A Survey on Mechanism Design Meets Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) increasingly operate in interactive and strategic settings such as negotiation, preference elicitation, evaluation, alignment, and multi-agent coordination. These settings are inherently incentive-driven, yet most existing approaches rely on heuristic designs with limited guarantees against manipulation or misalignment. This survey develops a unified, bidirectional view of the relationship between mechanism design and LLMs. We discuss (i) LLMs for mechanism design, where language models are used as strategic agent proxies, simulators, and natural-language interfaces for economic mechanisms, and (ii) mechanism design for LLMs, where incentive-aware principles are applied to LLM evaluation, alignment, and training. We conclude by identifying challenges in incentive-compatible evaluation, human-LLM interaction protocols, and emerging markets for LLM services.

1 Introduction

LLMs are increasingly deployed in interactive and strategic settings, including negotiation, preference elicitation, evaluation, alignment, and multi-agent coordination. Modern LLMs have evolved from passive text generators into agentic systems capable of planning, multi-step reasoning, negotiation, and interaction in complex environments (OpenAI et al., 2024; Guo et al., 2024b; Wang et al., 2023; Acharya et al., 2025). In these settings, LLMs do not operate in isolation; they respond to feedback, adapt to incentives, and interact with humans (Chen et al., 2025d; Frankel, 2014). Consequently, the behavior of LLM-based systems is shaped not only by model architectures and training data, but also by the rules governing interaction, feedback, and information flow (Pan et al., 2024; Ouyang et al., 2022). Despite this shift toward interactive deployment, most current approaches to LLM-based agents and alignment rely on heuristic design choices, including prompt engineering, reward shaping, and ad hoc evaluation protocols, with few formal guarantees against manipulation, strategic reporting, or incentive misalignment (Naik et al., 2025; Anwar et al., 2024). While these methods can perform well in static or single-agent settings, they often break down in interactive and decision-making pipelines, where LLMs adapt to feedback and exploit weaknesses in the design (Lorè and Heydari, 2024; Hao and Duan, 2025; Buening et al., 2025). A principled framework that explicitly accounts for incentives and strategic interaction is therefore essential.

Mechanism design, a branch of economics and game theory, studies how to design rules of interaction so that self-interested agents generate desirable outcomes (Börgers, 2015). It provides algorithmic tools for handling incentives, information asymmetry, and equilibrium behavior, and has played a central role in the design of auctions (Myerson, 1981), matching markets (Roth and Sotomayor, 1992), resource allocation mechanisms (Hurwicz, 1973), and broader market design applications (Myerson, 2013). While mechanism design has long guided the analysis of strategic human behavior, its relevance to modern language-based AI systems has only recently become apparent and remains fragmented across disciplines.

This convergence makes the connection between mechanism design and LLMs both natural and increasingly necessary. On one hand, LLM alignment and reinforcement learning fundamentally rely on the design of rewards and feedback, which is itself a mechanism design problem (Ji et al., 2025; Munos et al., 2024). When multiple LLMs interact, they form multi-agent systems that may exhibit strategic behavior, coordination failures, or manipulation (Li et al., 2024; Abdelnabi et al., 2024). Moreover, as LLMs participate in data collection, annotation, evaluation, and decision-making, human–LLM ecosystems require interaction protocols that are robust to incentives and strategic reporting (Li et al., 2024; Sun et al.,

2024a). Taken together, these developments suggest that the next generation of AI systems will not be shaped by model architectures alone, but by the mechanisms governing how AI agents interact with humans, with data, and with each other (Jordan, 2025). On the other hand, LLMs themselves are increasingly used for mechanism design. Language models can serve as strategic agent proxies, simulators of human behavior, and natural-language interfaces for economic mechanisms, enabling new approaches to mechanism analysis and design in natural language-mediated environments (Shah et al., 2025a; Li et al., 2023; Wu and Hartline, 2024).

In this survey, we develop a unified view of the emerging connection between mechanism design and LLMs. We organize the literature along a bidirectional perspective, as outlined in Figure 1. First, we review LLMs for mechanism design, where language models act as natural-language interfaces for economic mechanisms. Second, we survey mechanism design for LLMs, where incentive-aware principles inform evaluation, alignment, and training in strategic and multi-agent settings. Across both directions, we emphasize that language fundamentally changes how information and feedback are represented. Our goal is to unify these perspectives and use mechanism design as a lens for understanding next-generation AI systems as economic systems that emerge from interacting agents.

1.1 Comparison with Existing Surveys

There are several surveys that either focus the intersection of game theory and LLMs (Sun et al., 2025) or focus on specific domains such as LLM-based agents in games (Feng et al., 2025) and LLMs for strategic reasoning and intelligent decision making (Huang et al., 2025b; Zhang et al., 2024). Related overviews include reviews on reward design for LLM alignment (Ji et al., 2025), applications of generative AI in economics (Sajjadi Mohammadabadi et al., 2025) and the intersection of game theory and cybersecurity (Zhu, 2025). In contrast, this survey adopts a mechanism-design perspective and proposes a unified taxonomy of the bidirectional relationship between LLMs and incentive-driven systems. We review both how LLMs are used for mechanism design and how mechanism design principles are applied to LLM evaluation, alignment, and training, highlighting how the two areas inform each other.

The remainder of the paper is organized as follows: Section 2 reviews LLMs for mechanism design. Section 3 discusses mechanism design for LLMs. Section 4 concludes the paper by reaffirming our unified mechanism-design view and its implications for the analysis of LLM-centered systems. Finally, we summarize evaluation protocols, benchmarks, and open-source softwares in Appendix A and outline key open challenges and future directions in Appendix B.

2 LLM for Mechanism Design

2.1 LLMs as Strategic Agent Proxies

LLMs provide a low-cost, scalable way to simulate strategic agents (Bae et al., 2024; Patel et al., 2024). Their natural language capabilities enable rich communication behaviors (Horton, 2023; Bubeck et al., 2023). We organize this literature into two testbeds: (i) auctions, which study bidding behavior under explicit mechanisms and structured action spaces, and (ii) interactive games, which study bargaining, negotiation, and coordination among LLM agents.

2.1.1 Auctions

Auctions are central to efficient resource allocation (Vickrey, 1961), but experimental auction research is often costly, slow, and difficult to replicate (Shah et al., 2025a). These challenges are amplified in dynamic environments such as real-time bidding, where many promising designs remain under-tested or are validated only under simplified settings or limited datasets (Ou et al., 2023). State-of-the-art LLMs (e.g., GPT-4) have demonstrated key auction competencies (Chen et al., 2024a) and often display risk-averse behavior consistent with predictions in strategy-proof environments, enabling a valid proxy for human subjects (Shah et al., 2025a). In highly dynamic settings such as real-time bidding, LLM agents can also reason over context (page content, user segment, budget state), propose bids, and provide explanations for observed trade-offs, supporting both evaluation and design iteration (Cai et al., 2025; Yin, 2025). However, recent work demonstrate challenges and limitations in using LLMs as human proxies (Corrigan et al., 2025). For example, LLM-based economic agents can exhibit systematic behavioral biases that deviate from classical predictions (Fish et al., 2025) and are highly sensitive to model priors and reasoning artifacts (Manning et al., 2024), thereby motivating incentive-aware evaluation and alignment of LLM agents before deploying them in economic environments.

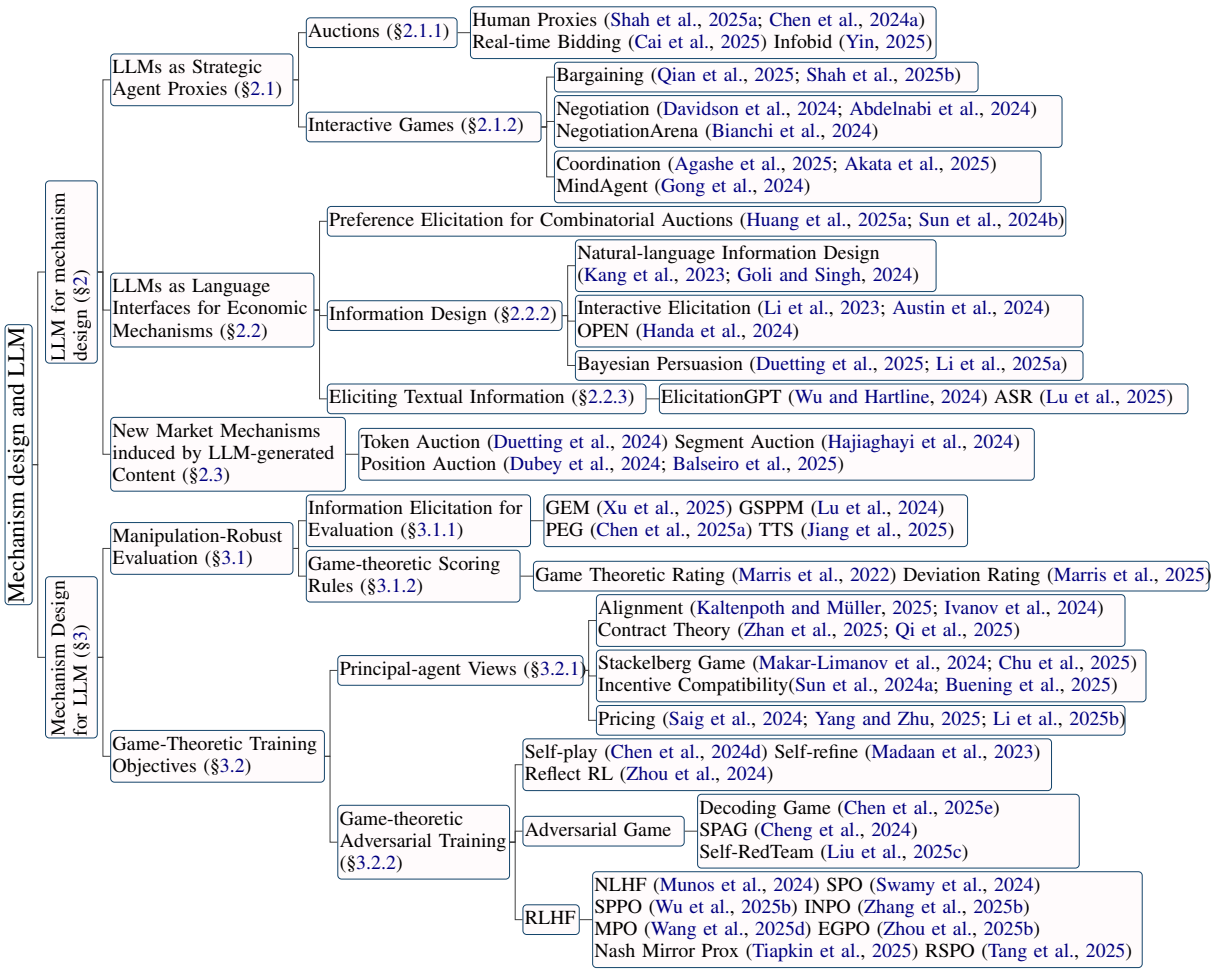

Figure 1: Taxonomy of mechanism design meets LLM.

2.1.2 Interactive Games: Bargaining, Negotiation and Coordination

LLMs can be modeled as economic agents as they can interpret and produce natural language, and they are trained on large corpora that implicitly encode many economic concepts such as incentives, trade-offs, bargaining norms, and market interactions (Horton, 2023; Bubeck et al., 2023). Compared with classical algorithmic agents that typically require an explicit utility function, structured action space, and carefully specified game rules (Roughgarden, 2010), LLM agents can operate directly in language-mediated environments where preferences and constraints are revealed through dialogue, descriptions, and feedback (Liu et al., 2024). Recent work uses LLMs to play bargaining (Qian et al., 2025; Xia et al., 2024), negotiation (Davidson et al., 2024; Mouri Zadeh Khaki et al., 2025), and coordination games (Agashe et al., 2025; Akata et al., 2025; Gong et al., 2024) in which agents must understand game instructions, maintain conversational context, justify decisions, and adapt strategies over multiple turns; these capabilities make LLMs a natural fit for studying strategic interaction when communication itself is part of the game (Zhang et al., 2024). Empirical evaluations further suggest that LLMs can exhibit elements of economic rationality and microeconomic reasoning (Raman et al., 2024, 2025), responding to incentives, budgets, and opportunity costs, and achieving near-optimal behavior in stylized dynamic market settings. As a result, LLM agents are increasingly used as computational proxies for economic research (e.g., simulating populations in bargaining or coordination tasks), as tools for generating and analyzing economic text, and as assistants for downstream economic decision making in interactive systems (Qian et al., 2025; Wu et al., 2025a; Immorlica et al., 2024). However, current LLMs are highly sensitive to prompt phrasing, yielding behavior that is not sufficiently robust (Sreedhar and Chilton, 2024). A key reason is that most models are not purpose-built for economic reasoning. Targeted post-training and data augmentation can strengthen economic reasoning and stability (Wang et al., 2025c; Zhou et al., 2025c). Moreover, when deploying LLMs in empirical studies, researchers should guard against

leakage and bias by isolating model training data from the study sample and by applying bias corrections alongside uncertainty quantification (Ludwig et al., 2025; Wan et al., 2023).

Discussion on LLMs as Strategic Agent Proxies

Key ideas. LLMs provide a low-cost and scalable proxy for human-subject experiments in auctions and for studying bargaining, negotiation, coordination, and other strategic interactions. This makes LLMs attractive as computational proxies for economic research, policy experiments, and decision-support tools, where they can simulate heterogeneous agents and counterfactual scenarios at scale.

Challenges. LLM behavior is highly prompt-sensitive and not purpose-built for economics, raising concerns about robustness, data leakage, and bias. This motivates domain-specialized economic LLMs with targeted post-training and richer economic benchmarks for training and evaluation.

2.2 LLMs as Language Interfaces for Economic Mechanisms

This section treats LLMs as language interfaces for economic mechanisms and groups the literature into four roles: (i) preference elicitation, where dialogue is mapped into structured reports (e.g., bids) for mechanisms such as combinatorial auctions; (ii) language-based information design/persuasion, where textual outputs shape beliefs and strategic behavior; and (iii) text elicitation with incentives, where private information is inherently free-form text and must be scored or aggregated reliably.

2.2.1 Preference Elicitation for Auctions

Combinatorial auctions face a fundamental expressiveness tension: fully expressive bidding allows bids on any subset of items and therefore requires reporting values for exponentially many bundles, namely 2^m bundles for m items, leading to computational intractability (Nisan, 2000). Recent work explores LLMs as natural-language interfaces for preference elicitation, aiming to reduce reporting frictions without sacrificing allocation quality (Soumalias et al., 2025). Rather than asking bidders to enumerate bundle values, LLMs can conduct interactive dialogues to extract constraints, priorities, and trade-offs, and then compile the resulting information into structured bids in a compact language that a mechanism can optimize over (Huang et al., 2025a). In parallel, LLMs can incorporate semantic context into bidder-specific valuation models to denoise imprecise descriptions and improve value estimates from partial reports (Sun et al., 2024b). Together, these approaches lower communication costs while targeting high-welfare outcomes in practical combinatorial allocation settings (Soumalias et al., 2025; Shah et al., 2025a).

2.2.2 Natural-Language Information Design and Bayesian Persuasion

Information design studies how a principal can strategically structure and disclose information to influence the decisions of Bayesian agents in a way that improves the designer's objective (Bergemann and Morris, 2019). A central prerequisite in such problems is how agents value different outcomes; without preference information, it is impossible to construct information structures that meaningfully shape behavior (Bergemann and Morris, 2005; Falk et al., 2018). For example, bidder preferences are pivotal in combinatorial auctions (Huang et al., 2025a); recommender systems infer user preferences to personalize content and support downstream decision-making (Austin et al., 2024; Kang et al., 2023); in modern LLM pipelines, preference data is foundational to reinforcement learning from human feedback (Ouyang et al., 2022; Cai and Li, 2025). Traditional preference elicitation methods rely on report-based mechanisms where agents disclose valuations or types and the system infers optimal outcomes (Chen and Pu, 2004; Kang et al., 2023). Although such elicitation mechanisms are theoretically query-efficient, the required communication can still be cognitively demanding (Huang et al., 2025a). LLMs offer new advantages through natural language understanding and generation, allowing them to capture implicit and context dependent preferences that elude traditional structured approaches (Kang et al., 2023; Goli and Singh, 2024; Ronanki et al., 2023).

To obtain more accurate preference information, LLMs can engage users in interactive multi turn dialogue to explore open questions and elicit task specific preferences (Li et al., 2023). Compared with traditional user written prompts or static labels, this conversational approach better captures the complexity and implicit intentions of human preferences (Kendapadi et al., 2025). For recommender systems, For recommender systems, LLMs

can generate a semantic summary after each interaction to reflect the system's current understanding of the user's preferences, allowing users to verify or correct this summary (Loepp and Ziegler, 2024) and can proactively ask clarifying questions to iteratively refine the preference estimate (Montazeralghaem et al., 2025). These approaches leverage the ability of LLMs to engage in interactive conversations, using in-context learning to accurately infer user preferences. Building on this foundation, another line of work focuses on optimizing the multi turn interaction to explore the preference space and fully learn user preferences. Using Bayesian optimization, the system selects the most informative natural language questions with uncertainty driven decision strategies, enabling more efficient exploration and better preference learning (Austin et al., 2024; Handa et al., 2024).

LLMs not only help elicit and interpret preferences, but also act as language-based information designers that structure and design information to shape beliefs and guide decisions (Lin et al., 2025).In particular, in Bayesian persuasion settings, LLMs can design information that is easier to understand and more effective at influencing beliefs and downstream choices (Duetting et al., 2025). Their linguistic fluency enables a systematic exploration of rich framing strategies that were previously inaccessible in traditional models (Cheng and You, 2025). Building on this perspective, Li et al. (2025a) design practical systems that implement Bayesian persuasion mechanisms directly in natural language, while Wu et al. (2025a) develop methods for generating persuasive marketing contents.

2.2.3 Eliciting Textual Information

LLMs can generate rich text such as comments, reviews, and summaries (Chen et al., 2024b). In many applications, however, the central challenge how to elicit informative and truthful textual reports: there is often no objective ground truth, and the underlying criteria are subjective, so naive incentives reward fluency, verbosity, or strategic hedging rather than genuine information (Schneider et al., 2024; Lu et al., 2025). Classical proper scoring rules provide a principled foundation for truthful information elicitation by making honest belief reporting a dominant strategy, but they rely on numerical forecasts and realized outcomes (Gneiting and Raftery, 2007). Directly applying these ideas to open-ended text is infeasible, motivating mechanisms that operate over semantic representations rather than surface forms. ElicitationGPT (Wu and Hartline, 2024) addresses this challenge by mapping textual reports into a learned semantic space and applying a semantic proper scoring rule under a know-it-or-not model. This design incentivizes agents to accurately express what they know and to explicitly acknowledge uncertainty, rather than optimizing for stylistic plausibility. Aligned Scoring Rules (ASR) (Lu et al., 2025) extend this approach by adapting the semantic scoring mechanism to task-specific alignment objectives and preference distributions, preserving truthful elicitation guarantees while improving robustness and flexibility across domains.

Discussion on LLMs as Language Interfaces

Key ideas. Since LLMs can parse free-form language and run interactive, g dialogues effectively, it is possible to elicit richer and more structured preferences and feed them into preference-learning pipelines. This enables more expressive mechanisms for resource-allocation problems.

Challenges. Current models struggle with preferences expressed over long, multi-turn interactions, making their inferred preferences noisy and challenging (Zhao et al., 2025). A key need is high-quality datasets of real human preferences that capture this complexity, to support reliable preference learning and mechanism design.

2.3 New Market Mechanisms Induced by LLM-Generated Contents

LLM-generated contents create two concrete scarce resources: generation budget during decoding and visibility in ranked outputs (Feizi et al., 2024). These constraints motivate auction designs that price either generation resources or exposure directly. Accordingly, we focus on two new-arisen mechanisms: token auctions (Duetting et al., 2024), which allocate and price tokens or decoding steps, and position auctions (Balseiro et al., 2025), which allocate visibility in LLM-mediated rankings such as answers, summaries, or recommendations.

2.3.1 Token Auctions

The rise of LLMs opens new avenues for auction design because the model can now participate directly in how ads and organic content are com-

posed. At one extreme, a token auction treats each advertiser as an LLM that proposes a next-token distribution; the mechanism aggregates these multiple LLM outputs into a single continuation using bid-dependent weights (Duetting et al., 2024). A related approach aggregates advertisers' LLM-encoded preferences while preserving incentive compatibility (Soumalias et al., 2024). A weaker form of integration keeps the LLM's organic answer fixed and fuses it with an advertisement. Advertisers then bid for this sponsored answer, and the platform selects the winner (Mordo et al., 2024). Moving from whole-answer fusion to finer granularity, segment auctions use retrieval-augmented generation (RAG) to score how relevant each ad is to individual sentences or paragraphs, and then choose one or more ads per segment via a truthful randomized rule that maximizes a logarithmic social-welfare objective (Hajiaghayi et al., 2024).

2.3.2 Position Auctions

Once ads are interwoven with generated content, even the logic of classic position auctions becomes unstable. Traditional positions have fixed prominence, but LLM-generated layouts make exposure stochastic and context-dependent (Varian, 2007; Dubey et al., 2024). The auction can no longer determine display outcomes on its own; it must coordinate with the LLM that renders the final arrangement (Dubey et al., 2024). Moreover, a position's value is no longer common across advertisers: the utility of an ad depends on surrounding text, and attention-based user models can predict clicks when multiple creatives are shown jointly, enabling mechanisms that optimize over full joint layouts (Balseiro et al., 2025).

Discussion on New Market Mechanisms

Key ideas. This section discusses two ways in which LLMs reshape mechanism design. First, token-level auctions redefine the unit of trade, shifting attention from whole items to finer-grained units such as tokens or generation segments. Second, position auctions become increasingly relevant in LLM-mediated settings, where ranked exposure and attention over generated content play a central role.

Challenges. Designing mechanisms for LLM generated content raises challenges beyond classical settings. LLM outputs are context dependent, stochastic, and uncertain in quality, so mechanisms should explicitly incorporate and adapt to this uncertainty. In addition, matching based designs that assign queries to models, prompts, or human oversight remain underexplored and offer a promising direction.

3 Mechanism Design for LLMs

3.1 Manipulation-Robust Evaluation

LLMs have exceptional text-generation ability, such as generating comments, reviews, and summaries (Chen et al., 2024b). Yet, evaluating these LLM outputs remains difficult primarily because ground truth is often unavailable: the tasks are inherently subjective, and obtaining consistent human judgments is costly (Schneider et al., 2024). LLMs-as-judges offer a scalable solution. However, their judgments can deviate from human references, and their evaluations usually lack provable guarantees, leaving them vulnerable to strategic manipulation (Lu et al., 2025).

3.1.1 Information Elicitation for Evaluation

Peer prediction mechanisms offer an alternative elicitation paradigm that guarantee more information is scored higher in expectation (Miller et al., 2005). A key challenge of applying this idea to textual data is how to predict a textual response X_i given another textual response X_j (e.g. X_i and X_j can be two reviews of the same paper). Lu et al. (2024) first propose a solution by using an open-source LLM as a textual predictor, where the probabilities $\Pr(X_j \mid X_i)$ can be estimated by applying the chain rule on the next-token-prediction probabilities of the open-source LLM. This probability can then be used to compute an information score for incentivizing high-effort signal X_i. Xu et al. (2025) then adapt the above idea and design an evaluator that directly estimates the pointwise mutual information between two agents' responses, which additionally estimates $\Pr(X_j)$ and removes the prior information. The pointwise mutual information score then serves as a verification-free and manipulation-resistant benchmark for evaluating LLMs' judgment abilities. The above two methods use an open-box LLM as a textual predictor, while there are other methods that use a strong LLM as an oracle (closed-form adaptation). For example, Chen et al. (2025a) adapt multi-task

peer prediction for answer evaluation in order to incentivize more informative and truthful responses. Such mechanisms can be further integrated into retrieval and generation pipelines to elicit higher-quality evidence and produce more informative textual reports (Jiang et al., 2025). Furthermore, Zhang et al. (2025a) show that peer prediction can also be used to score residual human information in addition to LLM, which can be used to incentivize human effort and detect LLM cheaters. However, their method primarily applies to labeling settings.

3.1.2 Game-theoretic Scoring Rules

Traditional AI evaluation methods typically assume that candidate models can be meaningfully ordered along a single dimension of strength. Representative approaches such as Elo (Elo and Sloan, 1978) and the Bradley–Terry model (Bradley and Terry, 1952) aggregate pairwise outcomes into a single global ranking, forming the foundation of popular crowdsourced LLM benchmarks such as Chatbot Arena (Chiang et al., 2024). However, the single-dimensional assumption is increasingly fragile in modern settings, where AI capabilities are highly multidimensional, interactions are strategic, and benchmarks are dynamic rather than fixed (Marris et al., 2022). To address these challenges, Liu et al. (2025d) reframes open-ended LLM evaluation as a multi-player game and replaces Elo with clone-invariant, equilibrium-based ratings, improving robustness to prompts redundancy. Complementing this perspective, deviation ratings (Marris et al., 2025) formalize redundancy robustness as a property of the scoring rule itself: duplicating or near-duplicating strategies or prompts should not change the resulting ratings, preventing redundancy from distorting evaluation.

Discussion on Mechanism for Evaluation

Key ideas. This section discusses how to adapt peer prediction and scoring rules for data evaluation. Together, these methods offer alternatives to align LLM-based evaluators with human judgments in settings such as grading, review aggregation, and content moderation.

Challenges. A key direction is to integrate these techniques into end-to-end LLM pipelines, covering human annotation, RAG-based retrieval and summarization, and downstream deployment so that truthfulness incentives and robustness guarantees hold throughout the full workflow.

3.2 Game-theoretic Training Objectives

The cornerstone of the success of LLMs is alignment, which enables them to generate text that reflects human preferences for both usability and safety (Köbis et al., 2025). At the same time, user preferences are heterogeneous across tasks and individuals (Yang et al., 2019; Conitzer et al., 2024). This heterogeneity motivates incentive-aware alignment frameworks, which we group into (i) principal–agent, contract-theoretic approaches that design explicit incentives for truthful, high-effort feedback, and (ii) game-theoretic adversarial training approaches that model strategic interactions to improve robustness and empirical performance.

3.2.1 Principal-Agent Views

Contract theory studies how a principal can optimally design rewarding schemes to align an agent's actions with the principal's objectives under information asymmetry (Bolton and Dewatripont, 2004), thus offering a principled toolkit for preference alignment and the incentivization of LLM agents toward desirable outcomes (Kaltenpoth and Müller, 2025). This can be modeled as a multi-turn reinforcement learning, where the principal updates a sequence of contracts to steer behavior, and the agent learns a policy that targets social welfare (Ivanov et al., 2024). Complementing this, Zhan et al. (2025) use online contract design to infer hidden provider choices and to compute bonus schemes that reliably induce high-quality AI-generated content. This framework also applies to multi-agent systems, where smart contracts reward truthful behavior and ensure transparent collaboration (Qi et al., 2025).

Overall, contract-theoretic work treats alignment as explicit incentive design to steer behavior and elicit reliable feedback. In contrast, Stackelberg formulations emphasize strategic response dynamics between the policy and a learned (or adversarial) reward model, casting alignment as a bilevel game (Makar-Limanov et al., 2024). In some variants of the framework, the follower is modeled as an adversarial preference distribution to increase robustness to annotation noise and distribution shift (Chu et al., 2025). In multi-objective alignment, diverse groups or labelers may strategically misreport to steer the outcome and reduce overall performance. A VCG-

style payment mechanism (Sun et al., 2024a) or an approximately strategyproof algorithm (Buening et al., 2025) can both encourage truthful reporting and prevent misreporting that would bias the aggregated objective (Manyika et al., 2025). In LLM systems, these mechanisms are implemented by aggregating user or labeler feedback into the fine-tuning objective (e.g., RLHF/DPO) and assigning rewards or credits based on each agent's marginal contribution or strategyproof aggregation rules, making misreporting unprofitable.

Another challenge is that LLM services are often black boxes priced per token or per call (Velasco et al., 2025). This pricing rewards verbosity and allows providers to quietly lower effort while charging the same, creating a moral hazard. A contract-theoretic remedy is to tie payment to verifiable outcomes, such as accuracy, which has been shown to improve generation quality (Saig et al., 2024). The same idea extends to multi-dimensional menus that price speed, accuracy, reliability, and liability, so users choose what they value and providers are incentivized to deliver it (Yang and Zhu, 2025). In parallel, Li et al. (2025b) design usage-based prices that incorporate users' heterogeneous prompt capabilities and strategic behaviors, thereby maximizing platform payoff while promoting efficient service utilization.

3.2.2 Game-Theoretic Adversarial Training

Game-theoretic modeling formalizes LLMs as agents in strategic interactions and motivates empirical game–based adversarial training for improved learning and empirical performance. LLM self play seeks to build strong agents with minimal human oversight via reflection, critique, and iterative improvement (Fu et al., 2023). One approach is self-reflection at the prompt level that repeatedly refines reasoning (Chen et al., 2024d; Madaan et al., 2023; Zhou et al., 2024). Another approach frames self-play as an adversarial game in which one agent generates challenging inputs while the other attempts to solve them (Chen et al., 2025b), often as attacker and defender pairings (Cheng et al., 2024; Liu et al., 2025c). These formulations can also help improve prompts through strategic interaction by generating optimal prompts under adversarial pressure (Ye et al., 2025; Chen et al., 2025e).

Motivated by this adversarial perspective, recent work frames alignment as a two-player min–max game between policies and proposes algorithms grounded in Nash equilibria (Ouyang et al., 2022; Ge et al., 2024; Rosset et al., 2024). For example, NLHF models alignment as a two-player zero-sum preference game where each policy aims to outperform the other under a learned preference function (Munos et al., 2024). Follow-up studies focus on training stability and robustness (Swamy et al., 2024; Wu et al., 2025b; Zhang et al., 2025b), as well as on providing theoretical guarantees and improved interpretability, for example by designing more robust gradient methods with better convergence properties(Wang et al., 2025d; Tiapkin et al., 2025; Zhou et al., 2025b). Additionally, KL-regularization during policy optimization further stabilizes learning and yields generalization benefits (Ye et al., 2024; Tang et al., 2025).

Discussion on Game-Theoretic Training

Key ideas. The principal–agent framework models LLM alignment as a game between a principal and an agent, enabling algorithms that learn equilibrium policies. Contract theory then provides incentive-compatible designs, strengthening both the theoretical foundations and interpretability of alignment procedures.

Challenges. Most existing work focuses on a single-principal, single-agent setting, while real deployments involve multiple principals and agents with heterogeneous objectives, competition, and externalities. Developing stronger guarantees and scalable algorithms for these multi-agent, multi-principal environments on equilibrium existence/uniqueness, convergence rates, and partial observability remains an important open problem.

4 Conclusions

In this survey, we develop a unified mechanism-design view of the bidirectional relationship between LLMs and incentive-driven systems. LLMs increasingly operate within economic mechanisms, while mechanism design provides principled tools to evaluate, align, and train LLMs in strategic multi-agent settings with heterogeneous, manipulable feedback. Together, we view next-generation AI systems as economic systems shaped by interactions among models, users, platforms, and data providers, making mechanism design a unifying lens for their analysis and design. Open challenges and future directions are deferred to Appendix B.

Limitations

This survey focuses on mechanism design and LLMs, with the goal of understanding how to design incentives, rules, and information structures for interactive, language-mediated settings. As a result, we do not aim to systematically cover work whose primary objective is game-theoretic analysis rather than design, such as equilibrium characterization in general or repeated games, learning dynamics in games, or the use of LLMs as solvers or approximators for abstract game-theoretic models. While such works are closely related and occasionally discussed, they fall outside the main scope of this survey.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation game. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEe Access*.

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8038–8057.

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

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, and 1 others. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.

David Austin, Anton Korikov, Armin Toroghi, and Scott Sanner. 2024. Bayesian optimization with llm-based acquisition functions for natural language preference elicitation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 74–83.

Henry Bae, Aghyad Deeb, Alex Fleury, and Kehang Zhu. 2024. Complexitynet: Increasing llm inference efficiency by learning task complexity. *Preprint*, arXiv:2312.11511.

Santiago Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Jieming Mao, Aranyak Mehta, Vahab Mirrokni, Renato Paes Leme, Di Wang, and Song Zuo. 2025. Position auctions in ai-generated content. *Preprint*, arXiv:2506.03309.

Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, Will Epperson, Tyler Payne, Jake M. Hofman, Brendan Lucier, Chinmay Singh, Markus Mobius, Akshay Nambi, Archana Yadav, Kevin Gao, David M. Rothschild, Aleksandrs Slivkins, Daniel G. Goldstein, Hussein Mozannar, Nicole Immorlica, and 5 others. 2025. Magentic marketplace: An open-source environment for studying agentic markets. *Preprint*, arXiv:2510.25779.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. 2025. The economics of large language models: Token allocation, fine-tuning, and optimal pricing.

Dirk Bergemann and Stephen Morris. 2005. Robust mechanism design. *Econometrica*, pages 1771–1813.

Dirk Bergemann and Stephen Morris. 2019. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. In *International Conference on Machine Learning*, pages 3935–3951. PMLR.

Patrick Bolton and Mathias Dewatripont. 2004. *Contract theory*. MIT press.

Tilman Börgers. 2015. *An introduction to the theory of mechanism design*. Oxford university press.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. 2025. Strategyproof reinforcement learning from human feedback. *Preprint*, arXiv:2503.09561.

Leng Cai, Junxuan He, Yikai Li, Junjie Liang, Yuanping Lin, Ziming Quan, Yawen Zeng, and Jin Xu. 2025. Rtbagent: A llm-based agent system for real-time bidding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 104–113.

Zhongze Cai and Xiaocheng Li. 2025. Users as annotators: Llm preference learning from comparison mode. *Preprint*, arXiv:2510.13830.

Raul Castro Fernandez. 2023. Data-sharing markets: model, protocol, and algorithms to incentivize the formation of data-sharing consortia. *Proceedings of the ACM on Management of Data*, 1(2):1–25.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219.

Baiting Chen, Tong Zhu, Jiale Han, Lexin Li, Gang Li, and Xiaowu Dai. 2025a. Incentivizing truthful language models via peer elicitation games. *Preprint*, arXiv:2505.13636.

Jiangjie Chen, Siyu Yuan, Rong Ye, Bodhisattwa Prasad Majumder, and Kyle Richardson. 2024a. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *Preprint*, arXiv:2310.05746.

Jiaqi Chen, Bang Zhang, Ruotian Ma, Peisong Wang, Xiaodan Liang, Zhaopeng Tu, Xiaolong Li, and Kwan-Yee K. Wong. 2025b. Spc: Evolving self-play critic via adversarial games for llm reasoning. *Preprint*, arXiv:2504.19162.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024b. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer.

Junhao Chen, Jingbo Sun, Xiang Li, Haidong Xin, Yuhao Xue, Yibin Xu, and Hao Zhao. 2025c. Llmspark: A benchmark for evaluating large language models in strategic gaming contexts. *Preprint*, arXiv:2509.16610.

Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024c. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077.

Li Chen and Pearl Pu. 2004. Survey of preference elicitation methods. *Prefer Elicit*, 8.

Lin Chen, Yunke Zhang, Jie Feng, Haoye Chai, Honglin Zhang, Bingbing Fan, Yibo Ma, Shiyuan Zhang, Nian Li, Tianhui Liu, Nicholas Sukiennik, Keyu Zhao, Yu Li, Ziyi Liu, Fengli Xu, and Yong Li. 2025d. Ai agent behavioral science. *Preprint*, arXiv:2506.06366.

Sijin Chen, Omar Hagrass, and Jason Matthew Klusowski. 2025e. Decoding game: On minimax optimality of heuristic text generation strategies. In *The Thirteenth International Conference on Learning Representations*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024d. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, pages 6621–6642. PMLR.

Pengyu Cheng, Yong Dai, Tianhao Hu, Han Xu, Zhisong Zhang, Lei Han, Nan Du, and Xiaolong Li. 2024. Self-playing adversarial language game enhances llm reasoning. *Advances in Neural Information Processing Systems*, 37:126515–126543.

Zirui Cheng and Jiaxuan You. 2025. Towards strategic persuasion with language models. *Preprint*, arXiv:2509.22989.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Xu Chu, Zhixin Zhang, Tianyu Jia, and Yujie Jin. 2025. Stackelberg game preference optimization for data-efficient alignment of language models. *Preprint*, arXiv:2502.18099.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, and 1 others. 2024. Social choice for ai alignment: Dealing with diverse human feedback. *CoRR*.

Jay R Corrigan, Carola Grebitus, and Matthew C Rousu. 2025. On the usefulness of using current llms for experimental auction valuation. *Available at SSRN*.

James C Cox, Vernon L Smith, and James M Walker. 1988. Theory and individual behavior of first-price auctions. *Journal of Risk and uncertainty*, 1(1):61–99.

Tim Ruben Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. 2024. Evaluating language model agency through negotiations. In *The Twelfth International Conference on Learning Representations*.

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations. *Advances in Neural Information Processing Systems*, 37:28219–28253.

Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. 2024. Auctions with llm summaries. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 713–722.

Paul Duetting, Safwan Hossain, Tao Lin, Renato Paes Leme, Sai Srivatsa Ravindranath, Haifeng Xu, and Song Zuo. 2025. Information design with large language models. *Preprint*, arXiv:2509.25565.

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

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)*.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, and 1 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. Global evidence on economic preferences. *The quarterly journal of economics*, 133(4):1645–1692.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. 2024. Online advertisements with llms: Opportunities and challenges. *Preprint*, arXiv:2311.07601.

Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. 2025. A survey on large language model-based social agents in game-theoretic scenarios. *Preprint*, arXiv:2412.03920.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. 2025. Algorithmic collusion by large language models. *Preprint*, arXiv:2404.00806.

Alexander Frankel. 2014. Aligned delegation. *American Economic Review*, 104(1):66–83.

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

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. 2024. Axioms for ai alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Ali Goli and Amandeep Singh. 2024. Can llms capture human preferences? *Preprint*, arXiv:2305.02531.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Yusuke Noda, Zane Durante, Zilong Zheng, Demetri Terzopoulos, Li Fei-Fei, Jianfeng Gao, and Hoi Vo. 2024. Mindagent: Emergent gaming interaction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3154–3183.

Shangmin Guo, Haochuan Wang, Haoran Bu, Yi Ren, Dianbo Sui, Yu-Ming Shang, and Siting Estee Lu. 2024a. Economics arena for large language models. In *Language Gamification-NeurIPS 2024 Workshop*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

MohammadTaghi Hajiaghayi, Sébastien Lahaie, Keivan Rezaei, and Suho Shin. 2024. Ad auctions for llms via retrieval augmented generation. *Advances in Neural Information Processing Systems*, 37:18445–18480.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. 2024. Bayesian preference elicitation with language models. *Preprint*, arXiv:2403.05534.

Shugang Hao and Lingjie Duan. 2025. Online learning from strategic human feedback in llm fine-tuning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Hadi Hosseini and Samarth Khanna. 2025. Distributive fairness in large language models: Evaluating alignment with human values. *Preprint*, arXiv:2502.00313.

Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P. Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. 2025. lmgame-bench: How good are llms at playing games? *Preprint*, arXiv:2505.15146.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. Game-theoretic llm: Agent workflow for negotiation games. *Preprint*, arXiv:2411.05990.

David Huang, Francisco Marmolejo-Cossío, Edwin Lock, and David Parkes. 2025a. Accelerated preference elicitation with llm-based proxies. *Preprint*, arXiv:2501.14625.

Jincai Huang, Yongjun Xu, Qi Wang, Qi Cheems Wang, Xingxing Liang, Fei Wang, Zhao Zhang, Wei Wei, Boxuan Zhang, Libo Huang, and 1 others. 2025b. Foundation models and intelligent decision-making: Progress, challenges, and perspectives. *The Innovation*.

Leonid Hurwicz. 1973. The design of mechanisms for resource allocation. *The American Economic Review*, 63(2):1–30.

Mateusz Idziejczak, Vasyl Korzavatykh, Mateusz Stawicki, Andrii Chmutov, Marcin Korcz, Iwo Błądek, and Dariusz Brzezinski. 2025. Among them: A game-based framework for assessing persuasion capabilities of llms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 183–195. Springer.

Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. 2024. Generative ai as economic agents. *ACM SIGecom Exchanges*, 22(1):93–109.

Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C. Parkes. 2024. Principal-agent reinforcement learning: Orchestrating ai agents with contracts. *Preprint*, arXiv:2407.18074.

Miaomiao Ji, Yanqiu Wu, Zhibin Wu, Shoujin Wang, Jian Yang, Mark Dras, and Usman Naseem. 2025. A survey on progress in llm alignment from the perspective of reward design. *Preprint*, arXiv:2505.02666.

Yanchen Jiang, Zhe Feng, and Aranyak Mehta. 2025. Incentive-aligned multi-source llm summaries. *Preprint*, arXiv:2509.25184.

Michael I. Jordan. 2025. A collectivist, economic perspective on ai. *Preprint*, arXiv:2507.06268.

Sascha Kaltenpoth and Oliver Müller. 2025. Getting in contract with large language models – an agency theory perspective on large language model alignment. *Preprint*, arXiv:2509.07642.

Inwon Kang, Sikai Ruan, Tyler Ho, Jui-Chien Lin, Farhad Mohsin, Oshani Seneviratne, and Lirong Xia. 2023. Llm-augmented preference learning from natural language. *Preprint*, arXiv:2310.08523.

Shirish Karande, V Santhosh, and Yash Bhatia. 2024. Persuasion games with large language models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 576–582.

Aum Kendapadi, Kerem Zaman, Rakesh R Menon, and Shashank Srivastava. 2025. Interact: Enabling interactive, question-driven learning in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8992–9024.

Nils Köbis, Zoe Rahwan, Raluca Rilla, Bramantyo Ibrahim Supriyatno, Clara Bersch, Tamer Ajaj, Jean-François Bonnefon, and Iyad Rahwan. 2025. Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, pages 1–9.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *Preprint*, arXiv:2310.11589.

Wenhao Li, Yue Lin, Xiangfeng Wang, Bo Jin, Hongyuan Zha, and Baoxiang Wang. 2025a. Verbalized bayesian persuasion. *Preprint*, arXiv:2502.01587.

Xiang Li, Bing Luo, Jianwei Huang, and Yuan Luo. 2025b. Strategic prompt pricing for aigc services: A user-centric approach. *Preprint*, arXiv:2503.18168.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Yunhao Liang, Yuan Qu, Jingyuan Yang, Shaochong Lin, and Zuo-Jun Max Shen. 2025. Everyone contributes! incentivizing strategic cooperation in multi-llm systems via sequential public goods games. *Preprint*, arXiv:2508.02076.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Yue Lin, Shuhui Zhu, William A Cunningham, Wenhao Li, Pascal Poupart, Hongyuan Zha, and Baoxiang Wang. 2025. Information bargaining: Bilateral commitment in bayesian persuasion. *Preprint*, arXiv:2506.05876.

Hongtao Liu, Zhicheng Du, Zihe Wang, and Weiran Shen. 2025a. Chbench: A cognitive hierarchy benchmark for evaluating strategic reasoning capability of llms. *Preprint*, arXiv:2508.11944.

Jiayuan Liu, Mingyu Guo, and Vincent Conitzer. 2025b. An interpretable automated mechanism design framework with large language models. *Preprint*, arXiv:2502.12203.

Mickel Liu, Liwei Jiang, Yancheng Liang, Simon Shaolei Du, Yejin Choi, Tim Althoff, and Natasha Jaques. 2025c. Chasing moving targets with online self-play reinforcement learning for safer language models. *Preprint*, arXiv:2506.07468.

Siqi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. 2025d. Re-evaluating open-ended evaluation of large language models. In *The Thirteenth International Conference on Learning Representations*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, and 1 others. 2024. Agentbench: Evaluating llms as agents. In *The Twelfth International Conference on Learning Representations*.

Benedikt Loepp and Jürgen Ziegler. 2024. Exploring the potential of generative ai for augmenting choice-based preference elicitation in recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 114–119.

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

Yuxuan Lu, Yifan Wu, Jason Hartline, and Michael J. Curry. 2025. Aligned textual scoring rules. *Preprint*, arXiv:2507.06221.

Yuxuan Lu, Shengwei Xu, Yichi Zhang, Yuqing Kong, and Grant Schoenebeck. 2024. Eliciting informative text evaluations with large language models. *Preprint*, arXiv:2405.15077.

Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2025. Large language models: An applied econometric framework. Technical report, National Bureau of Economic Research.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Jacob Makar-Limanov, Arjun Prakash, Denizalp Goktas, Amy Greenwald, and Nora Ayanian. 2024. STA-RLHF: Stackelberg aligned reinforcement learning with human feedback. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Julian Manyika, Michael J Wooldridge, and Jiarui Gan. 2025. Mechanism design for alignment via human feedback. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2024. Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents. *Preprint*, arXiv:2311.03220.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. Alympics: Llm agents meet game theory. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 2845–2866, Abu Dhabi, UAE. Association for Computational Linguistics.

Luke Marris, Marc Lanctot, Ian Gemp, Shayegan Omidshafiei, Stephen McAleer, Jerome Connor, Karl Tuyls, and Thore Graepel. 2022. Game theoretic rating in n-player general-sum games with equilibria. *Preprint*, arXiv:2210.02205.

Luke Marris, Siqi Liu, Ian Gemp, Georgios Piliouras, and Marc Lanctot. 2025. Deviation ratings: A general, clone-invariant rating method. *Preprint*, arXiv:2502.11645.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.

Ali Montazeralghaem, Guy Tennenholtz, Craig Boutilier, and Ofer Meshi. 2025. Asking clarifying questions for preference elicitation with large language models. *Preprint*, arXiv:2510.12015.

Tommy Mordo, Tomer Kordonsky, Haya Nachimovsky, Moshe Tennenholtz, and Oren Kurland. 2025. Csp: A simulator for multi-agent ranking competitions. *Preprint*, arXiv:2502.11197.

Tommy Mordo, Moshe Tennenholtz, and Oren Kurland. 2024. Sponsored question answering. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–173.

Ahmad Mouri Zadeh Khaki, Ahyoung Choi, and Laleh Seyyed-Kalantari. 2025. Simulating social behavior of llm-based autonomous negotiator agents in a game-theoretical framework using multi-agent systems. *International Journal of Human–Computer Interaction*, pages 1–10.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. 2023. Welfare diplomacy: Benchmarking language model cooperation. In *Socially Responsible Language Modelling Research*.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, and 1 others. 2024. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*.

Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Roger B Myerson. 2013. *Game theory*. Harvard university press.

Akshat Naik, Patrick Quinn, Guillermo Bosch, Emma Gouné, Francisco Javier Campos Zabala, Jason Ross Brown, and Edward James Young. 2025. Agent-misalignment: Measuring the propensity for misaligned behaviour in llm-based agents. *Preprint*, arXiv:2506.04018.

Noam Nisan. 2000. Bidding and allocation in combinatorial auctions. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pages 1–12.

Aidan O'Gara. 2023. Hoodwinked: Deception and cooperation in a text-based game for language models. *Preprint*, arXiv:2308.01404.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maayan Orner, Oleg Maksimov, Akiva Kleinerman, Charles Ortiz, and Sarit Kraus. 2025. Explaining decisions of agents in mixed-motive games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 22, pages 23267–23275.

Weitong Ou, Bo Chen, Xinyi Dai, Weinan Zhang, Weiwen Liu, Ruiming Tang, and Yong Yu. 2023. A survey on bid optimization in real-time bidding display advertising. *ACM Transactions on Knowledge Discovery from Data*, 18(3):1–31.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024. Feedback loops with language models drive in-context reward hacking. In *International Conference on Machine Learning*, pages 39154–39200. PMLR.

Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini. 2024. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE.

Minfeng Qi, Tianqing Zhu, Lefeng Zhang, Ningran Li, and Wanlei Zhou. 2025. Towards transparent and incentive-compatible collaboration in decentralized llm multi-agent systems: A blockchain-driven approach. *Preprint*, arXiv:2509.16736.

Crystal Qian, Kehang Zhu, John Horton, Benjamin S Manning, Vivian Tsai, James Wexler, and Nithum Thain. 2025. Understanding economic tradeoffs between human and ai agents in bargaining games.

Narun Raman, Taylor Lundy, Thiago Amin, Jesse Perla, and Kevin Leyton-Brown. 2025. Steer-me: Assessing the microeconomic reasoning of large language models. *Preprint*, arXiv:2502.13119.

Narun Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. 2024. Steer: assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 42026–42047.

Krishna Ronanki, Christian Berger, and Jennifer Horkoff. 2023. Investigating chatgpt's potential to assist in requirements elicitation processes. In *2023 49th Euromicro conference on software engineering and advanced applications (SEAA)*, pages 354–361. IEEE.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *Preprint*, arXiv:2404.03715.

Alvin E Roth and Marilda Sotomayor. 1992. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541.

Tim Roughgarden. 2010. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86.

Eden Saig, Ohad Einav, and Inbal Talgam-Cohen. 2024. Incentivizing quality text generation via statistical contracts. *Advances in Neural Information Processing Systems*, 37:51196–51222.

Seyed Mahmoud Sajjadi Mohammadabadi, Burak Cem Kara, Can Eyupoglu, Can Uzay, Mehmet Serkan Tosun, and Oktay Karakuş. 2025. A survey of large language models: evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, 14(18).

Olivier Schipper, Yudi Zhang, Yali Du, Mykola Pechenizkiy, and Meng Fang. 2025. Pillagerbench: Benchmarking llm-based agents in competitive minecraft team environments. In *2025 IEEE Conference on Games (CoG)*, pages 1–15. IEEE.

Johannes Schneider, Bernd Schenk, and Christina Niklaus. 2024. Towards llm-based autograding for short textual answers. *Preprint*, arXiv:2309.11508.

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

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

Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. 2025. Can llms replace economic choice prediction labs? the case of language-based persuasion games. In *Will Synthetic Data Finally Solve the Data Access Problem?*

Ermis Soumalias, Michael Curry, and Sven Seuken. 2024. Truthful aggregation of llms\\with an application to online advertising. In *Agentic Markets Workshop at ICML 2024*.

Ermis Soumalias, Yanchen Jiang, Kehang Zhu, Michael Curry, Sven Seuken, and David C. Parkes. 2025. Llm-powered preference elicitation in combinatorial assignment. *Preprint*, arXiv:2502.10308.

Karthik Sreedhar and Lydia Chilton. 2024. Simulating human strategic behavior: Comparing single and multi-agent llms. *Preprint*, arXiv:2402.08189.

Haoran Sun, Yurong Chen, Siwei Wang, Wei Chen, and Xiaotie Deng. 2024a. Mechanism design for llm fine-tuning with multiple reward models. *Preprint*, arXiv:2405.16276.

Haoran Sun, Yusen Wu, Peng Wang, Wei Chen, Yukun Cheng, Xiaotie Deng, and Xu Chu. 2025. Game theory meets large language models: A systematic survey with taxonomy and new frontiers. *Preprint*, arXiv:2502.09053.

Jie Sun, Tianyu Zhang, Houcheng Jiang, Kexin Huang, Chi Luo, Junkang Wu, Jiancan Wu, An Zhang, and Xiang Wang. 2024b. Large language models empower personalized valuation in auction. *Preprint*, arXiv:2410.15817.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pages 47345–47377.

Onat Arda Tamer and Abdurrahman Gümüş. 2025. Collabpersona: A framework for collaborative decision analysis in persona driven llm-based multi-agent systems. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. 2025. Game-theoretic regularized self-play alignment of large language models. In *Scaling Self-Improving Foundation Models without Human Supervision*.

Daniil Tiapkin, Daniele Calandriello, Denis Belomestny, Eric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, and Pierre Menard. 2025. Accelerating nash learning from human feedback via mirror prox. *Preprint*, arXiv:2505.19731.

Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. 2024. Evaluating large language models with grid-based game competitions: An extensible llm benchmark and leaderboard. *Preprint*, arXiv:2407.07796.

Hal R Varian. 2007. Position auctions. *international Journal of industrial Organization*, 25(6):1163–1178.

Ander Artola Velasco, Stratis Tsirtsis, Nastaran Okati, and Manuel Gomez-Rodriguez. 2025. Is your llm overcharging you? tokenization, transparency, and incentives. *Preprint*, arXiv:2505.21627.

William Vickrey. 1961. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2025a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

Haochuan Wang, Xiachong Feng, Lei Li, Yu Guo, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. 2025b. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *Preprint*, arXiv:2410.10479.

Mengxin Wang, Dennis J. Zhang, and Heng Zhang. 2025c. Large language models for market research: A data-augmentation approach. *Preprint*, arXiv:2412.19363.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. 2025d. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, Yitao Liang, and Team CraftJarvis. 2023. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34153–34189.

Jibang Wu, Chenghao Yang, Yi Wu, Simon Mahns, Chaoqi Wang, Hao Zhu, Fei Fang, and Haifeng Xu. 2025a. Ai realtor: Towards grounded persuasive language generation for automated copywriting. *Preprint*, arXiv:2502.16810.

Yifan Wu and Jason Hartline. 2024. Elicitationgpt: Text elicitation mechanisms via language models. *Preprint*, arXiv:2406.09363.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2025b. Self-play preference optimization for language model alignment. In *The Thirteenth International Conference on Learning Representations*.

Yue Wu, Xuan Tang, Tom Mitchell, and Yuanzhi Li. 2024a. Smartplay: A benchmark for llms as intelligent agents. In *The Twelfth International Conference on Learning Representations*.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024b. Smartplay: A benchmark for llms as intelligent agents. *Preprint*, arXiv:2310.01557.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. 2024. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3579–3602.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. 2025. Benchmarking llms' judgments with no gold standard. *Preprint*, arXiv:2411.07127.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024. Exploring large language models for communication games: An empirical study on werewolf. *Preprint*, arXiv:2309.04658.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. 2019. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *Advances in neural information processing systems*, 32.

Ya-Ting Yang and Quanyan Zhu. 2025. Pact: A contract-theoretic framework for pricing agentic ai services powered by large language models. *Preprint*, arXiv:2505.21286.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. 2024. Human vs. generative ai in content creation competition: Symbiosis or conflict? *Preprint*, arXiv:2402.15467.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. 2024. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807.

Ziyu Ye, Rishabh Agarwal, Tianqi Liu, Rishabh Joshi, Sarmishta Velury, Quoc V Le, Qijun Tan, and Yuan Liu. 2025. Reward-guided prompt evolving in reinforcement learning for llms. In *Forty-second International Conference on Machine Learning*.

Qiyue Yin, Pei Xu, Qiaozhe Li, Shengda Liu, Shengqi Shen, Tong Wang, Yihong Han, Xiaonan Zhao, Likun Yang, Shiyue Cao, Shiyu Qiu, Yuxuan Liu, Shizhao Yu, Lei Cui, Chengxin Yan, Jie Sun, Xiangquan Tang, and Kaiqi Huang. 2025. Wgsr-bench: Wargame-based game-theoretic strategic reasoning benchmark for large language models. *Preprint*, arXiv:2506.10264.

Yue Yin. 2025. Infobid: A simulation framework for studying information disclosure in auctions with large language model-based agents. *Preprint*, arXiv:2503.22726.

Zijun Zhan, Yaxian Dong, Daniel Mawunyo Doe, Yuqing Hu, Shuai Li, Shaohua Cao, and Zhu Han. 2025. Learning to incentivize: Llm-empowered contract for aigc offloading in teleoperation. *Preprint*, arXiv:2508.03464.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. In *First Conference on Language Modeling*.

Yichi Zhang, Jinlong Pang, Zhaowei Zhu, and Yang Liu. 2025a. Evaluating llm-corrupted crowdsourcing data without ground truth. *Preprint*, arXiv:2506.06991.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. 2025b. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. In *The Thirteenth International Conference on Learning Representations*.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do llms recognize your preferences? evaluating personalized preference following in llms. *Preprint*, arXiv:2502.09597.

Han Zhou, Xingchen Wan, Ruoxi Sun, Hamid Palangi, Shariq Iqbal, Ivan Vulić, Anna Korhonen, and Sercan Ö. Arık. 2025a. Multi-agent design: Optimizing agents with better prompts and topologies. *Preprint*, arXiv:2502.02533.

Runlong Zhou, Simon Du, and Beibin Li. 2024. Reflect-rl: Two-player online rl fine-tuning for lms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 995–1015.

Runlong Zhou, Maryam Fazel, and Simon S. Du. 2025b. Extragradient preference optimization (egpo): Beyond last-iterate convergence for nash learning from human feedback. *Preprint*, arXiv:2503.08942.

Yufa Zhou, Shaobo Wang, Xingyu Dong, Xiangqi Jin, Yifang Chen, Yue Min, Kexin Yang, Xingzhang Ren, Dayiheng Liu, and Linfeng Zhang. 2025c. Reasoning like an economist: Post-training on economic problems induces strategic generalization in llms. *Preprint*, arXiv:2506.00577.

Quanyan Zhu. 2025. Game theory meets llm and agentic ai: Reimagining cybersecurity for the age of intelligent threats. *Preprint*, arXiv:2507.10621.

A Benchmarks and Softwares

Using LLMs as tools for mechanism design requires understanding how well they engage in strategic reasoning. Current empirical evaluations come from a diverse set of benchmarks that emphasize different sources of strategic complexity, ranging from opponent modeling to social interaction to incentive-driven behavior. To clarify what is known and identify what remains untested, we review existing evaluations of LLM strategic behavior and organize them along three dimensions of strategic complexity: (1) individual strategic reasoning in simple games, (2) social strategic interaction in multi-agent settings involving cooperation, coordination, or deception, and (3) incentive-driven strategic behavior in economic environments such as bargaining, persuasion, and market reasoning. Table 1 provides a summary of representative benchmark results across three areas, and Table 2 summarizes representative software implementations and simulation platforms.

Given the rise of agentic AI, a growing line of benchmarks evaluates LLMs' strategic reasoning in simple games (Akata et al., 2025). Several works provide relatively comprehensive coverage over diverse games. For example, GTBench (Duan et al., 2024) tests LLMs on board games, card games, and a range of other classical games; Wu et al. (2024a) study performance on six games, including rock–paper–scissors and the MinEscort game; and Wang et al. (2025b) introduce a systematic benchmark covering 144 two-player games. Other efforts focus on specific game types. Hu et al. (2025) evaluate LLMs in video-game environments, Topsakal et al. (2024) consider grid-based games, and ClemBench (Chalamalasetti et al., 2023) focuses on dialogue games that test whether LLMs can reliably follow game instructions (Abdelnabi et al., 2024). A complementary line of work targets social deduction games, including Hoodwinked (O'Gara, 2023), Werewolf (Xu et al., 2024), and Avalon (Light et al., 2023).

The second line of benchmarks extends simple-game scenarios to multi-agent settings, where collaboration, coordination, competition, and deception can emerge (Liang et al., 2025; Tamer and Gümüş, 2025; Mordo et al., 2025). An important strand focuses on negotiation behavior, such as GameBench (Hua et al., 2024) and Negotiation-Arena (Bianchi et al., 2024). Other benchmarks emphasize cooperation, including Welfare Diplomacy (Mao et al., 2024; Mukobi et al., 2023; , FAIR), which builds on Diplomacy and requires alliance formation and joint planning, and Minecraft-based environments (Wang et al., 2025a; Gong et al., 2024; Fan et al., 2022; Schipper et al., 2025), which test LLMs' ability to coordinate with other agents on long-horizon, open-ended tasks. Complementing these environment-specific benchmarks, LLMArena (Chen et al., 2024c) and LLMsPark (Chen et al., 2025c) provide multi-agent, game-theoretic evaluation frameworks across several dynamic environments, targeting abilities such as opponent modeling, communication, team collaboration, and strategic depth (Yin et al., 2025).

Lastly, some empirical evaluations focus on economic behavior, revealing elements of economic rationality (Raman et al., 2024) , microeconomic reasoning (Raman et al., 2025) and bounded rationality (Liu et al., 2025a). There are also evaluations based on a variety of economic scenarios, including auction games (Chen et al., 2024a), which study how LLM agents bid and respond to auction rules; bargaining games (Qian et al., 2025), which model multi-round negotiation over surplus division; persuasion games (Idziejczak et al., 2025; Karande et al., 2024; Shapira et al., 2025), where an expert agent attempts to influence a decision-maker through messages; and marketplace environments such as Magentic-Marketplace (Bansal et al., 2025), which study two-sided agentic markets in which assistant agents represent consumers and service agents represent competing businesses, allowing the analysis of welfare, bias, and search frictions. Finally, Guo et al. (2024a) propose an economic arena based on competitive games and empirically analyze how LLM agents' play converges (or fails to converge) to Nash equilibria. Additionally, researchers also train domain-specific LLMs for economics and finance, using curated corpora and post-training to enhance economic reasoning and better align agent behavior with formal models (Zhou et al., 2025c).

B Open Challenges and Future Directions

Capability Mismatch in LLM-based Mechanism Design Despite their strong generative and reasoning capabilities, LLMs are not purpose-built for mechanism design tasks such as auctions, incentive alignment, or equilibrium reasoning, and may therefore miss key economic constraints without explicit structure or supervision (Zhou et al., 2025c).

Strategic Dimension	Benchmark	Environment / Game Type
Individual Strategic Reasoning	GTBench (Duan et al., 2024)	Game-theoretic games
	SmartPlay (Wu et al., 2024b)	Strategic, planning, and embodied games
	TMGBench (Wang et al., 2025b)	144 two-player games
	ClemBench (Chalamalasetti et al., 2023)	Dialogue-based instruction-following games
	LMGAME-Bench (Hu et al., 2025)	Video-game environments
	LLM-Deliberation (Abdelnabi et al., 2024)	Scorable multi-issue negotiation games
Social Strategic Interaction	GameBench (Hu et al., 2025)	Multi-agent negotiation games
	CuisineWorld (Gong et al., 2024)	Multi-agent collaboration games
	PillagerBench (Schipper et al., 2025)	Competitive multi-agent games
	LLMARENA (Chen et al., 2024c)	Multi-agent dynamic games
	LLMsPark (Chen et al., 2025c)	Game-theoretic multi-agent games
	WGSR-Bench (Yin et al., 2025)	Wargame-based strategic games
Incentive-Driven Economic Behavior	STEER (Raman et al., 2024)	Economic decision tasks
	STEER-ME (Raman et al., 2025)	Microeconomic reasoning tasks
	CHBench (Liu et al., 2025a)	Auction games
	Among Them (Idziejczak et al., 2025)	Social deception games

Table 1: Representative benchmarks for evaluating strategic reasoning in LLMs, organized by strategic dimension.

Moreover, mechanism design is inherently multi-agent: outcomes depend on induced incentives and strategic interaction networks, motivating structure-aware and game-theoretic modeling of LLM-based systems (Zhou et al., 2025a).

Human–AI Collaboration and Incentive Alignment As LLM agents move from tools to decision-making partners, human-AI systems must account for heterogeneous objectives, costs, and strategic behavior (Frankel, 2014). This motivates incentive-aware interaction protocols that elicit truthful information, allocate tasks efficiently, and mitigate manipulation or misalignment in real deployments (Wu and Hartline, 2024; Buening et al., 2025).

LLM for Explaining Mechanisms LLMs open new opportunities for explaining and interacting with mechanisms (Orner et al., 2025). Recent work uses LLMs to produce natural-language rationales for outcomes and to frame automated mechanism design as code generation of interpretable implementations (Liu et al., 2025b; Hosseini and Khanna, 2025). A key direction is to build faithful explanation frameworks for voting, matching, and auctions by grounding natural-language interfaces.

Market Design for LLMs. As LLMs become stand-alone products, platforms must design markets to allocate limited inference capacity and price access (Feizi et al., 2024). Emerging mechanisms in online advertising include position auctions and token auctions. Related questions arise in data markets (Castro Fernandez, 2023), recommender systems (Yao et al., 2024), and optimal pricing of AI services (Bergemann et al., 2025).

C Experiments

We evaluate the ability of using LLMs as proxies for human participants in sealed bid auction experiments. Following Shah et al. (2025a), we simulate auctions with three homogeneous LLM agents bidding against one another. Each simulation proceeds through six stages: (1) planning, where agents interpret the auction rules and objectives; (2) value realization, in which private valuations are drawn or assigned; (3) bidding, where agents submit bids based on their strategies; (4) outcome determination, including allocation and payments; (5) reflection, where agents receive feedback and update their reasoning; and (6) repetition, allowing behavior to evolve over multiple auction rounds. We evaluate three cutting-edges LLMs: GPT5.1, GPT4omini and Gemini2.5 (OpenAI et al., 2024; Comanici et al., 2025).

C.1 Setting

We study four sealed-bid auction formats: First-Price (FPSB), Second-Price (SPSB), Third-Price (TPSB), and All-Pay in an independent private values (IPV) environment. Each auction features three bidders. Bidder i's private valuation is drawn independently from a uniform distribution,

$$v_i \sim \mathrm{Unif}\{0, \ldots, 99\}.$$

After observing v_i, bidder i submits a sealed bid $b_i = \beta_i(v_i)$, where $\beta = (\beta_1, \beta_2, \beta_3)$ denotes the bidding strategy profile mapping values to bids.

In the FPSB auction, the highest bidder wins the object and pays her own bid, while all other bidders

Software / Framework	System Type	Primary Capability
Hoodwinked (O'Gara, 2023)	Social deduction environment	Deception and cooperation analysis in text-based social games
Werewolf (Xu et al., 2024)	Social deduction environment	Strategic reasoning and deception in multi-agent settings
Avalon (Light et al., 2023)	Social deduction environment	Trust and deception analysis in team-based games
Alympics (Mao et al., 2025)	Game-theoretic simulation framework	Strategic interaction and robustness analysis
NegotiationArena (Bianchi et al., 2024)	Multi-agent negotiation framework	Communication and negotiation strategy analysis
VOYAGER (Wang et al., 2025a)	Embodied agent framework	Lifelong skill learning and compositional generalization
MindAgent (Gong et al., 2024)	Multi-agent coordination framework	Planning and collaboration infrastructure
MINEDOJO (Fan et al., 2022)	Embodied agent framework	Open-ended task learning in Minecraft environments
TactiCrafter (Schipper et al., 2025)	Multi-agent system	Tactical coordination and adaptive learning
AucArena (Chen et al., 2024a)	Auction simulation environment	Strategic bidding and adaptive decision-making analysis
Persuasive Games with LLMs (Karande et al., 2024)	Multi-agent persuasion framework	Personalized persuasion and resistance-aware dialogue
Magentic Marketplace (Bansal et al., 2025)	Agentic marketplace simulation	Market dynamics and strategic behavior analysis
Economics Arena (Guo et al., 2024a)	Competitive game simulation	Strategic reasoning and rationality analysis

Table 2: Representative software frameworks and simulation environments used to study strategic, social, and economic behavior of LLM agents.

pay zero. The payment rule is

$$t_i(b) = \mathbf{1}\{i \text{ wins}\} \cdot b_i.$$

In the SPSB auction, the highest bidder wins and pays the second-highest bid:

$$t_i(b) = \mathbf{1}\{i \text{ wins}\} \cdot b_{(2)},$$

where $b_{(2)}$ denotes the second-order statistic (the second-highest bid).

In the TPSB auction, the highest bidder wins and pays the third-highest bid:

$$t_i(b) = \mathbf{1}\{i \text{ wins}\} \cdot b_{(3)},$$

where $b_{(3)}$ denotes the third-order statistic (the third-highest bid).

Finally, in the all-pay auction, all bidders pay their bids regardless of the outcome, and the highest bidder wins the object:

$$t_i(b) = b_i.$$

Bids are restricted to integer-dollar amounts (i.e., \$1 increments), and ties are broken uniformly at random. For each of the auction formats, we perform 3 experiments with 10 rounds and 3 LLM bidders competing over a prize.

C.2 Baselines

The risk-neutral equilibrium behavior in these four auction formats is well understood. Table 3 summarizes the equilibrium bidding functions specialized to our setting with $n = 3$ bidders and i.i.d. values $v_i \sim \text{Unif}\{0, \ldots, 99\}$. A single asterisk $(*)$ indicates a Bayes–Nash equilibrium, whereas a double asterisk $(**)$ denotes a dominant-strategy equilibrium. In terms of incentive compatibility, only the second-price auction is (dominant-strategy) incentive compatible: truthful bidding $b_i(v_i) = v_i$ is optimal regardless of others' bids.

C.3 Results

We compare each LLM's bidding behavior against the risk-neutral equilibrium and the incentive-compatible benchmark in Figures 2, 3, and 4. For FPSB, most bids exceed the Bayes-Nash equilibrium, consistent with empirical findings in (Shah et al., 2025a) and (Cox et al., 1988). Relative to GPT-5.1, GPT-4o mini shows more structured departures from the equilibrium curve, including non-monotonicity and locally inflated bids at lower valuation levels, plausibly reflecting its more limited capacity. For SPSB, both GPT-5.1 and GPT-4o mini tend to underbid, whereas Gemini 2.5 is closer to the equilibrium benchmark. All three models underbid substantially in TPSB, while the all-pay setting exhibits pronounced overbidding, reflecting its greater strategic complexity. We summarize the deviations between LLM bidding behavior and theoretical benchmarks in Table 4. Together, these results underscore the need to design LLM agents with domain-specific structure and supervision tailored to mechanism design.

Auction Type	Risk-Neutral Equilibrium Strategy
First-Price Sealed-Bid*	$b(v_i) = \frac{2}{3} v_i$
Second-Price Sealed-Bid**	$b(v_i) = v_i$
Third-Price Sealed-Bid*	$b(v_i) = 2v_i$
All-Pay Auction*	$b(v_i) = \frac{2}{3 \cdot 99^2} v_i^3$

Table 3: Risk-neutral equilibria in IPV auctions with three bidders and values $v_i \sim \mathrm{Unif}\{0, \ldots, 99\}$. A single asterisk $(*)$ denotes a Bayes–Nash equilibrium, while a double asterisk $(**)$ denotes a dominant-strategy equilibrium.

	Distance to Bayes-Nash Equilibrium		
Auction	GPT5.1	GPT4o-mini	Gemini2.5flash
First-price	$8.47_{\pm 7.64}$	$17.84_{\pm 15.17}$	$13.57_{\pm 10.32}$
Second-price	$11.37_{\pm 6.35}$	$11.98_{\pm 14.67}$	$6.98_{\pm 8.18}$
Third-price	$61.52_{\pm 27.90}$	$53.89_{\pm 30.28}$	$48.33_{\pm 25.58}$
All-pay	$15.80_{\pm 13.02}$	$31.14_{\pm 15.34}$	$24.15_{\pm 15.26}$
	Distance to Incentive-Compatible Behavior		
Auction	GPT5.1	GPT4o-mini	Gemini2.5flash
First-price	$17.00_{\pm 8.91}$	$19.90_{\pm 17.13}$	$8.42_{\pm 8.75}$
Second-price	$11.37_{\pm 6.35}$	$11.98_{\pm 14.67}$	$6.98_{\pm 8.18}$
Third-price	$11.82_{\pm 7.92}$	$15.72_{\pm 19.82}$	$10.06_{\pm 14.68}$
All-pay	$19.33_{\pm 10.41}$	$8.14_{\pm 9.09}$	$11.91_{\pm 10.75}$

Table 4: Distance between LLM bidding behavior and theoretical benchmarks.

D Use of AI Assistants

AI assistants (e.g., ChatGPT) were used to support writing and editing, including improving clarity, organization, and language. All technical content, interpretations, and conclusions were produced and verified by the authors.

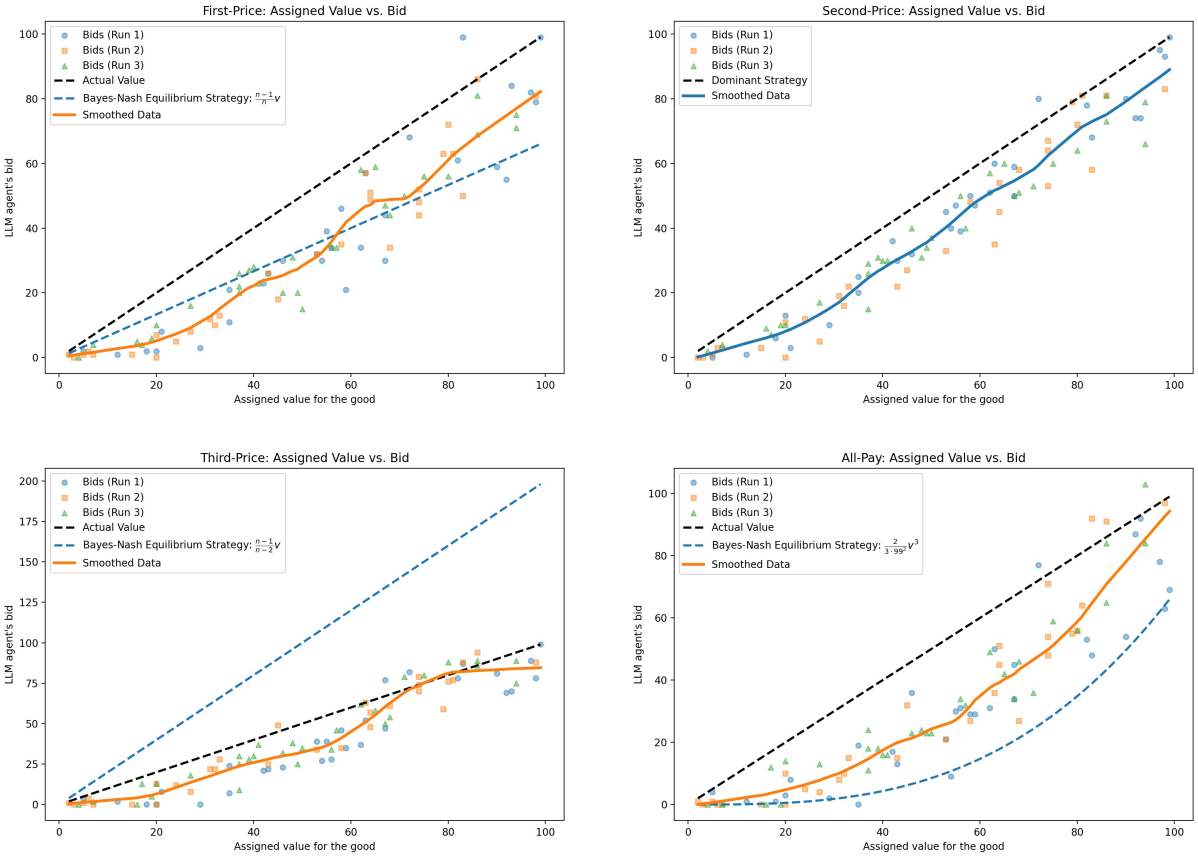

Figure 2: **GPT-5.1 bidding behavior across four auction formats under the IPV setting.** Bidders' private values are independently drawn from the uniform distribution on $\{0, \ldots, 99\}$, and GPT-5.1 bids are plotted against values (blue circles, orange squares, and green triangles). The black dotted line indicates truthful bidding $b(v) = v$ (the dominant strategy in second-price auctions). The blue dashed line shows the Bayes–Nash equilibrium bidding strategy reported in Table 3. The solid orange line is the LOESS-smoothed bidding curve.

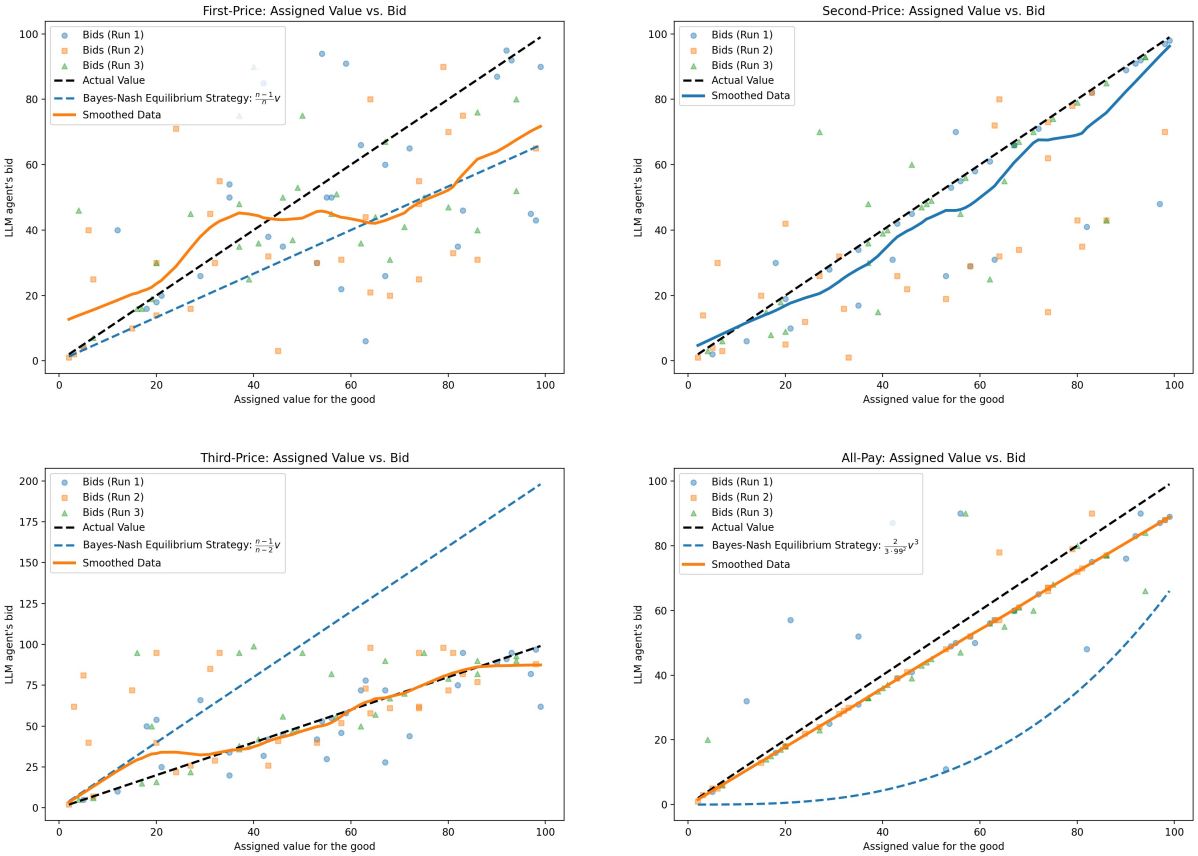

Figure 3: **GPT-4omini bidding behavior across four auction formats under the IPV setting.** Bidders' private values are independently drawn from the uniform distribution on $\{0, \ldots, 99\}$, and GPT-4omini bids are plotted against values (blue circles, orange squares, and green triangles). The black dotted line indicates truthful bidding $b(v) = v$ (the dominant strategy in second-price auctions). The blue dashed line shows the Bayes–Nash equilibrium bidding strategy reported in Table 3. The solid orange line is the LOESS-smoothed bidding curve.

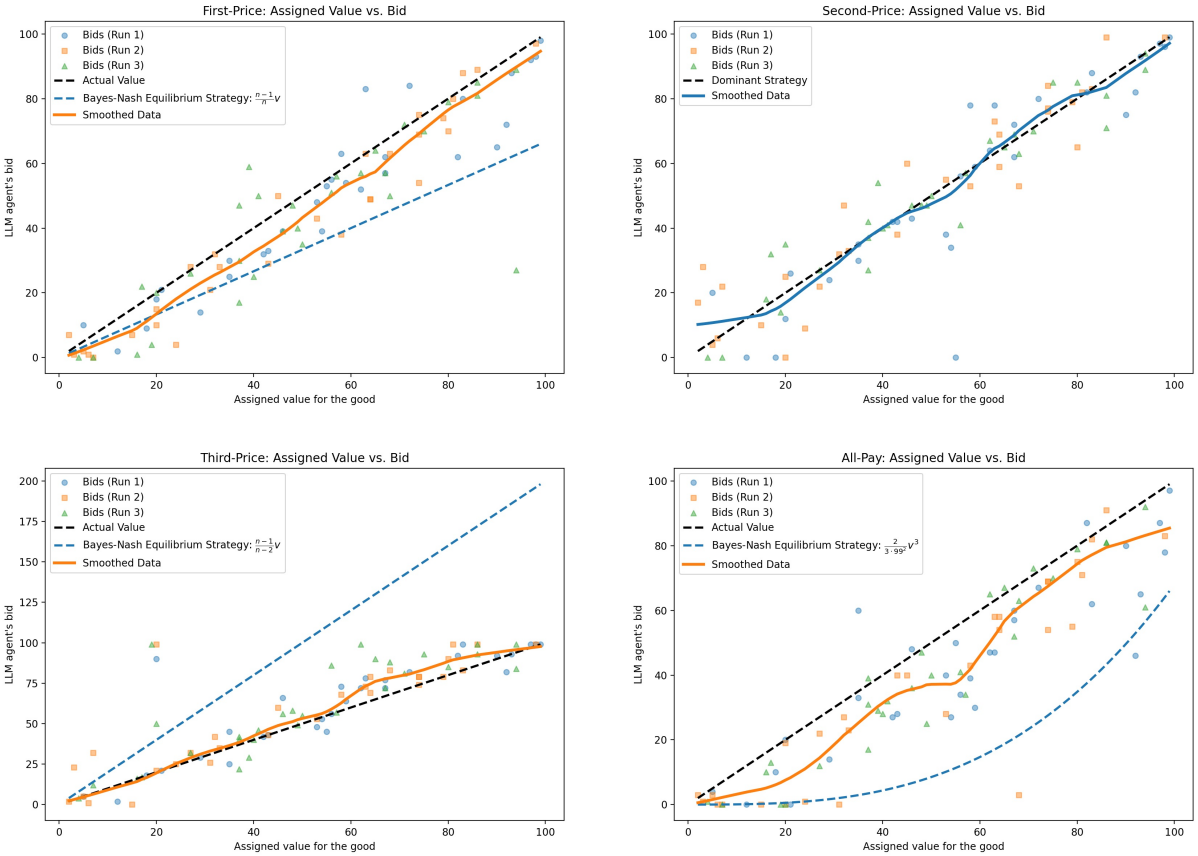

Figure 4: **Gemini2.5 bidding behavior across four auction formats under the IPV setting.** Bidders' private values are independently drawn from the uniform distribution on $\{0, \ldots, 99\}$, and Gemini2.5 bids are plotted against values (blue circles, orange squares, and green triangles). The black dotted line indicates truthful bidding $b(v) = v$ (the dominant strategy in second-price auctions). The blue dashed line shows the Bayes–Nash equilibrium bidding strategy reported in Table 3. The solid orange line is the LOESS-smoothed bidding curve.