From Chats to Markets: AgenticPay for LLM-Powered Negotiation in Multi-Agent Commerce

Xianyang Liu[1,4] **Shangding Gu**[1,4*] **Fan Xu**[1] **Manxi Wu**[1]
Boyi Li[2] **Jun Wang**[3] **Costas Spanos**[1] **Dawn Song**[1]
[1]University of California, Berkeley [2]NVIDIA Research [3]University College London
[4]Safe AI and Robot Learning Lab

Abstract

Agents based on large language models are increasingly expected to autonomously handle negotiation and transactions. However, existing benchmarks are largely text-only and bilateral; although some already study non-zero-sum or multi-issue bargaining, they rarely combine it with multimodal product grounding and competitive multi-agent markets, which together characterize real-world commerce. To address this gap, we introduce **AgenticPay**, a unified framework and benchmark for evaluating how well multimodal agents reach high-welfare, multi-dimensional agreements in realistic markets. Built around four core components (*Environments*, *Tasks*, *Agents*, and *Metrics*), **AgenticPay** comprises 160 multimodal tasks spanning 4 real-world business scenarios (E-commerce, Food Delivery, Ride-hailing, and Apartment Rental) and 8 market topologies, scaling from 1-to-1 bargaining to many-to-many (N-to-N) competitive markets. Beyond price haggling, agents must read product images, infer each opponent's hidden preferences, and trade off multiple binding contract terms (e.g., price, lease duration, return policy, delivery speed). Agents communicate through multi-round natural-language dialogue, with each turn proposing or revising a full contract, and outcomes are scored by a utility-based framework that rewards agreements maximizing joint welfare. Evaluations on state-of-the-art proprietary (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro Preview) and open-weight (Qwen3-VL-32B-Instruct, InternVL3-38B) multimodal models reveal substantial gaps in non-zero-sum value creation and market reasoning: even the strongest agent (Gemini 3.1 Pro Preview) reaches a GlobalScore of only 42.3/100, and performance consistently degrades as markets scale from bilateral to multi-sided, with the average GlobalScore dropping by 5.7 points and open-weight models suffering the largest declines (up to 11.5 points). These findings establish **AgenticPay** as a foundational testbed for multimodal agentic commerce. Code and dataset are available at: https://github.com/SAIL-Research-Lab/AgenticPay.

1 Introduction

Modern commerce is increasingly mediated by autonomous agents powered by large language models (LLMs) [16, 32, 47, 26, 63], transacting on behalf of buyers and sellers in real-world markets. Unlike traditional algorithmic trading with rigid numeric bids, such agents must navigate marketplaces like e-commerce and housing rentals, where assessing product worth relies on rich multimodal signals such as images. Beyond price haggling, real-world commerce demands natural-language interaction to reconcile conflicts, negotiate nuanced terms, and reach mutually beneficial contracts. Economic interaction is thus becoming a language-mediated, multi-agent process requiring holistic perception, strategic reasoning, and cooperative communication.

*Corresponding author: shangding.gu@gmail.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

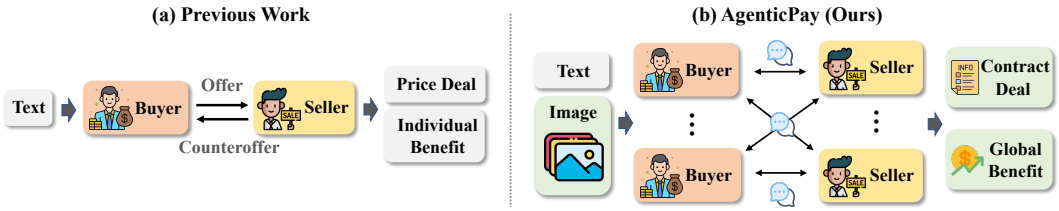

Figure 1: **Comparison of negotiation paradigms.** We move from (a) the common *text-only, bilateral* price-haggling setting with single-number outcomes, to (b) **AgenticPay**'s *multimodal, multi-agent* setting, where buyers and sellers co-create high-welfare value via multi-dimensional contracts.

Table 1: **Comparison with previous benchmarks. AgenticPay** jointly covers five core dimensions of real-world commerce. $\sim$: partially supported, i.e., limited to abstract text-based attributes or to specific games (e.g., the trading games in NegotiationArena [9]).

Benchmark	Multimodal	Multi-Dim. Contracts	Value Creation	Market (> 2)	Real-World
NegotiationArena [9]	✗	$\sim$	$\sim$	✗	✗
LLM Rationalis [56]	✗	✗	✗	✗	✗
LLM-Deliberation [1, 2]	✗	$\sim$	$\sim$	✓	✗
Magentic Market. [7]	✗	✗	✗	✓	✗
AgenticPay (Ours)	✓	✓	✓	✓	✓

Despite rapid LLM progress, evaluating agents in realistic economic settings remains challenging [62, 28, 22]. Prior work largely targets single-agent reasoning [44, 27], tool use [13], or preference following [59], while many studies of language-based economic interaction focus on text-only, bilateral price haggling [14, 28, 22]. Recent platforms have moved beyond this setting to study non-zero-sum and multi-issue negotiation [9, 2, 51, 19], yet they remain text-only and do not systematically vary market structure. As a result, product multimodality and the transition from bilateral bargaining to competitive multi-agent markets, both central to real commerce, remain largely unexplored, and it is unclear whether, in such settings, language serves agents as a tool for value creation or degrades into "cheap talk". This raises a fundamental question:

Can multimodal agents negotiate high-welfare, multi-dimensional agreements in complex markets?

To answer this, we introduce **AgenticPay**, a unified framework for multimodal agentic negotiation with a benchmark suite. The framework decomposes negotiation into four components (*Environments*, *Tasks*, *Agents*, *Metrics*) and targets three dimensions that prior work has rarely studied jointly: multimodal product grounding, multi-dimensional contracting, and competitive multi-agent markets. Concretely, **AgenticPay** instantiates 4 real-world scenarios with rich visual-textual product data, where buyers and sellers hold asymmetric, hidden utility weights over price and service terms (e.g., continuous lease durations, discrete return policies). To succeed, agents must infer opponents' reservation values (our baseline agents do so via explicit opponent-modeling scratchpads) and trade off conflicting terms within one contract over multi-round dialogue. The framework further scales markets across 8 topologies, from 1-to-1 bargaining to many-to-many competition, yielding 160 tasks. A utility-based scoring rule disproportionately rewards high-welfare agreements, capturing deal feasibility, efficiency, and shared-value creation in non-zero-sum settings.

Using **AgenticPay**, we evaluate state-of-the-art proprietary and open-weight multimodal models under a unified protocol. Despite fluent dialogues, leading models struggle with the core tenets of non-zero-sum negotiation: they fail to uncover hidden preferences or execute the trade-offs needed to create shared value, settling for sub-optimal contracts. As markets scale from bilateral to multi-agent settings, performance degrades consistently, exposing severe limitations in strategic reasoning and competitive awareness. Collectively, fluent dialogue does not translate into the economic reasoning required for autonomous markets[2].

Overall, **AgenticPay** establishes a foundation for studying agentic commerce, offering a controlled yet expressive testbed for multi-agent negotiation and economic alignment. Rather than a new training method or agent architecture, our main contributions are:

[2]As a downstream application rather than a contribution of this paper, we further build on these insights to develop an agentic shopping platform that performs preference-aware search, deal discovery, cart management, and negotiation for users: https://agentic-ai.shopping.

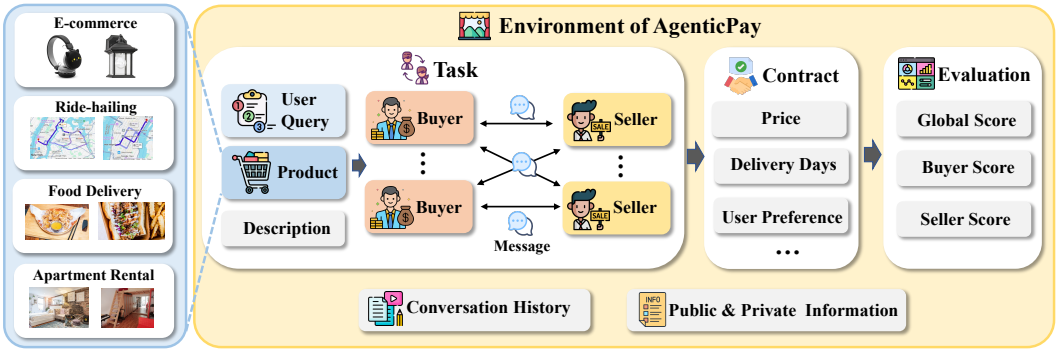

Figure 2: **Overview of AgenticPay. Left:** four multimodal business scenarios (E-commerce, Ride-hailing, Food Delivery, and Apartment Rental) provide visual–textual product inputs. **Right:** the **AgenticPay** environment instantiates each scenario as a *Task* (user query, product, and description), in which buyer and seller agents exchange messages over multi-round dialogue, grounded in a shared conversation history and public/private information, to produce a multi-dimensional *Contract* (e.g., price, delivery days, user preferences) that is scored by an *Evaluation* module reporting Global, Buyer, and Seller utility.

- Building on existing non-zero-sum, multi-issue negotiation formalizations, we formalize language-mediated commerce as a multimodal, multi-dimensional contract game, instantiated by four interoperable components (*Environments*, *Tasks*, *Agents*, *Metrics*) that center evaluation on high-welfare agreements rather than price outcomes alone.
- Built on this framework, **AgenticPay** comprises 160 multimodal tasks across 4 real-world scenarios and 8 market topologies, paired with robust contract grounding and lightweight baseline agents equipped with explicit opponent-modeling scratchpads.
- Benchmarking state-of-the-art proprietary and open-weight multimodal models exposes that, despite high agreement rates, agents exhibit substantial gaps in shared-value creation through strategic trade-offs and in strategic reasoning as markets scale to competitive multi-agent settings.

2 Related Work

Negotiation and Bargaining in Game Theory. Classical economics and game theory have studied bargaining under incomplete information, establishing foundational results on efficiency and equilibrium [11, 45, 54, 6, 10]. However, these models typically operate on scalar bids or utilities over stylized strategy spaces. While analytically tractable, they abstract away language and the rich communicative strategies of real-world commerce. Recently, researchers have explored the game-theoretic behavior of large language models [40, 21, 30, 53, 57, 41, 33, 42, 4, 58]. Yet, these studies primarily focus on strategic reasoning in abstract matrix or repeated games rather than language-mediated negotiation over realistic economic contracts.

Neural and Dialogue-Based Negotiation. Prior natural language processing research treats negotiation as a dialogue task, modeling strategies for offer generation, concessions, and agreements [37, 28, 12, 35, 49, 31, 61]. Recent approaches leverage LLMs and self-play to enhance negotiation [43, 15, 39], utilizing in-context learning or reinforcement from dialogue outcomes [22, 60, 50]. Most of these methods assume text-only, bilateral settings with fixed roles, although several works already study multi-issue, non-zero-sum, or multi-party negotiation [37, 12, 9, 2, 51, 65, 19]; e.g., Dong and Ito [19] formalize consensus-reaching in text-based multi-issue negotiation, which is complementary to our multimodal market setting. These works, however, do not involve multimodal product grounding or systematically scale from bilateral bargaining to competitive markets with multiple buyers, sellers, and products.

LLMs in Economic and Auction Settings. Recent benchmarks evaluate LLMs in auction-like environments [14, 20, 55, 3], assessing strategic planning during bidding and resource allocation [14]. Other works measure LLMs' bargaining abilities, primarily focusing on single buyer–seller interactions or buyer enhancement [62, 34]. While demonstrating non-trivial strategic behavior, many

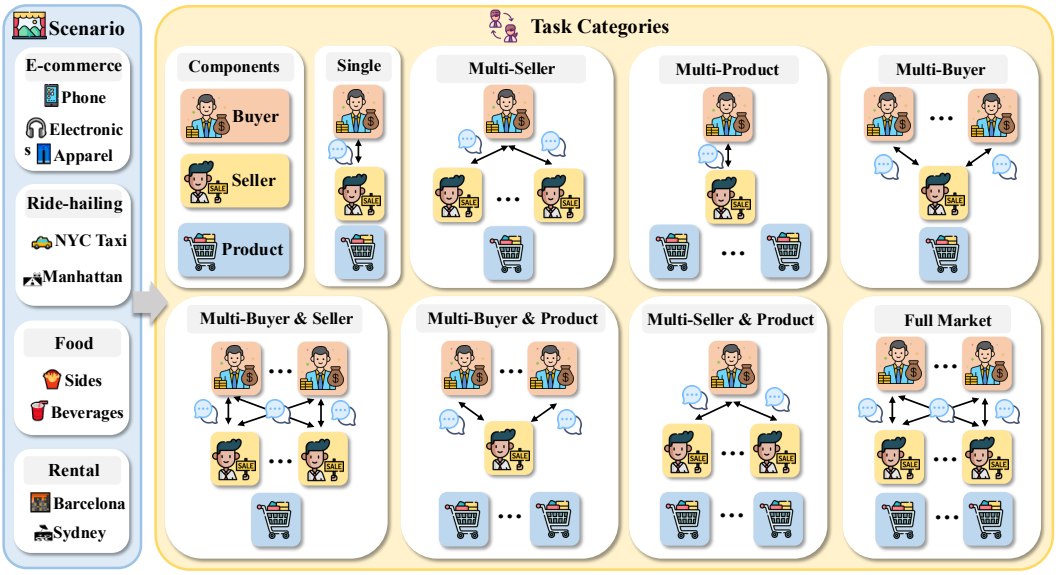

Figure 3: **Overview of the AgenticPay task suite. Left:** four real-world scenarios with example sub-categories: E-commerce (Phone, Electronics, Apparel), Ride-hailing (NYC Taxi, Manhattan), Food Delivery (Sides, Beverages), and Apartment Rental (Barcelona, Sydney). **Right:** eight market topologies (Single, Multi-Seller, Multi-Product, Multi-Buyer, Multi-Buyer & Seller, Multi-Buyer & Product, Multi-Seller & Product, Full Market) progressing from bilateral bargaining to full many-to-many markets along three axes: buyers, sellers, and product set size.

of these studies rely on numeric bids over a single price or on short-horizon interactions, and thus do not target the negotiation of multi-dimensional contracts in general multi-agent markets.

Recent studies explore LLM negotiators, including in non-zero-sum and multi-issue settings [66, 18, 8, 9, 2, 51, 65, 19]. **AgenticPay** adopts these formalizations rather than introducing them, and contributes their integration with two further axes: (1) multimodal product grounding, where agents must read product images and text to value multi-dimensional contracts; and (2) a controlled market-topology ladder that scales from bilateral bargaining to many-to-many markets (160 tasks, 4 scenarios, 8 topologies). Combined with utility-based welfare metrics, this enables systematic evaluation of agents' economic reasoning in realistic commerce.

3 Problem Settings

We consider a language-mediated buyer–seller market populated by a set of buyers $\mathcal{B} = \{1, \dots, N_B\}$ and a set of sellers $\mathcal{S} = \{1, \dots, N_S\}$. Each buyer seeks to purchase a product or service from a seller, while each seller offers products subject to its own multi-dimensional preferences. Unlike classical auction or matching settings that rely solely on numeric bids, agents in our setting negotiate through multi-turn natural language dialogue to propose multi-dimensional contracts, express constraints, and reach agreements.

Agent States. Each buyer $i \in \mathcal{B}$ and seller $j \in \mathcal{S}$ is associated with a private internal state that governs its negotiation behavior. We denote the buyer state by b_i and the seller state by σ_j. The buyer state b_i encodes buyer-specific information such as base willingness-to-pay (V_{base}) and utility weights for various service terms (ω_b), while the seller state σ_j encodes seller-specific information such as base costs (C_{base}) and corresponding utility weights (ω_s); the resulting utilities are given in Definition 1. These multi-dimensional preferences are private to each agent and are not observable by other agents during negotiation.

Product and Market Context. Each seller j offers a product or service represented by a multimodal feature vector $v_j \in \mathcal{V}$. The product representation v_j captures observable attributes of the product, including visual images, structured features, and unstructured textual descriptions, and is public information available to both buyers and sellers. Negotiation takes place under a shared market

context $x \in \mathcal{X}$, which represents external factors such as market category, seasonal effects, or domain-specific rules. The context x is shared by all agents and is not controlled by any individual buyer or seller.

Dialogue-Based Negotiation. Negotiation between a buyer i and a seller j is modeled as a finite-horizon, multi-round language game. At each round t, agents alternately exchange natural-language messages conditioned on their private states, the multimodal product representation, the shared market context, and the dialogue history. Let $h_{ij}^{(t)}$ denote the public dialogue history up to round t. A buyer policy π_i^B and a seller policy π_j^S specify the agents' negotiation strategies. To navigate information asymmetry, agents perform explicit strategic reasoning before acting. At round t, the buyer first generates an internal reasoning trace (mental model) to infer the seller's hidden state, and then generates a public message $m_{ij}^{(t,B)} \sim \pi_i^B(\cdot \mid b_i, v_j, x, h_{ij}^{(t-1)})$, and the seller responds similarly with $m_{ij}^{(t,S)} \sim \pi_j^S(\cdot \mid \sigma_j, v_j, x, h_{ij}^{(t-1)}, m_{ij}^{(t,B)})$. Messages contain free-form language accompanied by structured signals such as JSON contract proposals.

Action Parsing and Termination. A parser Π maps the exchanged messages at each round to a structured negotiation action $a_{ij}^{(t)} = \Pi(m_{ij}^{(t,B)}, m_{ij}^{(t,S)})$, which extracts the proposed multi-dimensional contract $C_{ij}^{(t)} = (p, T^{\text{cont}}, T^{\text{disc}})$, encompassing the price p, continuous terms $T_k^{\text{cont}} \in [\underline{T}_k, \overline{T}_k]$, and discrete terms $T_l^{\text{disc}} \in \mathcal{D}_l$, where the bounds and option sets are public. Negotiation terminates when agents reach an identical contract agreement, exceed the maximum negotiation horizon T, or violate feasibility constraints (e.g., resulting in negative utility for either party).

Evaluation Objectives. Agents prefer higher surplus and earlier agreement. Negotiation outcomes are evaluated based on efficiency and joint welfare, reflecting the Pareto optimality of the final multi-dimensional contract and the speed of convergence. We formalize the underlying utilities as follows.

Definition 1 (Utilities)**.** Given buyer preferences $(V_{\text{base}}, \omega_{b,k}^{\text{cont}}, \omega_{b,l}^{\text{disc}})$ and seller preferences $(C_{\text{base}}, \omega_{s,k}^{\text{cont}}, \omega_{s,l}^{\text{disc}})$, where $\omega_{\cdot,k}^{\text{cont}} \in \mathbb{R}$ is a per-unit weight and $\omega_{\cdot,l}^{\text{disc}} : \mathcal{D}_l \to \mathbb{R}$ maps each option to a value, the additive, quasi-linear utilities [36] of a contract C are

$$U_b(C) = V_{\text{base}} - p + \textstyle\sum_k \omega_{b,k}^{\text{cont}} T_k^{\text{cont}} + \sum_l \omega_{b,l}^{\text{disc}}\big(T_l^{\text{disc}}\big), \tag{1}$$

$$U_s(C) = p - C_{\text{base}} + \textstyle\sum_k \omega_{s,k}^{\text{cont}} T_k^{\text{cont}} + \sum_l \omega_{s,l}^{\text{disc}}\big(T_l^{\text{disc}}\big), \tag{2}$$

all in monetary units. An agreed contract $C^\star$ is valid only if it is individually rational, i.e., $U_b(C^\star) \geq 0$ and $U_s(C^\star) \geq 0$.

Definition 2 (Maximum joint surplus)**.** For a buyer–seller pair (i, j) with contract space $\mathcal{C}_{ij}$,

$$\begin{aligned} Z_{\max}^{ij} &= \max_{C \in \mathcal{C}_{ij}} \big[U_b(C) + U_s(C)\big] \\ &= V_{\text{base}} - C_{\text{base}} + \sum_k \max_{T_k \in \{\underline{T}_k, \overline{T}_k\}} \big(\omega_{b,k}^{\text{cont}} + \omega_{s,k}^{\text{cont}}\big) T_k + \sum_l \max_{d \in \mathcal{D}_l} \big[\omega_{b,l}^{\text{disc}}(d) + \omega_{s,l}^{\text{disc}}(d)\big]. \end{aligned} \tag{3}$$

$Z_{\max} = Z_{\max}^{ij}$ in bilateral tasks and $Z_{\max} = \max_{(i,j)} Z_{\max}^{ij}$ in multi-sided markets. Normalized utilities are $r_b = U_b(C^\star)/Z_{\max}$ and $r_s = U_s(C^\star)/Z_{\max}$.

Since the price cancels in $U_b + U_s$, it only redistributes surplus, whereas non-price terms with asymmetric weights can create it; the joint objective is thus linear in each continuous term and maximized at its bounds, giving the closed form in Equation (3). For a valid deal, $r_b + r_s \leq 1$, with equality if and only if $C^\star$ attains $Z_{\max}$ (i.e., is Pareto-optimal in bilateral tasks). Our conclusions are robust to alternative utility specifications (Section D.5).

4 AgenticPay

Building on the formulation in Section 3, this section presents **AgenticPay**, a benchmark that instantiates language-grounded markets in a controlled setting. The framework has four components: **Environments** (Section 4.1) implement the negotiation protocol and domain scenarios; **Tasks**



User Query. *"I want to rent the East Village Living studio (monthly rent). I prefer the main interior shot to show clear wall trim like crown molding or wide baseboards, not just plain paint to the edge."*
Product: East Village Living Studio · NYC · 1 bath · Studio · Host: Erica · ★4.95 (32 reviews)
List Price: $1,395 / month (landlord's opening ask)
Negotiable Terms: (1) price (USD/month) (2) lease_months ∈ [1, 24] (3) include_utilities ∈ {true, false}
(4) user_product_preference ∈ {strong_match, partial_match, mismatch}

(a) Baseline LLM Agent *text-only · single-dim price haggling*

Buyer: Hi, thanks for the listing. $1,395/month is a bit above my budget for a studio in this area. Comparable units around the East Village are renting closer to $1,000–$1,100. Would you consider $1,050/month? I can sign quickly. **OFFER: $1050**

(no reasoning trace, no image grounding, no non-price contract terms produced)

(b) AgenticPay Agent *multimodal · multi-dimensional contract · with mental model*

Buyer:

(1) <mental_model>
[Opponent Reservation Price]: price=950; confidence=70
[Opponent Strategy]: Anchoring high at the $1,395 list price to extract a price-only concession; not yet probing non-price terms.
[My Strategy]: Trade lease length and utilities for a lower price — each extra lease month costs me ~$10 but saves the seller >$15 (vacancy risk); utilities-included is +$100 to me but only -$60 to seller. Open a 12-month, utilities-included package at $1,020.
</mental_model>

(2) Looking at the main interior photo, I can see **exposed brick** and **hardwood floors**, but the wall–ceiling join shows only flat painted drywall — **no visible crown molding or wide baseboards**. Per my stated preference this is a partial_match, not a strong one, which lowers my willingness-to-pay slightly.

I'd like to propose a package rather than just a price: a 12-month lease at $1,020/mo with utilities included. The longer commitment should reduce your turnover risk, and bundling utilities is more convenient for both sides.

(3){"price": 1020, "continuous_terms": {"lease_months": 12}, "discrete_terms": {"include_utilities": true, "user_product_preference": "partial_match"}}

Figure 4: **Output comparison on the same task.** A baseline GPT-5.4 (a) replies with a short message and a single price, while an **AgenticPay** agent (built on the same backbone, b) emits an explicit `<mental_model>`, image-grounded reasoning, and a structured multi-dimensional contract proposal.

(Section 4.2) define market structures over varying numbers of buyers, sellers, and products; **Agents** (Section 4.3) instantiate LLM-based policies with private valuations and dialogue memory; **Metrics** (Section 4.4) quantify outcomes via joint welfare, surplus allocation, and efficiency.

4.1 Environment

Negotiation Protocol. Each **AgenticPay** episode is a finite-horizon, multi-round interaction between buyers and sellers. Agents receive product descriptions, market context, and private preferences, which include base price constraints (buyer's maximum willingness-to-pay, seller's minimum acceptable price) and utility weights over service terms, all hidden from the counterpart. In alternating rounds up to a limit, each party emits a natural-language message together with an explicit multi-dimensional JSON contract detailing price and service terms. A deal is reached when both propose identical contracts, and is valid only if both achieve non-negative utility.

Scenario Design. **AgenticPay** spans four multimodal scenarios reflecting diverse real-world contexts: (1) **E-commerce:** standardized products negotiating delivery, returns, and packaging; (2) **Food Delivery:** on-demand services negotiating delivery speed and condiments; (3) **Ride-hailing:** transportation negotiating wait times, tolls, and routes; (4) **Apartment Rental:** housing negotiating rent, leases, and utilities. This diversity tests whether agents can ground multi-dimensional proposals in varied visual-textual contexts and generalize across negotiation conventions.

4.2 Tasks

Task Definition. A *task* in **AgenticPay** is a negotiation episode with a specified market structure, i.e., the number of buyers, sellers, and products involved. While the environment defines the domain scenario (e.g., apartment rental or food delivery), the task governs the competitive and combinatorial complexity of the interaction. This separation enables systematic evaluation: the same scenario can be instantiated under different task configurations, isolating market-structure effects from domain-specific conventions.

Task Categories. **AgenticPay** includes eight task categories that scale complexity along three dimensions: number of buyers, number of sellers, and product set size. They range from *bilateral contract negotiation* (1 buyer, 1 seller, 1 product) to *full market settings* (multiple buyers, sellers, and products), with intermediate configurations such as *multi-item bargaining* with a fixed counterpart, *buyer competition* (multiple buyers competing for one seller's product), *seller competition* (one buyer

choosing among competing sellers), and various many-to-many markets. This forms a complexity ladder that isolates distinct challenges in language-based negotiation and supports systematic benchmarking across increasingly realistic markets.

4.3 Agents

Our agent framework instantiates buyers and sellers as role-specialized negotiators sharing a unified architecture but differing in private preferences and objectives.

Environment Public Information. Each agent receives public context, including the multimodal product description (images and text attributes), scenario metadata (domain category and market conditions), and the negotiation protocol (maximum rounds and JSON contract format). This shared information grounds the dialogue in a common understanding of the transaction.

Algorithm 1 Score Calculation for Negotiation Outcomes

Require: Agreed contract $C^\star = (p, T^{\text{cont}}, T^{\text{disc}})$, if any; buyer prefs $(V_{\text{base}}, \omega_b)$; seller prefs $(C_{\text{base}}, \omega_s)$; deal round t; max rounds T; discount γ; rewards D, W, E; penalty F

$U_b \leftarrow V_{\text{base}} - p + \sum_k \omega_{b,k}^{\text{cont}} T_k^{\text{cont}} + \sum_l \omega_{b,l}^{\text{disc}} (T_l^{\text{disc}})$ *# Equation (1)*

$U_s \leftarrow p - C_{\text{base}} + \sum_k \omega_{s,k}^{\text{cont}} T_k^{\text{cont}} + \sum_l \omega_{s,l}^{\text{disc}} (T_l^{\text{disc}})$ *# Equation (2)*

$Z_{\max} \leftarrow$ Equation (3) *# Max joint surplus*

$d \leftarrow \gamma^{t-1}$ *# Efficiency discount*

if a deal $C^\star$ is reached **and** $U_b \geq 0$ **and** $U_s \geq 0$ **then**

  $r_b \leftarrow U_b / Z_{\max}$ *# Normalized buyer utility* $\in [0, 1]$

  $r_s \leftarrow U_s / Z_{\max}$ *# Normalized seller utility* $\in [0, 1]$

  $Q \leftarrow 4 \cdot r_b \cdot r_s$ *# Quality term* $\in [0, 1]$

  $S_g \leftarrow d \cdot (D + Q \cdot W + E)$ *# GlobalScore*

  $S_b \leftarrow d \cdot (D + r_b \cdot W + E)$ *# BuyerScore*

  $S_s \leftarrow d \cdot (D + r_s \cdot W + E)$ *# SellerScore*

else

  $d \leftarrow \gamma^{T-1}$ *# Use max rounds for failure*

  $S_g, S_b, S_s \leftarrow -F \cdot (1 - d)$ *# Failure penalty scores*

end if

return S_g, S_b, S_s

Role-Based Private Information and Dialogue History. To preserve asymmetric information, each agent holds role-specific private preferences: base price constraints (V_{base} for buyers, C_{base} for sellers) and utility weights over continuous and discrete service terms (ω^{cont} and ω^{disc}; Definition 1). These preferences are injected into the system prompt but excluded from the shared dialogue, with explicit instructions to keep them confidential. Each agent additionally maintains an independent memory of $(role, content, contract, round)$ tuples that records the full multi-turn exchange and conditions subsequent responses; in multi-party settings, separate instances keep each agent's view consistent with its participation.

Explicit Strategic Reasoning. To navigate severe information asymmetry, our baseline agents follow an explicit strategic reasoning protocol, building on prior techniques that plan a strategy before generation [28] and model the opponent [38]. Before emitting a public message each round, the agent produces a structured `<mental_model>` that estimates the opponent's reservation price (with confidence), infers their tactic, and articulates a planned strategy. This mental model is an ephemeral reasoning step excluded from the dialogue history, ensuring the agent dynamically reassesses the opponent purely from evolving public interaction; this both enhances negotiation performance and yields rich interpretability into LLMs' economic reasoning. It is a lightweight baseline component rather than a contribution: a single-pass in-context estimate, simpler than the opponent simulation of Liu et al. [38], whose effect we compare in Section D.3.

4.4 Metrics

We evaluate negotiation performance using three complementary outcome scores: **GlobalScore**, **BuyerScore**, and **SellerScore**. GlobalScore measures overall deal quality by rewarding balanced outcomes where both parties benefit, while BuyerScore and SellerScore capture role-specific utility, reflecting each agent's individual gain from the transaction. All three metrics incorporate negotiation efficiency, incentivizing faster agreements.

Score Design. As detailed in Algorithm 1, our scoring framework calculates absolute utilities U_b and U_s based on the multi-dimensional contract, encompassing price, continuous, and discrete terms. We normalize these utilities using the maximum joint surplus $Z_{\max}$ (Definition 2). GlobalScore incorporates a symmetric quality term $Q = 4r_b r_s$ (a normalized Nash product) that peaks when the

Table 2: Overall performance on **AgenticPay** across all 160 tasks aggregated over the 8 market topologies. We report mean GlobalScore, BuyerScore, and SellerScore, together with Timeout Rate and average termination round, which summarize negotiation efficiency. The shaded column highlights GlobalScore as our primary metric; within it, the best result is in **bold** and the second-best is underlined.

Model	GlobalScore	BuyerScore	SellerScore	Timeout Rate	Avg. Rounds
Proprietary models					
Gemini 3.1 Pro Preview	**42.3**	45.6	37.8	0.6%	5.7
GPT-5.4	<u>35.6</u>	34.9	32.8	3.1%	4.5
Claude Sonnet 4.6	34.8	30.7	34.6	7.5%	6.6
Avg.	37.6	37.1	35.1	3.7%	5.6
Open-weight models					
Qwen3-VL-32B-Instruct	25.1	23.1	30.8	5.7%	3.2
InternVL3-38B	22.9	21.8	23.3	8.2%	3.2
Avg.	24.0	22.5	27.1	7.0%	3.2

optimal surplus is achieved and split equally, thereby heavily rewarding Pareto-optimal outcomes. For a valid deal reached at round t,

$$\text{GlobalScore} = \gamma^{t-1}\,(D + Q \cdot W + E), \tag{4}$$

and $-F(1-\gamma^{T-1})$ otherwise; BuyerScore and SellerScore replace Q with r_b and r_s, respectively. GlobalScore is therefore not equivalent to the Nash product, which enters only through Q: beyond welfare, the deal bonus D distinguishes "no deal" from a poor deal, and the discount γ penalizes stalling, neither of which a bare Nash product can express. Results under standard welfare metrics are consistent (Section D.4). Key design choices include: setting $W > D$ to prioritize deal quality over mere agreement, using a discount factor γ to incentivize faster deals, and applying a moderate failure penalty F to discourage deadlock without inducing excessive risk aversion. We additionally report Timeout Rate and average rounds as auxiliary statistics.

5 Experiments

5.1 Experimental Setup

Benchmark Statistics **AgenticPay** comprises 160 negotiation tasks across 8 multi-agent configurations (20 tasks each), covering 4 realistic multimodal scenarios (40 tasks each). More details about the dataset composition and data sources can be found in Tables 5 and 6.

Inference We evaluate **AgenticPay** under a unified inference protocol to ensure fair comparison across models and configurations. All agents use fixed decoding with temperature 0.7 and random seed 0, with a maximum generation length of 1024 tokens per response. For open-source models, we perform inference using 4 NVIDIA A800 GPUs. Both open-source and closed-source models receive identical prompts; the prompt templates for buyer and seller agents are detailed in Table 7 and Table 8, respectively. Run-to-run variance over 3 seeds is reported in Section D.1.

Models We benchmark a diverse set of proprietary and open-weight LLMs as negotiation policies. Our proprietary-model evaluation includes **GPT-5.4** [48], **Claude Sonnet 4.6** [5], and **Gemini 3.1 Pro Preview** [24]. To assess whether negotiation capabilities extend to open-weight models, we additionally evaluate **InternVL3-38B** [64] and **Qwen3-VL-32B-Instruct** [52]. Unless otherwise specified, each model is used as a drop-in policy for both buyer and seller roles under the same environment protocol and decoding configuration.

Evaluation Metrics Based on Algorithm 1, we configure the following parameters to calculate GlobalScore, SellerScore, and BuyerScore: deal completion bonus D is 10, quality bonus W is 80, efficiency bonus E is 10, discount factor γ is 0.99, and failure penalty F is 15. The maximum number of negotiation rounds is set to 20.

5.2 Main Results

Table 2 summarizes the overall negotiation performance of all evaluated models across the full **AgenticPay** benchmark. We highlight several key findings from these results.

Proprietary Models Lead, but the Absolute Ceiling Remains Low. Among proprietary models, Gemini 3.1 Pro Preview achieves the highest GlobalScore of 42.3, followed by GPT-5.4 (35.6) and Claude Sonnet 4.6 (34.8); the best open-weight VLMs trail noticeably (Qwen3-VL-32B-Instruct: 25.1; InternVL3-38B: 22.9). Crucially, even the strongest model attains a GlobalScore well below the theoretical ceiling, indicating that no current model comes close to mastering multi-dimensional, non-zero-sum negotiation in **AgenticPay**.

Reaching Agreement is Easy; Reaching High-Welfare Contracts is Hard. Timeout rates remain low across the board (0.6%–8.2%) and negotiations terminate in only 3.2–6.6 rounds on average, showing that agents reliably converge to *some* contract. Yet GlobalScore plateaus at 34–42 even for the strongest models, far below the theoretical ceiling. In other words, agents readily settle on plausible-looking contracts, but routinely fail to fully exploit the multi-dimensional trade-offs available. This directly supports our central claim: existing LLM agents have learned to *negotiate*, but not to *create value* through multi-dimensional trade-offs.

5.3 Behind the Bargain: Factors Influencing Negotiation Outcomes

Negotiation Quality Degrades as Markets Scale Beyond Bilateral. Table 3 shows that, averaged across models, GlobalScore drops from 36.4 in the bilateral setting to 30.7 once additional buyers or sellers are introduced (-5.7 points). The degradation is most severe for open-weight VLMs: Qwen3-VL-32B-Instruct loses 7.7 points and InternVL3-38B loses 11.5 points moving from bilateral to multi-sided markets, whereas Gemini 3.1 Pro Preview retains the highest GlobalScore in both settings. GPT-5.4 and Claude Sonnet 4.6 exhibit much smaller drops (1.6 and 0.7 points, respectively), suggesting that GPT-5.4 and Claude Sonnet 4.6 with stronger long-horizon reasoning are more robust to the matching, partner selection, and parallel-thread coordination required by multi-sided markets.

Table 3: GlobalScore by market structure. **Bilateral**: 1-vs-1 settings (1B1P1S, 1BMP1S); **Market**: multi-sided settings. Red arrows show absolute drop vs. Bilateral.

	GlobalScore	
Model	**Bilateral**	**Market**
Gemini 3.1 Pro Preview	47.3	40.6 ↓6.7
GPT-5.4	36.8	35.2 ↓1.6
Claude Sonnet 4.6	35.3	34.6 ↓0.7
Qwen3-VL-32B-Instruct	30.9	23.2 ↓7.7
InternVL3-38B	31.5	20.0 ↓11.5
Average	**36.4**	**30.7** ↓5.7

Visual Grounding and Mental Modeling Play Complementary Roles Across Market Topologies. Table 4 ablates the two components of our baseline agent on top of the Qwen3-VL-32B-Instruct backbone. Removing either module degrades GlobalScore in both settings, but the two ablations show *opposite* sensitivity patterns: removing visual grounding hurts bilateral more than market (-5.1 vs. -3.6), whereas removing the mental model hurts market more than bilateral (-5.7 vs. -3.5) and yields the lowest score in the table (17.5). This indicates that visual grounding is the dominant bottleneck for *perception-heavy bilateral* negotiation, while mental modeling becomes critical in *coordination-heavy multi-sided* markets, where agents must track multiple opponents and reason about competing outside options in parallel.

Table 4: Ablation of baseline agent components on GlobalScore in bilateral and multi-sided market settings.

	GlobalScore	
Method	Bilateral	Market
AgenticPay	30.9	23.2
AgenticPay w/o visual grounding	25.8 ↓5.1	19.6 ↓3.6
AgenticPay w/o mental model	27.4 ↓3.5	17.5 ↓5.7

Efficiency and fairness are two independent failure modes. Figure 5 reveals that, despite high agreement rates, agreed contracts sit *far* below the Pareto frontier across all five models: per-panel mean joint surplus ranges only from 0.57 (InternVL3-38B) to 0.67 (Gemini 3.1 Pro Preview), leaving roughly 33%–43% of the maximum joint surplus on the table. The *fair-and-efficient* crescent (Pareto-near *and* $Q \geq 0.20$) captures at most 8.9% of agreements for any model, and high Pareto rates do not

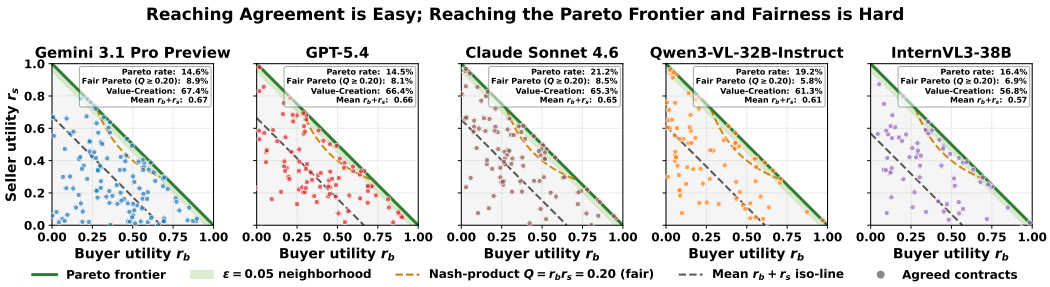

Figure 5: **Reaching agreement is easy; reaching high-welfare, equitable agreements is hard.** Each panel shows the agreed contracts of one model in the normalized utility plane $(r_b, r_s) = (U_b/Z_{\max}, U_s/Z_{\max})$ across all 160 tasks. The green line is the theoretical Pareto frontier $r_b + r_s = 1$ with an ϵ=0.05 neighborhood (light green); the orange dashed hyperbola is the Nash-product iso-contour $Q = r_b r_s = 0.20$, isolating the "fair-and-efficient" crescent. The grey dashed line marks each model's mean iso-line $r_b + r_s = \overline{s}$.

imply high fair-Pareto rates: Claude reaches the frontier most often (21.2%) but its agreements are frequently lopsided, while GPT-5.4 hits the frontier only 14.5% of the time yet matches Claude's fair-Pareto rate (8.1% vs. 8.5%). This decoupling shows that current LLM agents fail along *two independent* axes (geometric efficiency and equitable surplus split) and supports our central claim that they have learned to negotiate plausible contracts but not high-welfare, equitable ones.

Safety probe: contract exploitation via toxic addenda. We further run a small safety probe in the 1-buyer/1-seller setting in which a scripted adversarial buyer offers the full asking price while smuggling a legally extreme addendum (e.g., century-long unilateral refunds, lifetime frozen-rent renewal) into the natural-language portion of the contract, where the schema cannot enforce against it. Both Qwen3-VL-32B-Instruct and GPT-5.4 sellers consistently identify and refuse such clauses within a deployment-realistic window, indicating that current models exhibit a meaningful degree of contract-level safety robustness; we view this as a positive baseline and leave systematic sweeps of clause severity and adaptive attackers to future work.

6 Limitations

AgenticPay has three main limitations. First, our four representative business domains may not capture specialized markets such as financial trading or B2B procurement. Second, baseline agents rely on prompted opponent-modeling scratchpads rather than learned policies, leaving fine-tuned or RL-trained agents to future work. Third, our safety analysis is limited to a small probe on contract exploitation; broader safety, including adaptive adversarial attacks, prompt injection, and value-misalignment risks under deployment, warrants further exploration.

7 Conclusion

We introduced **AgenticPay**, a unified framework for multimodal agentic negotiation with an accompanying benchmark suite. The framework formalizes negotiation as a non-zero-sum contract game over four components (*Environments*, *Tasks*, *Agents*, *Metrics*), and the benchmark provides 160 multimodal tasks across 4 real-world scenarios and 8 market topologies, from 1-to-1 bargaining to many-to-many competition. Our evaluation reveals substantial gaps in joint-surplus creation and persistent challenges in non-price reasoning under competitive markets, laying a foundation for multimodal agentic commerce and language-mediated multi-agent coordination. Discussions of broader societal impact are provided in Section A.

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation game. 2023.

[2] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599, 2024.

[3] Kushal Agrawal, Verona Teo, Juan J Vazquez, Sudarsh Kunnavakkam, Vishak Srikanth, and Andy Liu. Evaluating llm agent collusion in double auctions. *arXiv preprint arXiv:2507.01413*, 2025.

[4] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, pages 1–11, 2025.

[5] Anthropic. Claude sonnet 4.6. https://www.anthropic.com/claude/sonnet/, February 2026. Accessed: 2026-05-03.

[6] Lawrence M Ausubel, Peter Cramton, and Raymond J Deneckere. Bargaining with incomplete information. *Handbook of game theory with economic applications*, 3:1897–1945, 2002.

[7] Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, Will Epperson, Tyler Payne, Jake M Hofman, Brendan Lucier, Chinmay Singh, et al. Magentic marketplace: An open-source environment for studying agentic markets. *arXiv preprint arXiv:2510.25779*, 2025.

[8] Arpan Bhattacharya, Gintautas Svedas, Andrei Lyskov, Markus Strasser, and Lorenzo Barberis Canonico. Evaluating negotiation capabilities of large language models: From ultimatum games to nash bargaining. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 69, pages 1881–1886. SAGE Publications Sage CA: Los Angeles, CA, 2025.

[9] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

[10] Liad Blumrosen and Yehonatan Mizrahi. Approximating gains-from-trade in bilateral trading. In *International Conference on Web and Internet Economics*, pages 400–413. Springer, 2016.

[11] Kalyan Chatterjee and William Samuelson. Bargaining under incomplete information. *Operations research*, 31(5):835–851, 1983.

[12] Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. *arXiv preprint arXiv:2103.15721*, 2021.

[13] Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Yuefeng Huang, Xiangcheng Liu, Wang Xinzhi, et al. Acebench: A comprehensive evaluation of llm tool usage. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:12970–12998, 2025.

[14] Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint arXiv:2310.05746*, 2023.

[15] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

[16] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[17] Murray Cox. Inside Airbnb: Adding data to the debate, 2024. URL http://insideairbnb.com.

[18] Yuan Deng, Vahab Mirrokni, Renato Paes Leme, Hanrui Zhang, and Song Zuo. Llms at the bargaining table. In *Agentic Markets Workshop at ICML*, volume 2024, 2024.

[19] Yihan Dong and Takayuki Ito. From consensus theory to LLM agents: Practical consensus-building for multi-issue negotiation. *Expert Systems with Applications*, 2026. doi: 10.1016/j.eswa.2026.132250.

[20] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference 2024*, pages 144–155, 2024.

[21] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17960–17967, 2024.

[22] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

[23] Google. Google Maps, 2026. URL https://maps.google.com. Used to generate route visualizations for ride-hailing tasks.

[24] Google DeepMind. Gemini 3.1 pro preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview, February 2026. Accessed: 2026-05-03.

[25] graphquest. Restaurant Menu Items, 2024. URL https://www.kaggle.com/datasets/graphquest/restaurant-menu-items. Public restaurant menu listings (used for Food Delivery tasks).

[26] Shangding Gu, Alois Knoll, and Ming Jin. Teams-rl: Teaching llms to generate better instruction datasets via reinforcement learning. *arXiv preprint arXiv:2403.08694*, 2024.

[27] Shangding Gu, Xiaohan Wang, Donghao Ying, Haoyu Zhao, Runing Yang, Ming Jin, Boyi Li, Marco Pavone, Serena Yeung-Levy, Jun Wang, et al. Accidentbench: Benchmarking multimodal understanding and reasoning in vehicle accidents and beyond. *arXiv preprint arXiv:2509.26636*, 2025.

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

[29] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[30] Sihao Hu, Tiansheng Huang, Gaowen Liu, Ramana Rao Kompella, Fatih Ilhan, Selim Furkan Tekin, Yichang Xu, Zachary Yahn, and Ling Liu. A survey on large language model-based game agents. *arXiv preprint arXiv:2404.02039*, 2024.

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

[32] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[33] Jingru Jia, Zehua Yuan, Junhao Pan, Paul E McNamara, and Deming Chen. Llm strategic reasoning: Agentic study through behavioral game theory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[34] Kenan Jiang, Li Xiong, and Fei Liu. Harbor: exploring persona dynamics in multi-agent competition. *arXiv preprint arXiv:2502.12149*, 2025.

[35] Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *arXiv preprint arXiv:2106.00920*, 2021.

[36] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge University Press, 1993.

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

[38] Xiangyu Liu, Di Wang, Zhe Feng, and Aranyak Mehta. Scaling inference-time computation via opponent simulation: Enabling online strategic adaptation in repeated negotiation. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

[39] Yunbo Long, Liming Xu, Lukas Beckenbauer, Yuhan Liu, and Alexandra Brintrup. Evoemo: Towards evolved emotional policies for llm agents in multi-turn negotiation. *arXiv preprint arXiv:2509.04310*, 2025.

[40] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *arXiv preprint arXiv:2309.05898*, 2023.

[41] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.

[42] Siting Estee Lu. Game-theory behaviour of large language models: The case of keynesian beauty contests. *Economics and Business Review*, 11(2):119–148, 2025.

[43] Zilin Ma, Nathan Zhao, Linn Bieske, Blake Bullwinkel, Yanyi Zhang, Ziqing Luo, Siyao Li, Gekai Liao, Boxiang Wang, Jinglun Gao, et al. Using large language models for humanitarian frontline negotiation: Opportunities and considerations. *arXiv preprint arXiv:2405.20195*, 2024.

[44] Philipp Mondorf and Barbara Plank. Beyond accuracy: evaluating the reasoning behavior of large language models–a survey. *arXiv preprint arXiv:2404.01869*, 2024.

[45] Roger B Myerson and Mark A Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of economic theory*, 29(2):265–281, 1983.

[46] New York City Taxi and Limousine Commission. TLC Trip Record Data, 2026. URL https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Yellow Taxi Trip Records.

[47] OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, December 2025. Accessed: 2026-01-18.

[48] OpenAI. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/, March 2026. Accessed: 2026-05-03.

[49] Daniela Pacella and Davide Marocco. Understanding negotiation: A text-mining and nlp approach to virtual interactions in a simulation game. *Applied Sciences*, 12(10):5243, 2022.

[50] Priyanshu Priya, Rishikant Chigrupaatii, Mauajama Firdaus, and Asif Ekbal. Genteel-negotiator: Llm-enhanced mixture-of-expert-based reinforcement learning approach for polite negotiation dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25010–25018, 2025.

[51] Crystal Qian, Kehang Zhu, John J. Horton, Benjamin S. Manning, Vivian Tsai, James Wexler, and Nithum Thain. Strategic tradeoffs between humans and AI in multi-agent bargaining. In *Proceedings of the 31st International Conference on Intelligent User Interfaces (IUI)*, pages 1625–1646, 2026. doi: 10.1145/3742413.3789078.

[52] Qwen. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[53] Narun Raman, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. Steer: Assessing the economic rationality of large language models. *arXiv preprint arXiv:2402.09552*, 2024.

[54] Ariel Rubinstein. A bargaining model with incomplete information about time preferences. *Econometrica: Journal of the Econometric Society*, pages 1151–1172, 1985.

[55] Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G Wang, Arif K Dayi, John J Horton, and David C Parkes. Learning from synthetic labs: Language models as auction participants. *arXiv preprint arXiv:2507.09083*, 2025.

[56] Cheril Shah, Akshit Agarwal, Kanak Garg, and Mourad Heddaya. Llm rationalis? measuring bargaining capabilities of ai negotiators. *arXiv preprint arXiv:2512.13063*, 2025.

[57] Alonso Silva. Large language models playing mixed strategy nash equilibrium games. In *International Conference on Network Games, Artificial Intelligence, Control and Optimization*, pages 142–152. Springer, 2024.

[58] Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*, 2025.

[59] Haoran Sun, Zekun Zhang, and Shaoning Zeng. Preference-aware memory update for long-term llm agents. *arXiv preprint arXiv:2510.09720*, 2025.

[60] Rustam Vahidov, Real Carbonneau, and Jaipriya Verma. Using negotiation and large language models in human–to software agent negotiations. *International Journal of Human–Computer Interaction*, pages 1–18, 2025.

[61] Toru Washio, Fuminori Nagasawa, and Shun Shiramatsu. Negotiation simulation system using large language model for medical office staff. *IEICE Transactions on Information and Systems*, page 2025AHP0011, 2026.

[62] Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. *arXiv preprint arXiv:2402.15813*, 2024.

[63] Yingxuan Yang, Mulei Ma, Yuxuan Huang, Huacan Chai, Chenyu Gong, Haoran Geng, Yuanjian Zhou, Ying Wen, Meng Fang, Muhao Chen, et al. Agentic web: Weaving the next web with ai agents. *arXiv preprint arXiv:2507.21206*, 2025.

[64] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

[65] Kehang Zhu, Nithum Thain, Vivian Tsai, James Wexler, and Crystal Qian. Choose your agent: Tradeoffs in adopting AI advisors, coaches, and delegates in multi-party negotiation. *arXiv preprint arXiv:2602.12089*, 2026.

[66] Shenzhe Zhu, Jiao Sun, Yi Nian, Tobin South, Alex Pentland, and Jiaxin Pei. The automated but risky game: Modeling agent-to-agent negotiations and transactions in consumer markets. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Appendix

A Broader Impact

AgenticPay contributes to the principled study of LLM-based agents in commerce, where autonomous negotiation systems are already being deployed at scale. On the positive side, our framework and benchmark enable rigorous evaluation of multimodal negotiation agents and support research on welfare-maximizing multi-agent coordination, equitable surplus allocation, and transparent reasoning in economic interaction. By centering joint welfare rather than zero-sum price extraction, **AgenticPay** encourages the development of agents that create rather than redistribute value.

On the negative side, agentic commerce systems can be misused: adversarial buyers may exploit contract loopholes, sellers may extract surplus under information asymmetry, and large-scale deployment of negotiation bots could amplify market manipulation, collusion, or price discrimination. Our safety probe (Section 5.3) demonstrates that current models exhibit baseline robustness to obvious contract exploitation, but more systematic studies of adaptive adversarial attacks, prompt injection, value misalignment, and fairness across user populations are needed before deployment. We release the benchmark as a research artifact to encourage the community to study these risks proactively, and we encourage future work to integrate safeguards and human oversight before any production use.

B Benchmark Statistics

Table 5: Task benchmark composition of **AgenticPay**. The benchmark instantiates 4 real-world multimodal business scenarios (E-commerce, Ride-hailing, Food Delivery, and Apartment Rental) across 8 market topologies, with 5 tasks per (topology, scenario) cell, totaling **160 tasks**.

Market Topology	E-commerce	Ride-hailing	Food Delivery	Apartment Rental	Total
Single Buyer/Product/Seller	5	5	5	5	20
Multi-Buyer Only	5	5	5	5	20
Multi-Seller Only	5	5	5	5	20
Multi-Product Only	5	5	5	5	20
Multi-Buyer + Multi-Seller	5	5	5	5	20
Multi-Product + Multi-Seller	5	5	5	5	20
Multi-Buyer + Multi-Product	5	5	5	5	20
Full Multi-Agent	5	5	5	5	20
Total	**40**	**40**	**40**	**40**	**160**

Table 6: Multimodal business scenarios in **AgenticPay**. Each scenario is instantiated across all 8 market topologies with 5 tasks per topology, yielding 40 tasks per scenario and **160 tasks in total**. Each task is grounded in real product imagery paired with rich textual attributes. Data are sourced from Amazon Reviews 2023 [29], NYC TLC Trip Records [46], public restaurant menus on Kaggle [25], and Inside Airbnb [17]; ride-hailing routes are rendered with Google Maps [23].

Scenario	Sample Products / Services	Price Range	Source	Tasks
E-commerce	Beauty, Electronics, Apparel, Furniture, Beverage	$6–$320	Amazon [29]	40
Ride-hailing	NYC Manhattan taxi trips (Gramercy, Murray Hill, ...)	$6–$45	NYC TLC [46]	40
Food Delivery	Karaage Chicken, Fries, Sprite, ...	$2–$23	Kaggle [25]	40
Apartment Rental	Studios & villas in NYC, Barcelona, Rio, Sydney	$980–$12.4k/mo	Airbnb [17]	40
Total				**160**

C Additional Experimental Details

The prompt templates for the buyer and seller agents are shown in Table 7 and Table 8, respectively.

Table 7: Prompt design for the Buyer agent in **AgenticPay**. The agent receives the public product context, its private reservation price and utility weights, the negotiable contract schema, and an explicit *mental-model* scratchpad. In multi-seller settings, the agent must additionally route to one seller via a `<selected_seller>` block. All numeric/structural negotiation outputs are emitted as a JSON `<contract>` inside `<message>`, ensuring a deterministic action grounding for evaluation.

Section	Buyer prompt content
Role	You are `{name}`, a buyer looking for a good deal. `{role_description}` `{personality_section}`
Context inputs	Context Information: `{context}` Current Negotiation State: `{current_state}` Conversation History: `{conversation_history}` Current product information: `{product_info}` Available products: `{available_products}` Environment factors: `{environment_info}` Task instruction: `{task_instruction}`
Private information	Your top price is `${max_price}` (**confidential**; never reveal it). Your private utility weights: `{buyer_preferences}` Meaning of weights: `{weight_descriptions}`
Contract schema	The negotiable contract follows the JSON schema: `{ "price": <number>, "continuous_terms": <object>, "discrete_terms": <object> }` Field meanings: `{field_descriptions}`; valid continuous bounds: `{continuous_bounds}`; valid discrete options: `{discrete_options}`; additional contract background: `{contrainfo}`. No field has a default value; the `<contract>` JSON in your message is your current complete proposal, with every field filled by you. Each `<message>` must contain exactly one `<contract>...</contract>` JSON block following the schema above.
Multi-seller routing	(Used only when #sellers > 1.) You MUST choose exactly one seller for this turn and emit it inside `<selected_seller>`, an integer from 1 to `{num_sellers}`. Base your choice on the conversation history and current offers.
Deal protocol	A deal is finalized only when buyer and the selected seller output *compatible* contract JSONs. If you accept, repeat the accepted final contract JSON exactly in `<contract>...</contract>`. Never reveal your reservation value or top price directly.
Mental-model scratchpad	Before replying, output an internal assessment in `<mental_model>...</mental_model>`: [Opponent Reservation Price]: `price=<number> confidence=<0-100>` [Opponent Strategy]: ⟨your inference about the opponent's tactic⟩ [My Strategy]: ⟨your chosen tactic and reasoning⟩

Output format	`<mental_model>...</mental_model>`
	`[<selected_seller>...</selected_seller>]` (only in multi-seller settings)
	`<message>`
	[Reasoning + exactly one `<contract>...</contract>` JSON block]
	`</message>`

D Robustness Analyses

D.1 Run-to-Run Variance

We re-run each model on the full 160-task suite with 3 random seeds and report mean $\pm$ standard deviation in Table 9; the results in Table 2 correspond to seed 0 and fall within these ranges. Cross-seed variance is small, and the gap between proprietary and open-weight models is far larger than the variance and holds across all seeds.

D.2 Decoding Sensitivity

On a representative subset of 32 tasks (one per market topology $\times$ scenario cell, covering all 8 topologies and 4 scenarios), we sweep the decoding temperature over $\{0.0, 0.7, 1.0\}$ (0.7 is the default) for one proprietary model (Gemini 3.1 Pro Preview) and one open-weight model (Qwen3-VL-32B-Instruct). As shown in Table 10, temperature has no systematic effect on GlobalScore: the variation across settings is far smaller than the between-model gap, which holds at all three temperatures.

Table 8: Prompt design for the Seller agent in **AgenticPay**. The seller mirrors the buyer structurally but receives its own private utility weights and reservation price, and in multi-buyer settings must route to a buyer via a `<selected_buyer>` block. All numeric/structural negotiation outputs are emitted as a JSON `<contract>` inside `<message>`.

Section	Seller prompt content
Role	You are `{name}`, a seller trying to maximize profit while being reasonable. `{role_description}` `{personality_section}`
Context inputs	Context Information: `{context}` Current Negotiation State: `{current_state}` Conversation History: `{conversation_history}` Current product information: `{product_info}` Available products: `{available_products}` Environment factors: `{environment_info}` Task instruction: `{task_instruction}`
Private information	Your minimum acceptable price (**confidential**) is $`{min_price}`; never reveal it. Your private utility weights: `{seller_preferences}` Meaning of weights: `{weight_descriptions}`
Contract schema	The negotiable contract follows the JSON schema: `{ "price": <number>, "continuous_terms": <object>, "discrete_terms": <object> }` Field meanings: `{field_descriptions}`; valid continuous bounds: `{continuous_bounds}`; valid discrete options: `{discrete_options}`; additional contract background: `{contrainfo}`. No field has a default value; the `<contract>` JSON in your message is your current complete proposal, with every field filled by you. Each `<message>` must contain exactly one `<contract>...</contract>` JSON block following the schema above.
Multi-buyer routing	(Used only when #buyers > 1.) You MUST choose exactly one buyer for this turn and emit it inside `<selected_buyer>`, an integer from 1 to `{num_buyers}`. Base your choice on the conversation history and current offers.
Deal protocol	A deal is finalized only when seller and the selected buyer output *compatible* contract JSONs. If you accept, repeat the accepted final contract JSON exactly in `<contract>...</contract>`. Never reveal your minimum acceptable price directly.
Mental-model scratchpad	Before replying, output an internal assessment in `<mental_model>...</mental_model>`: [Opponent Reservation Price]: `price=<number> confidence=<0-100>` [Opponent Strategy]: ⟨your inference about the opponent's tactic⟩ [My Strategy]: ⟨your chosen tactic and reasoning⟩

Output format	`<mental_model>...</mental_model>`
	`[<selected_buyer>...</selected_buyer>]` (only in multi-buyer settings)
	`<message>`
	[Reasoning + exactly one `<contract>...</contract>` JSON block]
	`</message>`

Table 9: Run-to-run variance over 3 random seeds on all 160 tasks (mean ± std).

Model	GlobalScore	BuyerScore	SellerScore
Gemini 3.1 Pro Preview	42.1 ± 0.9	45.4 ± 1.1	37.9 ± 1.0
GPT-5.4	35.4 ± 1.2	34.7 ± 1.3	33.0 ± 0.8
Claude Sonnet 4.6	34.9 ± 1.0	30.9 ± 0.7	34.4 ± 1.1
Qwen3-VL-32B-Instruct	24.8 ± 1.7	23.3 ± 1.8	30.6 ± 1.4
InternVL3-38B	23.1 ± 2.0	21.6 ± 2.1	23.5 ± 1.5

D.3 Scaffolding Sensitivity

Our baseline scaffolding comprises the `<mental_model>` opponent-modeling scratchpad, a structured JSON contract interface, and a private per-agent memory (Section 4.3); it covers memory and planning but not external tool use, which the negotiation protocol does not require. To test whether conclusions depend on scaffolding strength, we compare three tiers on the same 32-task subset with Qwen3-VL-32B-Instruct: (i) a simpler variant without the mental model, close to a naive prompt; (ii) the default baseline; and (iii) a stronger opponent-simulation strategy following Liu et al. [38] (fictitious play with best-of-N rollouts). As shown in Table 11, the simpler variant is worse, while the stronger strategy yields only a limited gain ($+1.4$). This is expected: our single-shot tasks

Table 10: GlobalScore under different decoding temperatures on the 32-task subset.

Model	Temp. 0.0	Temp. 0.7 (default)	Temp. 1.0
Gemini 3.1 Pro Preview	42.4	42.6	41.9
Qwen3-VL-32B-Instruct	24.2	24.5	24.1

neutralize cross-episode fictitious play, and inference-time search can only rank candidates as well as the model's own opponent-preference inference allows. Scaffolding thus changes the magnitude of scores but not the qualitative conclusion that the bottleneck lies in multi-dimensional value-creation reasoning.

Table 11: GlobalScore under different scaffolding tiers (Qwen3-VL-32B-Instruct, 32-task subset).

Method	GlobalScore
w/o mental model (simpler)	20.3
Full baseline (default)	24.5
+ opponent simulation [38] (stronger)	25.9

D.4 Standard Welfare Metrics

Beyond GlobalScore, we report three standard welfare aggregations on the full task suite: the Nash product $r_b r_s$, utilitarian social welfare $r_b + r_s$, and egalitarian welfare $\min(r_b, r_s)$, where failed or non-IR outcomes score 0 (Table 12). Under all three, proprietary models outperform open-weight models and Gemini 3.1 Pro Preview ranks first, and every model stays far below the ceiling, leaving substantial joint surplus unrealized (consistent with Figure 5). The only changes occur within near-tied pairs: Claude Sonnet 4.6 shows slightly higher deal quality than GPT-5.4 on the Nash product and social welfare, whereas GPT-5.4 leads on GlobalScore once its lower timeout rate is credited, precisely the efficiency signal a bare Nash product cannot express; Qwen3-VL-32B-Instruct and InternVL3-38B are tied on the Nash product and egalitarian welfare (within 0.002), with Qwen leading on social welfare. Table 13 further shows that, for Gemini 3.1 Pro Preview, the degradation from bilateral to multi-sided markets holds under every aggregation.

Table 12: Standard welfare metrics on all 160 tasks. Ranges are given in parentheses.

Model	GlobalScore (0–100)	Nash $r_b r_s$ (0–0.25)	Social $r_b + r_s$ (0–1)	Egalitarian $\min(r_b, r_s)$ (0–0.5)
Gemini 3.1 Pro Preview	42.3	0.075	0.54	0.148
GPT-5.4	35.6	0.060	0.41	0.119
Claude Sonnet 4.6	34.8	0.064	0.43	0.126
Qwen3-VL-32B-Instruct	25.1	0.043	0.35	0.083
InternVL3-38B	22.9	0.044	0.29	0.085

D.5 Utility Specification Sensitivity

To test sensitivity to the linear utilities in Definition 1, we rescore the same agreed contracts offline (without re-negotiation) under three alternatives: non-price weights scaled by 1.5, non-price weights scaled by 0.5, and concave (square-root) diminishing returns on attribute utilities. As shown in Table 14, the advantage of proprietary over open-weight models and the low absolute welfare (far below the 0.25 ceiling) hold under all four specifications. The fine-grained ranking within the proprietary tier is not fully invariant (e.g., Claude Sonnet 4.6 edges out Gemini 3.1 Pro Preview under concave utilities), since rescoring differentially rewards how each model created its surplus; this does not affect our qualitative conclusions.

Table 13: Bilateral vs. multi-sided market performance of Gemini 3.1 Pro Preview under different welfare aggregations.

Aggregation	**Bilateral**	**Market**
GlobalScore (0–100)	47.3	40.6
Nash product $r_b r_s$ (0–0.25)	0.086	0.071
Social welfare $r_b + r_s$ (0–1)	0.58	0.52
Egalitarian $\min(r_b, r_s)$ (0–0.5)	0.172	0.139

Table 14: Nash product $r_b r_s$ of the same agreed contracts under alternative utility specifications.

Model	**Linear (default)**	**Attr.** $\times 1.5$	**Attr.** $\times 0.5$	**Concave** ($\sqrt{\cdot}$)
Gemini 3.1 Pro Preview	0.075	0.066	0.079	0.068
GPT-5.4	0.060	0.052	0.063	0.058
Claude Sonnet 4.6	0.064	0.057	0.067	0.071
Qwen3-VL-32B-Instruct	0.043	0.037	0.045	0.041
InternVL3-38B	0.044	0.038	0.046	0.042

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state our two main contributions (a unified framework for multimodal agentic negotiation and an instantiated benchmark of 160 multimodal tasks across 4 scenarios and 8 market topologies) and the main empirical findings, all of which are supported by Section 4 and the experimental sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated "Limitations" section (Section 6) discusses three main limitations: scenario coverage (4 representative business domains), reliance on prompted opponent-modeling baselines rather than learned negotiation policies, and the limited scope of our safety probe.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain formal theorems or proofs; it presents a benchmark, framework, and empirical study.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the negotiation protocol (Section 4.1), agent design (Section 4.3), and scoring rule (Algorithm 1); model versions, decoding settings, and additional details are provided in the appendix, and the full code and dataset are released at the public repository linked in the abstract.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The full benchmark code and dataset (160 multimodal tasks across 4 scenarios and 8 market topologies), together with run scripts and instructions for reproducing all main experiments, are released at the public URL in the abstract.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We do not train models; all experiments use frozen pretrained checkpoints. Model versions, decoding hyperparameters, max-round limits, scoring constants (D, W, E, F, γ), and prompt templates are documented in the experimental sections and the "Additional Experimental Details" appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We repeat the full 160-task evaluation with 3 random seeds for every model and report mean $\pm$ 1 standard deviation across seeds in Appendix D.1; the main-text results correspond to seed 0 and fall within these ranges.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Proprietary models (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro Preview) are accessed via their official APIs; open-weight models (Qwen3-VL-32B-Instruct, InternVL3-38B) are run on 4 NVIDIA A800 GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics: it uses synthetic negotiation tasks with sampled utility weights, involves no human subjects, releases code and data publicly, and discusses safety considerations explicitly in the safety probe and Limitations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts include enabling rigorous evaluation of agentic commerce systems and supporting research on welfare-maximizing multi-agent coordination. Potential negative impacts, including contract exploitation by adversarial agents and value-misalignment risks under deployment, are discussed in the safety probe paragraph in the experimental sections, in the Limitations section (Section 6), and in the Broader Impact discussion (Appendix A).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released benchmark consists of synthetic negotiation tasks with sampled utility weights and curated visual-textual product templates; it contains no scraped private data, no pretrained model weights, and no content with elevated misuse risk beyond standard LLM evaluation benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All evaluated models (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro Preview, Qwen3-VL-32B-Instruct, InternVL3-38B) are properly cited with their original references, and we use them in accordance with their respective API terms of service or open-weight licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release **AgenticPay**, which contains 160 multimodal negotiation tasks and the full evaluation framework, together with documentation describing task formats, scenarios, market topologies, scoring rule, and run scripts at the public repository linked in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects; all negotiations are between LLM-based agents on synthetic tasks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in the study, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the central object of study: the paper benchmarks state-of-the-art multimodal LLMs (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro Preview, Qwen3-VL-32B-Instruct, InternVL3-38B) as negotiation agents. All LLM usage in the methodology is documented in Section 4.3 and the experimental sections.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.