

P3 Negotiation Arena - Interaction Regimes and Emergent Strategies in Multi-Agent LLM Negotiation

Ionut Bogdan Donici^{1,2*}

^{1*}Department of Computer Science, University of Milan, Via Celoria
18, Milano, Italy.

Corresponding author(s). E-mail(s):
ionutbogdan.donici@studenti.unimi.it;

Abstract

This project introduces a structured framework for studying negotiation dynamics among Large Language Model agents operating under different interaction regimes. The system simulates multi-round dialogues between autonomous agents with distinct objectives, private constraints, and scenario-specific incentives. Two negotiation settings are implemented: a partially zero-sum resource division task and a multi-issue salary negotiation scenario.

The central goal is to analyze how interaction framing (cooperative, competitive, mixed) and scenario structure influence agreement dynamics and emergent communicative strategies such as persuasion, concession, deception, and cooperation. To support systematic evaluation, a dual-judge architecture is introduced, combining incremental monitoring with global dialogue analysis.

1 Introduction

Large Language Models (LLMs) are increasingly used as autonomous agents in interactive settings. When multiple agents interact, they must negotiate, cooperate, or compete in order to reach agreements despite having different objectives or incomplete information. These environments allow us to observe how strategic communication emerges over multiple conversational turns.

Unlike single-turn benchmarks, negotiation requires sustained dialogue, proposal refinement, concession-making, and adaptation to the counterpart's moves. This

makes it a suitable setting for studying whether LLM agents behave cooperatively, adversarially, or strategically under different constraints.

This project designs and implements a multi-agent negotiation framework in which LLM-based agents interact across several rounds under controlled conditions. Agents are instantiated with distinct objectives and constraints and operate in three negotiation modes: cooperative, competitive, and mixed. Two structurally different scenarios are considered: a resource division task and a multi-issue salary negotiation.

The goal of the project is to analyze how interaction mode and scenario structure influence:

- agreement rate
- number of rounds to convergence
- and emergent behaviors such as persuasion, concession, deception, and cooperation.

To support systematic evaluation, this work introduces a dual-judge architecture that evaluates negotiations both incrementally (round by round) and globally (after termination). A total of 63 automated simulations are conducted, and the resulting behavioral patterns are analyzed both quantitatively and qualitatively.

The remainder of this report presents the methodological design of the framework, the experimental setup, and a discussion of the observed results.

2 Research Question and Methodology

2.1 Research Question

This project investigates how structural and procedural factors influence the behavior of LLM-based negotiation agents. More specifically, the analysis focuses on the effect of the **negotiation mode** — *cooperative*, *competitive*, or *mixed* — and the **scenario structure** — zero-sum resource division versus multi-issue negotiation — on a set of observable outcomes: agreement rate, number of rounds to convergence, and emergent strategic behaviors, including persuasion, deception, concession, and cooperation. The goal is not to optimize negotiation performance, but to understand how different interaction regimes shape agents' behavioral patterns.

2.2 Negotiation Framework

Each negotiation is defined through a structured configuration specifying the involved agents, the resources or issues under negotiation, the evaluation metrics, and the procedural rules. Two distinct scenarios were implemented. The first, **Resource Division**, simulates a negotiation between two startup co-founders over equity shares, budget allocation, and decision rights. This is a partially zero-sum setting with rigid structural constraints. The second, **Salary Negotiation**, involves a candidate and an HR manager in a multi-issue negotiation — covering salary, bonus, equity, remote work, and benefits — which allows for richer trade-offs and a more nuanced behavioral analysis. In both scenarios, each agent is instantiated as an LLM-based persona characterized

by a role and narrative identity, a private objective function, hard constraints (deal-breakers), and soft preferences. Agents operate autonomously, generating messages sequentially across rounds.

2.3 Dialogue Protocol

Negotiations unfold over discrete rounds: if a scenario defines k agents, each round consists of exactly k sequential messages, one per agent. The conversation history at step t is represented as:

$$H = (a_1, m_1), (a_2, m_2), \dots, (a_t, m_t) \quad (1)$$

where a_i denotes the agent and m_i the corresponding message.

The three negotiation modes differ in the orientation imposed on agents: in **cooperative** mode agents are encouraged to seek mutually beneficial outcomes, in **competitive** mode they prioritize individual utility maximization, while in **mixed** mode both orientations — cooperative and opportunistic — are permitted. A negotiation terminates upon reaching an agreement, upon judge-detected failure/impasse, or when the configured maximum number of rounds is reached. In the reported dataset, `max_rounds` varied across runs (10, 15, or 20).

2.4 Evaluation Architecture

To enable systematic analysis, a dual-judge evaluation mechanism was introduced. The **Round Judge** operates at the end of each round: it assigns numeric scores (0–10 scale) to scenario-defined metrics — such as fairness, cooperativeness, and satisfaction — and determines the current negotiation status (*ongoing*, *reached*, *failed*), enabling incremental monitoring and termination control. The **Final Judge** intervenes once the negotiation has ended, analyzing the entire dialogue trajectory. In addition to scenario metrics, it assigns diagnostic scores for persuasion, deception, concession, and cooperation, and classifies the overall interaction pattern as *scripted*, *adaptive*, or *mixed*. This separation allows real-time decision tracking to be combined with retrospective behavioral analysis.

2.5 Experimental Design

A total of 63 automated simulations were conducted for this report: 48 in the Resource Division setting and 15 in the Salary Negotiation setting, distributed across the three negotiation modes. All interactions were fully logged to enable both quantitative aggregation and qualitative inspection.

3 Results

3.1 Dataset Overview

The experimental corpus used in this report consists of 63 automated negotiation runs: 48 instances of the Resource Division scenario and 15 instances of the Salary Negotiation scenario.

Across all runs, 22 negotiations (34.9%) reached agreement, 8 (12.7%) were classified as failed, and 33 (52.4%) ended as ongoing at the round limit (treated as stalled for analysis). Successful negotiations converged earlier (mean 7.86 rounds) compared to stalled interactions (mean 12.27 rounds), suggesting that feasible agreements tend to emerge relatively quickly, whereas impasses persist until forced termination.

3.2 Effect of Negotiation Mode

Negotiation mode significantly influenced outcome distributions. Cooperative settings achieved agreement in 60% of cases, compared to 20% under competitive constraints and 27.8% in mixed mode. Competitive negotiations exhibited the highest stall rate (60%), indicating difficulty in reconciling adversarial objectives within the fixed round limit.

Table 1 Outcome distribution by negotiation mode (percentages)

Mode	Failed	Ongoing	Reached
Cooperative	4.8%	38.1%	57.1%
Competitive	23.1%	57.7%	19.2%
Mixed	16.7%	55.6%	27.8%

Diagnostic metrics from the Final Judge further highlight behavioral differences across modes. Cooperative negotiations displayed higher cooperation scores and lower deception values compared to competitive runs. This suggests that agents internalize mode-specific framing and adjust their strategic behavior accordingly.

3.3 Scenario Structure and Complexity

Agreement rates differed markedly between scenarios. Salary Negotiation reached agreement in 66.7% of runs, whereas Resource Division achieved agreement in only 25% of cases. The latter also exhibited a substantially higher stall rate.

This divergence appears to be linked to structural properties of the scenarios. Salary Negotiation spans multiple interdependent issues, allowing compensatory trade-offs across dimensions. In contrast, Resource Division involves partially zero-sum allocation, limiting the space of mutually beneficial solutions. These findings indicate that multi-issue settings facilitate convergence more effectively than constrained allocation problems.

3.4 Behavioral Correlates of Outcomes

Outcome-dependent behavioral patterns emerged consistently. Successful negotiations were characterized by high cooperation and concession scores combined with minimal deception. Failed negotiations displayed the inverse pattern, with lower cooperation and elevated deception indicators.

The Final Judge classified 95.2% of interactions as adaptive rather than scripted, suggesting that agents updated their strategies in response to counterpart behavior. Dominance dynamics were also frequently observed, with one agent exerting greater control over framing or proposal evolution. Such dominance was scenario-dependent and often linked to information asymmetry or stronger initial anchoring strategies.

Table 2 Mean diagnostic metrics by outcome (0–10 scale)

Outcome	Persuasion	Deception	Concession	Cooperation
Reached	7.41	0.64	8.41	8.50
Failed	5.20	2.10	5.10	3.30
Ongoing	6.72	1.88	6.44	6.25

Overall, the results show that interaction regime and scenario structure systematically shape both agreement likelihood and strategic behavior patterns in LLM-based negotiation.

4 Discussion and Limitations

Negotiation mode and scenario structure systematically affect agreement rates and behavior. Cooperative framing increases convergence, whereas competitive settings yield longer interactions and more stalls. Multi-issue negotiations also facilitate agreement more than partially zero-sum allocations, likely because they allow compensatory trade-offs.

Behaviorally, successful negotiations combine high cooperation and concessions with low deception, consistent with classical theory in which mutual adaptation and transparent trade-offs enable convergence. High deception and low concession correlate with impasse. These patterns indicate that LLM agents respond to structural incentives and framing, adjusting their strategies.

Several limitations apply.

First, evaluation relies on LLM-based judges. Although incremental monitoring and post-hoc analysis are separated, both use language models, introducing bias and circularity. The scores thus reflect model-internal norms rather than human validation.

Second, the scenarios are synthetic and narrow. They approximate realistic settings but omit key aspects of human negotiation, such as emotions, long-term reputation, and bounded rationality, limiting real-world generalization.

Third, the design does not establish strict causal effects. Differences across modes and scenarios are observed, but no formal significance tests or ablations were performed. Future work should vary model parameters, prompts, or temperature in controlled ways to clarify behavioral drivers.

Finally, labeling interactions as “adaptive” does not prove genuine strategic reasoning. Apparent adaptation may arise from probabilistic generation rather than explicit planning, making it hard to distinguish stochastic variation from deliberate strategy.

Despite these caveats, the project offers a structured, reproducible framework for analyzing LLM-based negotiation, with preliminary evidence that interaction regimes and structural constraints strongly shape agreement dynamics and communicative strategies in multi-agent LLM systems.

5 Potential Future Developments

- First, the evaluation process could be enriched through lexical-level analysis of negotiation transcripts. Instead of relying solely on the scores assigned by the LLM-based judges, future work could examine which words or linguistic constructions correlate with specific metric values (e.g., persuasion, concession, deception). By analyzing the distribution of terms across different score ranges, it would be possible to estimate the relative weight of linguistic features and better understand how specific language choices influence negotiation outcomes.
- Second, the framework could integrate external NLP tools to complement model-based evaluation. Libraries such as NLTK or spaCy could be used for sentence segmentation, lemmatization, syntactic parsing, and entity recognition. This would enable the extraction of structural and stylistic features — such as modality usage, sentence complexity, or polarity shifts — providing an additional layer of analysis and reducing potential circularity in LLM-as-judge assessment.
- **Special (proposed by AI)** A human-centered evaluation layer could be introduced to complement the current dual-judge architecture. While both the Round Judge and Final Judge enable scalable and structured assessment, they rely on LLM-based evaluation, which may introduce bias or internal normative assumptions. Future work could involve a controlled human annotation phase on a representative subset of negotiations, using the same diagnostic dimensions (e.g., persuasion, concession, deception, cooperation). Comparing human judgments with model-assigned scores would allow correlation analysis, identification of systematic divergences, and a more robust validation of the evaluation framework. This extension would strengthen the methodological reliability of the project and provide deeper insight into whether the observed behavioral patterns reflect genuinely interpretable strategic dynamics.

6 Conclusion

This work introduces a framework for simulating and evaluating negotiation dynamics among LLM-based agents. Two negotiation scenarios with different structures were tested under cooperative, competitive, and mixed regimes. In 63 automated simulations, both interaction framing and scenario complexity significantly affected agreement rates and behavioral patterns.

Cooperative settings yielded higher convergence, whereas competitive regimes increased stalls. Multi-issue negotiations produced agreements more reliably than partially zero-sum allocations, underscoring the value of structural flexibility for trade-offs. Successful negotiations showed higher cooperation and concession scores and lower deception indicators, indicating that incentives shape emergent communicative strategies.

Despite relying on model-based judges and synthetic environments, the architecture supports reproducible, scalable analysis of multi-agent LLM behavior. The results provide initial evidence that LLM agents adapt strategically to incentive structures and dialogue constraints.

Future work could add more agents, cross-model comparisons, human baselines, and more formal statistical analysis. Increasing negotiation complexity and modeling longitudinal interactions would help distinguish genuine strategic reasoning from probabilistic language generation.

Overall, controlled negotiation simulations offer a valuable experimental setting for studying cooperation, competition, and emergent strategy in multi-agent LLM systems.

Declarations

This project is the result of my independent work and intellectual effort. No part of the content has been copied from external sources without proper acknowledgment.

Given the technical complexity and scope of the project, generative AI systems were used as development support tools. Their contribution was limited to assisting with selected aspects of code implementation, exploratory analysis of negotiation scenarios, and structural refinement of the evaluation framework. All AI-generated content was critically reviewed, modified where necessary, and fully integrated under my direct supervision and understanding.

A paid subscription to external AI services (e.g., Claude) was used to enable extended experimentation with multi-agent simulations and to access models with sufficient context capacity for negotiation tasks.

The following models were employed during the project:

- Claude models (Haiku, Sonnet, Opus) for negotiation simulation and agent interaction experiments
- ChatGPT Codex 5.3 for partial assistance in code development
- Claude Sonnet 4.5 (Chat Console) for brainstorming and iterative refinement of experimental design

I take full responsibility for the final implementation, methodological decisions, experimental results, and interpretations presented in this work.