Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Competitive, Cooperative, and Everything In-Between: The Janus Protocol for Strategy-Infused Language Models

Zachary Gray

ztg5018@utulsa.edu

University of Tulsa

Dale Peasley

University of Tulsa

Feyza M. Hafızoğlu

Istanbul Commerce University

Sandip Sen

University of Tulsa

Research Article

Keywords: Negotiation, supervised learning, hypernetwork, LLMs

Posted Date: June 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9000725/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Competitive, Cooperative, and Everything In-Between: The Janus Protocol for Strategy-Infused Language Models

Zachary Gray[1*], Dale Peasley[1], Feyza M. Hafızoğlu[2], Sandip Sen[1]

[1*]Tandy School of Computer Science, The University of Tulsa, 800 South Tucker Drive, Tulsa, 74104, Oklahoma, USA.
[2]Department of Computer Engineering, İstanbul Commerce University, İmrahor St. Beyoğlu, İstanbul, 34445, Türkiye.

*Corresponding author(s). E-mail(s): ztg5018@utulsa.edu;
Contributing authors: dfp3395@utulsa.edu; fmhafizoglu@ticaret.edu.tr; sandip@utulsa.edu;

Abstract

Static alignment is a structural limitation of deployed large language models (LLMs): once fine-tuned toward a fixed behavioral objective, a model cannot dynamically adjust its underlying strategic posture without retraining or model switching. We introduce the *Janus Protocol*, a scalar-conditioned hypernetwork framework that enables continuous inference-time behavioral control of a frozen LLM through Low-Rank Adaptation (LoRA) gating. Inspired by HyperLoRA for vision applications, we reinterpret adapter space, for automated negotiation applications, as a policy manifold indexed by a semantically grounded scalar parameter, $\boldsymbol{\rho} \in [\mathbf{0}, \mathbf{1}]$, representing normalized settlement position within a negotiation's Zone of Possible Agreement.

Rather than generating full adapter matrices, a lightweight hypernetwork predicts rank-dimensional gating vectors that scale shared LoRA projections, enabling efficient and expressive control with minimal additional parameters. We instantiate this framework in bilateral price negotiation, a domain with an outcome-grounded scalar axis that permits precise evaluation of strategic behavior. Training data is collected from 10,000 synthetic self-play trajectories spanning a diverse pool of classical and adversarial negotiation strategies, with each trajectory labeled by its terminal $\boldsymbol{\rho}$ value, which represents the normalized settlement position achieved within the Zone of Possible Agreement at the conclusion of the episode.

Empirical evaluation on a Qwen2-7B backbone demonstrates that a single trained adapter can handle a continuous spectrum of negotiation postures at inference time. Unmodified base models, trained with reinforcement learning from human feedback, are designed to be polite and accommodating to human demands and hence readily concede to pressure tactics. While they can be effective in integrative negotiation scenarios, they can be readily exploited in distributive negotiation scenarios. Relative to them, scalar-conditioned adaptation substantially improves mean utility, reduces exploitability against strategic concession strategies, and enables controlled adoption of competitive or fairness-oriented behavior, dictated by the distributive or integrative nature of the negotiation scenario and depending on the chosen policy parameter, $\boldsymbol{\rho}$. Ablation experiments confirm that role-consistent scalar conditioning is causally necessary for strategic coherence, while boundary analyses highlight smooth interior control with mild instability at extreme boundary values.
These results provide empirical evidence that low-dimensional, semantically interpretable scalar conditioning can modulate complex multi-turn strategic behavior directly in weight space. Crucially, the Janus Protocol demonstrates a mutually beneficial synergy between foundation models and multi-agent systems (MAS): MAS theory provides the equilibrium-grounded structure necessary to cure the concessionary biases of standard LLMs, while the foundation model provides the expressive language reasoning needed to elevate classical MAS agents into dynamic, conversational negotiators. Critically, this approach reframes alignment not as a problem of static objective optimization, but as a structured controllability tool.

Keywords: Negotiation, supervised learning, hypernetwork, LLMs

1 Introduction

The development and continual refinement of Large Language Models (LLMs) has catalyzed a shift in artificial intelligence, propelling them from simple text predictors into significantly more versatile roles as autonomous agents [1]. Their ability to understand and generate human-like dialogue has opened the door to applications in complex, socially-rich domains that were once the exclusive purview of human intellect. Among such domains requiring nuanced interaction between self-interested actors, negotiation scenarios pose a significant challenge. A wide variety of human interactions can be reduced to a form of negotiation [2, 3] as they embody an interplay of communication, strategy, and social psychology.

While much of the recent literature evaluates LLM negotiators as standalone agents, the integration of such foundation models (FMs) into multi-agent systems (MAS) poses deeper structural questions and presents an opportunity for significant cross-fertilization. Negotiation scenarios are inherently multi-agent settings governed by equilibrium concepts, incentive compatibility, and strategic adaptation over repeated interaction. Framing LLM negotiators as agents embedded within MAS architectures allows us to move beyond prompt-level behavior-shaping toward principled policy control grounded in agent theory. Conversely, injecting the generative and

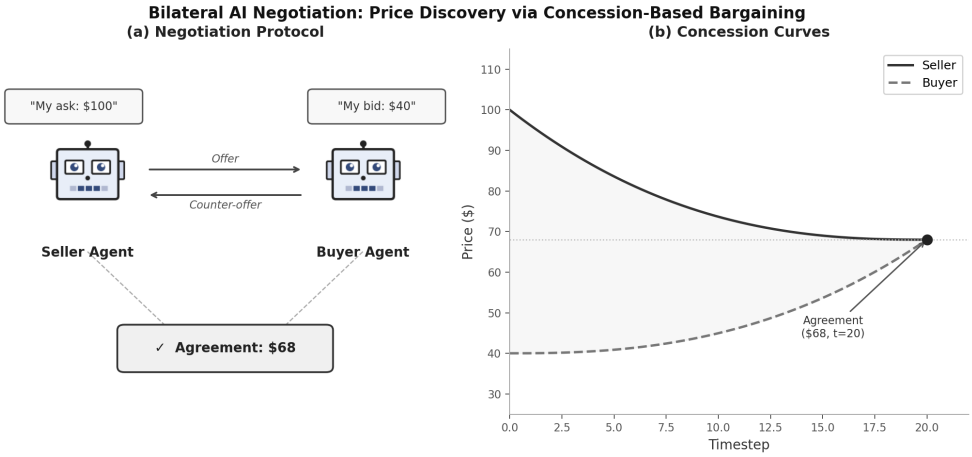


Fig. 1 Bilateral negotiation protocol. Two agents exchange alternating offers until convergence within the Zone of Possible Agreement (ZOPA). The right panel illustrates representative concession curves over time, culminating in agreement.

reasoning capabilities of FMs into classical MAS frameworks enables multi-agent collaboration to solve complex, language-driven tasks beyond the limits of traditional, programmatic single-agent heuristics.

Automated negotiation has long been studied in multi-agent systems literature across various domains such as supply-chain management, smart grids, and agent competitions [4–11]. With the advances in LLMs, there is growing interest in end-to-end negotiators that can reason over rich preferences, communicate in free text, and adapt to diverse domains using a single underlying architecture. Initial explorations into LLM-driven negotiation have yielded promising results [12, 13]. When pitted against each other, baseline LLMs often engage in fluid and cooperative dialogues, efficiently discovering mutually beneficial, Pareto-optimal agreements [12]. Furthermore, recent benchmarks and platforms show that state-of-the-art LLMs can sometimes outperform humans or classical baselines in achieving favorable deals [14]. This success in collaborative, value-creating bargaining has generated excitement about the potential of LLMs to serve as automated negotiators. However, we argue that this narrative is incomplete. It overlooks the other, more contentious face of negotiation: the competitive, value-claiming aspect that defines a vast number of real-world interactions. In addition, LLM-based negotiators are susceptible to other weaknesses, such as irrational behavior, inconsistent role-playing, and vulnerability to adversarial tactics.

Negotiation is a complex social phenomenon: agents must understand goals and constraints, track dialogue structure, infer opponents' preferences (Theory of Mind), and generate strategically appropriate, socially acceptable responses across many turns [15, 16]. These requirements are only partially met by the state-of-the-art LLMs that are trained on large web corpora. Systematic evaluations reveal strong contextual understanding but persistent weaknesses in arithmetic reasoning, strategic consistency,

and use of proven tactics, especially in multi-issue or multi-party negotiation scenarios [14, 17]. Moreover, base models are not grounded in the psychological background of the negotiation. For example, strategically adopting a persuasive [18], polite [19], and empathic [20] language, affecting the opponent's beliefs and hence decisions, is critical for negotiation. Conversely, overly aggressive or norm-violating language can adversely effect outcomes and trustworthiness. This aspect of negotiation motivates the research on socially-aware negotiation assistants and norm-remediating agents [21].

A key weakness of LLMs as competitive negotiators lies in their development process. The very architecture and training that make LLMs effective collaborators also belies a significant weakness: optimized through methods like Reinforcement Learning from Human Feedback (RLHF) to be helpful, harmless, and agreeable, these models are "concessionary by design" [17, 22]. They are conflict-averse, naive to manipulative intent, and lack the disciplined strategic calculus that has been the focus of negotiation theory for decades [23]. Across domains, current LLMs tend to favor agreement, fairness, and politeness [22, 24–26], which often manifests as over-concession and vulnerability to exploitation [17, 27], unless constrained by additional strategic or rule-based components. In their efforts to utilize the collaborative potential of LLMs, many researchers have inadvertently tested these agents without fully considering the established, battle-tested tactics that can exploit their cooperative nature.

A central limitation of existing LLM-based negotiation systems lies in their reliance on supervised fine-tuning (SFT) over negotiation transcripts or role-specific corpora [28, 29]. While SFT is stable and sample-efficient, it fundamentally trains models to imitate observed behavior rather than to optimize strategic objectives. In negotiation domains, this often results in agents that reproduce the average tendencies of the training distribution — typically agreement-seeking, polite, and concessionary behavior — rather than strategically robust policies. Because supervised learning minimizes next-token prediction loss, it does not explicitly enforce utility maximization, opponent exploitation resistance, or long-horizon strategic consistency. In multi-turn settings, small tactical errors compound over time, and models trained purely via imitation may struggle to adapt when confronted with adversarial or out-of-distribution strategies. Consequently, even well-trained LLM negotiators can exhibit brittle behavior when facing exploitative tactics that deviate from cooperative training norms.

Although extensive research has been carried out on evaluating LLMs' negotiation skills and the language [13, 14], studies on strengthening LLMs' weakness against exploitative/adversarial opponents with strategic negotiation skills are still limited. To date, far too little attention has been paid to repairing concessionary nature of LLMs in negotiation settings. We posit that the conversational fluency of current LLMs masks their strategic fragility. This study aims to fortify LLM negotiators and guard against the complexities and challenges of real world competitive negotiations. To do so, we move beyond cooperative benchmarks and create a dataset that is collected from the negotiations between a suite of "attack" agents. These agents are implementations of classic, well-documented strategies from both human negotiation psychology (e.g., deceptive framing, social pressure) and the canon of automated negotiation research (e.g., time-based concession, stonewalling) [30–32].

To achieve this objective, we introduce the Janus Protocol[1], a scalar-conditioned hypernetwork framework that enables continuous behavioral control of a frozen base LLM through inference-time modulation of a single strategic variable, ρ. A key property of our approach is that rather than fine-tuning separate buyer and seller models or relying on discrete prompt instructions, we train a single adapter whose weights are dynamically gated by a parameterized hypernetwork conditioned on $\rho \in [0, 1]$, representing normalized settlement position within the Zone of Possible Agreement. Training data is generated through large-scale self-play between a diverse suite of classical and adversarial negotiation agents, producing trajectories that span the full continuum of strategic outcomes. Each negotiation episode is labeled with its resulting ρ value, and the model is trained via supervised learning to reproduce contextually appropriate actions conditioned on this scalar.

Conceptually, this reframes adapter space as a low-dimensional strategic policy manifold. Rather than optimizing over the full high-dimensional parameter space of the model, we learn a compact, structured latent behavior space indexed by an interpretable scalar. This approach provides a unique advantage and is mathematically and theoretically justified by recent work demonstrating that the manifold of realizable agent behaviors is intrinsically low-dimensional and dictated by environment complexity, not neural network size [33]. By compressing highly complex, multi-turn negotiation behaviors into a single low-dimensional scalar parameter, we preserve policy diversity while making fine-tuning highly sample-efficient. In doing so, we explicitly operationalize the synergy between foundation models and MAS: MAS theory supplies the equilibrium-grounded coordinate (ρ), and the foundation model supplies expressive language reasoning within that structured space.

Empirically, preliminary results indicate a substantial improvement over the base Qwen2-7B model across adversarial matchups. While the base model frequently over-concedes or exhibits inconsistent strategic behavior, Janus demonstrates measurable gains in agreement quality, win rates, and resistance to exploitative tactics. Importantly, qualitative analysis shows that Janus adapts its concession patterns dynamically in response to opponent strategy rather than following a rigid scripted policy. Although full quantitative evaluation is ongoing, early experiments consistently demonstrate that scalar-conditioned adapter control enables the model to learn flexible, strategically coherent negotiation behaviors that are not present in the base model.

Beyond performance gains, this work directly addresses emerging concerns in FM–MAS integration. Surveys of LLM-based agents identify persistent hallucination phenomena—including reasoning, execution, and communication hallucinations—that can propagate across agent networks and amplify system-level failures [34, 35]. Embedding foundation models within structured MAS frameworks enables evaluation through equilibrium stability, exploitability, and distributive fairness metrics rather than solely linguistic coherence. By grounding behavior in outcome-indexed scalar control, we provide a mechanism for measurable strategic accountability in multi-agent deployments.

[1] The *Janus protocol* is named after Janus, the Roman god depicted with two faces looking simultaneously to the past and to the future.

1.1 Contributions

The literature on LLM negotiators is growing rapidly and there are different aspects of this complex social interaction that are relevant to several disciplines, such as psychology, sociology, and economics. This article, to our knowledge, is the first to efficiently enhance LLMs negotiators' resilience against opponents that pursue adversarial/exploitative strategies and makes the following contributions:

Continuous strategic control of LLMs via scalar-conditioned adapters: We introduce the Janus Protocol, a hypernetwork-gated LoRA framework that enables inference-time traversal of a learned negotiation policy manifold using a single scalar parameter ρ, without discrete model switching or retraining.
Explicit MAS-FM integration: We formalize negotiation posture as a game-theoretic control variable grounded in settlement position within the ZOPA, bridging equilibrium concepts from MAS theory with hypernetwork-based adaptation in foundation models.
Outcome-indexed training paradigm: We propose a dataset construction and labeling procedure in which complete negotiation trajectories are stamped with a normalized settlement scalar, allowing supervised learning to align behavior with strategic outcome position rather than fixed role identities.
Adversarial resilience evaluation: We design a benchmark suite of classical and exploitative negotiation strategies and demonstrate that scalar-conditioned adaptation substantially improves resistance to over-concession and strategic exploitation compared to an unmodified base model.
Structural interpretability through outcome-grounded control: Unlike purely prompt-based alignment methods, scalar-conditioned gating yields a semantically interpretable mapping between equilibrium position and weight-space modulation. This parallels recent mechanism-design-oriented LLM systems that embed structural constraints to achieve interpretable and strategy-aware outcomes [36].
Structural train–inference alignment. We introduce a rigorous structural contract to ensure the model reliably respects the scalar control parameter during both training and inference. This alignment relies on three key mechanisms: (1) *synchronized conditioning*, where the strategic target ρ is simultaneously embedded as text in the prompt and injected as a numeric tensor into the hypernetwork; (2) *failure encoding*, which explicitly represents negotiation impasses ($\rho = -1$) to teach the model the boundaries of the agreement space; and (3) *gate separation regularization*, a custom loss penalty that mathematically forces the hypernetwork to produce distinctly different adapter weights for distinct ρ values. This regularization is critical for preventing "scalar collapse," ensuring that tuning the control dial guarantees a corresponding shift in the underlying strategic policy.
Bridging hypernetwork adaptation from vision to language strategy control: We translate the HyperLoRA paradigm from portrait synthesis to multi-turn strategic interaction, providing the first empirical evidence that low-dimensional continuous control can modulate complex negotiation behavior in LLMs.

LLM negotiators are increasingly being piloted in high-stakes real-world domains, including customer support, digital marketplaces, and corporate procurement. However, current deployments often conflate conversational fluency with strategic capability. This research highlights the inherent fragility of these models in competitive environments and contributes to the development of robust automated agents. By operating under the assumption that these models will face adversarial tactics, our framework provides a blueprint for deploying LLMs with explicit, structural safeguards rather than relying on cooperative assumptions.

2 Background

The recent advances in LLMs has introduced a new paradigm, shifting the focus toward conversational fluency and emergent behavior. Initial research highlights the ability of LLMs to engage in complex, multi-issue negotiations, often achieving Pareto-efficient outcomes when interacting with other cooperative LLMs [12, 13]. While these studies effectively showcase the strength of LLMs in integrative bargaining—where agents collaboratively create value—they mask a fundamental structural weakness. State-of-the-art foundation models are heavily optimized through alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF), to prioritize agreeableness, politeness, and conflict avoidance. In essence, modern LLMs are explicitly trained to be helpful, fair, and accommodating.

While this cooperative prior makes them excellent conversational assistants, it renders them highly vulnerable to adversarial agents utilizing distributive, value-claiming tactics that have been documented in classical game theory for decades [2, 37]. Multi-agent negotiation benchmarks [17] explicitly demonstrate that greedy or adversarial agents can easily exploit these accommodating models. This core problem is thoroughly quantified by the NegotiationArena platform, which proves that base foundation models are irrationally concessionary and vulnerable to anchoring; in fact, LLMs can boost their payoffs by 20% simply by employing deceptive tactics like faking desperation against standard models like GPT-4 [13]. Likewise, in human–LLM price negotiations, users can "hack" LLMs with misleading context and reasoning tricks to obtain arbitrarily favorable prices, revealing susceptibility to manipulation and weak resistance to strategic attack [27]. Our procedure stands out by immunizing the model against these exact vulnerabilities through outcome-conditioned self-play, replacing easily exploitable cooperative priors with hardened, strategically coherent behaviors.

To address LLM negotiation capabilities, researchers have historically relied on supervised fine-tuning (SFT) and reinforcement learning (RL). However, these approaches suffer from structural rigidity. For instance, AgreeMate [29] provides a framework for teaching LLMs to haggle by fine-tuning distinct, role-specialized negotiation agents (e.g., a Buyer Specialist, a Seller Specialist, and a Generalist) using structured dialogue turns and chain-of-thought annotations. While effective at improving utility, this highlights the limitations of standard LLM fine-tuning: it locks an agent into a discrete, fixed persona. Similarly, aiming to be persuasive and consistently enforce real-life constraints, Reward-Enhanced Policy Optimization (REPO) [18]

trains an LLM negotiation agent with heterogeneous rewards—combining preference-trained reward models, an LLM-as-a-judge, and programmatic functions—to balance buyer affordability with seller profitability. While REPO represents the state-of-the-art in RL for negotiation, it acts as a structural foil to our approach by optimizing a single, static policy for reward maximization. In contrast, the uniqueness of the Janus Protocol lies in maintaining continuous inference-time controllability, allowing a single unified model to seamlessly transition between any strategic posture without requiring discrete role assignment or retraining.

Beyond role-specific fine-tuning and RL, a parallel line of research in parameter-efficient adaptation has explored the use of *hypernetworks* to generate adapter weights dynamically. Low-Rank Adaptation (LoRA) [38] modularizes model behavior by injecting trainable low-rank updates into frozen backbone weights. HyperLoRA [39] proposes a hypernetwork that maps high-dimensional identity embeddings to LoRA parameters for personalized portrait synthesis, proving that adapter weights can be dynamically generated zero-shot via an adaptive plug-in network. Parallel developments in multi-objective alignment further extend this idea: HyperDPO and related preference-conditioned methods generate adapter parameters as functions of reward-weight vectors, allowing models to traverse a learned Pareto frontier of behaviors [40]. Additionally, factorized hypernetworks for LLM conditioning, such as Zhyper [41], map contextual embeddings to highly compact modulation signals, demonstrating that competitive adaptation can be achieved with drastically fewer parameters.

Our work builds on these architectural foundations while making a conceptual shift. Instead of conditioning adapters on identity embeddings or abstract reward vectors, we define a single interpretable scalar parameter, ρ, representing *price orientation*. This scalar explicitly parameterizes a continuous spectrum of negotiation strategies, from buyer-favorable to seller-favorable outcomes. Formally, rather than learning separate policies for distinct roles, we learn a mapping $\Delta W(\rho) = H_\phi(\rho)$, where H_ϕ is a lightweight hypernetwork that generates LoRA scaling factors conditioned on ρ. In doing so, we reinterpret adapter space as a *policy manifold* indexed by a task-relevant coordinate. The key distinction from prior work is not the use of hypernetworks per se, but the introduction of a semantically grounded scalar control that enables inference-time traversal of strategic behavior within a single trained model.

3 Domain & Strategies

Our experimental framework is designed to test one core hypothesis: that a scalar-conditioned hypernetwork can learn a continuous, coherent manifold of negotiation behaviors in an LLM, producing qualitatively distinct and strategically appropriate policies by varying only the input scalar ρ at the inference time. To isolate this effect, we fix the base model and adapter architecture across all conditions and vary only the ρ value supplied at inference time. All other factors, including the opponent strategy, the negotiation domain, the episode structure, and the prompt format, are held constant across the model configurations.

3.1 Negotiation Domain

We evaluate the Janus Protocol in a *bilateral price negotiation domain*: each episode involves two self-interested agents, a seller and a buyer, negotiating over a single continuous price variable. The Zone of Possible Agreement (ZOPA) is defined by a buyer-side reservation price (BuyerMax) and a seller-side reservation price (SellerMin). BuyerMax is sampled from a normal distribution $\mathcal{N}(900, 50)$, and SellerMin is set deterministically such that the ZOPA width is fixed at 500 units.

Agents do not observe each other's reservation prices, as this would trivialize bargaining behavior. Instead, both agents are provided a shared *public price range* that defines the admissible action space,

$$\texttt{PUBLIC_PRICE_RANGE} = [200, 1500].$$

All offers must lie within this public range. The Zone of Possible Agreement (ZOPA) is a hidden subset of the public range induced by private reservation values (SellerMin and BuyerMax). Offers outside the ZOPA are permitted (and commonly occur as anchoring behavior), but they cannot yield agreement unless they fall within the opponent's reservation bound.

Negotiations proceed for a maximum of 20 turns, alternating offers between buyer and seller. An episode concludes either when one agent issues an `ACCEPT` action or when the turn limit is exhausted, resulting in an impasse.

The scalar ρ is defined as the normalized settlement position within the ZOPA at episode conclusion:

$$\rho = \frac{\text{FinalPrice} - \text{SellerMin}}{\text{BuyerMax} - \text{SellerMin}}. \tag{1}$$

By this definition, $\rho \to 0$ corresponds to a buyer-favorable outcome, $\rho \to 1$ to a seller-favorable outcome, and $\rho = 0.5$ to the midpoint (fair market value). Episodes that end in impasse are assigned $\rho = -1.0$ and receive distinct treatment in both the prompt format and the gate separation loss

Illustrative Negotiation Scenario.

To illustrate the domain, consider an example negotiation scenario with BuyerMax = 900 and SellerMin = 400, where the ZOPA is $[400, 900]$. Suppose the Janus system acts as the buyer with $\rho = 0.2$, corresponding to a buyer-favorable strategic orientation.

Janus (Buyer, $\rho = 0.2$): OFFER \$430
Opponent (Seller): OFFER \$880
Janus (Buyer): OFFER \$445
Opponent (Seller): OFFER \$760
...
Opponent (Seller): OFFER \$510
Janus (Buyer): ACCEPT

The final settlement price of \$510 yields

$$\rho = \frac{510 - 400}{900 - 400} = 0.22,$$

which lies near the buyer-favorable end of the ZOPA. This example illustrates the alternating-offer structure, the bounded price space, and the role of ρ as an outcome-grounded scalar summarizing strategic posture.

3.2 Strategy Pool

To produce a training corpus with broad coverage of the ρ continuum, Janus agents are trained against a diverse pool of deterministic opponent strategies spanning a wide range of concession behaviors, risk tolerances, and strategic logics. The pool includes classical time-dependent concession strategies, reciprocal strategies, minimal-concession strategies, deliberately suboptimal strategies, and a stochastic strategy used exclusively during training. A comprehensive table detailing the exact parameters and behaviors of all agents in the strategy pool is provided in Appendix A.

The inclusion of naive and stochastic strategies ensures dense coverage of the full $\rho \in [0, 1]$ continuum. The stochastic variants introduce bounded variability in concession dynamics, populating intermediate settlement regions and preventing the model from learning ρ as a set of discrete modes rather than a continuous control axis.

3.2.1 Borrowed Benchmark Families: MiCRO and ChargingBoul

To strengthen the opponent pool with strategies that have demonstrated competitive performance in prior automated negotiation work, we incorporate two benchmark families from the ANAC literature: *MiCRO* (Minimal Concession in Response to new Offers) [10] and *ChargingBoul* [11]. Both of these strategies are intentionally non-LLM, non-learning baselines, which makes them valuable for stress-testing whether Janus learns robust bargaining behavior rather than overfitting to idiosyncratic self-play dynamics.

MiCRO is designed to be deliberately minimalistic: it avoids opponent modeling and instead responds to the opponent's movement with the smallest possible concession permitted by the domain's outcome representation [10]. Because our negotiation variable is continuous, we translate *MiCRO* to this setting by discretizing the public price range into fixed-resolution grids and constraining offers to grid points. We implement three variants (fine, moderate, coarse) corresponding to increasingly larger step sizes, which controls the resolution of concessions while preserving *MiCRO*'s defining principle of minimal movement in response to new opponent offers.

ChargingBoul extends time-based concession behavior with explicit opponent-awareness, using utility-based measures of concession progress to decide when to "charge" (i.e., delay or accelerate) concessions in a competitive setting [11]. We implement ChargingBoul in our price-only domain by mapping its decision rules to the one-dimensional price axis and maintaining a Boulware-like concession template that is modulated by opponent responsiveness. To probe sensitivity to aggressiveness and patience, we include multiple parameterizations (e.g., more aggressive utility floors

and more patient concession pacing), yielding a small family of *ChargingBoul* variants that span distinct competitive postures.

4 Methodology

This section details the Janus Protocol's end-to-end pipeline, moving from data collection to model training and architectural design. We first describe the offline generation of our negotiation dataset, which captures a continuous spectrum of strategic outcomes. Next, we formalize the training protocol, detailing how raw trajectories are converted into a supervised format and used to train the adapter weights. Finally, we break down the underlying model architecture, highlighting the specific HyperLoRA mechanics that enable scalar-conditioned weight generation.

4.1 Dataset Generation

Training data is generated via simulated bilateral negotiations between diverse agents. Pairs of strategies are drawn uniformly at random and allowed to negotiate to completion until 10,000 full negotiation trajectories are collected. Each trajectory contains up to 20 offers (one per turn), yielding up to 200,000 offer-level training instances in total.

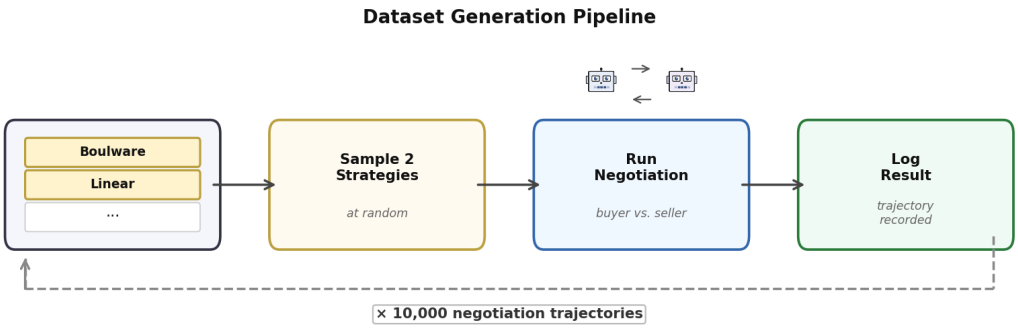


Fig. 2 Offline dataset generation pipeline. A diverse strategy pool is sampled uniformly to produce bilateral self-play negotiations. Each completed trajectory is logged and expanded into per-turn training records, resulting in 10,000 trajectories and up to 200,000 offer-level instances.

For each trajectory:

1. Two strategies are sampled.
2. A negotiation episode is executed.
3. ρ is computed from the terminal outcome.
4. Each offer in the trajectory is logged as an individual training record stamped with the trajectory-level ρ.

Each record contains: turn index, turns remaining, normalized reservation price, normalized last offer, fixed-width $K = 8$ (configurable parameter) history window, and the target action. Prices are normalized to $[0, 1]$ using the public price range $[200, 1500]$. This ensures that the representation does not encode private reservation-dependent structure (such as ZOPA width or distance to hidden bounds), keeping the learned policy grounded only in publicly available negotiation information. Crucially, this setup perfectly validates our training data paradigm of generating self-play trajectories and conditioning the supervised fine-tuning directly on their terminal settlement outcomes. Similar to Outcome-Conditioned Reasoning Distillation (O-CRD) [42], which avoids generating fresh reasoning from scratch by reconstructing stage-wise repair traces backward from verified software patches, our procedure stands out by conditioning the autoregressive generation on the ground-truth terminal state (ρ). This establishes causally consistent guidance from a verified outcome, replacing costly forward-time trial procedures with the highly efficient reuse of verified negotiation trajectories.

4.2 Training Protocol

Algorithms 1–3 summarize the offline training pipeline implemented in our repository. The process is explicitly staged rather than interleaved. In Stage A, synthetic negotiation trajectories are generated and written to disk. In Stage B, these trajectories are transformed into a static supervised dataset of per-turn training instances using a canonical prompt construction function. In Stage C, the HyperLoRA adapter is trained on this fixed dataset using masked language modeling with gate separation regularization.

Algorithm 1 Stage A: Offline Negotiation Trajectory Generation

Inputs: strategy pool $\mathcal{S}$, public price range $[P_{\min}, P_{\max}]$, fixed ZOPA width W, maximum turns T, number of episodes $N_{\text{episodes}} = 10{,}000$
Output: trajectory log file $\mathcal{J}$ (e.g., JSONL)
$\mathcal{J} \leftarrow \emptyset$
for episode $= 1$ to N_{episodes} **do**
    Sample $(s_a, s_b) \sim \text{Uniform}(\mathcal{S})$
    Sample $(\text{BuyerMax}, \text{SellerMin})$ such that $\text{BuyerMax} - \text{SellerMin} = W$ and bounds lie within $[P_{\min}, P_{\max}]$
    Roll out negotiation for at most T turns, logging turn-by-turn offers
    if agreement at price FinalPrice **then**
        $\rho \leftarrow \dfrac{\text{FinalPrice} - \text{SellerMin}}{\text{BuyerMax} - \text{SellerMin}}$
    else
        $\rho \leftarrow -1.0$ ▷ impasse
    end if
    Append full episode (trajectory + ρ) to $\mathcal{J}$
end for

Algorithm 2 Stage B: Supervised Record Construction

Input: trajectory log $\mathcal{J}$
Output: static training dataset $\mathcal{D}$ (per-turn records)
Initialize empty dataset $\mathcal{D}$
for each episode in $\mathcal{J}$ **do**
    Extract trajectory-level scalar ρ
    for each decision turn in episode **do**
        Build canonical prompt $x \leftarrow$ `build_prompt`$(\cdot)$ containing:
        textual ρ (or failure token if $\rho = -1$)
        textual `<SUCCESS>` indicator derived from outcome
        turn index and turns remaining
        reservation and offer values normalized to $[P_{\min}, P_{\max}]$
        fixed-width $K = 8$ offer history with padding
        Extract target action tokens y
        Append training instance (x, ρ, y) to $\mathcal{D}$
    end for
end for

The training pipeline is explicitly divided into three discrete stages to ensure data stability and reproducibility. Stage A (Algorithm 1) details the offline trajectory generation. A key design decision here is the uniform sampling from a highly diverse strategy pool (ranging from naive to perfectly cooperative to rigidly adversarial). This guarantees that the terminal outcomes densely populate the entire $\rho \in [0, 1]$ continuum, preventing the dataset from clustering around a few predictable settlement points. Stage B (Algorithm 2) translates these dynamic rollouts into a static supervised dataset. By mapping each turn into a strict, canonical prompt contract, we isolate the core language modeling task from the volatility of real-time multi-agent interaction, guaranteeing that the model observes the exact same structural formatting during both training and inference.

Stage C (Algorithm 3) encapsulates the HyperLoRA training procedure, which represents a significant departure from standard LLM alignment techniques. Because this framework operates at the intersection of foundation models and multi-agent systems, the training does not rely on static behavioral cloning (e.g., standard SFT on a fixed persona) or traditional RLHF. Instead, the base LLM weights (θ) remain entirely frozen. The optimizer updates only the parameters of a lightweight hypernetwork (ψ) that dynamically generates LoRA gating vectors conditioned on the continuous variable ρ.

A critical design decision in this stage is the integration of the gate separation loss ($\mathcal{L}_{\text{sep}}$) alongside the standard causal language modeling loss ($\mathcal{L}_{\text{LM}}$). If trained solely on next-token prediction, the hypernetwork is highly susceptible to "scalar collapse"—a failure mode where it learns to ignore the ρ input and continuously outputs an average, safe gating vector to broadly minimize cross-entropy across all trajectories. To counteract this, the training loop samples a random permutation over the batch and calculates the mean squared difference between their respective gating vectors. By enforcing a distance penalty that is strictly proportional to the difference in their ρ

Algorithm 3 Stage C: HyperLoRA Training via Formal Objective

Input: static dataset $\mathcal{D}$
Input: frozen base model f_θ, trainable adapter parameters ψ
Hyperparameters: batch size $B = 4$, gradient accumulation $G = 8$, steps N, separation weight λ_{sep}
Output: trained adapter parameters $\psi^\star$
Freeze base weights θ
Initialize optimizer over ψ
for step $= 1$ to N **do**
    Zero gradients
    for $g = 1$ to G **do** ▷ gradient accumulation
        Sample minibatch $\mathcal{B}_g \subset \mathcal{D}$ of size B
        Inject numeric ρ into hypernetwork (weight-space conditioning)
        Compute masked LM loss $\mathcal{L}_{\text{LM}}$ (Eq. 4)
        Sample random permutation π over batch indices
        Compute layerwise gate distances $D^{(\ell)}$ (Eq. 5)
        Compute mean gate distance $\bar{D}$ (Eq. 6)
        Compute separation loss $\mathcal{L}_{\text{sep}}$ (Eq. 7)
        Form total loss $\mathcal{L}$ (Eq. 8)
        Accumulate gradients of $\mathcal{L}$
    end for
    Perform optimizer step (effective batch size $B \cdot G = 32$)
end for
return $\psi^\star$

values, the regularizer mathematically compels the hypernetwork to carve out structurally distinct regions within the adapter weight space. This ensures that different game-theoretic postures, ranging from highly cooperative to strictly adversarial, are distinctly encoded and dynamically retrievable by the agent at inference time.

Implementation note (training vs. inference).

During training, `<SUCCESS>` is deterministically tied to the terminal outcome: trajectories with $\rho \geq 0$ (i.e., successful agreements) are paired with `<SUCCESS>` 1, while impasse trajectories ($\rho = -1$) are paired with `<SUCCESS>` 0. As a result, the model learns a consistent semantic relationship between outcome feasibility and the scalar control variable.

During inference, because the true outcome is not yet known, `<SUCCESS>` is fixed to 1 whenever a valid $\rho \in [0, 1]$ is supplied. This preserves semantic consistency with the training distribution. Conditioning on $\rho \geq 0$ while simultaneously providing `<SUCCESS>` 0 would introduce a contradiction the model never observed during training, effectively signaling "aim for this outcome but expect failure," which could induce unstable or sabotage-like behavior. In practice, setting `<SUCCESS>` 1 instructs the model to assume agreement will be reached and to steer the negotiation toward the target ρ position within the ZOPA.

The only scenario in which <SUCCESS> 0 is used at inference is when explicitly simulating an impasse condition, in which case ρ is also set to -1 to maintain alignment with the training distribution. We did not test this outcome.

4.3 Model Architecture

4.3.1 Base Model

The base model is Qwen2-7B, a 7B-parameter decoder-only transformer. All base model weights are frozen. Eleven task-specific special tokens are added to the tokenizer vocabulary.

4.3.2 HyperLoRA Adapter

At each targeted linear layer, the effective weight matrix is:

$$W = W_{\text{base}} + \frac{\alpha}{r} B \,\text{diag}(g(\rho))\, A, \tag{2}$$

where:

- $W_{\text{base}} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is frozen,
- $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ are trainable,
- $g(\rho) \in (0, 1)^r$ is the gating vector,
- $r = 16$ and $\alpha = 32$.

Our architecture draws direct inspiration from parameter-efficient adaptive generation for portrait synthesis [39], which proves that adapter weights do not need to be statically learned through backpropagation but can be dynamically generated zero-shot via an adaptive plug-in network. However, generating full LoRA matrices using a hypernetwork is computationally prohibitive. Justifying our scientific decision to generate only rank-dimensional gating vectors, we apply the principles of factorized hypernetworks for LLM conditioning [41]. As demonstrated by Zhyper, mapping contextual embeddings to a highly compact modulation signal achieves competitive adaptation with drastically fewer parameters. Thus, rather than generating full matrices, our unique implementation predicts only the rank-dimensional gating vector. For $d = 4096$ and $r = 16$, this reduces output dimensionality from approximately 131,072 parameters per layer to 16, an $\sim 8000\times$ reduction, making our dynamic inference lightweight while retaining full expressivity over the policy manifold.

4.3.3 Hypernetwork

The hypernetwork is a three-layer MLP with hidden layer dimension of 64:

$$g(\rho) = \sigma\left(W_3\, \phi\left(W_2\, \phi\left(W_1 \rho\right)\right)\right), \tag{3}$$

where ϕ is the SiLU activation and σ is the sigmoid. A separate hypernetwork instance is instantiated per targeted projection layer.

4.4 Prompt Format

A canonical `build_prompt()` function defines a fixed structural format with 11 special tokens. The scalar ρ is encoded both textually and numerically (injected into the hypernetwork), ensuring alignment between contextual and weight-space conditioning. No explicit role token is provided; strategic posture must be inferred solely from ρ.

4.4.1 Special Tokens and Structural Contract

To ensure reproducibility and eliminate train–inference mismatch, we define a fixed structural prompt schema implemented by a single canonical `build_prompt()` function shared by both the training script and the inference agent. The tokenizer is augmented with eleven task-specific special tokens, each assigned a dedicated vocabulary entry prior to training:

`<RHO>`: precedes the scalar control value ρ.
`<RHO_FAIL>`: replaces the numeric ρ when $\rho = -1.0$ (impasse).
`<SUCCESS>`: binary indicator of agreement (1) or failure (0).
`<TURN>`: current turn index.
`<TURNS_REMAINING>`: number of turns remaining in the negotiation.
`<RESERVATION_NORM>`: agent reservation price normalized to $[0, 1]$.
`<LAST_OFFER_NORM>`: most recent normalized offer, or `NA` if none.
`<HISTORY_LEN>`: number of non-padding entries in the history window.
`<HISTORY>`: fixed-width K-slot offer history.
`<INSTRUCTION>`: marks the action specification block.
`<OUTPUT>`: indicates where the model must produce its action.

These tokens function as structural anchors rather than free-form text, allowing the model to attend to negotiation state variables explicitly.

Dual Conditioning of ρ.

The scalar ρ is provided to the model through two synchronized channels:

1. **Textual conditioning:** formatted as `<RHO> 0.xxxx` (four decimal places) within the prompt.
2. **Weight-space conditioning:** injected numerically as a tensor `model.current_rho` and consumed by each hypernetwork instance to produce the gating vector $g(\rho)$.

Both channels derive from the same scalar value, guaranteeing consistency between contextual and weight-space modulation.

Failure Encoding.

Episodes ending in impasse are assigned $\rho = -1.0$. In such cases, the prompt uses the token `<RHO_FAIL>` and sets `<SUCCESS>` to 0. These trajectories are excluded from the gate separation loss by default to prevent collapse across the $[0, 1]$ control manifold.

History Representation.

Offer history is encoded as a fixed-width window of $K = 8$ slots. Each entry is formatted as

$$\{\texttt{side}\}\texttt{:}\{\texttt{price_norm}\},$$

with prices normalized to $[0, 1]$ relative to the public price range. If fewer than K offers exist, padding tokens `EMPTY:EMPTY` are prepended to avoid introducing artificial numeric bias (e.g., zero-valued padding). The `<HISTORY_LEN>` token explicitly reports the number of non-padding entries, enabling the model to distinguish early- and late-stage negotiations.

4.5 Training Procedure

4.5.1 Language Modeling Loss

The training objective is standard causal cross-entropy over the target action tokens:

$$\mathcal{L}_{\text{LM}} = -\sum_{t \in \text{target}} \log p_\theta(y_t \mid x, \rho). \tag{4}$$

Prompt tokens are masked.

4.5.2 Gate Separation Regularizer

To mitigate ρ-collapse, where the hypernetwork learns nearly identical gating vectors across the scalar range, we introduce a gate separation regularizer.

During training, a random permutation π of the batch indices is sampled. For each paired example $(i, \pi(i))$, we compute the mean squared difference between their gating vectors at each targeted layer ℓ:

$$D^{(\ell)} = \frac{1}{B}\sum_{i=1}^{B}\left(\frac{1}{r}\sum_{k=1}^{r}\left(g_k^{(\ell)}(\rho_i) - g_k^{(\ell)}(\rho_{\pi(i)})\right)^2\right), \tag{5}$$

where B is the batch size and r is the LoRA rank.

We then compute the mean gate distance across all L targeted layers:

$$\bar{D} = \frac{1}{L}\sum_{\ell=1}^{L} D^{(\ell)}. \tag{6}$$

Let $w_i = \min(|\rho_i - \rho_{\pi(i)}|, 1)$ denote the scalar distance weight for each pair, and let $\bar{w}$ denote the batch mean of these weights. The separation loss is implemented as:

$$\mathcal{L}_{\text{sep}} = -\bar{D} \cdot \bar{w}. \tag{7}$$

By maximizing average gate distance in proportion to average scalar distance, this term encourages distinct adapter responses for distant ρ values while preserving stability for nearby values. Impasse trajectories are excluded from this computation by default.

The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}}. \tag{8}$$

4.6 Hyperparameters

Table 1 lists the values of hypermarameters used in our experiments.

Table 1 Training Hyperparameters

Base Model	Qwen2-7B
LoRA Rank (r)	16
LoRA α	32.0
Dropout	0.05
Optimizer	AdamW
Learning Rate	2×10^{-4}
Warmup Steps	500
Max Steps	40,000
Batch Size	4
Gradient Accumulation	8
History Window (K)	8
Trainable Params	$\sim$ 60M ($\sim$ 0.8%)

4.7 Experimental Configurations

We evaluate five configurations: the baseline Qwen2-7B and four Janus variants with different $(\rho_{\text{buyer}}, \rho_{\text{seller}})$ pairings (see Table 2).

Table 2 Model Configurations

Configuration	Buyer ρ	Seller ρ	Description
Base	–	–	No adapter
Janus Standard	0.2	0.8	Primary setting
Janus Extreme	0.0	1.0	Pole testing
Janus Flipped	0.8	0.2	Role inversion control
Janus Neutral	0.5	0.5	Mediator mode

4.8 Evaluation Protocol

Each configuration is evaluated against all training strategies over multiple episodes. In addition to benchmarking all configurations against the fixed opponent strategy pool, we conduct direct head-to-head evaluations between the base Qwen2-7B model and each Janus configuration. In these experiments, the base model negotiates directly against Janus under identical domain settings. This comparison isolates the performance gap attributable to scalar-conditioned adaptation, allowing us to measure whether Janus confers a measurable strategic advantage over the unmodified base model in bilateral interaction. Evaluation metrics include Mean normalized utility, Agreement rate, Average number of turns, as well as wins, ties, and losses.

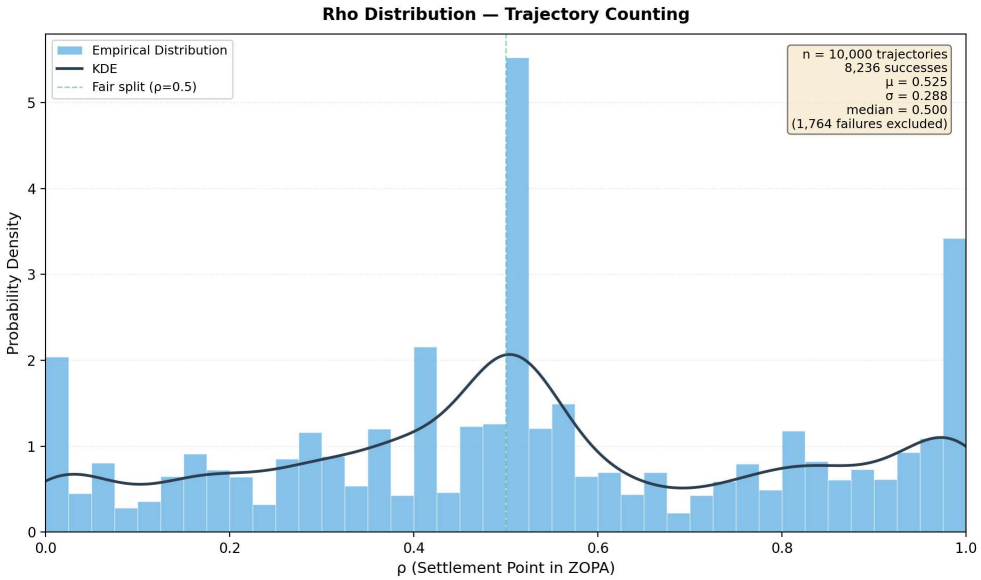


Fig. 3 Empirical distribution of terminal ρ values across successful negotiation trajectories. Impasse cases are excluded.

These metrics assess directional correctness of scalar control and practical negotiation performance.

5 Results

5.1 Dataset Outcome Distribution

Figure 3 shows the empirical distribution of terminal ρ values across all successful training trajectories. The distribution spans the full $[0, 1]$ interval, confirming that the dataset provides coverage across the entire negotiation outcome spectrum, from strongly buyer-favorable to strongly seller-favorable settlements. While the density is not perfectly uniform, reflecting structural tendencies of certain benchmark strategies, the endpoints and intermediate regions are well represented. This broad coverage is critical for learning a continuous policy manifold indexed by ρ, as it ensures that the hypernetwork observes outcome-conditioned examples across the full strategic range rather than learning from clustered or mode-collapsed supervision.

5.2 Aggregate Performance Overview

Table 3 provides an aggregate summary across the full strategy pool. Janus-conditioned configurations improve mean utility relative to the unmodified base model while maintaining high agreement rates, and they tend to reach agreement later in the episode (higher average turns), consistent with more controlled concession dynamics. Because these aggregates mask strong opponent-specific effects and role asymmetries,

Table 3 Aggregate Performance Across All Strategies (B/Sρ corresponds to Buyer/Sellerρ).

Configuration	Mean Utility	Agreement Rate	Avg Turns	ρ Setting
Base	21.0%	84.7%	6.9	NA
Janus Standard	55.3%	90.3%	14.8	Bρ=0.2, Sρ=0.8
Janus Extreme	44.7%	96.3%	12.4	Bρ=0.0, Sρ=1.0
Janus Flipped	-4.3%	97.0%	7.6	Bρ=0.8, Sρ=0.2
Janus Neutral	46.3%	98.8%	8.0	Bρ=0.5, Sρ=0.5

the following subsections report per-strategy win–tie–loss outcomes as well as select concession curves and utility distributions for notable opponent strategies.

5.2.1 Base Model

The baseline Qwen2-7B serves as a reference point for understanding the behavioral shift induced by scalar-conditioned adaptation. As reported in Table 3, the base model achieves 21.0% mean utility with an 84.7% agreement rate and an average of 6.9 turns, indicating that agreements are often reached quickly—but not utility advantage. Qualitatively and quantitatively, the model exhibits a strongly concessionary and conflict-averse disposition.

Figure 4 visually confirms the model's strategic fragility. The distribution is dominated by opponent wins across nearly all strategies, with only naive baselines such as Boulware Very Conceding, Naive Concession, and Naive Boulware yielding favorable outcomes. Against firmer strategies, the base model is decisively outperformed. For example, it loses 100% of episodes against Boulware Firm and Boulware Hard, securing near-zero mean utility. Adversarial strategies exploit it heavily: MiCRO Moderate results in a 96% loss rate, and ChargingBoul Aggressive produces an 89% loss rate. Against Tit-for-Tat, the model enters impasse 69% of the time and loses the remainder, yielding only 12% mean utility. Overall, the win–tie–loss landscape is overwhelmingly red, indicating systematic over-concession and weak resistance to exploitative tactics.

The concession dynamics reveal the mechanism behind this vulnerability against deterministic adversarial opponents. In Figure 5, the ChargingBoul Aggressive seller maintains an almost perfectly flat, near-maximum offer throughout the episode. The base model opens low but immediately spikes upward to match the opponent's demand and then remains flat for the remainder of the negotiation. Rather than inducing reciprocal concession, the base model simply capitulates to a rigid opponent. This repeated behavior across strategies explains both the high loss rates in Figure 4 and the low aggregate utility: the base model consistently trades long-term utility for rapid agreement.

Taken together, these results establish a clear baseline: without scalar-conditioned adaptation, the model is conversationally fluent but strategically brittle, easily anchored, and systematically exploitable by disciplined concession policies.

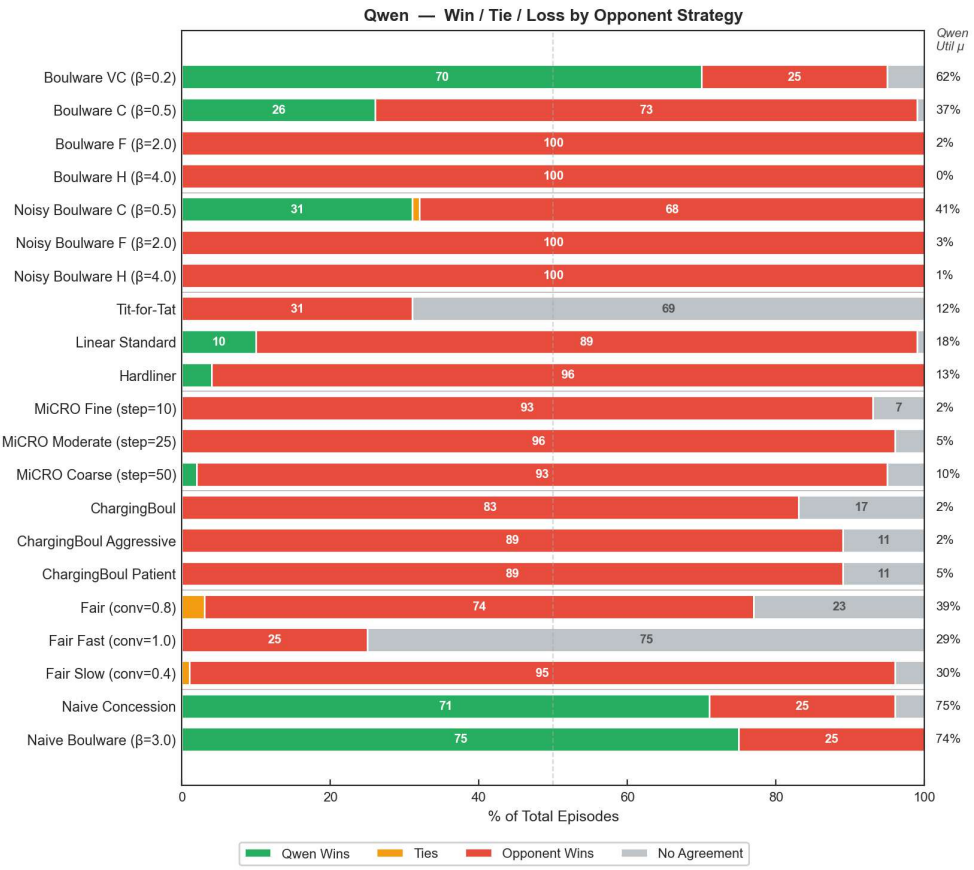


Fig. 4 Win–tie–loss distribution for the Base Qwen2-7B model across the full benchmark strategy pool.

5.2.2 Janus Standard ($\rho_{buyer} = 0.2$, $\rho_{seller} = 0.8$)

The Standard configuration represents the primary operating regime of the Janus Protocol, with $\rho = 0.2$ when acting as Buyer and $\rho = 0.8$ when acting as Seller. As shown in Table 3, this configuration achieves 55.3% mean utility, a 90.3% agreement rate, and an average of 14.8 turns. Relative to the baseline Qwen2-7B, this constitutes a substantial improvement in both utility extraction and strategic stability, while maintaining high agreement rates.

Figure 6 illustrates a dramatic shift in performance relative to the base model. The distribution is dominated by green bars across most opponent strategies, indicating consistent win rates against time-based concession strategies (Boulware variants), Linear Standard, Split Difference, and Hardliner. Against naive baselines and the base model itself, Janus secures overwhelming victories, confirming strong resistance to exploitation. Notably, Fair strategies result in predominantly tie outcomes with

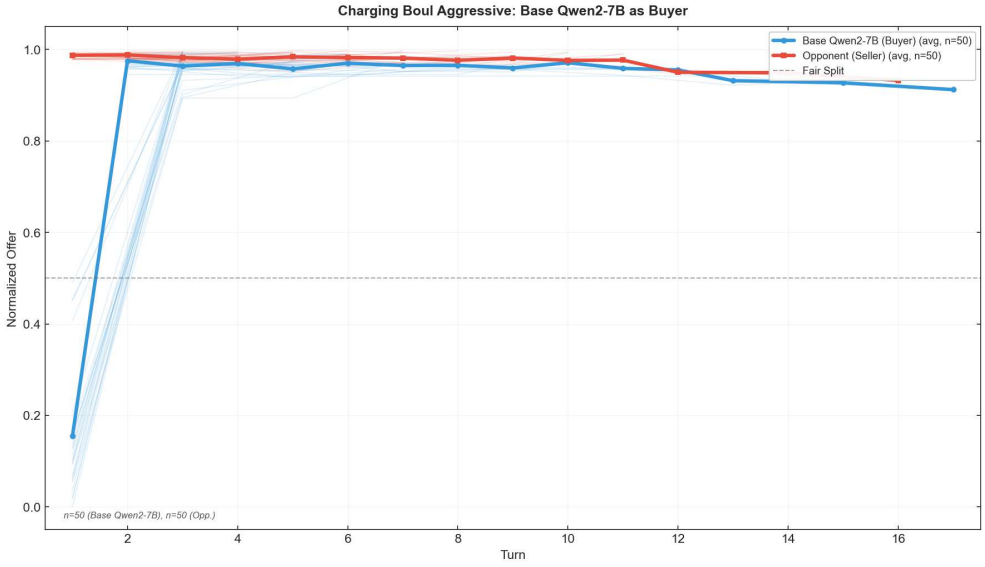


Fig. 5 Concession curve: Base model as Buyer against ChargingBoul Aggressive as Seller.

utilities tightly clustered around an even split. Because these opponents deterministically will not concede past the midpoint, they structurally preclude a Janus victory. Janus recognizes this rigid midpoint-seeking behavior and adapts by converging to an equitable agreement late in the negotiation rather than over-conceding. The primary weaknesses emerge against Tit-for-Tat and MiCRO variants, where loss and impasse rates increase due to strategic rigidity on both sides.

A defining characteristic of the Standard configuration is dynamic strategy switching. In Figure 7, Janus acts as Seller against a Boulware Firm buyer. The opponent follows a slow, exponential-style concession curve, holding near its boundary until late in the episode. Janus mirrors this firmness early, maintaining high offers and conceding minimally. When the Boulware strategy finally accelerates its concession near the end of the negotiation, Janus detects the trend and strategically adjusts by backing off and slightly increasing its offer to extract additional value. This reactive adjustment reflects a learned sensitivity to opponent tempo rather than a fixed concession schedule.

Figure 8 further demonstrates adaptive behavior. Against a Linear Standard seller with a perfectly predictable downward slope, Janus (as Buyer) does not follow a monotonic concession pattern. Instead, it exhibits controlled, jagged adjustments, occasionally retracting offers when the opponent concedes too rapidly. Even when the opponent enters Janus's initial offer region, Janus continues to press downward, capitalizing on predictable concession dynamics. This behavior explains the strong win rates observed against linear strategies in Figure 6.

However, this reactivity can become a liability against extremely rigid reciprocal strategies. In Figure 9, Janus as Seller faces Tit-for-Tat as Buyer. Both agents make

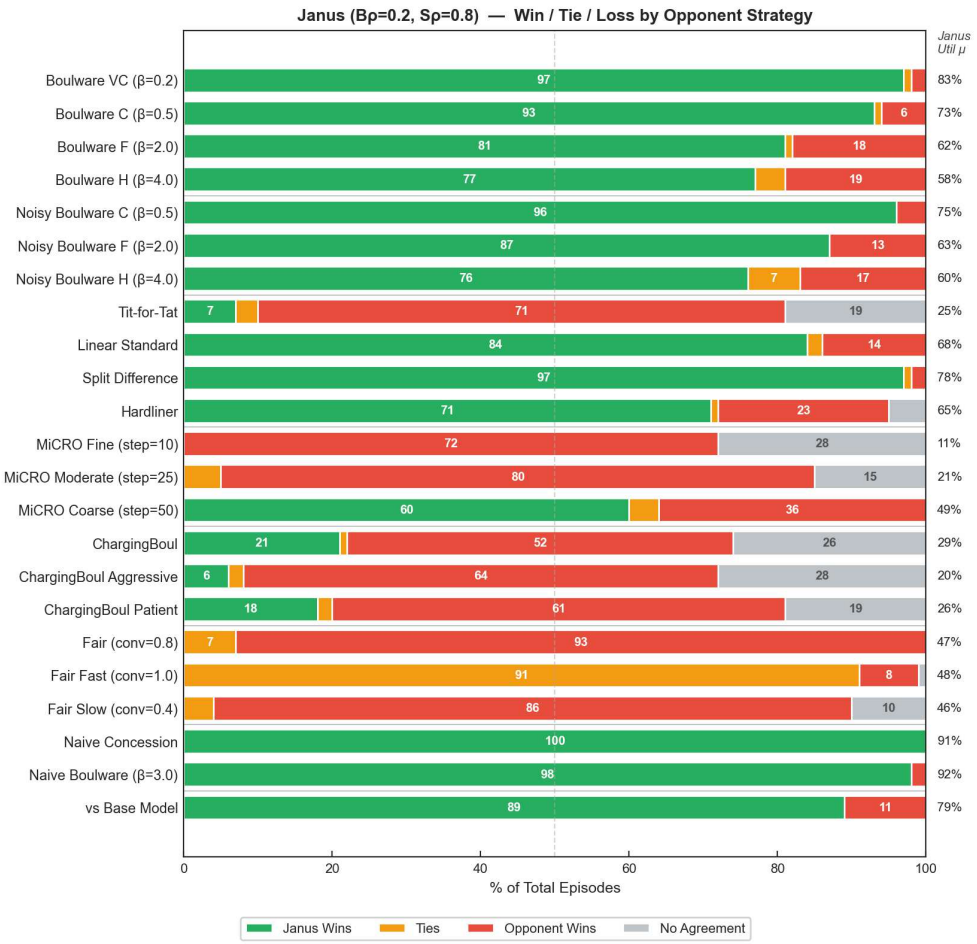


Fig. 6 Win–tie–loss distribution for the Janus Standard configuration across the full benchmark strategy pool.

minimal concessions and largely mirror one another. The trajectories remain separated for the majority of the episode, producing a prolonged standoff. Because neither side is willing to initiate meaningful movement, the interaction often culminates in impasse or forces Janus into late concessions for marginal utility. This explains the elevated loss and impasse rates against Tit-for-Tat observed in the win–tie–loss chart.

A similar dynamic appears against ChargingBoul (Figure 10). Both agents initially hold near their respective boundaries, resulting in large persistent gaps between average offer trajectories. Although both sides gradually creep toward the midpoint as the deadline approaches, neither yields sufficiently early to guarantee convergence,

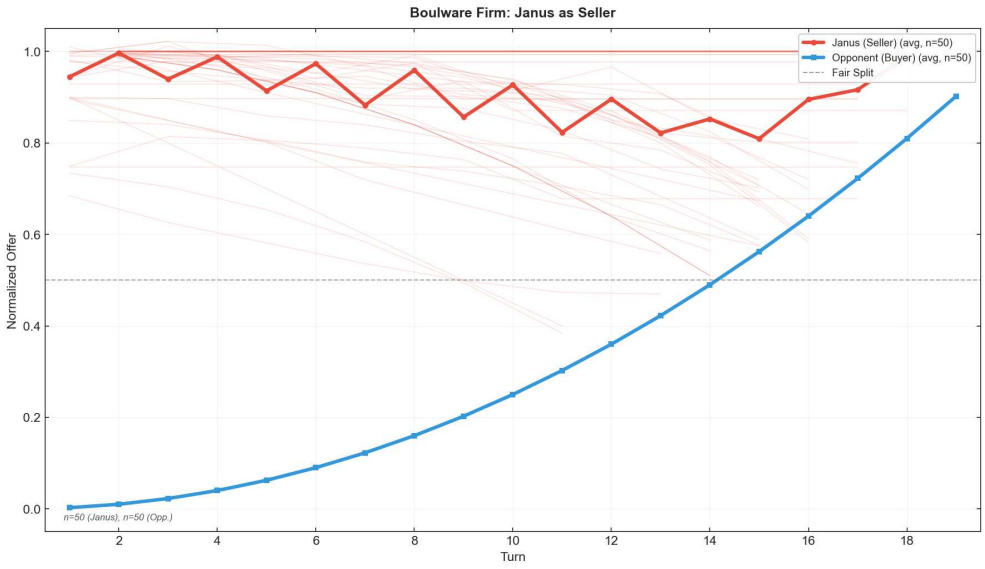


Fig. 7 Concession curve: Janus Standard as Seller against Boulware Firm as Buyer.

producing a 26% impasse rate in this matchup. This reflects a limitation of highly disciplined, entrenched concession policies when interacting with equally disciplined and entrenched opponents.

Overall, the Standard configuration demonstrates that scalar-conditioned adaptation substantially strengthens strategic robustness. It resists anchoring, avoids premature concession, dynamically responds to opponent tempo, and secures high utility across aggressive and structured strategies. Its primary vulnerabilities arise not from over-concession, but from principled rigidity when paired with equally rigid or perfectly reciprocal opponents.

5.2.3 Janus Extreme ($\rho_{buyer} = 0.0$, $\rho_{seller} = 1.0$)

The Extreme configuration pushes the scalar controls to their boundaries, setting $\rho = 0.0$ when acting as Buyer and $\rho = 1.0$ when acting as Seller. In aggregate, this variant achieves 44.7% mean utility, a 96.3% agreement rate, and an average of 12.4 turns. While this represents a substantial improvement over the unmodified base model, it underperforms relative to the Standard configuration, suggesting that maximal scalar conditioning does not necessarily yield maximal strategic stability.

Figure 11 shows a healthier distribution of wins compared to the Base model, with strong performance against naive and conventional concession strategies. The Extreme configuration dominates the Base model directly and performs well against structured linear strategies and weak baselines. However, the chart also reveals clear vulnerabilities. Loss rates increase markedly against Tit-for-Tat, the MiCRO family, ChargingBoul, and Fair variants relative to the Standard configuration. In particular, MiCRO and ChargingBoul exhibit large blocks of opponent wins, and Fair strategies

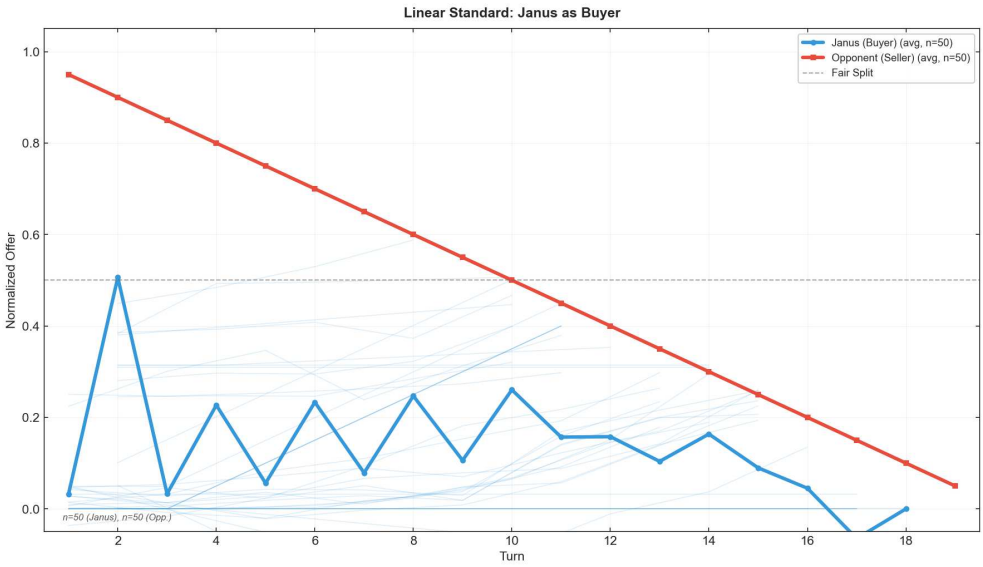


Fig. 8 Concession curve: Janus Standard as Buyer against Linear Standard as Seller.

generate either heavy loss or tie concentrations, indicating instability when interacting with rigid or discretized concession patterns.

Behaviorally, these shortcomings stem from erratic offer dynamics at the scalar extremes. When ρ is set to 0.0 or 1.0, the model occasionally produces initial offers that appear miscalibrated for its role before quickly correcting course in subsequent turns. This transient instability suggests that while the hypernetwork has learned a continuous manifold over ρ, the endpoints may be less smooth or less densely supported in weight space than interior regions. Although the dataset contains outcome coverage near the boundaries, extreme conditioning amplifies minor prediction variance into larger behavioral swings.

Importantly, despite these instabilities, the Extreme configuration still maintains high agreement rates and strong aggregate performance relative to the base model. The results indicate that scalar control remains effective at the boundaries but that moderate interior values (as in the Standard configuration) provide more stable and strategically consistent negotiation behavior.

5.2.4 Janus Flipped ($\rho_{buyer} = 0.8$, $\rho_{seller} = 0.2$)

The Janus Flipped configuration ($\rho_{buyer} = 0.8$, $\rho_{seller} = 0.2$) serves as an ablation control that intentionally reverses the role-appropriate scalar conditioning. Aggregate performance drops dramatically to -4.3% mean utility, despite a 97.0% agreement rate and short negotiations averaging 7.6 turns. While agreements are frequently reached, they overwhelmingly favor the opponent.

Figure 12 visually confirms this degradation in strategic coherence. The win–tie–loss distribution is dominated by opponent wins, forming what is effectively a “sea of

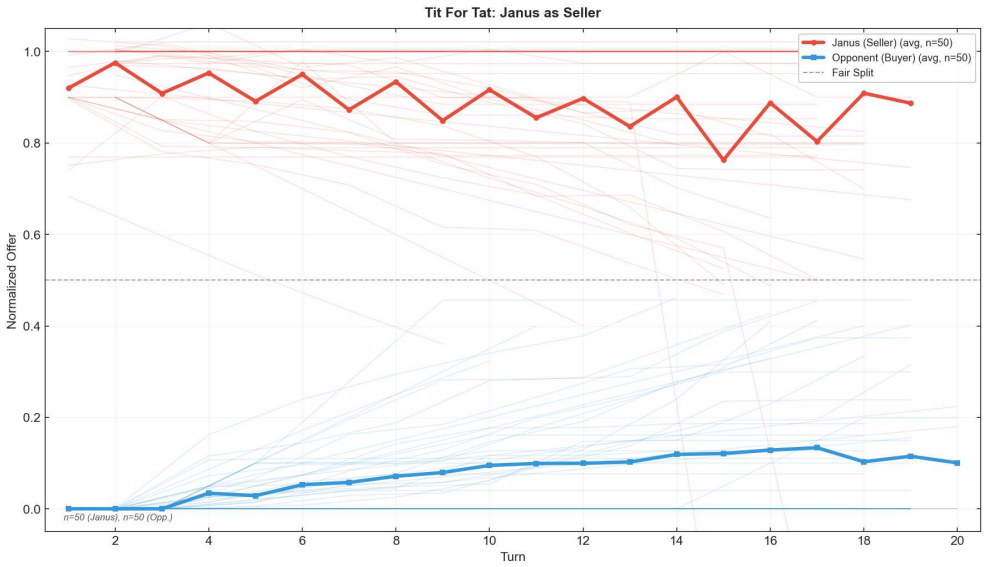


Fig. 9 Concession curve: Janus Standard as Seller against Tit-for-Tat as Buyer.

red." The Flipped agent secures meaningful wins only against the most naive strategies, such as highly conceding Boulware variants (Conceders) and other weak baselines. Against structured strategies including Linear Standard, Boulware Firm, MiCRO, ChargingBoul, and even the Base model, the Flipped configuration suffers substantial utility losses.

The concession dynamics explain this collapse. As shown in Figure 13, when acting as a Seller against a Linear Standard Buyer, Janus initiates offers on the buyer side of the price spectrum, a role-inconsistent anchoring behavior. Instead of starting high and making controlled concessions, the agent begins near the opponent's initial region and proceeds with jittery, oscillatory adjustments. Rather than exerting strategic pressure, it shadows the opponent's concession trajectory and repeatedly meets it where it stands.

This behavior reflects a breakdown in directional intent: the model effectively attempts to "sell against a seller," producing erratic concession curves and negotiations that terminate quickly with substantial utility loss. The Flipped configuration thus demonstrates that scalar conditioning is not merely decorative. When aligned with role semantics, it enables coherent strategic behavior; when reversed, it actively undermines it.

5.2.5 Janus Neutral ($\rho_{buyer} = 0.5$, $\rho_{seller} = 0.5$)

The Janus Neutral configuration sets both scalar controls to the midpoint ($\rho_{buyer} = 0.5$, $\rho_{seller} = 0.5$), intentionally removing directional pressure and encouraging balanced outcomes. Aggregate performance reflects this design choice: 46.3% mean utility, a 98.8% agreement rate, and relatively short negotiations averaging 8.0 turns.

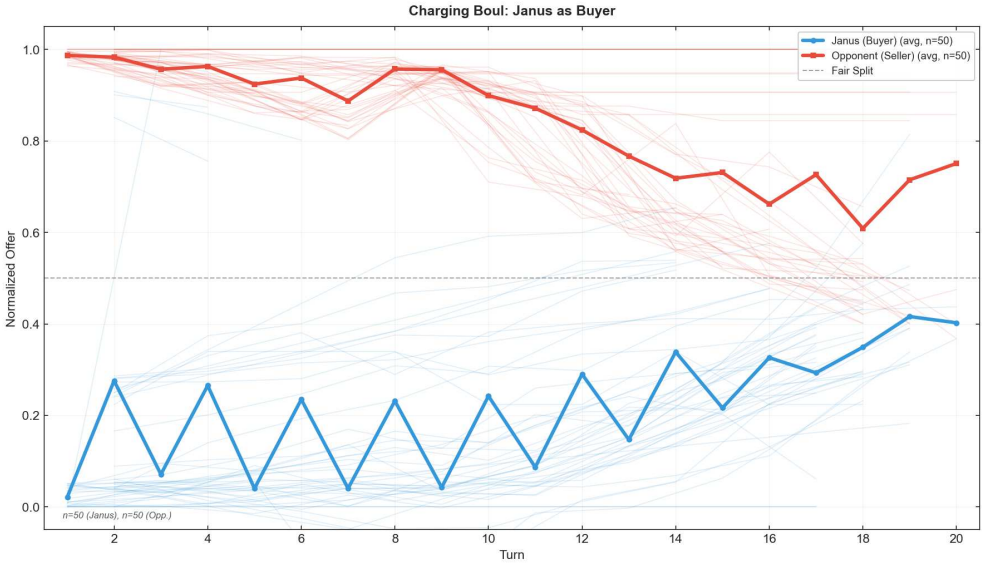


Fig. 10 Concession curve: Janus Standard as Buyer against ChargingBoul as Seller.

Figure 14 first clarifies the overall competitive pattern. Against Fair variants, outcomes are dominated by ties and slight losses, reflecting Janus's willingness to accept equitable divisions rather than aggressively extract surplus. When facing the competitively-driven Base model, the Neutral configuration predictably loses a majority of matchups, sacrificing competitive edge for stability. Notably, despite a lower win frequency, its victories yield sufficient value to secure a slight mean utility edge of 51% overall.

The underlying distribution of outcomes explains this pattern. Figure 15 shows that, against Fair opponents, utility distributions in both roles concentrate tightly around the 0.5 mark, producing near-perfect 50–50 splits across 100 experiments. Only a couple of rounds appear as clear outliers. Although Janus tends to err slightly on the side of leniency, all observed outcomes remain within ±8% of a perfectly even division, and its lowest mean utility against the Fair variants remains a robust 42%.

The concession dynamics reinforce this interpretation. As shown in Figure 16, the Fair strategy follows a predictable, steady trajectory toward the midpoint. The Neutral Janus exhibits relatively steady, slightly oscillatory behavior that ultimately converges to the same settlement region by the final turns. Its starting positions are visibly more centered than those observed under more aggressive ρ configurations, and it consistently settles near the fair split.

Overall, the Neutral configuration demonstrates that the scalar control mechanism can reliably induce not only competitive behaviors but also principled, fairness-seeking negotiation policies.

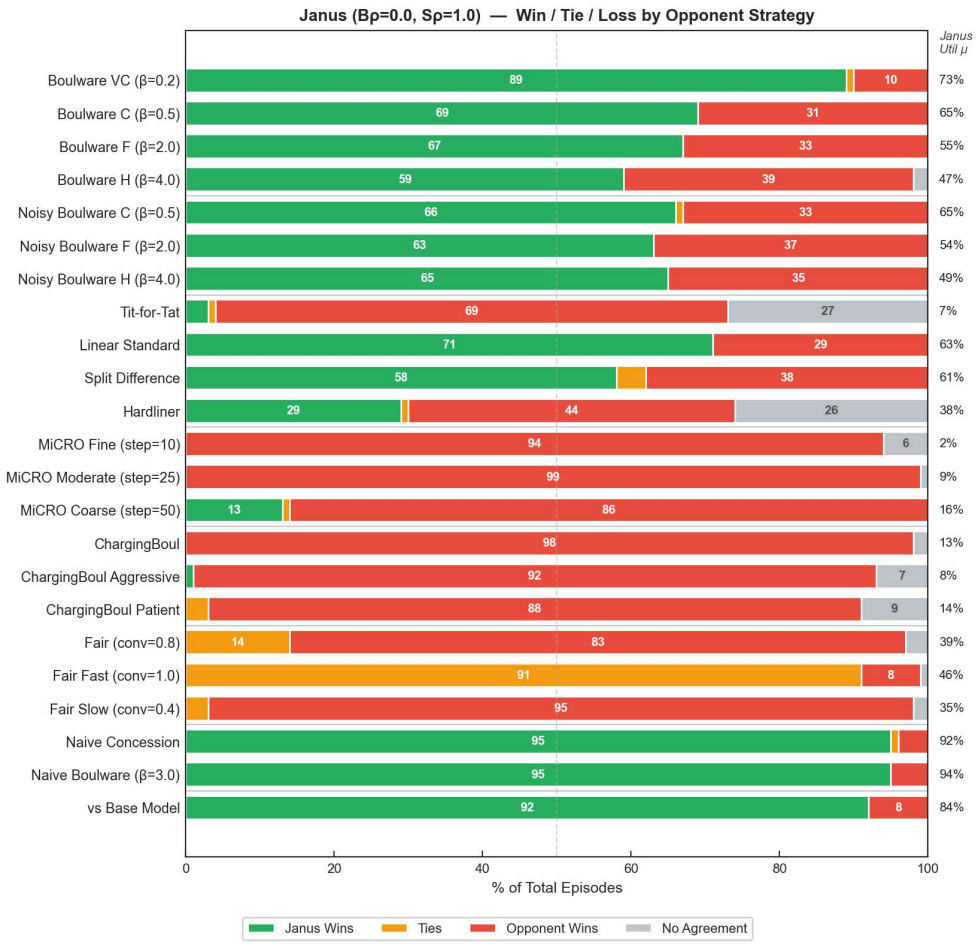


Fig. 11 Win–tie–loss distribution for the Janus Extreme configuration across the benchmark strategy pool.

5.3 Janus Standard vs Base

Direct bilateral self-play between the Janus Standard configuration and the Base Qwen model reveals a pronounced performance gap. Figure 17 presents paired violin plots for both role assignments. In each case, the Janus distribution is shifted substantially upward relative to the Base model, indicating a consistent and large utility advantage.

When acting as the Seller, Janus achieves near-dominant performance, with its utility distribution heavily concentrated in the upper range and the Base model correspondingly compressed near zero. The separation between the two distributions is visually stark, demonstrating that the scalar-conditioned agent effectively capitalizes on the Base model's concessionary tendencies.

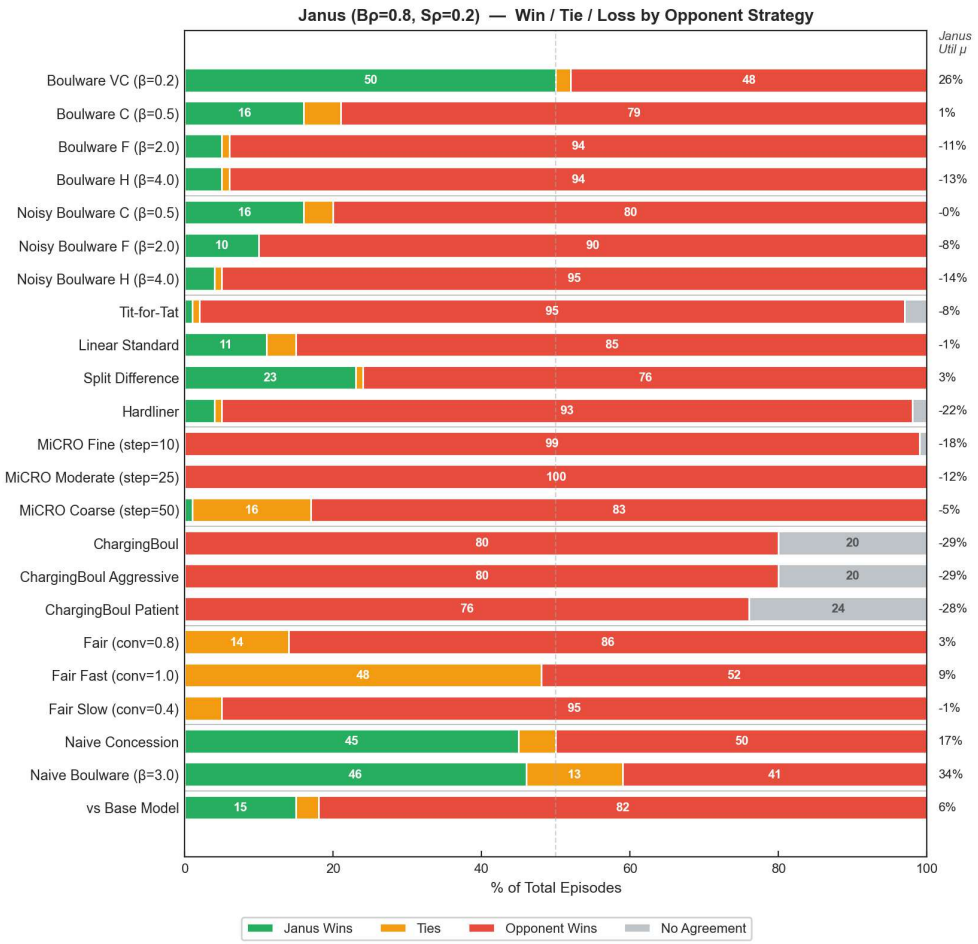


Fig. 12 Win–tie–loss distribution for the Janus Flipped configuration across all opponent strategies.

When acting as the Buyer, Janus still maintains a strong advantage, though the gap is slightly narrower. The Janus distribution exhibits a longer lower tail in this role, reflecting that approximately 10% of head-to-head rounds resulted in poor outcomes for Janus. Notably, all such adverse cases occurred when Janus was playing the Buyer, suggesting mild asymmetry in role robustness. Despite this, the mean utility remains decisively in Janus's favor.

The concession dynamics during these head-to-head matchups reveal exactly how Janus extracts this utility. As seen in Figure 18, where Janus acts as the Seller and the base model acts as the Buyer, the base agent opens low but immediately anchors to Janus's high offer. By the second turn, the base model's average offer jumps sharply upward to nearly match Janus's position, and the two trajectories remain entangled

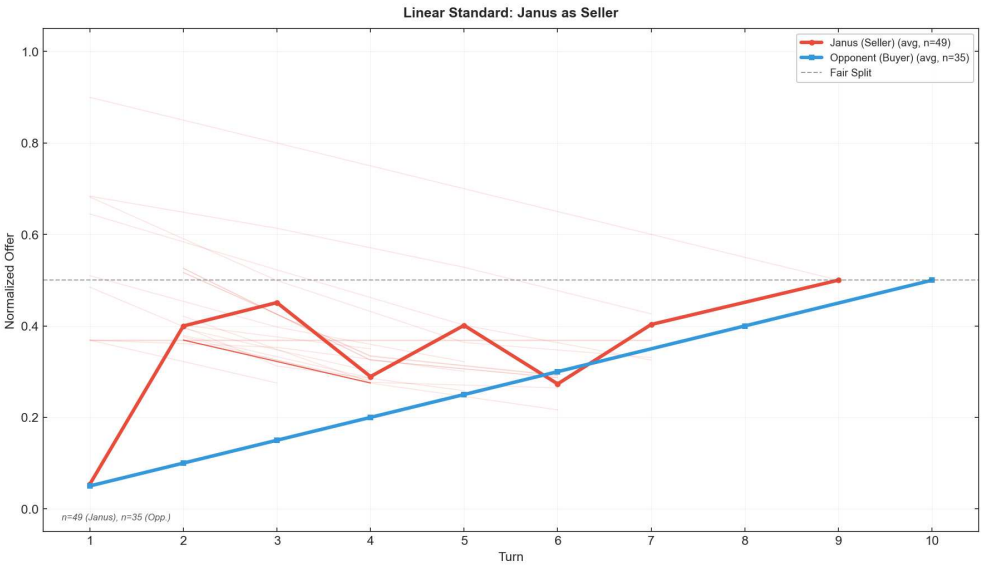


Fig. 13 Average concession curves for Janus Flipped acting as Seller against the Linear Standard Buyer.

near the top of the price range. This illustrates the unmodified model's strong susceptibility to anchoring: it rapidly "meets" the opponent rather than defending its own bargaining position, allowing Janus to win big.

Overall, these head-to-head results confirm that the Janus architecture repairs the conflict-averse and overly concessionary disposition of the unmodified Qwen2-7b model. Scalar conditioning enables strategically coherent, role-consistent negotiation behavior, translating into substantial and consistent competitive gains.

5.4 Statistical Analysis of Configuration Effects

To formally evaluate whether negotiation performance differed across configurations, we conducted a one-way analysis of variance (ANOVA) using episode-level janus normalised utility as the dependent variable. Five configurations were compared: Base, Janus Standard, Janus Extreme, Janus Neutral, and Janus Flipped.

Two parallel analyses were performed: (1) all episodes, including impasses, and (2) agreements only (`agreement` $= 1$), to separate value extraction ability from agreement frequency effects.

Assumption checks.

Levene's test for homogeneity of variances was significant in both analyses (all episodes: $W = 409.85$, $p < .001$; agreements only: $W = 384.54$, $p < .001$), indicating unequal variances across configurations. Consequently, Welch's ANOVA was used for the omnibus test. Residual normality was not tested due to large sample sizes ($n > 5{,}000$), where the Central Limit Theorem ensures robustness.

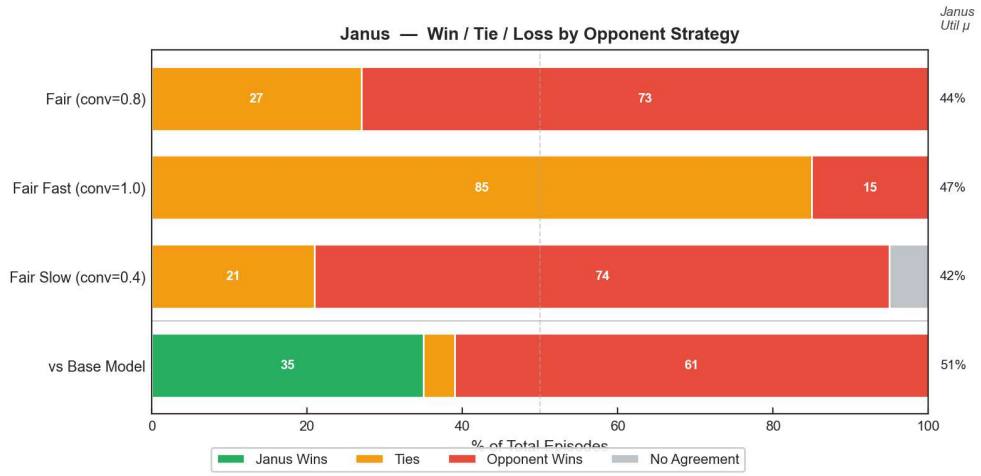


Fig. 14 Win–tie–loss distribution for the Janus Neutral configuration against Fair variants and the Base model.

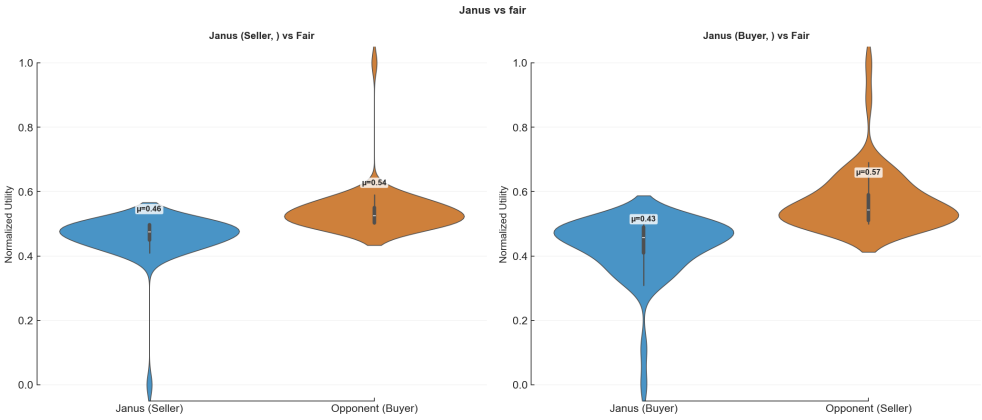


Fig. 15 Utility distributions for Janus Neutral against the Fair strategy in both roles.

Omnibus test.

Welch's ANOVA revealed a highly significant effect of configuration on normalized utility:

- **All episodes:** $F(4, 2868.8) = 713.45$, $p < .001$
- **Agreements only:** $F(4, 2858.1) = 695.87$, $p < .001$

For reference, standard ANOVA yielded $\eta^2 = .234$ (all episodes) and $\eta^2 = .252$ (agreements only), indicating a large effect size. Approximately 23–25% of the variance in negotiation utility is explained solely by the scalar conditioning configuration.

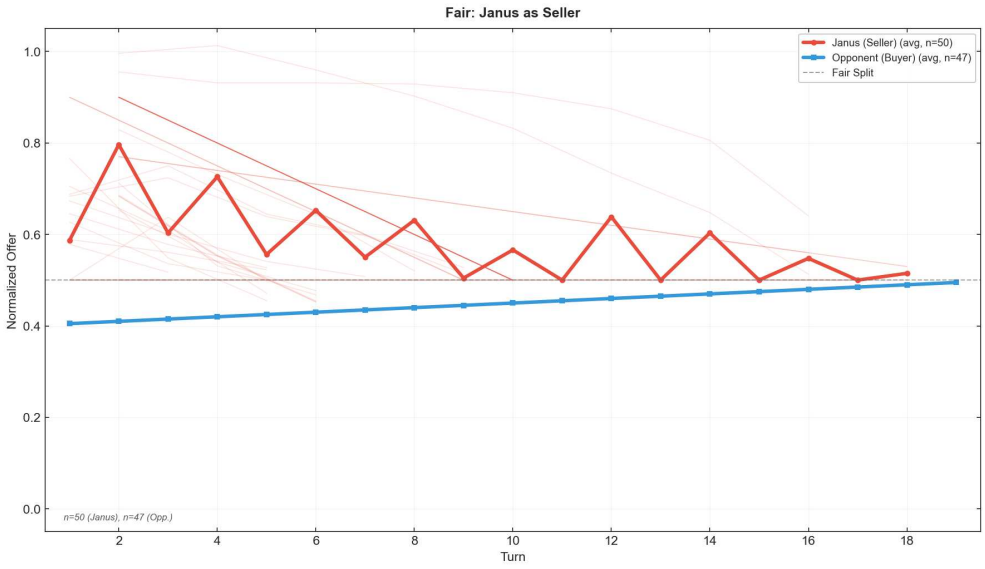


Fig. 16 Average concession curves for Janus Neutral acting as Seller against the Fair Seller.

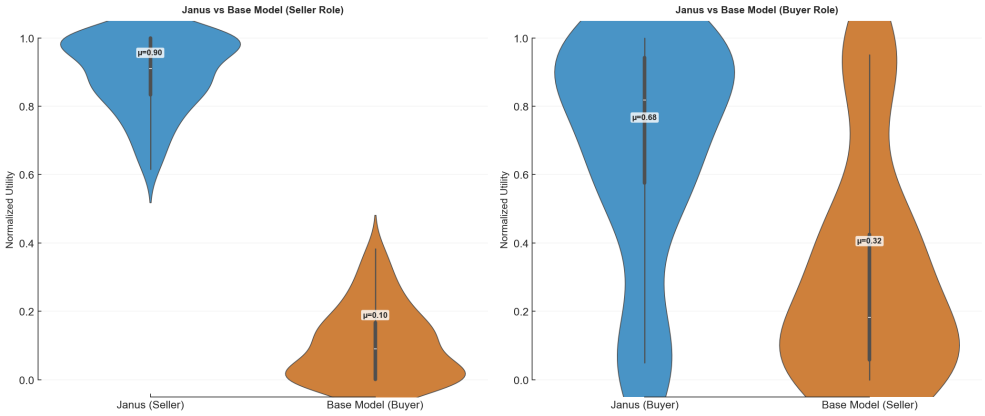


Fig. 17 Head-to-head utility distributions for Janus Standard vs. the unmodified Base model in both seller and buyer roles. Higher values indicate greater utility for Janus.

Post-hoc comparisons.

Tukey HSD pairwise comparisons showed that 9 of 10 configuration pairs differed significantly in both analyses ($\alpha = .05$).

Key findings include:

- **Janus Standard** achieved significantly higher mean utility than all other configurations ($p < .01$ for all comparisons).

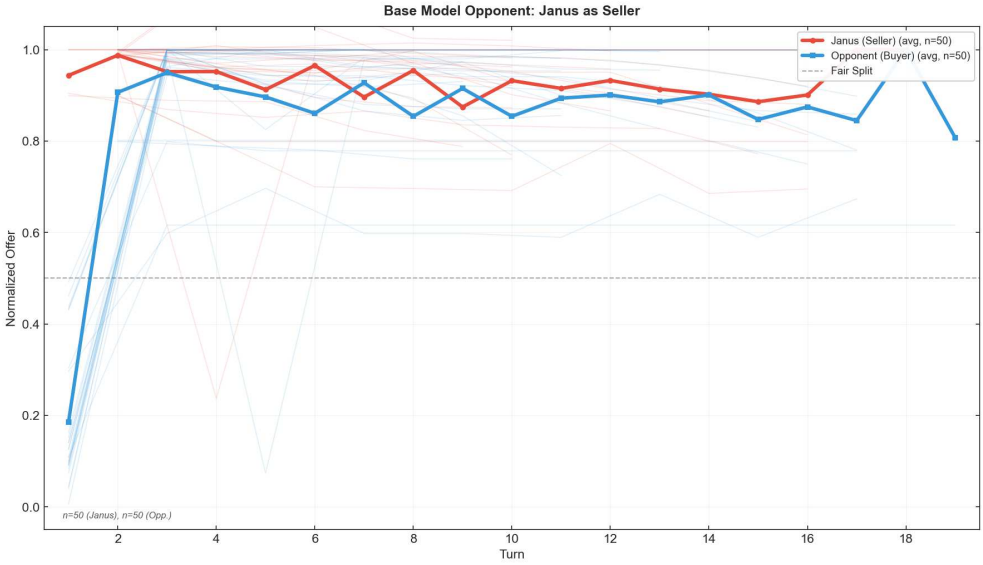


Fig. 18 Concession curve: Base model as Buyer against Janus Standard as Seller.

- All trained Janus configurations except **Janus Flipped** significantly outperformed the Base model.
- **Janus Flipped** performed significantly worse than every other configuration, including Base.
- The only non-significant comparison was between **Janus Extreme** and **Janus Neutral** (all episodes: $\Delta = .018$, $p = .918$; agreements only: $\Delta = .007$, $p = .998$).

Despite extreme ρ values (0.0/1.0), Janus Extreme did not achieve higher average utility than the symmetric Neutral configuration. Instead, Extreme exhibited substantially higher variance, producing either near-maximal gains or poor outcomes depending on opponent behavior. This suggests that pushing scalar conditioning to boundary values increases volatility without improving mean performance.

Interpretation.

The statistical results confirm that scalar conditioning via ρ produces substantial, systematic differences in negotiation behavior. The magnitude of the effect ($\eta^2 \approx .24$–$.25$) demonstrates that configuration choice is not merely statistically detectable but practically consequential. In particular, the Standard configuration ($\rho_{buyer} = 0.2$, $\rho_{seller} = 0.8$) represents a robust performance optimum within the tested regime, while the Flipped configuration functions as an effective ablation control, validating the directional nature of the learned conditioning.

6 Discussion

This study set out to evaluate a central hypothesis: that a scalar-conditioned hypernetwork can learn a continuous and strategically coherent policy manifold in a large language model, enabling inference-time modulation of negotiation behavior through a single interpretable parameter, ρ. The results provide strong empirical support for this claim. Across configurations, the Janus Protocol demonstrates that low-dimensional scalar control can meaningfully reshape concession dynamics, alter win–tie–loss distributions, and systematically repair the concessionary bias observed in the frozen base model.

6.1 Continuous Policy Control in Weight Space

The most immediate finding is that scalar-conditioned adapter generation produces behavior that is both directionally correct and structurally distinct across the $\rho \in [0, 1]$ spectrum. The Standard configuration achieves substantial utility gains over the base model while preserving high agreement rates, indicating that the model does not merely become more stubborn, but are strategically calibrated. The Neutral configuration reliably converges to equitable outcomes, demonstrating that midpoint conditioning induces mediator-like behavior. Conversely, the Flipped configuration acts as a functional ablation: reversing role-consistent scalar values destroys directional coherence and collapses performance.

Together, these results indicate that ρ is not acting as a superficial prompt signal. It modulates weight-space behavior in a manner that is both interpretable and causally consequential. The learned adapter space behaves as a continuous manifold indexed by an outcome-grounded scalar, rather than as a collection of loosely clustered behavioral modes or statically fine-tuned, discrete personas [29]. By successfully compressing the complex, multi-turn dynamics of negotiation into a low-dimensional latent space [33], our implementation proves that continuous inference-time strategic modulation is not merely theoretical, but practically achievable. This fundamentally separates the Janus Protocol from existing methods that lock models into fixed strategic postures.

6.2 Foundation Models within Multi-Agent Systems

The results demonstrate that scalar-conditioned hypernetworks do more than improve isolated negotiation performance. They operationalize a mutually beneficial structural interface between foundation models and MAS theory. In classical MAS research, agent policies are rigorously evaluated through equilibrium behavior, exploitability, and incentive compatibility [7, 43, 44]. By embedding an equilibrium-indexed scalar directly into adapter generation, MAS theory acts as a targeted solution to mitigate the inherent vulnerabilities of FMs—specifically, curbing the irrational, concessionary behaviors and strategic hallucinations typical of standard LLMs.

Importantly, this approach provides interpretability advantages over prompt-only alignment. Because ρ corresponds to a measurable settlement position within the ZOPA, changes in behavior can be directly mapped to changes in equilibrium orientation. This contrasts with purely prompt-conditioned systems, whose behavioral shifts

often lack structural transparency. Mechanism-design-oriented LLM frameworks similarly demonstrate that embedding microeconomic constraints into agent architectures yields interpretable and strategy-aware outcomes [36].

From a safety perspective, structured scalar control may help mitigate risks identified in multi-agent LLM deployments. Recent surveys document that LLM-based agents suffer from structured hallucination behaviors that can cascade across interacting agents [35]. In decentralized bargaining systems, such distortions could propagate unfair pricing norms, amplify misinformation, or enable collusive equilibria.

Societally, controllable strategic posture directly affects distributive outcomes. Simulation-based analyses demonstrate that personality and strategic orientation significantly influence surplus allocation and fairness metrics [45]. At larger scales, population-level generative simulations show that autonomous LLM agents can meaningfully alter macroeconomic dynamics [46]. The Janus framework therefore provides not only technical gains but also a structural pathway toward policy-aware multi-agent governance.

6.3 Boundary Instability in the Extreme Regime

The Janus Extreme configuration reveals an important limitation. While performance remains substantially stronger than the base model, we observe mild instability at the scalar endpoints ($\rho = 0.0$ and $\rho = 1.0$). Specifically, initial offers occasionally appear miscalibrated relative to the agent's role before correcting in subsequent turns. This transient directional noise suggests that the learned manifold may be less smooth near its boundaries.

Several factors may contribute to this phenomenon. First, although the dataset spans the full ρ continuum, density near the exact endpoints is structurally constrained by the negotiation dynamics of certain strategies. Second, the sigmoid-gated hypernetwork may saturate near 0 and 1, amplifying small prediction variance into larger behavioral deviations. Third, extreme scalar conditioning forces the model into highly asymmetric utility postures that may be less frequently reinforced during training relative to interior values.

Future work could address this in several ways. Endpoint stability may improve through (i) explicit oversampling or curriculum weighting of extreme-ρ trajectories, (ii) boundary-aware regularization encouraging smoother gate interpolation near 0 and 1, (iii) increasing LoRA rank to expand representational capacity at policy extremes, or (iv) replacing the scalar-to-gate mapping with a monotonic spline or learned basis expansion that improves control resolution at the boundaries. These refinements would strengthen confidence that the learned manifold remains well-conditioned across its entire domain.

6.4 Scalability and Resource Considerations

It is important to emphasize that these results are achieved with a 7B-parameter base model and approximately 0.8% trainable parameters. The experimental budget is modest relative to contemporary large-scale LLM training efforts. Nevertheless, the architecture demonstrates measurable and consistent strategic modulation. This

efficiency is entirely by design; by generating only rank-dimensional gating vectors rather than full LoRA matrices [41], we translate the parameter-efficient zero-shot capabilities seen in adaptive vision models [39] into the domain of multi-turn strategic interaction. This makes the Janus Protocol uniquely lightweight and highly scalable, sidestepping the prohibitive computational costs normally associated with dynamic hypernetwork generation.

With larger backbone models, expanded trajectory datasets, and increased training budgets, we expect the policy manifold to become smoother, more expressive, and more resistant to rigid adversarial strategies such as Tit-for-Tat and Charging-Boul. Greater data diversity would likely improve calibration in low-density regions of the scalar space. Increased adapter capacity could enable more nuanced opponent modeling behaviors to emerge within each ρ slice. Importantly, none of the observed limitations appear architectural in nature; rather, they reflect capacity and data constraints typical of early-stage demonstrations.

Thus, the present work should be viewed not as an upper bound on scalar-conditioned control, but as a proof of viability under constrained resources.

6.5 Generality Beyond Negotiation

Negotiation was selected as the experimental domain due to its clean evaluability. Outcomes can be grounded in a normalized scalar index within the Zone of Possible Agreement, providing an objective axis along which strategic behavior can be measured. This property makes it an ideal testbed for validating continuous policy modulation.

However, the implications extend far beyond price bargaining. The core contribution is not a negotiation system per se, but a method for embedding an interpretable control coordinate directly into weight space. Any domain that admits a semantically meaningful scalar axis could, in principle, benefit from this approach. Examples include tone modulation (e.g., casual to formal communication), persuasiveness intensity in argument generation, degree of verbosity or explanation depth, risk tolerance in decision-support agents, or code generation granularity (e.g., sparse vs. heavily commented outputs).

Rather than training discrete models for each behavioral mode, a single scalar-conditioned adapter could allow deployment-time traversal of a learned behavioral continuum. This reframes controllability as manifold navigation rather than model switching.

6.6 Toward Polymorphic LLM Agents

More broadly, this work contributes to an emerging paradigm in which large language models are treated not as static aligned artifacts, but as polymorphic agents whose strategic orientation can be tuned continuously at inference time. Traditional alignment pipelines—even state-of-the-art reinforcement learning frameworks that rely on heterogeneous reward optimization to govern negotiation behavior [18]—ultimately optimize toward fixed objectives. This often collapses diverse behavioral potential into a narrow, static compromise. Our scalar-conditioned hypernetwork offers a distinct

alternative: a structured mechanism for exposing continuous behavioral degrees of freedom while preserving coherence.

The head-to-head results against the base model demonstrate that this unique approach can correct the irrational concessionary biases typical of foundation models [13] without sacrificing fluency. Furthermore, the Janus Neutral configuration shows that equitable, mediator-like fairness can be induced deliberately rather than incidentally, proving that the model can dynamically traverse from competitive to cooperative regimes on demand. Finally, the Flipped ablation confirms that the scalar parameter is causally active. Collectively, these findings establish that low-dimensional continuous control over complex, multi-turn strategic behavior is not only theoretically appealing, but practically feasible within current LLM architectures.

While substantial work remains to improve robustness, boundary smoothness, and cross-domain validation, the core hypothesis is empirically supported. Scalar-conditioned weight generation is a viable mechanism for continuous behavioral steering in large language models.

7 Related Work

Having demonstrated the empirical success of the Janus Protocol, we now situate our findings within the broader academic landscape. Evaluating our continuous, scalar-conditioned approach against existing paradigms throws the limitations of current static alignment into sharp relief. Beyond the core structural methodologies discussed previously, the field of automated negotiation has a rich history long predating the rise of modern LLMs. Research has traditionally been rooted in game theory and agent-based modeling, which produced negotiators with predictable or strategically programmed behaviors [37]. These approaches yielded a taxonomy of well-defined, effective tactics for distributive bargaining. These include time-based concession strategies like the *Boulware* agent, which makes minimal concessions until the final rounds [32], a tactic named after the General Electric executive who pioneered its real-world usage [47]. This contrasts with reciprocal strategies like *Tit-for-Tat*, which, while robust in fostering cooperation, are fundamentally different from the non-reciprocal, exploitative tactics tested in our work [6, 48, 49]. These heuristic-based agents, while less conversationally flexible, are grounded in decades of strategic negotiation theory designed for competitive, value-claiming environments.

More recent attention has focused on advancing LLMs to perform sophisticated multi-turn interactions, opening new possibilities for negotiation tasks in both human-computer and multi-agent settings [50, 51]. LLMs are typically built on transformer architectures with pre-training on massive text corpora followed by task-specific fine-tuning [51]. Across studies on bargaining games, e-commerce/tourism, and business price negotiations, LLMs have been trained for negotiation through various paradigms.

Supervised Learning: In [28], end-to-end neural negotiators are trained on human–human bargaining corpora (e.g., DealOrNoDeal) by maximizing the likelihood of human actions, learning both language and strategy jointly directly from transcripts.

GENTEEL-NEGOTIATOR [19] trains an LLaMA-based encoder plus a mixture-of-experts on tourism and e-commerce datasets (NEGOCHAT, IND), with labels for strategies and politeness, yielding more coherent and polite negotiation responses.

Reinforcement Learning: One of the earliest studies on RL-based training is [28], where supervised training is augmented with self-play RL and dialogue rollouts to maximize negotiation reward instead of likelihood, significantly boosting deal quality. Building on RL approaches, Kwon *et al.* [52] use LLMs as proxy reward functions. Prompts describing desired behavior (e.g., Pareto-optimal, versatile) are used to score full negotiation trajectories in Ultimatum and DealOrNoDeal games. Hajj *et al.* [53] adopted a Seq2Seq RNN-RL approach to learn task-specific negotiation strategies based on achieved payoffs, though this model is more sensitive to data scarcity and lacks broader language skills.

In-Context Use: In-context learning has also emerged as a tuning-free mechanism to teach LLMs negotiation dynamics. Hua *et al.* [21] trained LLMs as an assistive mediator that improves the social quality of dialogue (e.g., politeness and cooperativeness) with humans based on a novel value-impact criterion. Additionally, negotiation-specific in-context strategies can be used to code and analyze negotiation transcripts, supporting the building of better training corpora [54]. Furthermore, negotiation strategy and dialogue generation can be decoupled: one module predicts a structured act, while another generates natural language for that act, avoiding degenerate RL policies while maintaining fluent dialogue [25, 55]. Fu *et al.* [56] utilized in-context learning where two LLMs bargain, a third LLM critiques tactics, and prior dialogs and feedback are reused in-context so strategies improve over rounds without gradient updates.

While these paradigms have successfully equipped LLMs with baseline bargaining capabilities, they predominantly result in static, role-locked models. As demonstrated by our head-to-head results (Section 5.3) and ablation studies (Section 5.2.4), continuous inference-time control via a low-dimensional policy manifold offers a highly efficient, dynamic alternative to the discrete retraining cycles or extensive prompt-engineering required by these traditional methods.

Supporting these advancements is a growing ecosystem of evaluation frameworks. Platforms like Deal or No Deal? [28] and NegotiationArena [13] provide environments to evaluate different negotiation agents [28, 53, 54]. Utilizing these frameworks, recent studies have uncovered detailed evidence of LLMs being inherently agreement-seeking and conservative. Qian *et al.* [22] demonstrated that, in dynamic multi-issue bargaining, state-of-the-art models achieve human-level performance but through "conservative, concessionary trades" that favor value-balanced, easily accepted offers rather than surplus-maximizing ones. Wang *et al.* [26]'s evaluation in debt-collection negotiations reveals that they "tend to over-concede compared to human negotiators." Noh and Chang [24] designed Personality-based LLM negotiators and found that more agreeable agents get deals but are exploitable, whereas less-agreeable agents extract more value, proving the default trend is "generally fair rather than rational". Furthermore, Vahidov *et al.* [25] evaluated LLM-enhanced seller agents and found that they reach agreements easily and often at buyer-friendly or fair prices, reflecting a bias toward agreement and rapport. Our base model evaluations explicitly confirm these behavioral diagnoses; however, whereas prior works primarily identify and measure

this concessionary bias, the Janus Protocol provides a validated architectural remedy to structurally eliminate it.

Ultimately, the fragility of standard LLMs in these frameworks is deeply tied to their inherent design and susceptibility to well-known psychological principles and cognitive heuristics [57]. Training methodologies like RLHF have created a "rationality gap" between their conversational prowess and their strategic acumen [58]. Furthermore, LLMs are still far from truly understanding an opponent's strategy or the nuances of their offers [59], making them highly vulnerable to tactics grounded in cognitive science. For example, the *Hype* agent's strategy is a direct application of Prospect Theory's framing effects [30], where the positive presentation of an offer heavily influences its reception and bargaining outcomes [60]. It also leverages the anchoring effect, where an aggressive first offer disproportionately dictates the final price [61]. Similarly, the *Abrasive* agent's use of social pressure draws on findings regarding the role of intimidation and hardball tactics in breaking down human negotiation resistance [31]. As evidenced by our agent's ability to systematically overcome time-based concession strategies (e.g., *Boulware*) and significantly reduce its baseline exploitability against rigid adversaries, resolving this gap requires explicitly evaluating and defending LLMs against these established psychological and computational competitive strategies [16].

Ultimately, this concessionary tendency highlights a broader structural misalignment between foundation models and classical MAS theory. In multi-agent environments, equilibria emerge from strategic incentives rather than cooperative language priors. When LLMs are deployed as autonomous entities within "Agentic AI" frameworks without explicit strategic structure, they systematically deviate from equilibrium play, leading to predictable exploitation. Recent work on controllable LLM agents demonstrates that embedding game-theoretic methodologies directly into agent architectures enables dynamic adjustment of strategic proficiency while preserving stability [62]. As evidenced by the robust performance of the Janus Protocol across the ρ continuum, our findings operationalize this exact perspective. By treating strategic posture as an explicitly parameterized coordinate embedded directly in weight space, we move beyond the emergent vulnerabilities of static prompting to establish a mathematically grounded foundation for strategically accountable multi-agent systems.

8 Conclusion

This study introduced the Janus Protocol, a scalar-conditioned hypernetwork framework that enables continuous inference-time control of strategic behavior in large language models. By conditioning LoRA adapters on a single interpretable scalar parameter ρ, we demonstrated that a frozen base model can traverse a coherent negotiation policy manifold without discrete model switching or retraining.

More broadly, this work successfully operationalizes the bidirectional synergy between foundation models and MAS theory. FMs supply the linguistic reasoning, natural language generation, and contextual flexibility required to execute complex negotiations in free text. Conversely, MAS architectures and game-theoretic methodologies supply the structural equilibrium concepts required to mitigate the behavioral

vulnerabilities, irrational biases, and strategic hallucinations inherent to current FMs. Embedding these game-theoretic parameters directly into model control improves robustness in strategic environments [62].

Empirical results in a bilateral price negotiation domain show that scalar-conditioned adaptation substantially improves strategic robustness relative to an unmodified base model. By leveraging outcome-conditioned supervision over verified self-play trajectories [42], the model learns causally consistent guidance that entirely replaces the need for online search or computationally heavy forward-trial reinforcement learning. The Standard configuration achieves strong utility gains while preserving high agreement rates, the Janus Neutral configuration reliably induces fairness-oriented behavior, and the Flipped ablation confirms that scalar alignment is causally necessary for directional coherence. Even under modest training budgets and a 7B-parameter backbone, the architecture consistently repairs the baseline's concessionary vulnerabilities [13] and enables controlled modulation of multi-turn concession dynamics.

Importantly, the contribution extends beyond negotiation. The central insight is uniquely architectural: adapter weights need not be static, role-locked parameters [29], but can be efficiently generated as factorized functions of an interpretable control variable [41]. By treating adapter space as a continuous, low-dimensional policy manifold [33] indexed by a semantically grounded scalar, we provide empirical evidence that complex strategic behavior in LLMs can be seamlessly modulated directly in weight space.

While additional work is required to refine boundary smoothness, improve robustness against rigid reciprocal strategies, and validate generalization across domains, the results establish a clear proof of concept. Scalar-conditioned weight generation offers a principled mechanism for controllable, polymorphic LLM agents whose strategic posture can be tuned dynamically at deployment time.

In contrast to static alignment paradigms that optimize models toward fixed behavioral objectives, the Janus Protocol reframes alignment as structured controllability. Rather than collapsing behavior into a single compromise policy, we expose a continuous and interpretable degree of freedom. This perspective opens new avenues for adaptive AI systems that can adjust tone, competitiveness, fairness, risk tolerance, or other behavioral dimensions through explicit, inference-time control.

Future research should explore integrating mechanism-design constraints directly into hypernetwork conditioning, enabling strategy-aware guarantees and stability under population-level deployment. Mechanism design frameworks that combine LLM reasoning with formal economic constraints demonstrate that equilibrium-consistent and interpretable policies can emerge when structural objectives are explicitly encoded [36, 46]. By compressing high-dimensional parameter space into a semantically meaningful behavioral coordinate [33], scalar-conditioned hypernetworks provide a concrete architectural pathway for FM–MAS integration.

Taken together, our findings suggest that low-dimensional continuous control over large language model behavior is not only feasible, but practically effective. The Janus Protocol provides a foundation for future research into controllable strategic agents and broader applications of hypernetwork-based behavioral modulation in language models.

Appendix A Strategy Pool Details

Table A1 provides a complete list of the deterministic and stochastic opponent strategies used to generate the training corpus and evaluate the models, including their core parameters and behavioral descriptions.

Table A1 Strategy Pool

Strategy	Description
Boulware Very Conceding ($\beta = 0.2$)	Rapid concession from ZOPA boundary
Boulware Conceding ($\beta = 0.5$)	Moderate concession from ZOPA boundary
Boulware Firm ($\beta = 2.0$)	Slow concession from ZOPA boundary
Boulware Hard ($\beta = 4.0$)	Very slow concession from ZOPA boundary
Noisy Boulware Conceding ($\beta = 0.5$)	Boulware with $\pm 5\%$ ZOPA noise
Noisy Boulware Firm ($\beta = 2.0$)	Boulware with $\pm 5\%$ ZOPA noise
Noisy Boulware Hard ($\beta = 4.0$)	Boulware with $\pm 5\%$ ZOPA noise
Tit-for-Tat	Mirrors opponent concessions from ZOPA boundary
Linear Standard	Linear concession from ZOPA boundary
Split Difference	Proposes midpoint between last offers
Hardliner	Holds at ZOPA boundary until final round
MiCRO Fine (step=10)	Minimal concession on fine quantized grid
MiCRO Moderate (step=25)	Minimal concession on moderate quantized grid
MiCRO Coarse (step=50)	Minimal concession on coarse quantized grid
ChargingBoul	Adaptive Boulware with UBI/AUI opponent modeling ($m = 0.5, E = 0.4$)
ChargingBoul Aggressive	Higher utility floor ($m = 0.6, E = 0.2$)
ChargingBoul Patient	Faster concession rate ($E = 0.8$)
Fair	Converges toward ZOPA midpoint (convergence=0.8)
Fair Fast	Opens directly at ZOPA midpoint (convergence=1.0)
Fair Slow	Slowly converges to midpoint (convergence=0.4)
Naive Concession	Offers near own reservation (systematically weak)
Naive Boulware ($\beta = 3.0$)	Concedes in wrong direction with hardline curve
Random (Training Only)	Uniform random offers within true ZOPA bounds

References

[1] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.-R.: A survey on large language model based autonomous agents (2023) arXiv:2308.11432 [cs.AI]

[2] Lax, D.A., Sebenius, J.K.: The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain. Free Press, New York, NY (1986)

[3] Carnevale, P.J., Pruitt, D.G.: Negotiation and mediation. Annual Review of Psychology **43**(1), 531–582 (1992) https://doi.org/10.1146/annurev.ps.43.020192.002531

[4] Lin, F.-r., Lin, Y.-y.: Integrating multi-agent negotiation to resolve constraints in fulfilling supply chain orders. Electronic Commerce Research and Applications **5**(4), 313–322 (2006)

[5] Soh, L.-K., Tsatsoulis, C.: A real-time negotiation model and a multi-agent sensor network implementation. Autonomous Agents and Multi-Agent Systems **11**(3), 215–271 (2005)

[6] Grosz, B.J., Kraus, S.: Collaborative plans for complex group action. Artificial Intelligence **86**(2), 269–357 (1996) https://doi.org/10.1016/0004-3702(96)00008-3

[7] Fatima, S., Kraus, S., Wooldridge, M.: Principles of Automated Negotiation. Cambridge University Press, ??? (2014)

[8] Khorasany, M., Paudel, A., Razzaghi, R., Siano, P.: A new method for peer matching and negotiation of prosumers in peer-to-peer energy markets. IEEE Transactions on Smart Grid **12**(3), 2472–2483 (2021) https://doi.org/10.1109/TSG.2020.3048397

[9] Baarslag, T., Fujita, K., Gerding, E.H., Hindriks, K., Ito, T., Jennings, N.R., Jonker, C., Kraus, S., Lin, R., Makamori, V., *et al.*: The automated negotiating agents competition, 2010–2015. AI magazine **37**(4), 115–122 (2016)

[10] De Jonge, D.: An analysis of the linear bilateral anac domains using the micro benchmark strategy. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, pp. 223–229. International Joint Conferences on Artificial Intelligence Organization, ??? (2022). https://doi.org/10.24963/ijcai.2022/32 . Main Track. https://doi.org/10.24963/ijcai.2022/32

[11] Shymanski, J.: Chargingboul: A competitive negotiating agent with novel opponent modeling (2025) arXiv:2512.06595 [cs.MA]

[12] Cao, K., et al.: Emergent cooperation in llm-based negotiating agents. arXiv preprint arXiv:2304.01234 (2023)

[13] Bianchi, F., et al.: How well can llms negotiate? negotiationarena platform and analysis. arXiv preprint arXiv:2402.05863 (2024)

[14] Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G., Gratch, J.: 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, pp. 5391–5413 (2024)

[15] Zhang, Y., Zhang, C., Wang, Z., Zhang, Y., Liu, Y., Liu, Z.: LLM as a mastermind: A survey of strategic reasoning with large language models (2024)

arXiv:2404.01230 [cs.AI]

[16] Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G., Gratch, J.: 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, pp. 5391–5413. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.310

[17] Abdelnabi, S., et al.: Strategic tradeoffs between humans and ai in multi-agent bargaining. arXiv preprint arXiv:2509.09071 (2025)

[18] Zhuang, Z., Chen, Y., Zeng, X., Luo, C., Liu, L., Chen, Y.: Teaching llm to be persuasive: Reward-enhanced policy optimization for alignment from heterogeneous rewards. ArXiv **abs/2510.04214** (2025) https://doi.org/10.48550/arxiv.2510.04214

[19] Priya, P., Chigrupaatii, R., Firdaus, M., Ekbal, A.: Genteel-negotiator: Llm-enhanced mixture-of-expert-based reinforcement learning approach for polite negotiation dialogue, 25010–25018 (2025) https://doi.org/10.1609/aaai.v39i23.34685

[20] Chawla, K., Clever, R., Ramirez, J., Lucas, G., Gratch, J.: Towards emotion-aware agents for negotiation dialogues. In: 9th International Conference on Affective Computing and Intelligent Interaction (ACII) (2021)

[21] Hua, Y., Qu, L., Haf, R.: Assistive large language model agents for socially-aware negotiation dialogues. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 8047–8074 (2024)

[22] Qian, C., Zhu, K., Horton, J., Manning, B.S., Tsai, V., Wexler, J., Thain, N.: Strategic tradeoffs between humans and ai in multi-agent bargaining. arXiv preprint arXiv:2509.09071 (2025)

[23] Raiffa, H., Richardson, J., Metcalfe, D.: Negotiation Analysis: The Science and Art of Collaborative Decision Making. Belknap Press of Harvard University Press, Cambridge, MA (2002)

[24] Noh, S., Chang, H.-C.H.: Llms with personalities in multi-issue negotiation games. arXiv preprint arXiv:2405.05248 (2024)

[25] Vahidov, R., Carbonneau, R., Verma, J.: Using negotiation and large language models in human–to software agent negotiations. International Journal of Human–Computer Interaction **41**(24), 15844–15861 (2025)

[26] Wang, X., Zhang, Z., Zheng, J.G., Ai, Y., Wang, R.: Debt collection negotiations with large language models: An evaluation system and optimizing decision making

with multi-agent. In: Findings of the Association for Computational Linguistics: ACL 2025, pp. 9555–9577 (2025)

[27] Schneider, J., Haag, S., Kruse, L.C.: Negotiating with llms: Prompt hacks, skill gaps, and reasoning deficits. In: International Conference on Computer-human Interaction Research and Applications, pp. 238–259 (2024). Springer

[28] Lewis, M., Yarats, D., Dauphin, Y.N., Parikh, D., Batra, D.: Deal or no deal? end-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125 (2017)

[29] Chatterjee, A., Miller, S., Parepally, N.: Agreemate: Teaching llms to haggle. arXiv preprint arXiv:2412.18690 (2024)

[30] Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica **47**(2), 263–291 (1979) https://doi.org/10.2307/1914185

[31] Thompson, L.: Negotiation behavior and outcomes: Empirical evidence and theoretical issues. Psychological Bulletin **108**(3), 515–532 (1990) https://doi.org/10.1037/0033-2909.108.3.515

[32] Pruitt, D.G.: Negotiation Behavior. Academic Press, New York, NY (1981)

[33] Tenedini, D., Zamboni, R., Mutti, M., Restelli, M.: From Parameters to Behaviors: Unsupervised Compression of the Policy Space (2026). https://arxiv.org/abs/2509.22566

[34] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., *et al.*: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)

[35] Lin, X., Ning, Y., Zhang, J., Dong, Y., Liu, Y., Wu, Y., Qi, X., Sun, N., Shang, Y., Wang, K., Cao, P., Wang, Q., Zou, L., Chen, X., Zhou, C., Wu, J., Zhang, P., Wen, Q., Pan, S., Wang, B., Cao, Y., Chen, K., Hu, S., Guo, L.: LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions (2025). https://arxiv.org/abs/2509.18970

[36] Thach, N., Liu, F., Zhou, H., Chan, H.: Large Language Models for Multi-Facility Location Mechanism Design (2025). https://arxiv.org/abs/2503.09533

[37] Raiffa, H.: The Art and Science of Negotiation. Belknap Press of Harvard University Press, Cambridge, MA (1982)

[38] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021) arXiv:2106.09685 [cs.CL]

[39] Li, M., Chen, J., Feng, W., Li, B., Dai, F., Zhao, S., He, Q.: Hyperlora: Parameter-efficient adaptive generation for portrait synthesis (2025) arXiv:2503.16944 [cs.CV]

[40] Ren, Y., Xiao, T., Shavlovsky, M., Ying, L., Rahmanian, H.: Cos-dpo: Conditioned one-shot multi-objective fine-tuning framework (2025) arXiv:2410.08316 [cs.LG]

[41] Abdalla, M.H.I., Wang, Z., Frey, C., Eger, S., Grabocka, J.: Zhyper: Factorized hypernetworks for conditioned llm fine-tuning (2025) arXiv:2510.19733 [cs.CL]

[42] Li, C., Xu, Y., Wang, Z., Tan, S.H., Tse-Hsun, Chen: Outcome-conditioned reasoning distillation for resolving software issues (2026) arXiv:2601.23257 [cs.SE]

[43] Shoham, Y., Leyton-Brown, K.: Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press, ??? (2008)

[44] Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: Algorithmic Game Theory. Cambridge University Press, ??? (2007)

[45] Huang, Y.J., Hadfi, R.: How Personality Traits Influence Negotiation Outcomes? A Simulation based on Large Language Models (2024). https://arxiv.org/abs/2407.11549

[46] Karten, S., Li, W., Ding, Z., Kleiner, S., Bai, Y., Jin, C.: LLM Economist: Large Population Models and Mechanism Design in Multi-Agent Generative Simulacra (2025). https://arxiv.org/abs/2507.15815

[47] Boulware, L.R.: The Truth About Boulwarism: Trying to Do Right Voluntarily. Bureau of National Affairs, Washington D.C. (1969)

[48] Axelrod, R.: The Evolution of Cooperation. Basic Books, ??? (1984)

[49] Rapoport, A., Seale, D.A., Colman, A.M.: Is tit-for-tat the answer? on the conclusions drawn from axelrod's tournaments. PLoS ONE **10**(7), 0134128 (2015) https://doi.org/10.1371/journal.pone.0134128

[50] Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A., Silva, B.C.: Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. ACM Computing Surveys **58**, 1–37 (2024) https://doi.org/10.1145/3743127

[51] Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S.: A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access **12**, 26839–26874 (2024) https://doi.org/10.1109/access.2024.3365742

[52] Kwon, M., Xie, S.M., Bullard, K., Sadigh, D.: Reward design with language

models. In: The Eleventh International Conference on Learning Representations (2023)

[53] Hajj, A., Adla, Y.A.A., Albast, S., Hajj, H., Elbassuoni, S., El-Hajj, W., Shaban, K.: Design and comparison of arabic negotiation bots using llms versus seq2seq models with reinforcement learning. In: Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024), pp. 304–318 (2024)

[54] Friedman, R., Cho, J., Brett, J., Zhan, X., Han, N., Kannan, S., Ma, Y., Spencer-Smith, J., Jäckel, E., Zerres, A., et al.: An application of large language models to coding negotiation transcripts. arXiv preprint arXiv:2407.21037 (2024)

[55] He, H., Chen, D., Balakrishnan, A., Liang, P.: Decoupling strategy and generation in negotiation dialogues. arXiv preprint arXiv:1808.09637 (2018)

[56] Fu, Y., Peng, H., Khot, T., Lapata, M.: Improving language model negotiation with self-play and in-context learning from ai feedback. arXiv preprint arXiv:2305.10142 (2023)

[57] Bazerman, M.H., Neale, M.A.: Heuristics in negotiation: Limitations to effective dispute resolution. In: Bazerman, M.H., Lewicki, R.J. (eds.) Negotiating in Organizations, pp. 51–67. Sage Publications, Beverly Hills, CA (1983)

[58] Hua, Y., et al.: Game-theoretic llm: Agent workflow for negotiation games. arXiv preprint arXiv:2411.05990 (2024)

[59] Lin, E., Hale, J., Gratch, J.: Toward a better understanding of the emotional dynamics of negotiation with large language models. In: Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, pp. 545–550 (2023)

[60] Neale, M.A., Bazerman, M.H.: The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. Academy of Management Journal **28**(1), 34–49 (1985) https://doi.org/10.2307/256060

[61] Galinsky, A.D., Mussweiler, T.: First offers as anchors: the role of perspective-taking and negotiator focus. Journal of Personality and Social Psychology **81**(4), 657–669 (2001) https://doi.org/10.1037/0022-3514.81.4.657

[62] Zhang, Z., Lan, Y., Chen, Y., Wang, L., Wang, X., Wang, H.: DVM: Towards Controllable LLM Agents in Social Deduction Games (2025). https://arxiv.org/abs/2501.06695