Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Survey Track

Game Theory Meets Large Language Models: A Systematic Survey

Haoran Sun[1] , **Yusen Wu**[1] , **Yukun Cheng**[2*] and **Xu Chu**[1,3*]

[1]Center on Frontiers of Computing Studies, School of Computer Science, Peking University
[2]School of Business, Jiangnan University
[3]Key Laboratory of High Confidence Software Technologies, Ministry of Education
{sunhaoran0301, 2401112067}@stu.pku.edu.cn, ykcheng@amss.ac.cn, chu_xu@pku.edu.cn

Abstract

Game theory establishes a fundamental framework for analyzing strategic interactions among rational decision-makers. The rapid advancement of large language models (LLMs) has sparked extensive research exploring the intersection of these two fields. Specifically, game-theoretic methods are being applied to evaluate and enhance LLM capabilities, while LLMs themselves are reshaping classic game models. This paper presents a comprehensive survey of the intersection of these fields, exploring a bidirectional relationship from three perspectives: (1) Establishing standardized game-based benchmarks for evaluating LLM behavior; (2) Leveraging game-theoretic methods to improve LLM performance through algorithmic innovations; (3) Characterizing the societal impacts of LLMs through game modeling. Among these three aspects, we also highlight how the equilibrium analysis for traditional game models is impacted by LLMs' advanced language understanding, which in turn extends the study of game theory. Finally, we identify key challenges and future research directions, assessing their feasibility based on the current state of the field. By bridging theoretical rigor with emerging AI capabilities, this survey aims to foster interdisciplinary collaboration and drive progress in this evolving research area.

1 Introduction

Game theory provides a mathematical framework for analyzing strategic interactions among rational agents and has evolved significantly since its seminal work [Von Neumann and Morgenstern, 2007]. Over the decades, it has established robust methodological foundations, including equilibrium analysis [Nash Jr, 1950] and mechanism design [Vickrey, 1961], which serve as essential analytical tools across disciplines such as economics and computer science.

With the rapid advancement of large language models (LLMs), researchers have increasingly explored the intersection of game theory and LLMs. A growing body of work investigates how game-theoretic principles can be used to evaluate and enhance LLMs and how LLMs can contribute to game theory. Specifically, existing studies apply game theory to develop theoretical frameworks for assessing LLMs' strategic reasoning. This approach optimizes their training methodologies and analyzes their societal implications. Key research directions include:

- Standardized Game-Based Evaluation: Researchers are constructing benchmark environments, such as matrix games [Akata *et al.*, 2023] and auctions [Chen *et al.*, 2023], to evaluate the strategic reasoning capabilities of LLMs systematically.
- Game-Theoretic Algorithmic Innovation: Concepts from cooperative and non-cooperative game theory, such as Shapley Value [Enouen *et al.*, 2024] and max-min equilibria [Munos *et al.*, 2024], are inspiring novel approaches to model interpretability and training optimization.
- Societal Impact Modeling: As LLMs transform information ecosystems, new theoretical frameworks are emerging to predict the societal consequences of human-AI interactions [Yao *et al.*, 2024], particularly in domains like advertising markets [Duetting *et al.*, 2024] and content creation [Fish *et al.*, 2024a].

Beyond these applications, recent research suggests that LLMs can also contribute to game theory by facilitating equilibrium analysis in complex text-based scenarios and extending classical models to more realistic settings.

Existing surveys [Zhang *et al.*, 2024b; Feng *et al.*, 2024; Hu *et al.*, 2024] primarily examine how game theory can be used to build evaluation environments and assess LLMs' strategic performance. For instance, [Zhang *et al.*, 2024b] classifies studies based on the game scenarios used to test LLM capabilities and methods for improving their reasoning. Meanwhile, [Feng *et al.*, 2024] and [Hu *et al.*, 2024] categorize the core competencies required for LLM-based agents in games, such as perception, memory, role-playing, and reasoning. While these surveys provide valuable insights, they primarily focus on the role of game theory in standardized evaluation frameworks, overlooking its broader potential for advancing LLM development. Moreover, they adopt a unidirectional perspective, treating game theory as a tool for assessing LLMs rather than exploring the reciprocal influence between the two fields.

*Corresponding Authors.

This paper aims to *bridge this gap by examining the bidirectional relationship between game theory and LLMs*. *We categorize the work in the intersection between game theory and LLMs into three key perspectives*, as illustrated in Figure 1. To the best of our knowledge, this is *the first comprehensive analysis of the bidirectional relationship between these two fields*.

In Section 2, we review studies that apply game models to evaluate LLMs' decision-making capabilities. Experiments conducted on both canonical matrix games and complex strategic scenarios reveal LLMs' strengths and limitations as game players. Beyond behavioral evaluations, we identify key strategies for enhancing LLMs' strategic decision-making, such as recursive reasoning frameworks and the integration of LLMs with auxiliary modules. Moreover, LLMs demonstrate the ability to formalize real-world scenarios into structured game models, extending game-theoretic analysis to broader and more complex contexts.

Section 3 examines how game-theoretic principles address key challenges in LLM development. We categorize existing research into two main areas: (1) using game theory to understand LLMs' text generation and training dynamics and (2) leveraging game-theoretic mechanisms to enhance LLM training algorithms. The first category explores how the Shapley Value improves model interpretability and how social choice theory facilitates preference alignment in human-AI interactions. The second category introduces studies that incorporate game-theoretic objectives to tackle challenges like heterogeneity and complexity in human preferences. The objectives include minimizing regret in multi-agent interactions and evaluation metrics, including Nash equilibrium convergence,

In Section 4, we discuss how game theory is used to predict and characterize the societal impact of LLMs. The human-AI interaction game model predicts the impact of competition between humans and AI. Emerging game models highlight the growing business and economic implications where LLMs are treated as products or platforms. Meanwhile, classic game theory models are also generalized to more realistic settings with LLMs' unique capabilities, such as natural language manipulation.

Finally, we identify key research challenges and future directions across these dimensions. By systematically analyzing the intersection of game theory and LLMs, we highlight their mutual influence and how they drive progress in both fields, contributing to the advancement of this interdisciplinary domain.

2 Game Theory for LLM Evaluation

In this section, we explore the integration of LLMs within the context of game theory, focusing on their evaluation as game players. Behavioral evaluations reveal that LLMs face challenges in identifying optimal actions in classic matrix games, yet they can demonstrate human-like strategies in more complex game scenarios. Several studies have explored methods to enhance LLMs' performance as game players, and two significant points can be identified from that: recursive thinking and auxiliary modules. Finally, we also discuss the role of LLMs in games beyond their function as players.

2.1 Evaluation of LLMs' Behavioral Performance

Struggles of LLM in Matrix Games. Matrix games are a fundamental concept in game theory. In a matrix game, two players make simultaneous decisions, and the outcomes can be represented by a finite payoff matrix. Recent studies have investigated how LLMs respond to these games by converting them into natural language prompts. Despite significant advancements, their results show that LLMs such as GPT-4 struggle to consistently select the optimal strategy in 2×2 matrix games [Akata *et al.*, 2023; Herr *et al.*, 2024; Lorè and Heydari, 2024; Wang *et al.*, 2024].

For instance, [Akata *et al.*, 2023] states that LLMs frequently fail to choose the optimal action in coordination games, like the Battle of the Sexes. Similarly, [Lorè and Heydari, 2024] examines how contextual framing and utility matrices influence LLM decision-making, revealing significant biases. Furthermore, [Herr *et al.*, 2024] explores the impact of game descriptions, player positioning, and payoffs on LLM performance, highlighting consistent behavioral biases. In more dynamic settings, [Fan *et al.*, 2024] observes that LLMs struggle to predict optimal strategies in ring-network games. Additionally, the TMGBench benchmark, which evaluates LLMs across 144 distinct 2×2 matrix games, further confirms these limitations [Wang *et al.*, 2024].

The matrix game is a cornerstone of game theory and is the foundation for more intricate strategic challenges. Studying LLMs' behaviors in such games provides valuable insights into their broader limitations in complex reasoning tasks.

Human-like Strategies of LLM in Realistic Game Scenarios. Beyond classic matrix games, numerous studies have analyzed LLM performance in more realistic game settings. While these games feature greater contextual complexity, they are not necessarily more challenging for LLMs. This is because strategic inference based on textual content can sometimes replace explicit computation.

Research indicates that LLMs can exhibit strategic behavior in communication-based games. In deception and negotiation games, including Werewolf [Xu *et al.*, 2023; Du and Zhang, 2024] and Avalon [Wang *et al.*, 2023; Lan *et al.*, 2024], LLMs demonstrate behaviors such as deception, trust-building, and leadership—traits typically associated with human strategic thinking. These findings suggest that LLMs can function as sophisticated communication agents in games.

LLMs have also demonstrated strategic reasoning in economically significant scenarios such as bargaining and pricing games. For instance, [Deng *et al.*, 2024] finds that LLMs possess advanced negotiation skills, while [Fish *et al.*, 2024b] shows that LLM-based pricing agents can autonomously engage in collusion to set prices above competitive levels. In auction contexts, [Guo *et al.*, 2024] finds that LLMs can formulate rational bidding strategies based on historical data, often converging toward a Nash equilibrium. Similarly, [Chen *et al.*, 2023] introduces AucArena, a platform where LLMs effectively manage budgets and optimize auction strategies.

Comprehensive Benchmarks for Game Performance. Several benchmarks that cover a diverse range of game scenarios have been developed to make a systematic assessment of LLM. Notable examples include GTBench [Duan *et al.*,

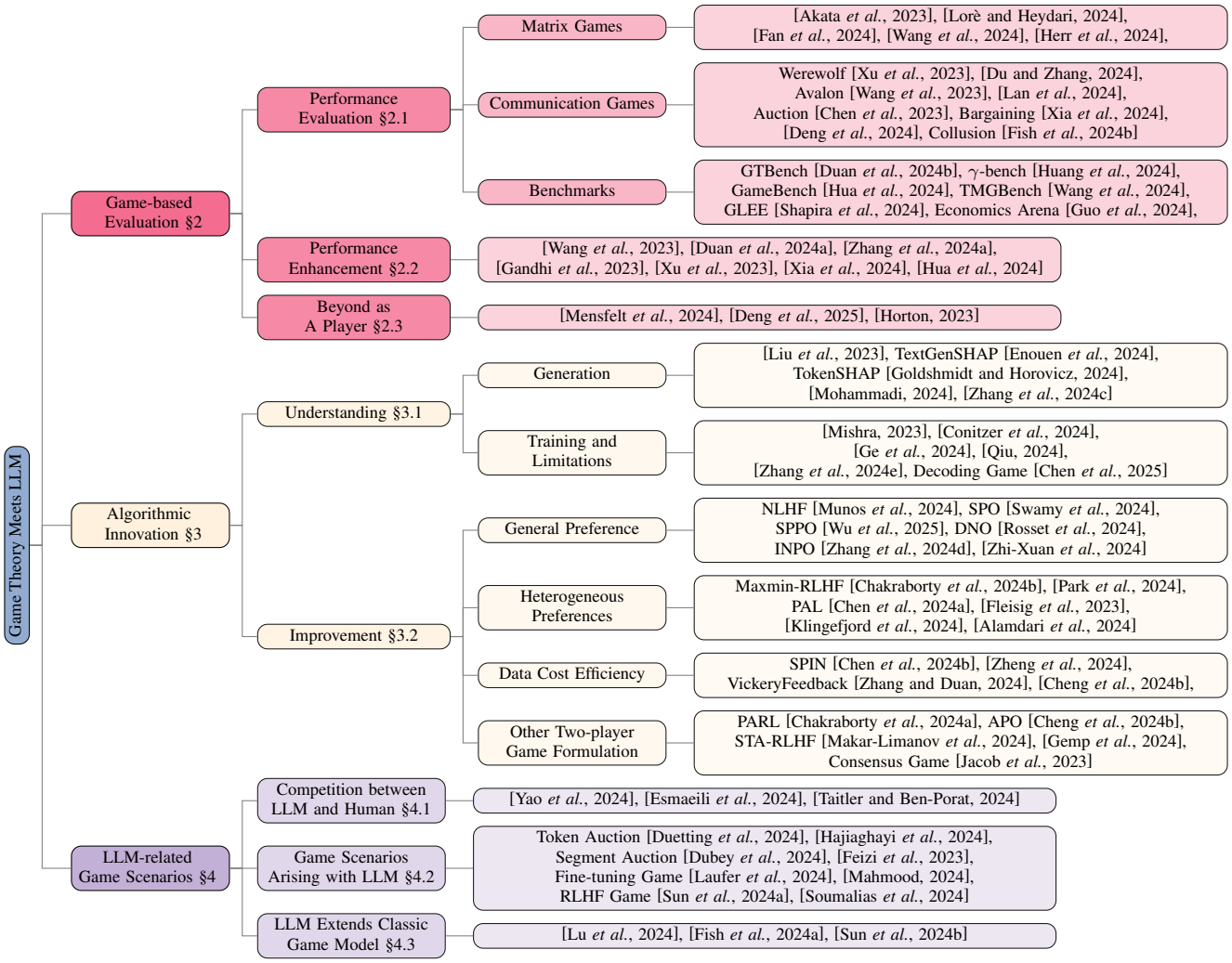

Figure 1: A taxonomy of the intersection between game theory and Large Language Models.

2024b], γ-bench [Huang *et al.*, 2024], GameBench [Hua *et al.*, 2024], GLEE [Shapira *et al.*, 2024], and TMGBench [Wang *et al.*, 2024]. These benchmarks serve to identify both the strengths and weaknesses of LLMs in different strategic environments, offering valuable insights into their potential improvements and real-world applications.

2.2 Enhancing LLMs' Game Performance

Building on the evaluation of LLMs' performance in various games, numerous studies have explored methods to enhance their strategic reasoning and decision-making. These works address key challenges LLMs face in gameplay and propose general frameworks for improving their capabilities. Below, we outline two significant approaches.

Recursively Thinking. In games requiring long-term or multi-level reasoning, LLMs often struggle to retain and build upon previous information, leading to suboptimal decision-making. To mitigate this, researchers have developed techniques that encourage LLMs to engage in recursive thinking, enabling them to leverage past information better when formulating strategies.

For instance, [Wang *et al.*, 2023] introduces the Recursive Contemplation (ReCon) framework. The framework prompts LLMs to engage in first-order and second-order perspective-taking during Avalon gameplay. This helps them avoid common pitfalls, such as deception. Similarly, [Duan *et al.*, 2024a] proposes a method where LLMs predict future moves in multi-turn games, improving their ability to anticipate opponents' strategies. Additionally, [Zhang *et al.*, 2024a] advances LLM reasoning through k-level rationality, which enhances multi-level thinking and significantly increases their win rates in competitive settings. These findings suggest that recursive reasoning can substantially improve LLMs' strategic capabilities.

Auxiliary Modules. As language models, LLMs often struggle in games that require complex mathematical calculations or historical data retrieval. Several studies have proposed integrating auxiliary modules that assist LLMs during gameplay to overcome these limitations.

For example, [Gandhi *et al.*, 2023] introduces a "prompt

compiler", which systematically guides LLMs in evaluating actions and forming beliefs, enabling them to generalize to new scenarios with minimal in-context learning. In the game Werewolf, [Xu *et al.*, 2023] integrates an additional BERT model to encode both historical and current game states, allowing LLMs to make more informed decisions.

In bargaining games, the OG-Narrator framework [Xia *et al.*, 2024] generates external offers, allowing the LLM to focus solely on negotiation language. More recently, [Hua *et al.*, 2024] developed a structured workflow to assist LLMs in solving game-theoretic problems, including computing Nash equilibria and optimizing strategies in complex negotiation tasks.

These auxiliary modules significantly improve LLMs' performance in various game settings, demonstrating that integrating additional computational tools can enhance their strategic decision-making.

2.3 Beyond as a Game Player

While much of the discussion has been centered on utilizing game-based scenarios to evaluate LLMs, work has also shown that LLM capability in games can, in turn, contribute to game theory. This section explores alternative roles for LLMs within game-theoretic contexts, broadening their applications.

In section 2.1, we noted that LLMs often struggle to compute optimal strategies in classical matrix games. However, some studies take an alternative approach by leveraging LLMs' natural language understanding instead of their ability to compute equilibria directly. For example, [Mensfelt *et al.*, 2024] utilizes LLMs to formalize game descriptions into a Game Description Language (GDL), allowing external solvers to process them. Similarly, [Deng *et al.*, 2025] introduces a two-stage framework for translating extensive-form games: first, the LLM identifies the information set, and then it constructs the complete game tree using in-context learning. These studies suggest that LLMs can act as intermediaries in converting natural language into formal game structures, a capability beyond traditional models.

Additionally, [Horton, 2023] explores the use of LLMs as substitutes for human participants in behavioral economic experiments. Their findings indicate that LLMs can replicate classic behavioral economics results, providing a scalable and cost-effective alternative for conducting social science research. This underscores the potential of LLMs as valuable tools in experimental economics and social science studies, facilitating large-scale simulations and deeper insights into human decision-making.

3 Game Theory for Algorithmic Innovation

This section investigates how game-theoretic principles contribute to developing LLMs by informing algorithmic innovation. Game theory has proven instrumental in enhancing our understanding of LLMs, mainly through the use of tools like the Shapley Value and social choice models. These methods offer valuable insights into model interpretability, enabling a deeper understanding of how LLMs process and respond to input. Beyond interpretability, game theory also provides a framework for developing training objectives and evaluation metrics that address key challenges in LLM development, such as model heterogeneity and alignment with human preferences.

3.1 Game Theory for Phenomenological Understanding LLMs

This line of research applies classical game theory concepts to explain observable phenomena in LLMs, including patterns in text generation and the inherent limitations of training within specific frameworks. Such studies are particularly valuable given that LLMs are often treated as "black boxes" due to their proprietary nature and large-scale complexity.

One approach connects cooperative game theory to LLMs, as these models perform parallel computations on input tokens and are structured around transformer layers. The Shapley Value [Shapley, 1953], a method for attributing contributions to individual players in cooperative games, has been adapted to assess the influence of specific tokens and layers on LLM-generated outputs. Several studies leverage the Shapley Value to evaluate token significance in prompts [Goldshmidt and Horovicz, 2024; Mohammadi, 2024]. For example, [Mohammadi, 2024] demonstrates that LLMs often assign disproportionately high weights to less informative input components, a behavior strongly correlated with incorrect responses. TokenSHAP [Goldshmidt and Horovicz, 2024] enhances Shapley Value computation using Monte Carlo sampling for efficiency, while TextGenSHAP [Enouen *et al.*, 2024] extends the approach to longer, structured input-output scenarios. [Liu *et al.*, 2023] applies the Shapley Value to multi-prompt learning, identifying the most impactful prompts for ensemble generation. Similarly, [Zhang *et al.*, 2024c] analyzes LLM layer contributions, finding that earlier layers exert a more significant influence on output generation.

Another research direction models LLM alignment with diverse human preferences using social choice theory. This framework helps address challenges aligning LLMs with human values and decision-making processes [Mishra, 2023]. For instance, [Conitzer *et al.*, 2024] analyzes the role of Reinforcement Learning from Human Feedback (RLHF) in expressing human preferences, identifying fundamental issues arising from preference conflicts, and advocating for social choice principles in LLM alignment. [Ge *et al.*, 2024] examines RLHF reward modeling as a social choice process, demonstrating that Bradley-Terry-based approaches suffer from intrinsic limitations that violate key axioms. [Qiu, 2024] proposes a representative social choice framework, which extracts a small but representative subset of opinions to manage large-scale preference aggregation effectively.

Additionally, some studies apply game theory to model alignment and decoding strategies. [Zhang *et al.*, 2024e] examine sociotechnical implications of real-world LLM applications, advocating for incentive compatibility to ensure AI systems align with societal goals while maintaining technical robustness. [Chen *et al.*, 2025] models the LLM decoding process as a Stackelberg game, where the decoder moves first, and an adversarial entity follows. By analyzing optimal strategies for both players, their study provides a theoretical basis for why heuristic sampling strategies perform well in practice.

3.2 Game Theory for Stimulating LLM Algorithms

In addition to enhancing our understanding of LLMs, game theory plays a crucial role in designing algorithms that improve their capabilities. This section highlights several key challenges in LLM training and illustrates how game theory has been applied to address these issues.

General Human Preference. Standard reward-based RLHF is limited to capturing only transitive preferences [Munos *et al.*, 2024]. Preference models, however, can express more general preferences by comparing two policies rather than assigning a reward for each response. This introduces new challenges in optimizing an LLM based on preference models. Nash Learning from Human Feedback (NLHF) aims to optimize the von Neumann winner of a game defined by the preference model, offering a feasible and robust direction for policy optimization.

Based on NLHF, SPO [Swamy *et al.*, 2024] introduces methods to express more complex preferences, such as non-transitive, stochastic, and non-Markovian preferences. SPPO [Wu *et al.*, 2025] designs an algorithm that efficiently implements SPO-like algorithms in large-scale language models. DNO [Rosset *et al.*, 2024] improves LLM optimization using a regression-based objective for more efficient and direct training. INPO [Zhang *et al.*, 2024d] introduces a loss function that can be directly minimized on preference datasets, further reducing the time overhead associated with calculating win rates in NLHF.

However, recent work by [Zhi-Xuan *et al.*, 2024] points out that preference-based approaches oversimplify human values, neglecting their complexity, incommensurability, and dynamic nature. As a result, designing more robust methods for aligning human preferences remains an ongoing scientific challenge.

Heterogeneity in Human Preferences. Capturing heterogeneity in human-annotated datasets remains a significant challenge in LLM alignment. Ignoring this heterogeneity often results in models that reflect only the preferences of the majority [Fleisig *et al.*, 2023]. Several studies have developed more inclusive training and alignment algorithms using social choice theory [Chakraborty *et al.*, 2024b; Park *et al.*, 2024; Alamdari *et al.*, 2024; Chen *et al.*, 2024a]. [Chakraborty *et al.*, 2024b] demonstrates the impracticality of using a single reward model and proposes the Egalitarian principle to learn preference distributions. [Park *et al.*, 2024] suggests clustering preferences and proposes a scalable, incentive-compatible framework for preference alignment. [Alamdari *et al.*, 2024] employs Borda count and quantile fairness for preference aggregation, ensuring fairness and computational feasibility. [Chen *et al.*, 2024a] introduces a mixture modeling framework for aggregating heterogeneous preferences. Additionally, [Klingefjord *et al.*, 2024] takes a macro perspective to examine the gap between human preferences and training objectives, offering solutions from a philosophical standpoint.

Data Cost Efficiency. Game theory has also been applied to enhance the cost efficiency of LLM training. Collecting a dataset with guaranteed quality and coverage is often challenging, so several works have used the self-play framework to improve data utilization, reducing the amount of data required while maintaining performance. [Chen *et al.*, 2024b] addresses the problem of fine-tuning a model with only a tiny amount of gold-standard data. Drawing from Generative Adversarial Networks [Goodfellow *et al.*, 2020], it allows the LLM to improve the quality of its answers while learning to distinguish between its responses and those of the gold-standard answers, ultimately converging to the distribution of the gold-standard data. [Cheng *et al.*, 2024a; Zheng *et al.*, 2024] models a game between an attacker and a defender, both of which are LLMs. [Zheng *et al.*, 2024] employs the attacker to propose prompts that the defender is less skilled at while the defender continuously improves. [Cheng *et al.*, 2024a] considers a classic game, Adversarial Taboo, to enhance model knowledge acquisition without introducing new data, leading to better performance in experiments. Furthermore, [Zhang and Duan, 2024] improves the efficiency of preference data collection by incorporating an auction model into the LLM fine-tuning process, demonstrating how this approach can enhance fine-tuning efficiency while maintaining strong performance.

Other Two-Player Game Formulations. In addition to the literature discussed above, several studies have formulated other two-player game models in specific phases of LLMs to enhance particular capabilities. [Chakraborty *et al.*, 2024a; Makar-Limanov *et al.*, 2024; Cheng *et al.*, 2024b] model the interaction between the reward model and the LLM as a two-player game. They aim to address the problem where a static reward model cannot handle the distribution shift of the evolving LLM policy. Their game-theoretic modeling captures the co-evolution of the reward model and the LLM, and equilibrium-solving algorithms are used to provide theoretically guaranteed LLM training methods.

[Jacob *et al.*, 2023] observes that generative and discriminative answers to a question by the same LLM are often inconsistent. It models the Consensus Game, where these two types of answers act as players seeking a consensus answer. Using equilibrium-solving algorithms, this approach significantly improves the LLM's accuracy across various datasets. Furthermore, [Gemp *et al.*, 2024] models the process of LLMs generating long-text conversations as a sequential game, using game-theoretic tools to enhance the model's ability to understand conversations and develop appropriate responses.

Remark. Game theory is essential in addressing the challenges in LLM development by offering clear principles for optimization and well-defined metrics for evaluating models' performance. Through its systematic approach, game theory helps refine LLM policies by aligning model behaviors with complex human preferences while providing a framework to measure and track model effectiveness improvements. This makes game theory a powerful tool for optimizing LLMs, ensuring training processes are both theoretically grounded and practically applicable.

4 Game Theory for LLM-Related Modeling

This section provides an overview of research on game-theoretic models that involve LLMs. The theoretical analysis of these models provides evidence of LLMs' impact on human

society. We categorize the literature into three main areas. The first area explores game-theoretic models that include both LLMs and humans, aiming to explain or predict the phenomena resulting from the development of LLMs. The second area examines scenarios where LLMs function as products or platforms. This creates competitive environments that exhibit game-theoretic dynamics like ad auctions. The third area extends classical game theory models, investigating how LLMs' unique capabilities can generalize and refine these models for more complex and realistic settings.

4.1 Competitions between LLM and Human

This body of work introduces several competition models, treating LLMs as players in the game [Yao *et al.*, 2024; Esmaeili *et al.*, 2024; Taitler and Ben-Porat, 2024]. These models generally arise from a recognition: modern LLMs possess powerful content generation capabilities and, compared to human creators, are characterized by lower costs and faster evolutionary rates.

[Yao *et al.*, 2024] investigates the impact of LLMs on human creators by proposing a competition model based on the Tullock contest. This model explores the dynamics between human-generated and LLM-generated content, modeling LLMs as cost-free players whose output quality improves as human content quality increases. Through equilibrium analysis, the study concludes that LLMs do not fundamentally conflict with or replace human creators but instead reduce the volume of human-generated content, ultimately pushing out less efficient creators. [Esmaeili *et al.*, 2024] extends this model into a repeated game setting, focusing on how humans can optimize their utility in dynamic competition with AI across various content domains. The study highlights the computational complexity of determining optimal strategies and proposes practical algorithms that offer near-optimal solutions.

[Taitler and Ben-Porat, 2024] examines the competitive dynamics between LLM-based generative AI and human-operated platforms, such as Stack Overflow, and their implications for social welfare. The study models the revenue-maximization problem of LLMs and uncovers phenomena akin to Braess's paradox: as human users increasingly rely on LLMs, the original platforms suffer from a lack of quality-enhancing data. Furthermore, generative AI models rarely undergo training aimed at quality improvement due to cost-saving incentives. The study also suggests theoretical regulatory frameworks to address these issues.

The development of LLMs has led to diverse societal effects, and game theory provides a robust theoretical framework for studying these effects. By employing appropriate models depicting optimal behaviors and equilibrium strategies, we can derive properties with theoretical guarantees.

4.2 Game Scenarios Emerging with LLMs

This section explores game-theoretic scenarios that arise from LLMs as products or platforms. In these scenarios, LLMs do not participate in the games; instead, they revolve around them.

As LLMs gain global attention, industries related to LLMs are creating substantial commercial value. [Laufer *et al.*, 2024] explores the feasibility of fine-tuning general-purpose models as a market service. The study models the bargaining process between generalists developing the model and domain specialists adapting it. By analyzing the sub-game perfect equilibria, the paper demonstrates that a profit-sharing outcome is possible and offers methods for determining Pareto-optimal equilibria. [Sun *et al.*, 2024a] investigates potential economic scenarios in which a platform offers fine-tuning services for LLMs through an auction-like process for multiple groups with different preferences. The study proposes an incentive-compatible payment scheme that guarantees social welfare maximization. [Mahmood, 2024] analyzes the competitive dynamics of LLM deployment, highlighting the value of market information and demonstrating that a first-to-market strategy may be cost-ineffective for all tasks when those tasks are sufficiently similar. [Saig *et al.*, 2024] proposes a pay-for-preference payment with a contract design model to address the potential moral hazard in the current pay-per-token pricing scheme.

In addition to their role as commodities, LLMs also offer potential commercial value through advertising revenue, similar to search engines. The emergence of LLMs has made traditional fixed-slot advertisements obsolete, prompting several studies on integrating LLMs into advertising auctions [Feizi *et al.*, 2023]. [Duetting *et al.*, 2024] models a scenario where each advertiser owns an agent LLM and bids to influence the probability distribution of the next token generated. The study derives an auction mechanism that ensures incentive compatibility by modifying the second-price auction. [Dubey *et al.*, 2024] assumes each advertiser provides a fixed advertisement copy, influencing LLM summaries through bidding. Their auction mechanism determines the prominence of each advertiser in the summary and the price they pay, ensuring incentive compatibility. [Hajiaghayi *et al.*, 2024] also assumes each advertiser possesses a document representing their content but models the advertisement insertion process in a Retrieval-Augmented Generation (RAG) framework. The mechanism probabilistically retrieves and allocates ads within each discourse segment of LLM-generated content, optimizing for logarithmic social welfare based on bids and relevance. [Soumalias *et al.*, 2024] investigates a scenario where each advertiser bids via a reward function for LLM-generated content. Their mechanism incentivizes truthful reporting of reward functions and demonstrates operational viability in a tuning-free setting.

4.3 LLM Extending Classic Game Models

In addition to these two areas, this section examines works that enhance traditional game theory models using LLMs, extending their applicability to more realistic scenarios.

LLMs' text comprehension and generation capabilities make them valuable tools for aggregating and eliciting opinions. [Lu *et al.*, 2024] explores using LLMs to assist peer review, noting that traditional peer prediction mechanisms are limited to simple reports, such as multiple-choice or scalar numbers. The study proposes peer prediction mechanisms that leverage LLMs' powerful text-processing capabilities to incentivize high-quality, truthful feedback. These mechanisms are shown to distinguish between human-written and LLM-generated reviews in empirical experiments. [Fish *et*

al., 2024a] uses LLMs to address the limitations of traditional social choice theory, which is constrained to choices among a few predetermined alternatives. The study employs LLMs to generate text and extrapolate preferences, providing a method for designing AI-augmented democratic processes with rigorous representation guarantees. [Sun *et al.*, 2024b] investigates how LLMs can provide richer information in traditional auctions. The study introduces the Semantic-enhanced Personalized Valuation in the Auction framework, which leverages LLMs to incorporate bidders' preferences and semantic item information into the valuation process. The framework integrates fine-tuned LLMs with the Vickrey auction mechanism to improve valuation accuracy and bidding strategies.

5 Conclusion and Future Directions

This survey provides a comprehensive overview of the research progress at the intersection of LLMs and game theory. We summarized the role that game theory has been playing in developing LLMs from three key perspectives: providing standardized game-based evaluation, driving game-theoretic algorithmic innovations, and modeling the societal impact of LLMs. Furthermore, we highlighted the bidirectional relationship between LLMs and game theory, exploring how LLMs influence traditional game models. Building on this review, we have identified several promising future directions.

LLM-based Agents with Comprehensive Game Abilities. Existing research has explored the evaluation of LLM agents across various game scenarios and developed methods to enhance their reasoning capabilities. However, while some of these methods demonstrate general applicability, their validation remains highly scenario-specific. A key future direction is to develop LLM agents proficient in game-theoretic reasoning, capable of applying their knowledge across diverse game settings without explicit customization. Achieving this requires advancements in rule comprehension, external environment modeling, and multi-agent reasoning. Key technical aspects include constructing a game-theoretic corpus, refining fine-tuning strategies, and incorporating tool-learning techniques.

Moving Beyond Human-Oriented Evaluation Frameworks. Game theory provides well-established evaluation criteria for rationality and strategic reasoning, such as K-level rationality, which has been widely adopted to assess LLM intelligence. However, these evaluation methods were originally designed for human cognition and may not fully capture the reasoning processes of next-token prediction models. To assess LLMs comprehensively from a game-theoretic perspective, it is crucial to move beyond existing human-oriented metrics and develop evaluation frameworks tailored to neural network-based models. This remains an underexplored area with the potential to improve our understanding of LLMs' decision-making significantly.

Theoretical Understanding of LLMs' Strategic Behavior. The application of game-theoretic concepts, such as the Shapley Value, to understanding LLMs' text generation behavior is still in its early stages. Most studies on LLMs' strategic behavior in real-world scenarios rely on empirical observations rather than systematic theoretical interpretations. For example, [Park *et al.*, 2025] has introduced hypothetical models to explain why LLMs struggle to achieve the performance level of no-regret learners in repeated games. Extending such theoretical investigations to more complex scenarios, including games like Werewolf, Avalon, or bargaining games, is essential. A deeper theoretical understanding of LLM strategic behavior will help to define their capability boundaries and provide insights for further improving their reasoning abilities.

Capturing Cooperative Games in LLM Optimization. Many studies leveraging game theory for optimizing LLM training, as discussed in Section 3.2, primarily focus on non-cooperative game settings. While non-cooperative approaches are a natural fit, cooperative game-theoretic methods offer additional insights into LLM optimization. For instance, different expert networks can be seen as participants in a cooperative game in the Mixture of Expert models. Adopting suitable payoff allocation mechanisms, such as the Shapley Value or the core concept, could optimize expert selection and task allocation, reducing redundancy while enhancing computational efficiency. Similarly, in ensemble learning and knowledge distillation, different sub-models can be treated as cooperative agents working together to refine decision boundaries or transfer knowledge. Effective reward allocation and weight adjustment strategies could enhance collaboration among sub-models, reducing redundant computation while improving generalization. Integrating cooperative game-theoretic methods into LLM training and optimization could offer new theoretical insights and practical solutions.

Modeling Cooperation Between Multi-LLMs and Humans. As discussed in Section 4.1, previous studies have primarily been focused on modeling competitive interactions between LLMs and humans, yielding valuable insights into their societal implications. However, beyond competition, understanding cooperative dynamics between multiple LLMs and humans remains an important research direction. One key challenge is to design mechanisms that incentivize LLMs to collaborate in fulfilling human-assigned tasks while considering their objectives. A theoretical characterization of LLM agents' goals and behaviors is essential for bridging the gap between game-theoretic mechanism design and real-world deployment. Advancing this line of research could facilitate the development of LLMs that align more effectively with human objectives and contribute positively to society.

Leveraging LLMs as Oracles to Extend Theoretical Game Models. As discussed in Section 4.3, several studies have explored how LLMs can be used to extend classical game-theoretic models. The key insight behind these works is that LLMs, with their strong language understanding and generative capabilities, can serve as oracles with specific functionalities in game-theoretic frameworks. This perspective opens up new opportunities to relax idealized assumptions or replace theoretical oracles in various game models using LLMs. By doing so, models that previously remained purely theoretical can now be practically implemented while preserving approximate theoretical properties. Systematic exploration of how LLMs can function as adaptable oracles in different theoretical models could bridge the gap between abstract game-theoretic concepts and real-world applications.

Acknowledgments

Supported by the National Natural Science Foundation of China (Grant No. 62172012, No. 12471339).

References

[Akata *et al.*, 2023] Elif Akata, Lion Schulz, et al. Playing repeated games with large language models. *arXiv preprint*, 2023.

[Alamdari *et al.*, 2024] Parand A Alamdari, Soroush Ebadian, and Ariel D Procaccia. Policy aggregation. *arXiv preprint*, 2024.

[Chakraborty *et al.*, 2024a] Souradip Chakraborty, Amrit Bedi, et al. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *ICLR*, 2024.

[Chakraborty *et al.*, 2024b] Souradip Chakraborty, Jiahao Qiu, et al. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. In *ICML*, 2024.

[Chen *et al.*, 2023] Jiangjie Chen, Siyu Yuan, et al. Put your money where your mouth is: Evaluating strategic planning and execution of llm agents in an auction arena. *arXiv preprint*, 2023.

[Chen *et al.*, 2024a] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint*, 2024.

[Chen *et al.*, 2024b] Zixiang Chen, Yihe Deng, et al. Self-play fine-tuning converts weak language models to strong language models. In *ICML*, 2024.

[Chen *et al.*, 2025] Sijin Chen, Omar Hagrass, and Jason M Klusowski. Decoding game: On minimax optimality of heuristic text generation strategies. In *ICLR*, 2025.

[Cheng *et al.*, 2024a] Pengyu Cheng, Tianhao Hu, et al. Self-playing adversarial language game enhances llm reasoning. In *NeurIPS*, 2024.

[Cheng *et al.*, 2024b] Pengyu Cheng, Yifan Yang, et al. Adversarial preference optimization: Enhancing your alignment via rm-llm game. In *ACL 2024 Findings*, 2024.

[Conitzer *et al.*, 2024] Vincent Conitzer, Rachel Freedman, et al. Position: Social choice should guide ai alignment in dealing with diverse human feedback. In *ICML*, 2024.

[Deng *et al.*, 2024] Yuan Deng, Vahab Mirrokni, et al. Llms at the bargaining table. In *ICML Workshop*, volume 2024, 2024.

[Deng *et al.*, 2025] Shilong Deng, Yongzhao Wang, and Rahul Savani. From natural language to extensive-form game representations. *arXiv preprint*, 2025.

[Du and Zhang, 2024] Silin Du and Xiaowei Zhang. Helmsman of the masses? evaluate the opinion leadership of large language models in the werewolf game. *arXiv preprint*, 2024.

[Duan *et al.*, 2024a] Jinhao Duan, Shiqi Wang, et al. Reta: Recursively thinking ahead to improve the strategic reasoning of large language models. In *NAACL*, 2024.

[Duan *et al.*, 2024b] Jinhao Duan, Renming Zhang, et al. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. In *NeurIPS*, 2024.

[Dubey *et al.*, 2024] Avinava Dubey, Zhe Feng, et al. Auctions with llm summaries. In *KDD*, 2024.

[Duetting *et al.*, 2024] Paul Duetting, Vahab Mirrokni, et al. Mechanism design for large language models. In *WWW*, 2024.

[Enouen *et al.*, 2024] James Enouen, Hootan Nakhost, et al. Textgenshap: Scalable post-hoc explanations in text generation with long documents. In *ACL Findings*, 2024.

[Esmaeili *et al.*, 2024] Seyed A Esmaeili, Kshipra Bhawalkar, et al. How to strategize human content creation in the era of genai? *arXiv preprint*, 2024.

[Fan *et al.*, 2024] Caoyun Fan, Jindou Chen, Yaohui Jin, and Hao He. Can large language models serve as rational players in game theory? a systematic analysis. In *AAAI*, 2024.

[Feizi *et al.*, 2023] Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges. *arXiv preprint*, 2023.

[Feng *et al.*, 2024] Xiachong Feng, Longxu Dou, et al. A survey on large language model-based social agents in game-theoretic scenarios. *arXiv preprint*, 2024.

[Fish *et al.*, 2024a] Sara Fish, Paul Gölz, et al. Generative social choice. In *EC*, 2024.

[Fish *et al.*, 2024b] Sara Fish, Yannai A Gonczarowski, and Ran I Shorrer. Algorithmic collusion by large language models. *arXiv preprint*, 2024.

[Fleisig *et al.*, 2023] Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *EMNLP*, 2023.

[Gandhi *et al.*, 2023] Kanishk Gandhi, Dorsa Sadigh, and Noah D Goodman. Strategic reasoning with language models. In *NeurIPS*, 2023.

[Ge *et al.*, 2024] Luise Ge, Daniel Halpern, et al. Axioms for ai alignment from human feedback. In *NeurIPS*, 2024.

[Gemp *et al.*, 2024] Ian Gemp, Yoram Bachrach, et al. States as strings as strategies: Steering language models with game-theoretic solvers. *arXiv preprint*, 2024.

[Goldshmidt and Horovicz, 2024] Roni Goldshmidt and Miriam Horovicz. Tokenshap: Interpreting large language models with monte carlo shapley value estimation. In *ACL Workshop NLP4Science*, 2024.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, et al. Generative adversarial networks. *Communications of the ACM*, 2020.

[Guo *et al.*, 2024] Shangmin Guo, Haoran Bu, et al. Economics arena for large language models. *arXiv preprint*, 2024.

[Hajiaghayi *et al.*, 2024] MohammadTaghi Hajiaghayi, Sébastien Lahaie, et al. Ad auctions for llms via retrieval augmented generation. In *NeurIPS*, 2024.

[Herr *et al.*, 2024] Nathan Herr, Fernando Acero, et al. Are large language models strategic decision makers? a study of performance and bias in two-player non-zero-sum games. *arXiv preprint*, 2024.

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

[Hu *et al.*, 2024] Sihao Hu, Tiansheng Huang, et al. A survey on large language model-based game agents. *arXiv preprint*, 2024.

[Hua *et al.*, 2024] Wenyue Hua, Ollie Liu, et al. Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint*, 2024.

[Huang *et al.*, 2024] Jen-tse Huang, Eric John Li, et al. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint*, 2024.

[Jacob *et al.*, 2023] Athul Paul Jacob, Yikang Shen, et al. The consensus game: Language model generation via equilibrium search. In *ICLR*, 2023.

[Klingefjord *et al.*, 2024] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *arXiv preprint*, 2024.

[Lan *et al.*, 2024] Yihuai Lan, Zhiqiang Hu, et al. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *EMNLP*, 2024.

[Laufer *et al.*, 2024] Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models. In *WWW*, 2024.

[Liu *et al.*, 2023] Hanxi Liu, Xiaokai Mao, et al. Prompt valuation based on shapley values. *arXiv preprint*, 2023.

[Lorè and Heydari, 2024] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 2024.

[Lu *et al.*, 2024] Yuxuan Lu, Shengwei Xu, et al. Eliciting informative text evaluations with large language models. In *EC*, 2024.

[Mahmood, 2024] Rafid Mahmood. Pricing and competition for generative ai. In *NeurIPS*, 2024.

[Makar-Limanov *et al.*, 2024] Jacob Makar-Limanov, Arjun Prakash, et al. Sta-rlhf: Stackelberg aligned reinforcement learning with human feedback. In *CoCoMARL*, 2024.

[Mensfelt *et al.*, 2024] Agnieszka Mensfelt, Kostas Stathis, and Vince Trencsenyi. Autoformalizing and simulating game-theoretic scenarios using llm-augmented agents. *arXiv preprint*, 2024.

[Mishra, 2023] Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint*, 2023.

[Mohammadi, 2024] Behnam Mohammadi. Wait, it's all token noise? always has been: interpreting llm behavior using shapley value. *Available at SSRN 4759713*, 2024.

[Munos *et al.*, 2024] Remi Munos, Michal Valko, et al. Nash learning from human feedback. In *ICML*, 2024.

[Nash Jr, 1950] John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 1950.

[Park *et al.*, 2024] Chanwoo Park, Mingyang Liu, et al. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML Workshop*, 2024.

[Park *et al.*, 2025] Chanwoo Park, Xiangyu Liu, et al. Do llm agents have regret? a case study in online learning and games. In *ICLR*, 2025.

[Qiu, 2024] Tianyi Qiu. Representative social choice: From learning theory to ai alignment. *NeurIPS 2024 Workshop*, 2024.

[Rosset *et al.*, 2024] Corby Rosset, Ching-An Cheng, et al. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint*, 2024.

[Saig *et al.*, 2024] Eden Saig, Ohad Einav, and Inbal Talgam-Cohen. Incentivizing quality text generation via statistical contracts. In *NeurIPS*, 2024.

[Shapira *et al.*, 2024] Eilam Shapira, Omer Madmon, et al. Glee: A unified framework and benchmark for language-based economic environments. *arXiv preprint*, 2024.

[Shapley, 1953] Lloyd S Shapley. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[Soumalias *et al.*, 2024] Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. *arXiv preprint*, 2024.

[Sun *et al.*, 2024a] Haoran Sun, Yurong Chen, et al. Mechanism design for llm fine-tuning with multiple reward models. *NeurIPS 2024 Workshop*, 2024.

[Sun *et al.*, 2024b] Jie Sun, Tianyu Zhang, et al. Large language models empower personalized valuation in auction. *arXiv preprint*, 2024.

[Swamy *et al.*, 2024] Gokul Swamy, Christoph Dann, et al. A minimaximalist approach to reinforcement learning from human feedback. In *ICML*, 2024.

[Taitler and Ben-Porat, 2024] Boaz Taitler and Omer Ben-Porat. Braess's paradox of generative ai. *arXiv preprint*, 2024.

[Vickrey, 1961] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 1961.

[Von Neumann and Morgenstern, 2007] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.

[Wang *et al.*, 2023] Shenzhi Wang, Chang Liu, et al. Avalon's game of thoughts: Battle against deception through recursive contemplation. *arXiv preprint*, 2023.

[Wang *et al.*, 2024] Haochuan Wang, Xiachong Feng, et al. Tmg-bench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *arXiv preprint*, 2024.

[Wu *et al.*, 2025] Yue Wu, Zhiqing Sun, et al. Self-play preference optimization for language model alignment. In *ICLR*, 2025.

[Xia *et al.*, 2024] Tian Xia, Zhiwei He, et al. Measuring bargaining abilities of llms: A benchmark and a buyer-enhancement method. In *ACL Findings*, 2024.

[Xu *et al.*, 2023] Yuzhuang Xu, Shuo Wang, et al. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint*, 2023.

[Yao *et al.*, 2024] Fan Yao, Chuanhao Li, et al. Human vs. generative ai in content creation competition: Symbiosis or conflict? In *ICML*, 2024.

[Zhang and Duan, 2024] Guoxi Zhang and Jiuding Duan. Vickreyfeedback: Cost-efficient data construction for reinforcement learning from human feedback. In *PRIMA*. Springer, 2024.

[Zhang *et al.*, 2024a] Yadong Zhang, Shaoguang Mao, et al. K-level reasoning with large language models. *arXiv preprint*, 2024.

[Zhang *et al.*, 2024b] Yadong Zhang, Shaoguang Mao, et al. Llm as a mastermind: A survey of strategic reasoning with large language models. In *COLM*, 2024.

[Zhang *et al.*, 2024c] Yang Zhang, Yanfei Dong, and Kenji Kawaguchi. Investigating layer importance in large language models. In *ACL Workshop BlackboxNLP*, 2024.

[Zhang *et al.*, 2024d] Yuheng Zhang, Dian Yu, et al. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint*, 2024.

[Zhang *et al.*, 2024e] Zhaowei Zhang, Fengshuo Bai, et al. Incentive compatibility for ai alignment in sociotechnical systems: Positions and prospects. *arXiv preprint*, 2024.

[Zheng *et al.*, 2024] Rui Zheng, Hongyi Guo, et al. Toward optimal llm alignments using two-player games. *arXiv preprint*, 2024.

[Zhi-Xuan *et al.*, 2024] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, 2024.