati vs Punjabi	-4.66	57412.0	0.006761	0.040569	*
	Hindi vs Punjabi	-3.87	58609.0	0.023011	0.052375	ns
P2 Payoff (H=5.42, p=0.143, ns)						
	Baseline vs Gujarati	0.99	64262.5	0.842791	0.842791	ns
	Baseline vs Hindi	-0.72	63148.0	0.541194	0.811790	ns
	Baseline vs Punjabi	2.87	69765.5	0.067414	0.202242	ns
	Gujarati vs Hindi	-1.71	63883.0	0.735474	0.842791	ns
	Gujarati vs Punjabi	1.88	68884.0	0.133563	0.267125	ns
	Hindi vs Punjabi	3.59	70756.0	0.028565	0.171392	ns
Initial Offer (H=49.27, p=1.14e-10, *)**						
	Baseline vs Gujarati	0.11	62883.0	0.735124	0.735124	ns
	Baseline vs Hindi	-1.77	60194.5	0.134966	0.202448	ns
	Baseline vs Punjabi	5.36	77324.0	0.000000	0.000000	***
	Gujarati vs Hindi	-1.88	60569.5	0.287276	0.344731	ns
	Gujarati vs Punjabi	5.25	75896.5	0.000000	0.000001	***
	Hindi vs Punjabi	7.13	79118.5	0.000000	0.000000	***
Total Turns (H=45.40, p=7.60e-10, *)**						
	Baseline vs Gujarati	-0.55	50565.0	0.000000	0.000000	***
	Baseline vs Hindi	-0.49	51253.0	0.000000	0.000000	***
	Baseline vs Punjabi	-0.59	50900.5	0.000000	0.000000	***
	Gujarati vs Hindi	0.06	65987.5	0.649083	0.778899	ns
	Gujarati vs Punjabi	-0.04	64483.5	0.903086	0.903086	ns
	Hindi vs Punjabi	-0.10	63510.0	0.618610	0.778899	ns
Acceptance Rate (χ^2=13.77, p=0.0032, **)						
	Baseline vs Gujarati	0.086	3.656	0.000256	0.001536	**
	Baseline vs Hindi	0.047	2.181	0.029196	0.058392	ns
	Baseline vs Punjabi	0.064	2.840	0.004515	0.013545	*
	Gujarati vs Hindi	-0.039	-1.522	0.128122	0.192183	ns
	Gujarati vs Punjabi	-0.022	-0.848	0.396380	0.475656	ns
	Hindi vs Punjabi	0.017	0.676	0.498962	0.498962	ns

Table 4: Statistical comparison across languages for the **Ultimatum** Game. Global tests (Kruskal-Wallis or Chi-square) are reported per metric. Pairwise comparisons use Mann-Whitney U tests (or proportion z-tests for acceptance rate) with Benjamini-Hochberg correction.

Metric	Combination	Mean Diff	U	p	p_corr	Sig
Seller Advantage (H=1.52, p=0.678, ns)						
	English vs Gujarati	-0.81	56867.5	0.3373	0.8758	ns
	English vs Hindi	-1.10	53857.5	0.8758	0.8758	ns
	English vs Punjabi	-1.05	52309.0	0.8430	0.8758	ns
	Gujarati vs Hindi	-0.29	48463.0	0.2300	0.8758	ns
	Gujarati vs Punjabi	-0.24	47596.0	0.5408	0.8758	ns
	Hindi vs Punjabi	0.05	49752.5	0.6166	0.8758	ns
Buyer Advantage (H=1.52, p=0.678, ns)						
	English vs Gujarati	0.81	52272.5	0.3373	0.8758	ns
	English vs Hindi	1.10	54602.5	0.8758	0.8758	ns
	English vs Punjabi	1.05	51391.0	0.8430	0.8758	ns
	Gujarati vs Hindi	0.29	53936.0	0.2300	0.8758	ns
	Gujarati vs Punjabi	0.24	50309.0	0.5408	0.8758	ns
	Hindi vs Punjabi	-0.05	47542.5	0.6166	0.8758	ns
Negotiation Rounds (H=14.33, p=0.0025, **)						
	English vs Gujarati	-0.22	52309.5	0.0481	0.0962	ns
	English vs Hindi	0.05	52706.5	0.1704	0.2045	ns
	English vs Punjabi	-0.33	45490.0	0.0002	0.0012	**
	Gujarati vs Hindi	0.28	53571.0	0.4272	0.4272	ns
	Gujarati vs Punjabi	-0.11	46711.0	0.1225	0.1838	ns
	Hindi vs Punjabi	-0.38	43977.5	0.0156	0.0468	*
Acceptance Rate						
English: 340/350 (97.1%)						
Gujarati: 321/325 (98.8%)						
Hindi: 319/319 (100.0%)						
Punjabi: 306/308 (99.4%)						
Chi-square test skipped (degenerate case)						

Table 5: Statistical comparison across languages for the **BuySell** Game. Global tests (Kruskal-Wallis or Chi-square) are shown alongside each metric. Pairwise comparisons use Mann-Whitney U tests (or proportion tests) with Benjamini-Hochberg correction.

Metric	Comparison	p	p_corr	Sig
Trade Volume (H=30.42, p=1.13e-06, *)**				
	English vs Gujarati	0.0001	0.0001	***
	English vs Hindi	0.0000	0.0000	***
	English vs Punjabi	0.0000	0.0000	***
	Gujarati vs Hindi	0.5743	0.8079	ns
	Gujarati vs Punjabi	0.7728	0.8079	ns
	Hindi vs Punjabi	0.8079	0.8079	ns
P1 Payoff (H=1.09, p=0.780, ns)				
	English vs Gujarati	0.5929	0.8249	ns
	English vs Hindi	0.5954	0.8249	ns
	English vs Punjabi	0.9096	0.9096	ns
	Gujarati vs Hindi	0.3222	0.8249	ns
	Gujarati vs Punjabi	0.5075	0.8249	ns
	Hindi vs Punjabi	0.6874	0.8249	ns
P2 Payoff (H=1.09, p=0.780, ns)				
	English vs Gujarati	0.5929	0.8249	ns
	English vs Hindi	0.5954	0.8249	ns
	English vs Punjabi	0.9096	0.9096	ns
	Gujarati vs Hindi	0.3222	0.8249	ns
	Gujarati vs Punjabi	0.5075	0.8249	ns
	Hindi vs Punjabi	0.6874	0.8249	ns
Negotiation Rounds (H=4.95, p=0.175, ns)				
	English vs Gujarati	0.0590	0.2381	ns
	English vs Hindi	0.0794	0.2381	ns
	English vs Punjabi	0.4542	0.5450	ns
	Gujarati vs Hindi	0.9426	0.9426	ns
	Gujarati vs Punjabi	0.2148	0.3911	ns
	Hindi vs Punjabi	0.2608	0.3911	ns
Win Rate (P1) (χ^2=2.26, p=0.520, ns)				
	English vs Gujarati	0.8521	0.9947	ns
	English vs Hindi	0.2746	0.5492	ns
	English vs Punjabi	0.9947	0.9947	ns
	Gujarati vs Hindi	0.1838	0.5492	ns
	Gujarati vs Punjabi	0.8418	0.9947	ns
	Hindi vs Punjabi	0.2607	0.5492	ns
Acceptance Rate (χ^2 test: not applicable, degenerate case)				

Table 6: Statistical comparison across languages for the **Resource Exchange** Game. Global tests (Kruskal-Wallis or Chi-square) are shown alongside each metric. Pairwise comparisons use Mann-Whitney U tests (or proportion tests) with Benjamini-Hochberg correction.

C Language Compliance

Game	Lang.	Rate	Conf.
Ultimatum	Gujarati	97.02%	0.999
	Hindi	92.18%	0.966
	Punjabi	94.84%	0.998
Buy-Sell	Gujarati	93.22%	0.998
	Hindi	94.47%	0.981
	Punjabi	89.65%	0.997
Res. Exchange	Gujarati	97.85%	0.999
	Hindi	94.97%	0.971
	Punjabi	90.70%	0.997

Table 7: Language adherence rates and average confidence scores per game and language.

D Prompt Sensitivity Ablation Details

D.1 Prompt Ablation 1

See Figure 7 for model comparison heatmaps, Table 9 for the data summarized per language and Table 8 for the statistical analysis.

D.2 Prompt Ablation 2

See Figure 8 for model comparison heatmaps, Table 11 for the data summarized per language and Table 10 for the statistical analysis.

D.3 Native Language Prompt Ablation

See Figure 9 for model comparison heatmaps, Table 13 for the data summarized per language and Table 12 for the statistical analysis.

D.4 Comparing all Prompt Ablations

Metric	Comparison	p	p_corr	Sig
Player 1 Payoff (H=26.252878, p=8.4426e-06, *)**				
	English vs Gujarati	0.581387	0.581387	ns
	English vs Hindi	0.188140	0.225768	ns
	English vs Punjabi	0.000028	0.000083	***
	Gujarati vs Hindi	0.145061	0.217591	ns
	Gujarati vs Punjabi	0.003272	0.006545	**
	Hindi vs Punjabi	0.000002	0.000014	***
Player 2 Payoff (H=31.897117, p=5.5014e-07, *)**				
	English vs Gujarati	0.058127	0.069752	ns
	English vs Hindi	0.278217	0.278217	ns
	English vs Punjabi	0.000003	0.000008	***
	Gujarati vs Hindi	0.010207	0.020414	*
	Gujarati vs Punjabi	0.022788	0.034182	*
	Hindi vs Punjabi	0.000000	0.000003	***
Initial Offer (H=7.791326, p=5.0527e-02, ns)				
	English vs Gujarati	0.555858	0.667030	ns
	English vs Hindi	0.319284	0.478926	ns
	English vs Punjabi	0.075029	0.150059	ns
	Gujarati vs Hindi	0.709829	0.709829	ns
	Gujarati vs Punjabi	0.036438	0.109315	ns
	Hindi vs Punjabi	0.011628	0.069769	ns
Total Turns (H=1.805516, p=6.1374e-01, ns)				
	English vs Gujarati	0.475385	0.713077	ns
	English vs Hindi	0.636916	0.764299	ns
	English vs Punjabi	0.406831	0.713077	ns
	Gujarati vs Hindi	0.348166	0.713077	ns
	Gujarati vs Punjabi	0.885536	0.885536	ns
	Hindi vs Punjabi	0.245411	0.713077	ns
Acceptance Rate (χ^2=26.005063, p=9.5142e-06, *)**				
	English vs Gujarati	0.000862	0.001842	**
	English vs Hindi	0.124783	0.140297	ns
	English vs Punjabi	0.000003	0.000018	***
	Gujarati vs Hindi	0.059214	0.088821	ns
	Gujarati vs Punjabi	0.140297	0.140297	ns
	Hindi vs Punjabi	0.000921	0.001842	**

Table 8: Statistical comparison across languages for prompt ablation 1 of the **Ultimatum Game**. Global tests (Kruskal-Wallis or Chi-square) are shown alongside each metric. Pairwise comparisons use Mann-Whitney U tests (or proportion tests) with Benjamini-Hochberg correction.

Language	Acceptance Rate	Initial Offer	P1 Payoff	P2 Payoff	P1 Win Rate	Conversation Rounds
English	**93.33% ± 24.94%**	41.43 ± 11.69	63.50 ± 19.16	34.83 ± 17.86	55.83% ± 49.66%	2.63 ± 1.17
Gujarati	78.33% ± 41.20%	41.65 ± 14.43	64.25 ± 28.35	29.08 ± 23.86	54.17% ± 49.83%	**2.82 ± 1.55**
Hindi	87.50% ± 33.07%	**42.69 ± 13.79**	60.54 ± 23.25	**36.96 ± 21.95**	44.17% ± 49.66%	2.71 ± 1.40
Punjabi	70.00% ± 45.83%	37.36 ± 16.11	**74.71 ± 25.75**	21.96 ± 22.07	**70.83% ± 45.45%**	2.87 ± 1.46

Table 9: Metrics for prompt ablation 1 of the **Ultimatum Game** aggregated across all model combinations (mean ± std).

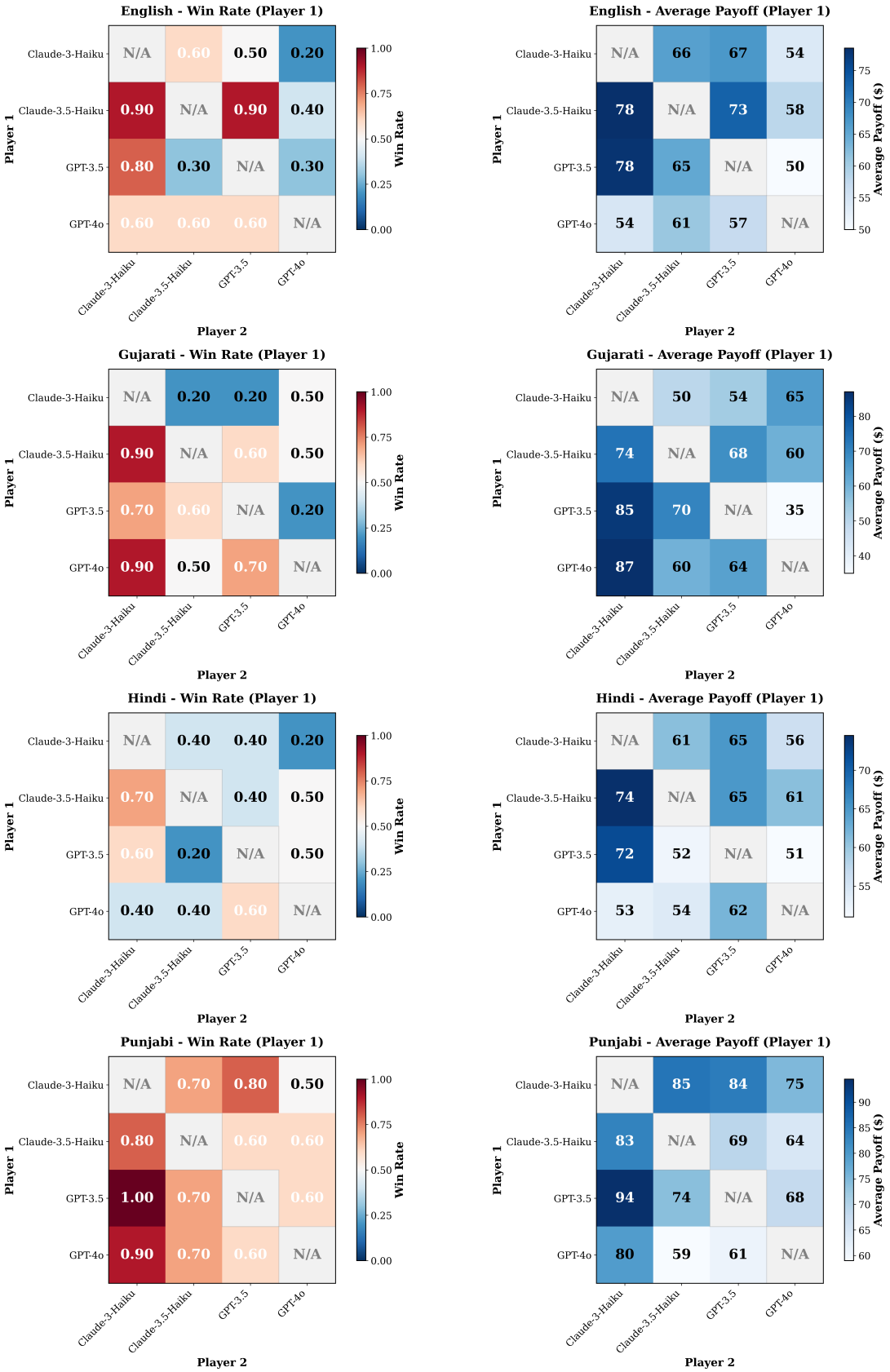

Figure 7: Heatmaps for **Ultimatum Game** prompt sensitivity ablation 1 comparing model combination outcomes across all languages.

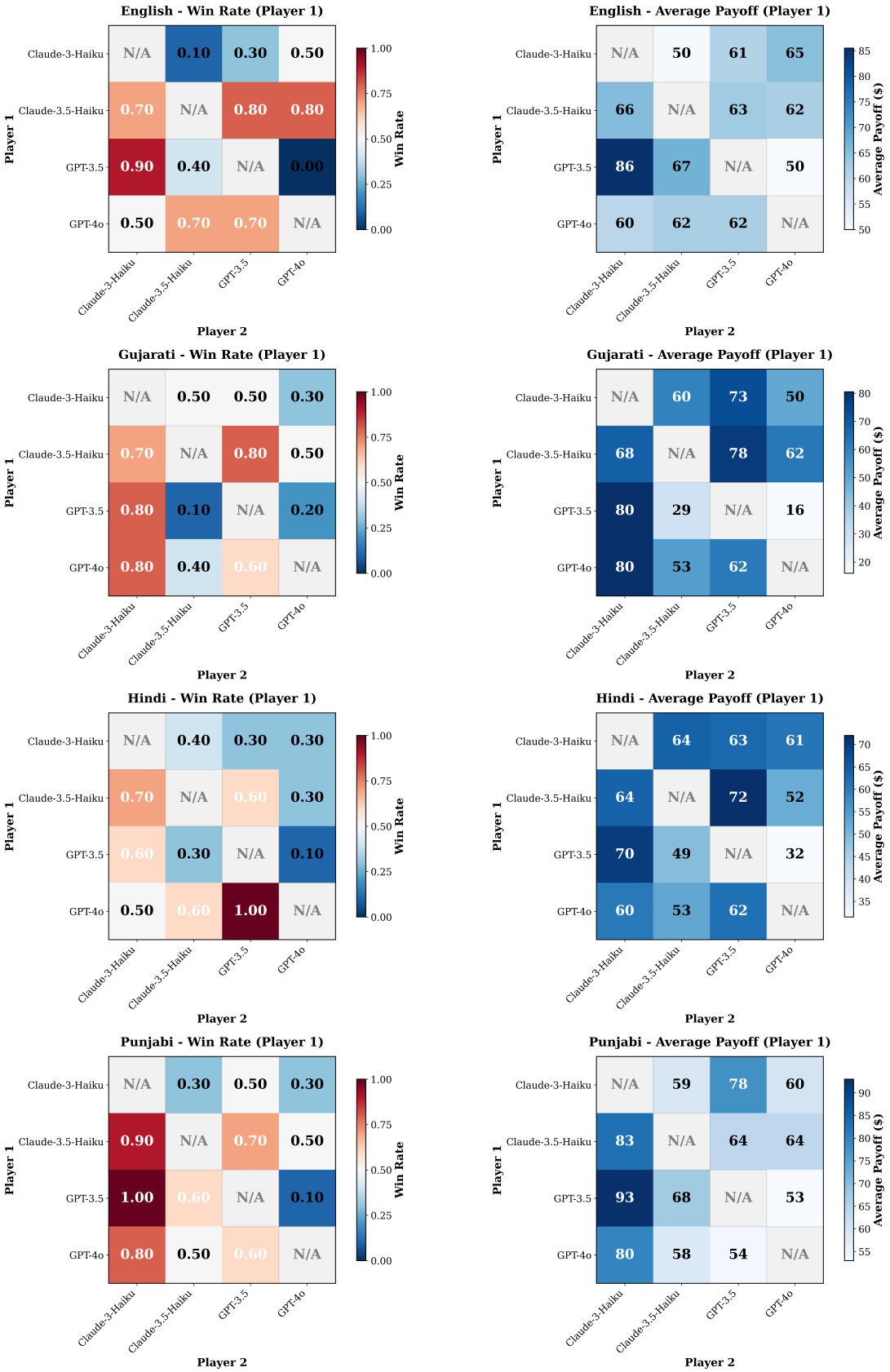

Figure 8: Heatmaps for **Ultimatum Game** prompt sensitivity ablation 2 comparing model combination outcomes across all languages.

Metric	Comparison	p	p_corr	Sig
Player 1 Payoff (H=6.429618, p=9.2480e-02, ns)				
	English vs Gujarati	0.791180	0.791180	ns
	English vs Hindi	0.188787	0.283180	ns
	English vs Punjabi	0.146649	0.283180	ns
	Gujarati vs Hindi	0.484535	0.581443	ns
	Gujarati vs Punjabi	0.093983	0.281948	ns
	Hindi vs Punjabi	0.016924	0.101542	ns
Player 2 Payoff (H=13.150719, p=4.3217e-03, **)				
	English vs Gujarati	0.018953	0.037907	*
	English vs Hindi	0.276479	0.331774	ns
	English vs Punjabi	0.056495	0.084743	ns
	Gujarati vs Hindi	0.002647	0.015882	*
	Gujarati vs Punjabi	0.799788	0.799788	ns
	Hindi vs Punjabi	0.008657	0.025970	*
Initial Offer (H=12.815051, p=5.0541e-03, **)				
	English vs Gujarati	0.468504	0.468504	ns
	English vs Hindi	0.253647	0.304377	ns
	English vs Punjabi	0.009133	0.027400	*
	Gujarati vs Hindi	0.099666	0.149499	ns
	Gujarati vs Punjabi	0.084277	0.149499	ns
	Hindi vs Punjabi	0.001100	0.006599	**
Total Turns (H=10.875555, p=1.2418e-02, *)				
	English vs Gujarati	0.002190	0.013143	*
	English vs Hindi	0.277313	0.415969	ns
	English vs Punjabi	0.478605	0.574326	ns
	Gujarati vs Hindi	0.036887	0.073773	ns
	Gujarati vs Punjabi	0.030010	0.073773	ns
	Hindi vs Punjabi	0.767648	0.767648	ns
Acceptance Rate (χ^2=24.532418, p=1.9337e-05, *)**				
	English vs Gujarati	0.000032	0.000194	***
	English vs Hindi	0.253369	0.304043	ns
	English vs Punjabi	0.000174	0.000522	***
	Gujarati vs Hindi	0.001677	0.003354	**
	Gujarati vs Punjabi	0.656486	0.656486	ns
	Hindi vs Punjabi	0.006565	0.009848	**

Table 10: Statistical comparison across languages for prompt ablation 2 of the **Ultimatum Game**. Global tests (Kruskal-Wallis or Chi-square) are shown alongside each metric. Pairwise comparisons use Mann-Whitney U tests (or proportion tests) with Benjamini-Hochberg correction.

Language	Acceptance Rate	Initial Offer	P1 Payoff	P2 Payoff	P1 Win Rate	Conversation Rounds
English	**93.33% ± 24.94%**	43.04 ± 10.61	62.79 ± 18.12	36.38 ± 17.50	53.33% ± 49.89%	2.51 ± 1.03
Gujarati	73.33% ± 44.22%	41.56 ± 12.41	59.34 ± 30.32	29.82 ± 24.49	51.67% ± 49.97%	**2.93 ± 1.61**
Hindi	89.17% ± 31.08%	**46.38 ± 17.14**	58.52 ± 24.04	**39.82 ± 23.38**	47.50% ± 49.94%	2.66 ± 1.19
Punjabi	75.83% ± 42.81%	39.13 ± 15.79	**67.83 ± 26.12**	29.67 ± 24.23	**56.67% ± 49.55%**	2.68 ± 1.29

Table 11: Metrics for prompt ablation 2 of the **Ultimatum Game** aggregated across all model combinations (mean ± std).

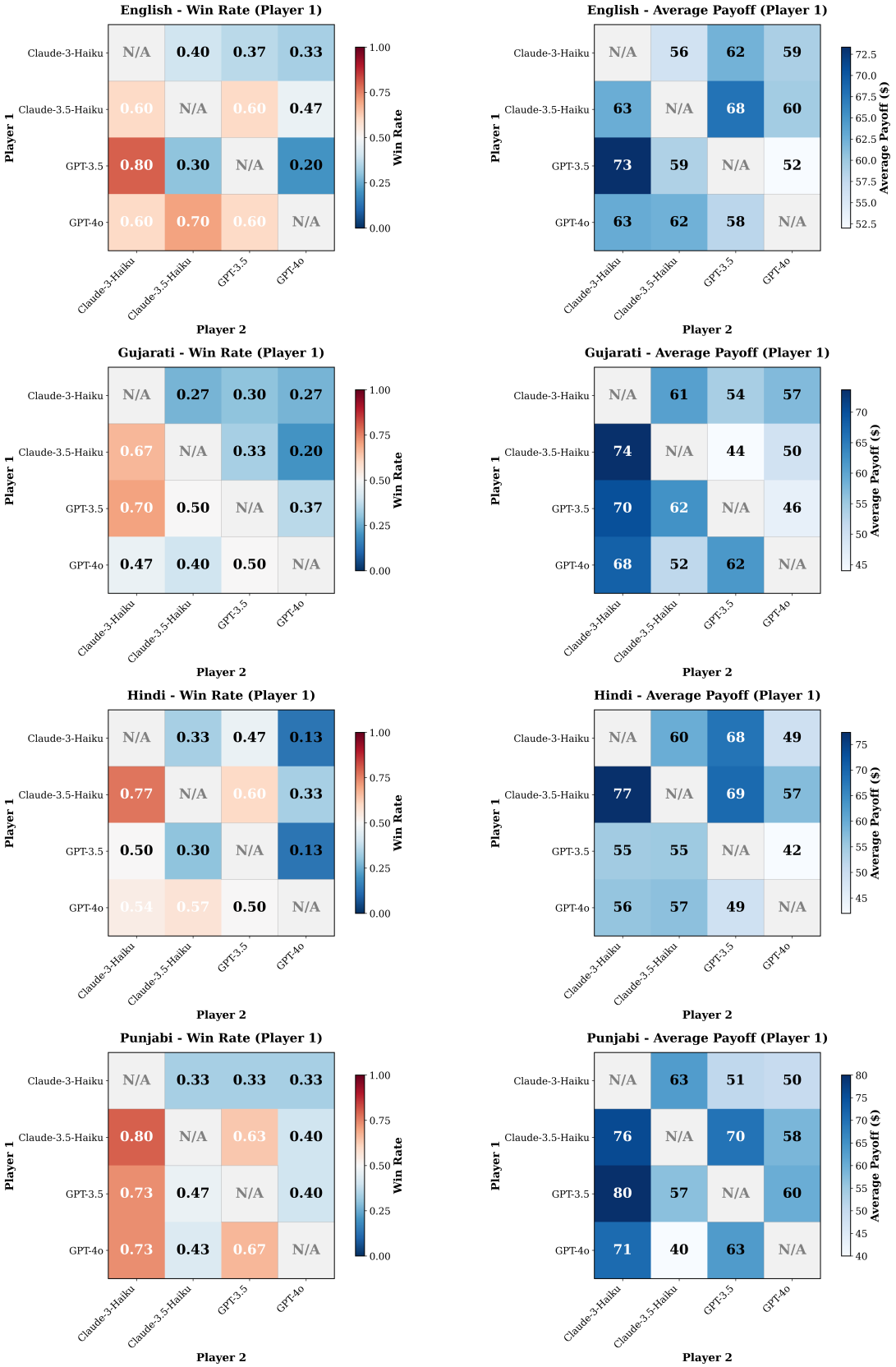

Figure 9: Heatmaps for **Ultimatum Game** with prompt in native language comparing model combination outcomes across all languages.

Metric	Comparison	p	p_corr	Sig
Player 1 Payoff (H=7.105005, p=6.8625e-02, ns)				
	English vs Gujarati	0.219148	0.269483	ns
	English vs Hindi	0.146722	0.269483	ns
	English vs Punjabi	0.224569	0.269483	ns
	Gujarati vs Hindi	0.838844	0.838844	ns
	Gujarati vs Punjabi	0.054615	0.163845	ns
	Hindi vs Punjabi	0.019696	0.118175	ns
Player 2 Payoff (H=40.474844, p=8.4506e-09, *)**				
	English vs Gujarati	0.005886	0.008829	**
	English vs Hindi	0.887990	0.887990	ns
	English vs Punjabi	0.000000	0.000000	***
	Gujarati vs Hindi	0.004693	0.008829	**
	Gujarati vs Punjabi	0.017205	0.020646	*
	Hindi vs Punjabi	0.000000	0.000000	***
Initial Offer (H=17.798184, p=4.8408e-04, *)**				
	English vs Gujarati	0.020541	0.030812	*
	English vs Hindi	0.006024	0.012048	*
	English vs Punjabi	0.257185	0.308622	ns
	Gujarati vs Hindi	0.550578	0.550578	ns
	Gujarati vs Punjabi	0.002303	0.006910	**
	Hindi vs Punjabi	0.000654	0.003923	**
Total Turns (H=22.020259, p=6.4601e-05, *)**				
	English vs Gujarati	0.000008	0.000049	***
	English vs Hindi	0.000395	0.001184	**
	English vs Punjabi	0.002921	0.005843	**
	Gujarati vs Hindi	0.219151	0.328726	ns
	Gujarati vs Punjabi	0.278274	0.333929	ns
	Hindi vs Punjabi	0.987905	0.987905	ns
Acceptance Rate (χ^2=67.810503, p=1.2560e-14, *)**				
	English vs Gujarati	0.000000	0.000000	***
	English vs Hindi	0.000744	0.001116	**
	English vs Punjabi	0.000000	0.000000	***
	Gujarati vs Hindi	0.005235	0.006282	**
	Gujarati vs Punjabi	0.056032	0.056032	ns
	Hindi vs Punjabi	0.000003	0.000006	***

Table 12: Statistical comparison across languages for the native language prompt ablation of the **Ultimatum Game**. Global tests (Kruskal-Wallis or Chi-square) are shown alongside each metric. Pairwise comparisons use Mann-Whitney U tests (or proportion tests) with Benjamini-Hochberg correction.

Language	Acceptance Rate	Initial Offer	P1 Payoff	P2 Payoff	P1 Win Rate	Conversation Rounds
English	**91.94% ± 27.22%**	42.75 ± 11.54	61.25 ± 20.05	37.08 ± 19.02	49.72% ± 50.00%	**2.75 ± 1.28**
Gujarati	75.28% ± 43.14%	44.55 ± 13.89	58.32 ± 29.66	31.40 ± 24.56	41.39% ± 49.25%	2.39 ± 1.31
Hindi	83.71% ± 36.93%	**46.00 ± 18.27**	57.92 ± 25.70	**37.58 ± 23.85**	42.98% ± 49.50%	2.45 ± 1.17
Punjabi	68.89% ± 46.29%	41.12 ± 16.12	**61.51 ± 30.45**	27.38 ± 23.40	**52.22% ± 49.95%**	2.58 ± 1.52

Table 13: Metrics for the native language prompt ablation of the **Ultimatum Game** aggregated across all model combinations (mean ± std).

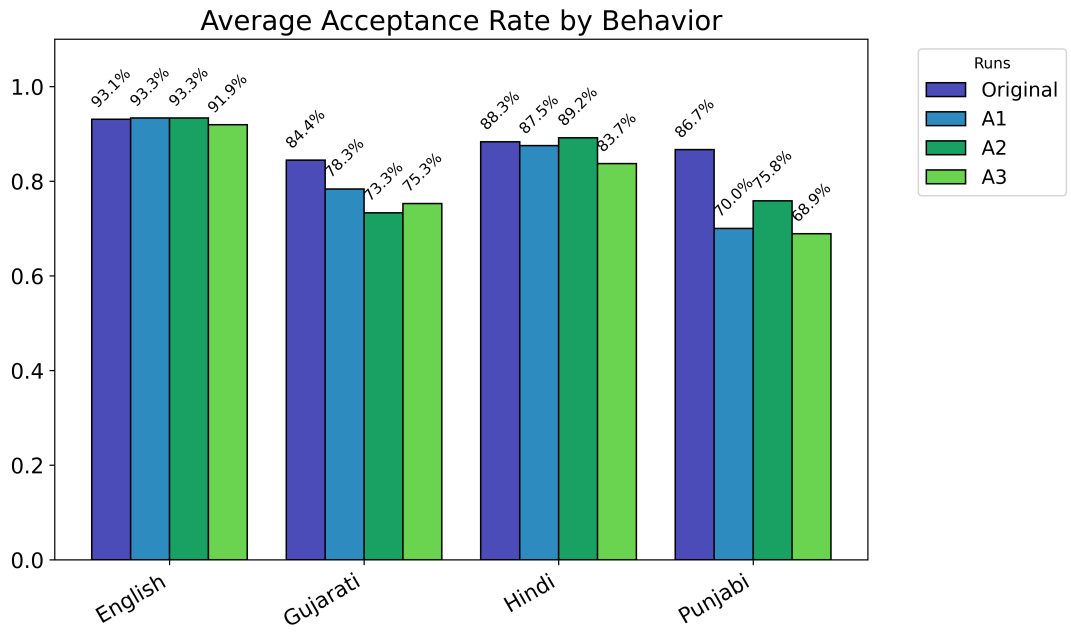

Figure 10: Acceptance Rate Comparison for **Ultimatum Game** comparing prompt ablation outcomes across all languages.

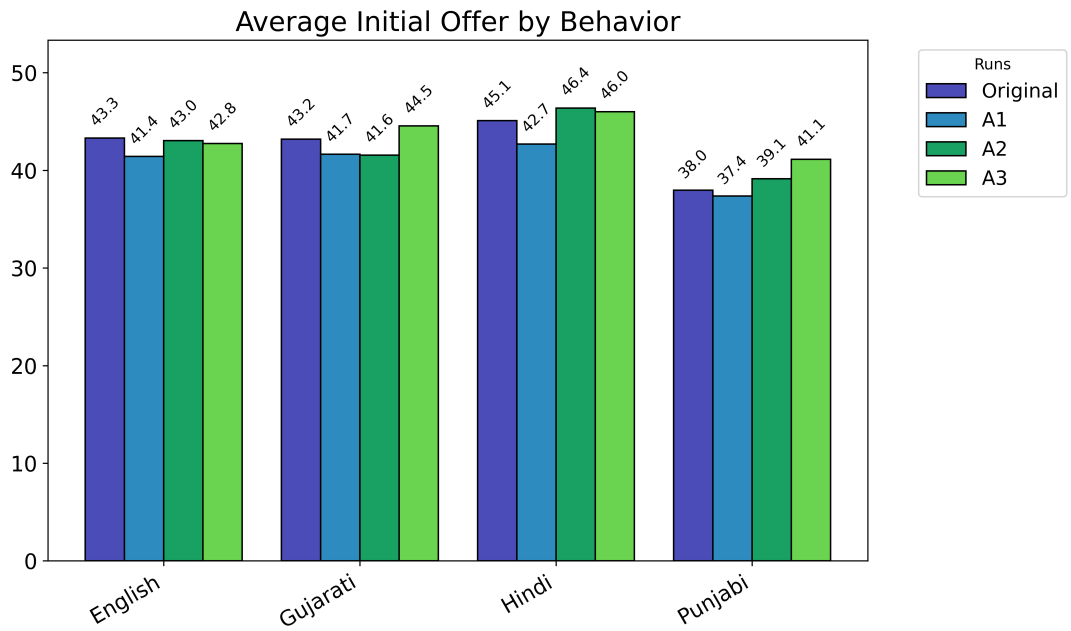

Figure 11: Initial Offer Comparison for **Ultimatum Game** comparing prompt ablation outcomes across all languages.

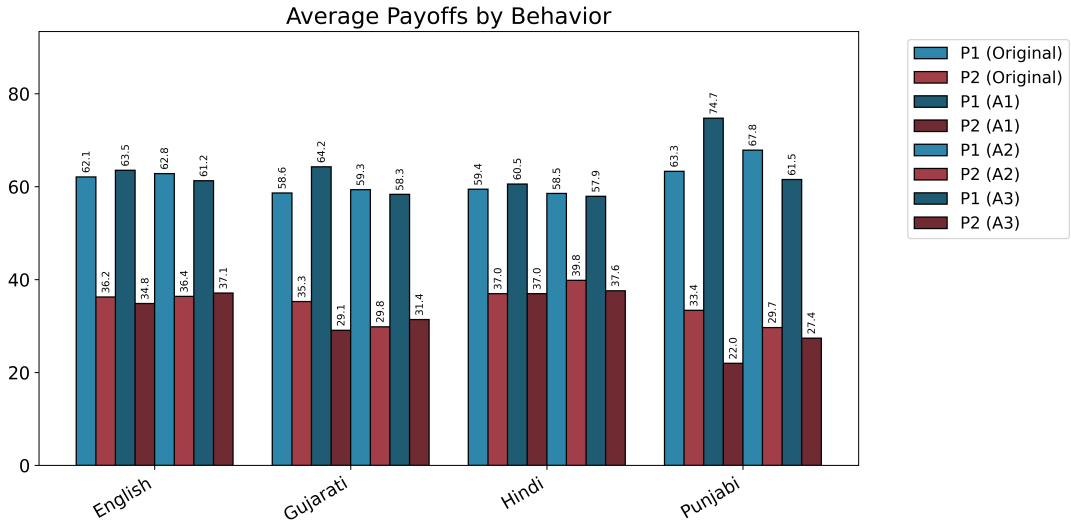

Figure 12: Payoffs Comparison for **Ultimatum Game** comparing prompt ablation outcomes across all languages.

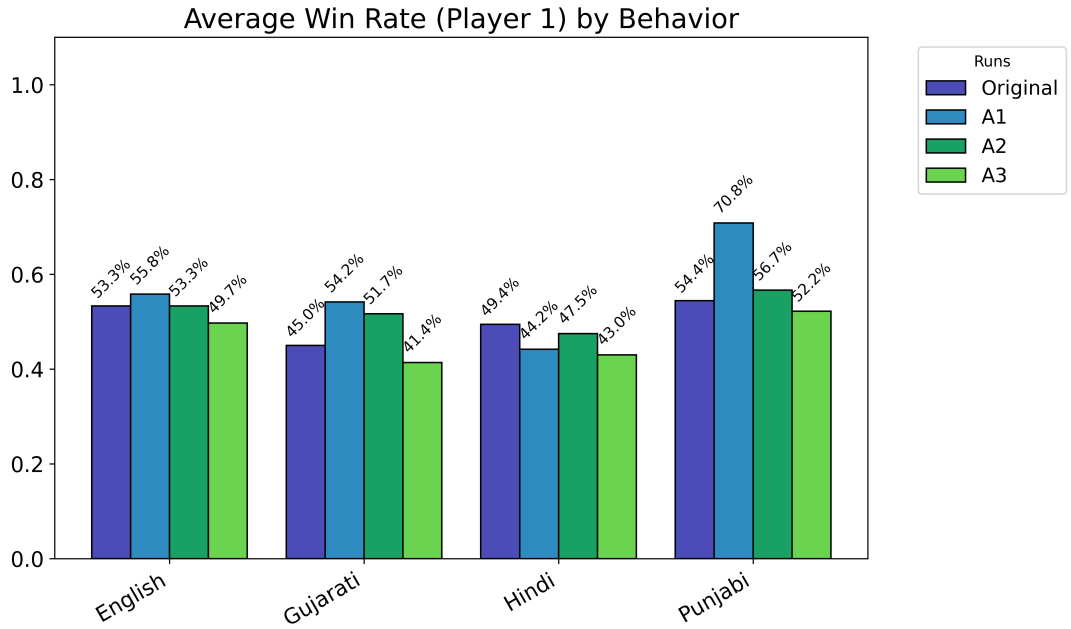

Figure 13: Payoffs Comparison for **Ultimatum Game** comparing prompt ablation outcomes across all languages.