AI Alignment through a Game-theoretic Lens: A Survey

Anonymous ACL submission

Abstract

As large language models and increasingly capable AI agents are deployed in high-risk settings, aligning them with complex human values has become a central challenge. Existing alignment methods, while effective in improving helpfulness, harmlessness, and controllability, often struggle to capture real-world preferences that are context-dependent, non-transitive, and shaped by dynamic multi-party interactions. This survey reviews AI alignment through a game-theoretic lens. Specifically, it organizes recent progress around key game-theoretic elements and synthesizes the literature along three complementary aspects: preference structure, interaction structure, and temporal dynamics. This perspective clarifies where current alignment methods genuinely benefit from game-theoretic analysis, where the framework is looser, and what challenges remain in building robust, adaptive, and verifiable AI systems.

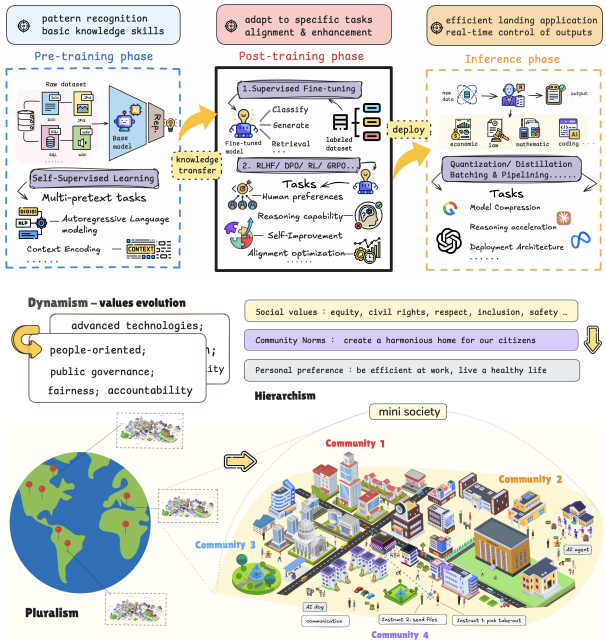

Figure 1: Post-training alignment (solid box in the top panel) and alignment challenges (bottom).

1 Introduction

As large language models (LLMs) and increasingly capable AI agents are deployed more broadly in real-world settings (Xi et al., 2025), alignment has become a central concern. AI systems are expected not only to perform tasks effectively, but also to act in ways that remain consistent with human intentions, social norms, and ethical constraints (Shen et al., 2024). The urgency of alignment is reflected in failures already observed in generative models, including bias, toxicity, manipulation, and misrepresentation (Duan et al., 2025; Gallegos et al., 2024; Guo, 2024), as well as concerns that more advanced systems may exhibit deception, reward hacking, or power-seeking behavior under imperfect oversight (Pan et al., 2022; Park et al., 2024; Turner and Tadepalli, 2022). Alignment is therefore not merely an objective, but also a safety and governance problem for increasingly agentic AI systems (Duenas and Ruiz, 2024; Milano and Nyholm, 2024).

Existing alignment approaches operate at both training and inference phase. In particular, post-training methods such as reinforcement learning from human feedback (RLHF) and direct preference optimization (DPO) have substantially improved helpfulness, harmlessness, and controllability (Ouyang et al., 2022; Rafailov et al., 2023). However, these methods still struggle when proxy rewards diverge from intended goals, when stakeholders hold conflicting preferences, and when static objectives fail under repeated interaction and environmental drift (Munos et al., 2024; Tang et al., 2025; Chu et al., 2025; Ye et al., 2024a). We attribute these limitations to three core challenges in alignment: *pluralism*, *hierarchism*, and *dynamism*. *Pluralism* arises because human values vary across cultures, groups, and contexts, making it difficult to specify a single objective that is both tractable and socially legitimate (Gabriel, 2020; Floridi and

Cowls, 2022). *Hierarchism* reflects the fact that alignment spans multiple levels, from optimization targets and reward proxies to user preferences and institutional or legal constraints, which often do not fully coincide (Amodei et al., 2016; Kleinberg et al., 2016). *Dynamism* captures temporal shifts in preferences, norms, environments, and model capabilities, especially when LLM-based agents interact repeatedly with humans, tools, and other agents (Carroll et al., 2024; Qiu et al., 2024).

These challenges motivate a more structural perspective on alignment. Rather than treating alignment solely as reward fitting or preference matching, a growing line of work models it as strategic interaction among humans, models, evaluators, institutions, and environments (Tang et al., 2025; Chu et al., 2025; Munos et al., 2024; Ye et al., 2024a). This perspective is well motivated because many alignment failures arise when systems optimize proxy objectives under imperfect oversight, interact with humans or other agents, and induce long-term incentive and feedback loops (Nayebi, 2025; Sun et al., 2025). In this sense, alignment is not only an optimization problem, but also an interaction problem under uncertainty, involving core game-theoretic elements such as players, utilities, information, equilibrium, and dynamics (Gemp et al., 2024; Mozikov et al., 2024). Recent research has begun to apply this perspective to preference conflict, equilibrium-based alignment, cooperative AI, online adaptation, and pluralistic evaluation (Tang et al., 2025; Chu et al., 2025; Ye et al., 2024a; Mao et al., 2025), but these efforts remain scattered across different settings and communities.

Some surveys have reviewed AI alignment broadly, emphasizing objectives, training pipelines, and governance (Ji et al., 2023; Sun et al., 2024b; Shen et al., 2024; Kayabay et al., 2025). Others focus on LLM agents and multi-agent systems, organizing the literature around architectures, workflows, and deployment challenges (Wang et al., 2024e; Li et al., 2024c; Zhou et al., 2025a). Recent work on game theory and LLMs, meanwhile, highlights the broader intersection between language models and games, including benchmarking, strategic behavior, and social simulation (Sun et al., 2025). In contrast, our survey is differentiated by both scope and organizing principle: we focus on alignment-relevant work for which game-theoretic notions materially clarify incentives, equilibrium, mechanism design, or temporal adaptation. Rather than surveying all alignment methods or all game-related LLM research, we ask when game theory changes the formulation, guarantee, or evaluation of alignment methods. Based on this focus, we make three contributions. First, we introduce a three-part challenge framework organized around *pluralism*, *hierarchism*, and *dynamism*. Second, we connect representative work to shared interaction structures and evaluation settings, making explicit where game-theoretic modeling is central and where it serves only as a looser interpretive lens. Third, we synthesize recent progress along three complementary analytical aspects: preference structure, interaction structure, and temporal dynamics, instantiated in the survey as preference alignment, interaction-driven capability enhancement, and dynamic evolution. This framing helps sharpen both the usefulness and the limits of game-theoretic analysis for alignment.

2 Taxonomy: A Game-theoretic Lens on Alignment Frameworks

2.1 Theoretical Clarification

To avoid overstating the role of game theory in AI alignment, we clarify what game-theoretic analysis can formally guarantee and under which assumptions. Our goal is not to claim that game theory universally solves alignment. Rather, we use game-theoretic language to characterize when alignment can be formulated as a strategic optimization problem, when an equilibrium can be interpreted as an aligned policy, and when convergence guarantees and equilibrium gaps can serve as formal certificates of approximate alignment (Munos et al., 2024; Ye et al., 2024a; Wu et al., 2024; Zhang et al.; Pásztor et al., 2025; Shi et al., 2025). We illustrate these ideas through alignment with general human preferences, which serves as a representative foundation for other alignment settings.

2.1.1 Problem Formulation

As noted earlier, aligning LLMs typically relies on learning from human preference feedback. Rather than providing a scalar reward signal $r(x, y)$, human feedback more often takes the form of a pairwise comparison signal $\mathbb{P}(y_1 \succ y_2 \mid x)$, where $\succ$ denotes a preference relation. Accordingly, preference learning more closely matches practical RLHF settings and is better suited to tasks that are difficult to specify using a single numerical reward. In this setting, we seek a policy that performs robustly under human pairwise comparisons against alterna-

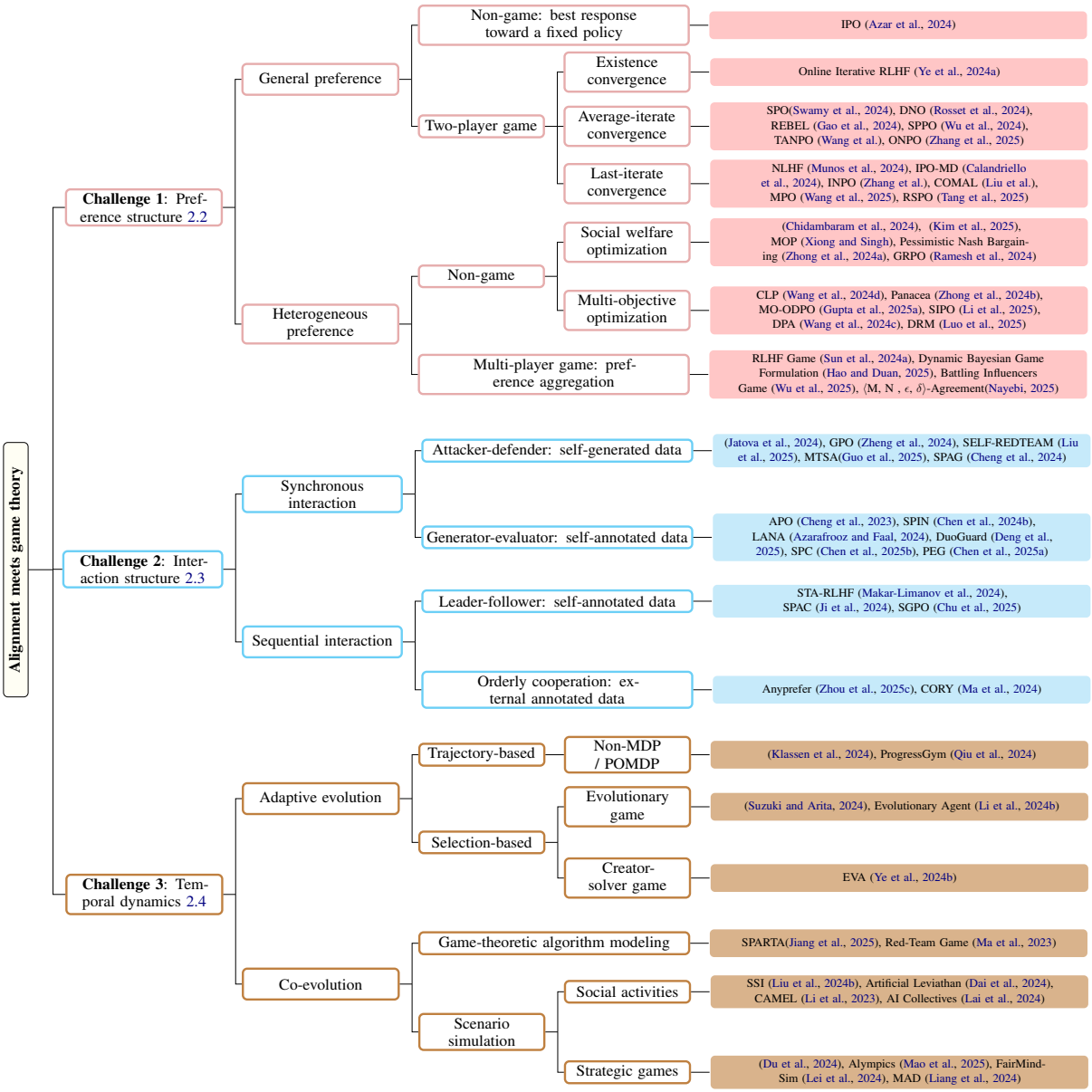

Figure 2: Taxonomy of alignment based on a game theory lens.

tive policies. Specifically, consider the objective

$$J(\pi_1, \pi_2) = \mathbb{E}_{x\sim d_0}\mathbb{E}_{y_1\sim\pi_1,\, y_2\sim\pi_2}\Big[\mathbb{P}(x, y_1, y_2) - \tau D_{\mathrm{KL}}\big(\pi_1(\cdot \mid x) \,\|\, \pi_{ref}(\cdot \mid x)\big) + \tau D_{\mathrm{KL}}\big(\pi_2(\cdot \mid x) \,\|\, \pi_{ref}(\cdot \mid x)\big)\Big], \tag{1}$$

where d_0 is a prompt distribution. The KL terms regularize both policies toward a reference policy π_{ref}, thereby stabilizing learning and controlling deviation from the reference model.

This objective defines a two-player zero-sum game in which the players choose policies π_1 and π_2. The first player maximizes $J(\pi_1, \pi_2)$, while the second player minimizes it. Intuitively, π_1 seeks a policy that is preferred under human comparisons, whereas π_2 acts as an adversarial comparator. The alignment problem can therefore be formulated as the saddle-point problem

$$(\pi_1^*, \pi_2^*) \in \arg\max_{\pi_1\in\Pi} \arg\min_{\pi_2\in\Pi} J(\pi_1, \pi_2), \tag{2}$$

where both players optimize over the same policy class Π.

2.1.2 Nash Equilibrium as Strategic Alignment

For analysis, we restrict attention to policies in Π that have the same support as π_{ref}. Under this assumption, (Ye et al., 2024a) shows that the game admits a unique Nash equilibrium. By symmetry, the

equilibrium policies coincide, i.e., $\pi_1^* = \pi_2^* = \pi^*$. We interpret π^* as the aligned policy because it satisfies the saddle-point property of the equilibrium. For any alternative policy $\pi \in \Pi$, we have

$$J(\pi, \pi^*) \leq J(\pi^*, \pi^*) \leq J(\pi^*, \pi).$$

Hence, no policy can systematically obtain a better payoff against π^* than π^* obtains against itself, while π^* performs at least as well against any opponent as it does against itself. In the unregularized case, or when J is interpreted as a regularized preference value, this further suggests that the win rate of π^* against any π is at least one half. Under this view, alignment is not defined as maximizing an unobserved utility function; rather, it is defined as non-domination under the induced preference relation. In this sense, an aligned policy is strategically stable under pairwise comparison and resistant to exploitation by competing policies.

2.1.3 Approximation and Convergence

To assess how well a candidate policy approximates π^*, we use the following duality gap (Zhang et al.):

$$\mathrm{DualGap}(\pi) := \max_{\pi_1 \in \Pi} J(\pi_1, \pi) - \min_{\pi_2 \in \Pi} J(\pi, \pi_2). \tag{3}$$

The duality gap is always nonnegative, and $\mathrm{DualGap}(\pi) = 0$ if and only if $\pi = \pi^*$. Thus, $\mathrm{DualGap}(\pi)$ quantifies how far π is from the aligned Nash policy. A small duality gap means that π is only weakly exploitable in the induced preference game, providing a formal certificate of approximate alignment. When $\mathrm{DualGap}(\pi) \leq \epsilon$, π may be regarded as an ϵ-approximate Nash policy. Related definitions and guarantees also appear in (Ye et al., 2024a; Ma et al., 2023).

The practical significance of this equilibrium formulation depends on whether the learning dynamics converge to such a stable point. Recent works (Munos et al., 2024; Wang et al., 2024f; Liu et al.) establish convergence guarantees for several algorithms, including average-iterate convergence of $\frac{1}{T}\sum_{t=1}^{T} \pi_t$. Moreover, if an algorithm enjoys last-iterate convergence, then the final output policy π_T itself can be interpreted as a strategically aligned solution, and the resulting model inherits the same stability interpretation (Tiapkin et al., 2025). That implies that alignment is not merely observed empirically, but rather the limit of a strategic learning process, as will be explained below.

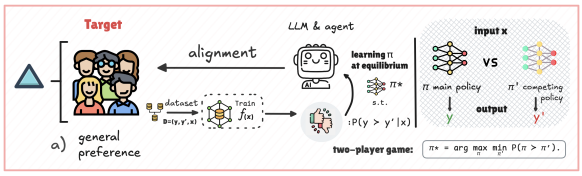

Figure 3: General preference alignment: pairwise optimization and game-theoretic formulation.

2.2 Preference Alignment

Preference alignment studies how AI systems are optimized from feedback. In practice, this process usually involves three components: the AI system to be aligned, the feedback source, and a proxy that models such feedback for learning, e.g., a reward or preference model (Ji et al., 2023). This section focuses on the structure of preferences rather than the timing of interaction. A game-theoretic lens becomes especially useful when alignment depends either on strategic robustness against alternative policies or on fair aggregation across stakeholders with conflicting values. We therefore discuss two related settings: general preference, where the key question is robustness under comparison, and heterogeneous preferences, where the main issue is how to aggregate diverse human interests in a fair and stable way.

2.2.1 General preference

General preference alignment aims to optimize model behavior toward broadly shared human judgments, as shown in Figure 3. Early work mainly provided theoretical perspectives on learning from human preferences (Wang et al., 2023b), but did not directly capture practical LLM alignment under pairwise feedback. A first line of work is non-game optimization against a fixed policy. The representative example is IPO (Azar et al., 2024), which directly optimizes pairwise preferences without assuming an explicit reward model:

$$\max_{\pi} \mathbb{E}_{y \sim \pi,\, y' \sim \mu} \left[\mathbb{P}(y \succ y') \right] - \tau\, \mathrm{KL}(\pi \| \pi_{\mathrm{ref}}). \tag{4}$$

Compared with point-wise reward optimization, IPO better matches preference learning, but it only learns a best response to a fixed reference distribution μ and therefore does not guarantee robustness against stronger competing policies.

This limitation motivates a second line of work that formulates alignment as a two-player game. Representative studies such as NLHF (Munos

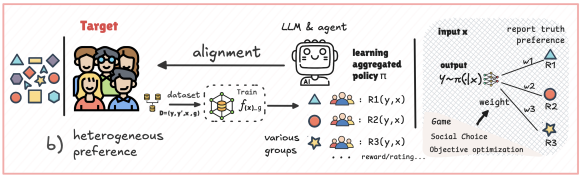

Figure 4: Heterogeneous preference alignment: multi-group feedback and preference aggregation.

et al., 2024), IPO-MD (Calandriello et al., 2024), and iterative RLHF (Ye et al., 2024a) cast alignment as a constant-sum or minimax game, where the target policy is optimized against a competing policy under a learned preference function. Early results mainly established the existence of Nash-style aligned policies. Subsequent work focused on stronger convergence guarantees. One branch studies average-iterate convergence, including SPO (Swamy et al., 2024), DNO (Rosset et al., 2024), REBEL (Gao et al., 2024), SPPO (Wu et al., 2024), TANPO (Wang et al.), and ONPO (Zhang et al., 2025). These methods often replace explicit win-rate estimation with simpler regression-style objectives or incorporate online exploration to improve sample efficiency. Another branch studies last-iterate convergence, which is more desirable because the final learned policy itself inherits the equilibrium interpretation. Representative examples include INPO (Zhang et al.), COMAL (Liu et al.), MPO (Wang et al., 2025), and RSPO (Tang et al., 2025). Overall, general preference alignment has evolved from fixed-opponent optimization to self-play Nash learning, with increasing emphasis on convergence, robustness, and the role of regularization.

2.2.2 Heterogeneous preferences

Heterogeneous preference alignment concerns settings where stakeholders hold overlapping or conflicting values, as shown in Figure 4. Unlike general preference alignment, the goal is not to fit a single dominant preference, but to balance multiple groups, objectives, and minority concerns.

A first category is non-game alignment. One line formulates the problem as social welfare optimization, where group utilities are aggregated through welfare or bargaining objectives (Mishra, 2023; Conitzer et al., 2024; Qiu). Representative methods include MOP (Xiong and Singh), Pessimistic Nash Bargaining (Zhong et al., 2024a), and group-robust approaches (Chidambaram et al., 2024; Ramesh et al., 2024; Kim et al., 2025). These methods improve fairness and reduce dominance, but their behavior depends strongly on the chosen welfare function. Another line formulates alignment as multi-objective optimization. Methods such as CLP (Wang et al., 2024d), Panacea (Zhong et al., 2024b), MO-ODPO (Gupta et al., 2025a), and SIPO (Li et al., 2025) jointly optimize multiple objectives and often support controllable trade-offs through preference vectors or conditional signals. Related work also studies decomposed or vector-valued reward representations for more interpretable preference control (Wang et al., 2024c; Luo et al., 2025).

A second category treats heterogeneous alignment as a multi-player game over preference aggregation. Here the main issue is that stakeholders may strategically report preferences to influence the final model. RLHF Game (Sun et al., 2024a) and Dynamic Bayesian Game (Hao and Duan, 2025) explicitly study truthful reporting and weighting mechanisms, while BIG (Wu et al., 2025) analyzes strategic influence among multiple parties. Other work further examines multi-preference optimization and the complexity of alignment across multiple objectives and agents under weak assumptions (Gupta et al., 2025b; Nayebi, 2025). Together, these studies show that heterogeneous preference alignment is not only an optimization problem, but also a mechanism-design problem shaped by strategic human behavior.

2.3 Enhancement via Interaction

Whereas the previous subsection organizes the literature by preference structure, this section shifts the focus to interaction structure. AI alignment is increasingly understood as an interactive process rather than a one-shot transfer of human preferences to a model (Shen et al., 2024). Many methods improve alignment not only by fitting feedback, but also by generating informative feedback through play, critique, or role-dependent adaptation. The main distinction here is whether strategic updates occur within a shared decision round or under an explicit order of moves. We therefore distinguish between synchronous interaction, which is especially useful for adversarial stress testing and self-improvement, and sequential interaction, which introduces commitment, asymmetry, and bilevel optimization into alignment.

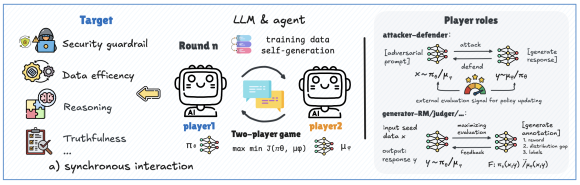

Figure 5: Synchronous interaction for alignment: simultaneous play, adversarial roles, and alignment targets.

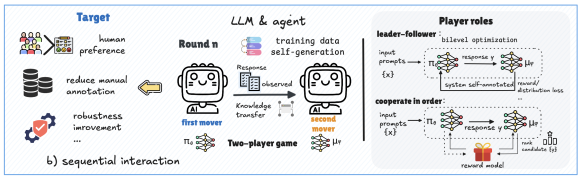

Figure 6: Sequential interaction for alignment: ordered roles and commitment-based optimization.

2.3.1 Synchronous interaction

When interaction is synchronous, both sides act within the same round, and alignment is improved through repeated exposure to hard cases generated during play. This setting has been particularly influential in safety alignment. GPO (Zheng et al., 2024) formulates alignment as a max–min game between an attacker and a defender, showing that adversarial prompt generation can uncover vulnerabilities while improving robustness through iterative updates. Closely related work applies the same intuition to toxicity mitigation, guardrail training, and jailbreak robustness, either by training separate attacker–defender models or by alternating these roles within a single model (Jatova et al., 2024; Deng et al., 2025; Liu et al., 2025; Guo et al., 2025). Across these studies, synchronous play functions less as an abstract game metaphor than as a practical mechanism for generating stress cases that standard alignment pipelines would otherwise miss.

The same interaction pattern has also been repurposed to reduce dependence on human annotation. APO (Cheng et al., 2023) couples the policy with a reward model in an adversarial loop to address the mismatch between evolving model outputs and static preference supervision. SPIN (Chen et al., 2024b) instead treats self-improvement as iterative self-distillation, while LANA (Azarafrooz and Faal, 2024) further shows that self-rewarding can remain effective even under noisy or suboptimal demonstrations. What connects these methods is that interaction does not merely refine the policy; it also becomes a means of producing or filtering the supervision signal itself.

This interactive template has further extended to reasoning and truthfulness. SPC (Chen et al., 2025b) uses adversarial self-play between a generator and critic to improve step-level reasoning and support test-time search, whereas SPAG (Cheng et al., 2024) relies on adversarial language games to strengthen reasoning-oriented dialogue. PEG (Chen et al., 2025a) pushes the idea toward truthful evaluation by casting peer assessment as a game in which truthful reporting is incentivized. Taken together, these studies suggest that synchronous interaction is most useful when alignment requires continual pressure from challenging inputs, self-generated supervision, or competing evaluations. Its limitations, however, remain clear: many methods are computationally intensive, sensitive to training dynamics, and still rely on assumptions that are difficult to satisfy in realistic black-box settings.

2.3.2 Sequential interaction

Once the order of moves becomes explicit, interaction changes character. The earlier player can commit to a strategy that reshapes the response of the later player, making this setting closer to bilevel or Stackelberg optimization than to simultaneous self-play. This perspective has been especially useful for preference learning. STA-RLHF (Makar-Limanov et al., 2024) models the language model and reward model as leader and follower, respectively, and searches for a Stackelberg equilibrium under nested optimization. SPAC (Ji et al., 2024) preserves the same leader–follower structure but replaces multi-stage updates with a single-timescale procedure, yielding stronger practical appeal in offline settings with sparse preference data.

A similar logic also appears when sequential interaction is used to improve data quality and robustness. SGPO (Chu et al., 2025) studies alignment under scarce and noisy preferences through a Stackelberg-style formulation with regret guarantees under bounded distribution shift. Anyprefer (Zhou et al., 2025c) combines automatic synthesis, external tools, and a cooperative Markov game to mitigate self-reward bias while increasing sample diversity. CORY (Ma et al., 2024) goes a step further by using sequential cooperative multi-agent reinforcement learning and role exchange to stabilize fine-tuning. In these settings, ordered interaction is valuable not only because it captures asymmetric roles more faithfully, but also because

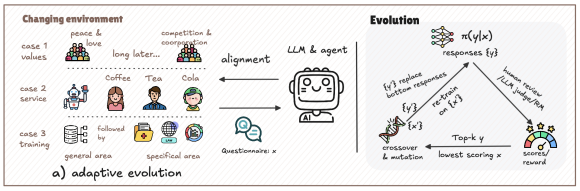

Figure 7: Adaptive evolution in alignment: changing social environments and a typical update framework.

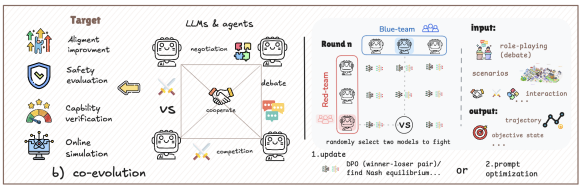

Figure 8: Co-evolution in alignment: multiple agents evolving through competition, cooperation, and social interaction.

it offers a natural way to control how evaluators, generators, and auxiliary tools influence one another over time.

These interaction-based approaches move alignment beyond static supervision, but they also expose a broader challenge: alignment objectives, evaluators, and environments may all evolve over time. We therefore turn next to dynamic evolution in alignment.

2.4 Dynamic Evolution

The previous two sections focus on preference structure and interaction structure at a given stage of learning or deployment. We now turn to a third aspect that cuts across both: temporal change. Once systems interact repeatedly with humans, other agents, and changing environments, alignment can no longer be treated as a one-shot objective fixed at training time. Instead, the central challenge is to maintain behavioral consistency as preferences, incentives, and external conditions evolve. In this sense, dynamic evolution is not a separate alignment category so much as a temporal extension of the preceding discussion. This has motivated a growing body of work that studies alignment through adaptive evolution and co-evolution.

2.4.1 Adaptive evolution

Adaptive evolution focuses on how AI systems update under temporally varying social environments, as illustrated in Figure 7. One line of work formulates this problem through sequential decision-making or trajectory-based evaluation. ProgressGym (Qiu et al., 2024), for example, frames alignment as moral progress over time and studies how historical moral data can guide future decisions. Relatedly, Klassen et al. (Klassen et al., 2024) propose a temporal pluralistic alignment framework that evaluates stakeholder satisfaction over trajectories rather than isolated outputs, allowing alignment to depend on long-horizon and non-Markovian requirements.

A complementary line introduces evolutionary mechanisms such as mutation, selection, and environmental variation. Li et al. (Li et al., 2024b) embed social norms into a genetic process so that agents adapt through repeated selection under changing group behavior, while Suzuki and Arita (Suzuki and Arita, 2024) use evolutionary games to study the emergence of cooperative traits in LLM-based agents. Ye et al. (Ye et al., 2024b) further model prompt mutation as a way to simulate novel environments and formulate alignment as a creator–solver game. Across these studies, the common intuition is that alignment should remain responsive to gradual changes in norms, tasks, and social expectations, rather than being anchored to a static preference snapshot.

2.4.2 Co-evolution

Whereas adaptive evolution emphasizes how a model tracks environmental change, co-evolution focuses on settings in which the environment itself is partly generated by other evolving agents, as shown in Figure 8. This perspective is particularly useful when alignment depends on multi-agent interaction, population diversity, or open-ended strategic adaptation. SPARTA (Jiang et al., 2025) studies collective alignment through repeated combat, evaluation, and learning among multiple models, while Ma et al. (Ma et al., 2023) cast safety alignment as a red–blue team game in which attack and defense strategies evolve together. These approaches suggest that robustness may emerge not only from stronger objectives, but also from continued exposure to adaptive opponents and evaluators.

The same co-evolutionary perspective also underlies a broader set of social simulation environments. Recent work uses agent societies, role-playing communities, and sandbox worlds to study how norms, cooperation, and governance emerge through interaction (Liu et al., 2024b; Lai et al., 2024; Li et al., 2023; Dai et al., 2024). In parallel, strategically structured environments such as

Alympics (Mao et al., 2025) and FairMindSim (Lei et al., 2024) are used to probe bargaining, fairness, and collective behavior under resource constraints or ethical dilemmas. Debate-based methods extend this logic to inference-time improvement: by allowing multiple agents to critique and challenge one another, they can improve reasoning and factuality without retraining (Liang et al., 2024; Du et al., 2024). Together, these studies broaden alignment from single-model optimization to the analysis of collective behavior in evolving social systems, but many of these environments remain behaviorally rich while still offering limited visibility into the internal mechanisms that drive adaptation.

3 Future Directions

Fairness guarantees Pluralistic societies typically do not rely on a single value standard, so alignment should not only search for a "true" value objective but also design fair procedures for deciding which values guide model behavior (MacIntyre, 2013; Gabriel, 2020). Existing multi-objective methods introduce fairness constraints, such as demographic parity, equalized odds, group-DRO, and worst-group lower bounds (Agnihotri et al., 2025; Chakraborty et al., 2024). However, soft regularization often gives weak guarantees, while hard constraints can reduce utility or destabilize training (Tran et al., 2024). A promising direction is to model fairness-utility trade-offs as games with social-welfare objectives (Chen et al., 2024a), including Nash welfare, Leximin (Barman and Suzuki, 2024), and Pareto-optimal solution families (Censor, 1977). Future work should further combine adaptive weighting with procedural safeguards from Constitutional AI, social choice, voting, and public participation, so that both outcomes and decision procedures remain legitimate and adaptable (Zhao et al., 2024; Huang et al., 2024b).

Continual alignment under drift. Online tool use, continual preference learning, and scenario migration can update useful capabilities while gradually overwriting aligned behavior. This causes catastrophic forgetting and value drift, especially in adversarial testing and long multi-turn interaction. From a game-theoretic view, alignment maintenance can be modeled as a dynamic interaction among the model, deployment environment, and strategic auditors or adversaries, where feedback streams may expose or induce degradation. Existing remedies include continual or online RLHF (Zhang et al., 2024), trust-region or strongly regularized updates (Li et al., 2024a), and generative replay under data-access limits (Huang et al., 2024a). However, most methods optimize general retention rather than treating alignment-critical knowledge as the main retention target (Wang et al., 2024h). Future work should combine continual learning with hierarchical or retrieval-augmented memory (Wang et al., 2023a; Borgeaud et al., 2022; Meng et al., 2022), classical anti-forgetting mechanisms (Dokania et al., 2019; Kirkpatrick et al., 2017), and adversarially robust updates (Khan et al., 2022; Mirzadeh et al., 2020; Ebrahimi et al., 2020), so that safety-relevant knowledge remains stable under changing or manipulated data flows.

Visual simulation for explainable AI. AI alignment remains difficult to inspect when simulations only report external behavior. Controlled social simulations, agent societies, and policy games, as discussed in Section 2.4, can reveal dynamic failures in interaction, adaptation, and social feedback. Yet many such environments remain behaviorally rich but mechanistically shallow. Future work should connect these environments with mechanistic interpretability, rule-based dynamic models, intervention logging, and counterfactual analysis. This would turn simulation from outcome-level evaluation into process-level diagnosis, making long-term alignment easier to inspect, explain, and steer.

Conclusion

This survey examined AI alignment through the lens of game theory, reframing alignment as a problem of strategic dependence rather than static objective fitting. By organizing the literature around preference structure, interaction structure, and temporal dynamics, this work clarifies how values are aggregated, how models or agents influence one another, and how aligned behavior may emerge, change, or remain stable over time. Furthermore, we point out that existing game-theoretic alignment methods still focus on the training phase, while their role in inference control, social game-based benchmarking, practical verification, and interpretability remains to be developed. Therefore, researchers need to further expand game-theoretic tools to achieve, diagnose, and manage AI alignment, strengthening the connection between theoretical guarantees and the real world, thereby supporting the development of safer and more reliable AI systems.

Limitations

This survey focuses primarily on post-training alignment work for LLMs and LLM-based agents that can be meaningfully analyzed through game-theoretic concepts. It therefore does not aim to exhaustively cover adjacent literature on pre-training objectives, inference-time defenses, mechanistic interpretability, or governance frameworks that do not directly adopt this lens. In addition, the reviewed areas are uneven in theoretical maturity: preference-learning work currently admits stronger formal guarantees than many studies on social interaction, simulation, and long-term evolution. As a result, some parts of the survey necessarily provide a looser synthesis than others.

Ethics Statement

While game-theoretic frameworks can help clarify incentives and interaction structures in alignment, they may also simplify complex moral, cultural, and institutional factors. Therefore, the methods reviewed in this paper should not be regarded as complete solutions to alignment or as substitutes for human oversight and governance. We encourage readers to interpret the reviewed approaches with caution, especially in safety-critical or socially sensitive settings.

References

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

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. 2025. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*.

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.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Ari Azarafrooz and Farshid Faal. 2024. Language alignment via nash-learning and adaptive feedback. *arXiv preprint arXiv:2406.15890*.

Siddharth Barman and Mashbat Suzuki. 2024. Compatibility of fairness and nash welfare under subadditive valuations. *arXiv preprint arXiv:2407.12461*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, and 1 others. 2024. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*.

Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. Ai alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5706–5756.

Yair Censor. 1977. Pareto optimality in multiobjective problems. *Applied Mathematics and Optimization*, 4(1):41–59.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Singh Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6116–6135.

Baiting Chen, Tong Zhu, Jiale Han, Lexin Li, Gang Li, and Xiaowu Dai. 2025a. Incentivizing truthful language models via peer elicitation games. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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. *arXiv preprint arXiv:2504.19162*.

Violet Chen, JN Hooker, and Derek Leben. 2024a. Assessing group fairness with social welfare optimization. In *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 208–225.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024b. 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.

Pengyu Cheng, Yifan Yang, Jian Li, Yong Dai, Tianhao Hu, Peixin Cao, Nan Du, and Xiaolong Li. 2023. Adversarial preference optimization: Enhancing your alignment via rm-llm game. *arXiv preprint arXiv:2311.08045*.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. 2024. Direct preference optimization with unobserved preference heterogeneity. *arXiv preprint arXiv:2405.15065*.

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

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 should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Pedro MP Curvo. 2025. The traitors: Deception and trust in multi-agent language model simulations. *arXiv preprint arXiv:2505.12923*.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Srihas Rao, Arthur Caetano, Misha Sra, and 1 others. 2024. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory. *arXiv preprint arXiv:2406.14373*.

Tim R Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. 2024. Evaluating language model agency through negotiations. In *ICLR 2024*.

Yihe Deng, Yu Yang, Junkai Zhang, Wei Wang, and Bo Li. 2025. Duoguard: A two-player rl-driven framework for multilingual llm guardrails. *arXiv preprint arXiv:2502.05163*.

P Dokania, P Torr, and M Ranzato. 2019. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, pages 11733–11763. PMLR.

Wen Duan, Nathan McNeese, and Lingyuan Li. 2025. Gender stereotypes toward non-gendered generative ai: The role of gendered expertise and gendered linguistic cues. *Proceedings of the ACM on Human-Computer Interaction*, 9(GROUP):1–35.

Tom Duenas and Diana Ruiz. 2024. The frontier of ai alignment: Challenges and strategies for future ai systems. *ResearchGate, 2024a. doi*, 10.

Sayna Ebrahimi, Franziska Meier, Roberto Calandra, Trevor Darrell, and Marcus Rohrbach. 2020. Adversarial continual learning. In *European Conference on Computer Vision*, pages 386–402.

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

Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for ai in society. *Machine learning and the city: Applications in architecture and urban design*, pages 535–545.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2025. Nicer than humans: How do large language models behave in the prisoner's dilemma? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 522–535.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Zhaolin Gao, Jonathan Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, Drew Bagnell, Jason D Lee, and Wen Sun. 2024. Rebel: Reinforcement learning via regressing relative rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400.

Ian Gemp, Yoram Bachrach, Marc Lanctot, Roma Patel, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, Siqi Liu, and Karl Tuyls. 2024. States as strings as strategies: Steering language models with game-theoretic solvers. *arXiv preprint arXiv:2402.01704*.

Weiyang Guo, Jing Li, Wenya Wang, Yu Li, Daojing He, Jun Yu, and Min Zhang. 2025. Mtsa: Multi-turn safety alignment for llms through multi-round red-teaming. *arXiv preprint arXiv:2505.17147*.

Ziyi Guo. 2024. Online disinformation and generative language models: Motivations, challenges, and mitigations. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1174–1177.

Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, and Abhinav Rastogi. 2025a. Robust multi-objective preference alignment with online dpo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27321–27329.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Chetan Bansal, and Saravan Rajmohan. 2025b. AMPO: Active multi preference optimization for self-play preference selection. In *Scaling Self-Improving Foundation Models without Human Supervision*.

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.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024a. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. 2024b. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417.

Lucas Jatova, Jacob Smith, and Alexander Wilson. 2024. Employing game theory for mitigating adversarial-induced content toxicity in generative large language models. *Authorea Preprints*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. 2024. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv preprint arXiv:2406.04274*.

Yuru Jiang, Wenxuan Ding, Shangbin Feng, Greg Durrett, and Yulia Tsvetkov. 2025. Sparta alignment: Collectively aligning multiple language models through combat. *arXiv preprint arXiv:2506.04721*.

Kerem Kayabay, Mert Onuralp Gökalp, Atilla Kılınç, Ebru Gökalp, and Tugrul Daim. 2025. Data science roadmapping for ai alignment: Insights from a multi-vocal literature review and a retrospective case study. *IEEE Transactions on Engineering Management*.

Hikmat Khan, Nidhal Carla Bouaynaya, and Ghulam Rasool. 2022. Adversarially robust continual learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, and Pablo A Parrilo. 2025. Population-proportional preference learning from human feedback: An axiomatic approach. *arXiv preprint arXiv:2506.05619*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526.

Toryn Q Klassen, Parand A Alamdari, and Sheila A McIlraith. 2024. Pluralistic alignment over time. *arXiv preprint arXiv:2411.10654*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. 2024. Position: Evolving ai collectives enhance human diversity and enable self-regulation. In *Forty-first International Conference on Machine Learning*.

Yu Lei, Hao Liu, Chengxing Xie, Songjia Liu, Zhiyu Yin, Canyu Chen, Guohao Li, Philip Torr, and Zhen Wu. 2024. Fairmindsim: Alignment of behavior, emotion, and belief in humans and llm agents amid ethical dilemmas. *arXiv preprint arXiv:2410.10398*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308.

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. 2025. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. *arXiv preprint arXiv:2502.14354*.

Shimin Li, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024b. Agent alignment in evolving social norms. *arXiv preprint arXiv:2401.04620*.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*.

Jiaheng Liu, Zehao Ni, Haoran Que, Tao Sun, Zekun Wang, Jian Yang, Jiakai Wang, Hongcheng Guo, Zhongyuan Peng, Ge Zhang, and 1 others. 2024a.

Roleagent: Building, interacting, and benchmarking high-quality role-playing agents from scripts. *Advances in Neural Information Processing Systems*, 37:49403–49428.

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

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2024b. Training socially aligned language models on simulated social interactions. In *The Twelfth International Conference on Learning Representations*.

Yixin Liu, Argyris Oikonomou, Weiqiang Zheng, Yang Cai, and Arman Cohan. Comal: A convergent meta-algorithm for aligning llms with general preferences. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

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.

Siting Estee Lu. 2024. Strategic interactions between large language models-based agents in beauty contests. *arXiv preprint arXiv:2404.08492*.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. 2025. Rethinking diverse human preference learning through principal component analysis. *arXiv preprint arXiv:2502.13131*.

Chengdong Ma, Ziran Yang, Hai Ci, Jun Gao, Minquan Gao, Xuehai Pan, and Yaodong Yang. 2023. Evolving diverse red-team language models in multi-round multi-agent games. *arXiv preprint arXiv:2310.00322*.

Hao Ma, Tianyi Hu, Zhiqiang Pu, Liu Boyin, Xiaolin Ai, Yanyan Liang, and Min Chen. 2024. Coevolving with the other you: Fine-tuning llm with sequential cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 37:15497–15525.

Alasdair MacIntyre. 2013. *After virtue*. A&C Black.

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

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*, pages 2845–2866.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Silvia Milano and Sven Nyholm. 2024. Advanced ai assistants that act on our behalf may not be ethically or legally feasible. *Nature Machine Intelligence*, 6(8):846–847.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. 2020. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems*, 33:7308–7320.

Abhilash Mishra. 2023. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. 2024. Secret collusion among AI agents: Multi-agent deception via steganography. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Pekhotin Vladislav, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, and 1 others. 2024. Eai: Emotional decision-making of llms in strategic games and ethical dilemmas. *Advances in Neural Information Processing Systems*, 37:53969–54002.

Remi 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 *International Conference on Machine Learning*, pages 36743–36768. PMLR.

Aran Nayebi. 2025. Barriers and pathways to human-ai alignment: A game-theoretic approach. *arXiv preprint arXiv:2502.05934*.

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, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Barna Pásztor, Thomas Kleine Buening, and Andreas Krause. 2025. Stackelberg learning from human feedback: Preference optimization as a sequential game. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*.

Tianyi Qiu. Representative social choice: From learning theory to ai alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Tianyi Alex Qiu, Yang Zhang, Xuchuan Huang, Jasmine Li, Jiaming Ji, and Yaodong Yang. 2024. Progressgym: Alignment with a millennium of moral progress. *Advances in Neural Information Processing Systems*, 37:14570–14607.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137.

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. *arXiv preprint arXiv:2404.03715*.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, and 1 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*.

Zhekun Shi, Kaizhao Liu, Qi Long, Weijie J Su, and Jiancong Xiao. 2025. Fundamental limits of game-theoretic llm alignment: Smith consistency and preference matching. *arXiv preprint arXiv:2505.20627*.

Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V Nickerson, and Lydia B Chilton. 2025. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1272–1286.

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

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

Qiyang Sun, Yupei Li, Emran Alturki, Sunil Munthumoduku Krishna Murthy, and Björn W Schuller. 2024b. Towards friendly ai: A comprehensive review and new perspectives on human-ai alignment. *arXiv preprint arXiv:2412.15114*.

Reiji Suzuki and Takaya Arita. 2024. An evolutionary model of personality traits related to cooperative behavior using a large language model. *Scientific Reports*, 14(1):5989.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. 2025. Game-theoretic regularized self-play alignment of large language models. *arXiv preprint arXiv:2503.00030*.

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. *arXiv 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. *arXiv preprint arXiv:2407.07796*.

Cuong Tran, Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. 2024. On the effects of fairness to adversarial vulnerability. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 521–529.

Alex Turner and Prasad Tadepalli. 2022. Parametrically retargetable decision-makers tend to seek power. *Advances in Neural Information Processing Systems*, 35:31391–31401.

Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024a. Towards objectively benchmarking social intelligence of language agents at the action level. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8885–8897.

Haochuan Wang, Xiachong Feng, Lei Li, Yu Guo, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. 2024b. Tmgbench: A systematic game benchmark for evaluating strategic reasoning abilities of llms. *arXiv preprint arXiv:2410.10479*.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. 2024c. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655.

Kaiwen Wang, Rahul Kidambi, Ryan Sullivan, Alekh Agarwal, Christoph Dann, Andrea Michi, Marco Gelmi, Yunxuan Li, Raghav Gupta, Avinava Dubey, and 1 others. 2024d. Conditional language policy: A general framework for steerable multi-objective finetuning. *arXiv preprint arXiv:2407.15762*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024e. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. 2024f. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. *arXiv preprint arXiv:2410.16714*.

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

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024g. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023a. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543.

Yibo Wang, Zikun Zhang, Zhihan Liu, Shenao Zhang, and Zhaoran Wang. Provably efficient and practical self-play for better llm alignment.

Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023b. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, and 1 others. 2024h. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*.

Young Wu, Yancheng Zhu, Jin-Yi Cai, and Xiaojin Zhu. 2025. The battling influencers game: Nash equilibria structure of a potential game and implications to value alignment. *arXiv preprint arXiv:2502.01127*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment, 2024b. *URL https://arxiv. org/abs/2405.00675*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

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.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sixiong Xie, Zhuofan Shi, Haiyang Shen, Gang Huang, Yun Ma, and Xiang Jing. 2026. M3-bench: Process-aware evaluation of llm agents social behaviors in mixed-motive games. *arXiv preprint arXiv:2601.08462*.

Nuoya Xiong and Aarti Singh. Projection optimization: A general framework for multi-objective and multi-group rlhf. In *Forty-second International Conference on Machine Learning*.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. 2024a. 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. 2024b. Evolving alignment via asymmetric self-play. In *Language Gamification-NeurIPS 2024 Workshop*.

Ivan Zakazov, Mikolaj Boronski, Lorenzo Drudi, and Robert West. 2024. Assessing social alignment: Do personality-prompted large language models behave like humans? *arXiv preprint arXiv:2412.16772*.

Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. 2024. CPPO: Continual learning for reinforcement learning with human feedback. In *The Twelfth International Conference on Learning Representations*.

Yuheng Zhang, Dian Yu, Tao Ge, Linfeng Song, Zhichen Zeng, Haitao Mi, Nan Jiang, and Dong Yu. 2025. Improving llm general preference alignment via optimistic online mirror descent. *arXiv preprint arXiv:2502.16852*.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and

Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. In *The Thirteenth International Conference on Learning Representations*.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. An electoral approach to diversify llm-based multi-agent collective decision-making. In *EMNLP*, pages 2712–2727.

Rui Zheng, Hongyi Guo, Zhihan Liu, Xiaoying Zhang, Yuanshun Yao, Xiaojun Xu, Zhaoran Wang, Zhiheng Xi, Tao Gui, Qi Zhang, and 1 others. 2024. Toward optimal llm alignments using two-player games. *arXiv preprint arXiv:2406.10977*.

Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. 2024a. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024b. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558.

Duo Zhou, Junyu Zhang, Tao Feng, and Yifan Sun. 2025a. A survey on alignment for large language model agents. In *UIUC Spring 2025 CS598 LLM Agent Workshop; Urbana, IL, USA*.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, and 1 others. 2025b. Socialeval: Evaluating social intelligence of large language models. *arXiv preprint arXiv:2506.00900*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, and 1 others. 2025c. Anyprefer: An agentic framework for preference data synthesis. *arXiv preprint arXiv:2504.19276*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and 1 others. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*.

A Summary Tables

This appendix provides representative methods and evaluation benchmarks for large language models or agent alignment, summarizing them in tabular form (see Tables 1- 4) to facilitate the presentation of details such as equilibrium convergence, self-game, and evaluation metrics, thereby enabling comparisons with similar methods.

Table 1, Table 2, and Table 3 correspond to the three branches of the taxonomy framework in Section 2, with the themes of preference structure, interaction structure, and temporal dynamics, respectively, demonstrating game-based alignment methods. Additionally, Game-based evaluation complements alignment methods by providing controlled settings for probing strategic reasoning, social behavior, and system-level interaction. Table 4 therefore lists representative benchmarks that are most directly related to the survey's main themes. These benchmarks are included as supplementary reference material rather than as a separate organizing axis of the survey.

Table 1: Overview of alignment frameworks based on preference alignment.

	Preference alignment							
Direction	**Framework**	**Regularization**	**Formalization**	**Convergence**	**Preference modeling**	**Self-play?**	**Robust alignment?**	**Online?**
General preference	IPO(Azar et al., 2024)	Explicit	Best response (fixed strategy)	Average-iterate (ϵ-OPT)	○	○	○	○
	NLHF(Munos et al., 2024)	Explicit	Two-player game (constant-sum)	Last-iterate (NE)	●	⊙	○	●
	IPO-MD(Calandriello et al., 2024)	Explicit	Two-player game	Exist (NE)	○	⊙	○	●
	Iterative RLHF(Ye et al., 2024a)	Explicit	Two-player game (mini-max)	Exist (ϵ-NE)	●	●	○	●
	SPO(Swamy et al., 2024)	-	Two-player game (zero-sum)	Average-iterate (ϵ-Minimum Winner)	●	●	○	●
	DNO(Rosset et al., 2024)	-	Two-player game (zero-sum)	Average-iterate (NE)	●	●	○	⊙
	REBEL(Gao et al., 2024)	-	Two-player game (zero-sum)	Average-iterate (ϵ-OPT)	●	●	○	⊙
	SPPO(Wu et al., 2024)	-	Two-player game (constant-sum)	Average-iterate (ϵ-NE)	●	●	○	⊙
	TANPO(Wang et al.)	Explicit	Two-player game (zero-sum)	Average-iterate (ϵ-NE)	○	●	○	●
	ONPO(Zhang et al., 2025)	-	two-player game (zero-sum)	Average-iterate (ϵ-NE)	○	●	○	●
	INPO(Zhang et al.)	Explicit	Two-player game	Last-iterate (ϵ-NE)	○	●	○	●
	COMAL(Liu et al.)	-	Two-player game (zero-sum)	Last-iterate (NE)	●	○	●	●
	MPO(Wang et al., 2025)	-	Two-player game (constant-sum)	Last-iterate (NE)	●	●	○	●
	RSPO (Tang et al., 2025)	Explicit	Two-player game	Last-iterate (NE)	●	●	○	●
Heterogeneous preference	(Chidambaram et al., 2024)	Explicit	Social welfare optimization	Average-iterate (ϵ-OPT)	●	○	●	○
	(Kim et al., 2025)	Explicit	Social welfare optimization	-	●	○	○	○
	MOP(Xiong and Singh)	Explicit	Social welfare optimization	Average-iterate (OPT)	●	○	○	⊙
	Pessimistic Nash Bargaining(Zhong et al., 2024a)	Implicit	Social welfare optimization	Exist (ϵ-OPT)	●	○	○	○
	GRPO(Ramesh et al., 2024)	Implicit	Social welfare optimization	Average-iterate (ϵ-OPT)	●	○	●	○
	CLP (Wang et al., 2024d)	Explicit	Multi-objective optimization	Exist (ϵ-OPT)	●	○	○	●
	Panacea(Zhong et al., 2024b)	Both	Multi-objective optimization	Exist (Pareto-OPT)	●	○	○	●
	MO-ODPO(Gupta et al., 2025a)	Explicit	Multi-objective optimization	-	●	○	○	●
	SIPO(Li et al., 2025)	Explicit	Multi-objective optimization	-	●	○	○	⊙
	DPA(Wang et al., 2024c)	-	Multi-objective optimization	-	●	○	○	⊙
	DRM(Luo et al., 2025)	Explicit	Multi-objective optimization	-	●	○	○	○
	RLHF Game(Sun et al., 2024a)	Explicit	Multi-player game (mechanism design)	-	●	○	○	○
	Dynamic Bayesian Game(Hao and Duan, 2025)	Explicit	Multi-player game (dynamic Bayesian)	Average-iterate (ϵ-OPT)	●	○	○	●
	BIG(Wu et al., 2025)	-	Multi-player game (cooperative)	Last-iterate (NE)	●	○	○	○
	$\langle M, N, \epsilon, \delta \rangle$-Agreement(Nayebi, 2025)	-	Multi-player game (theoretical)	-	○	○	○	●

Notes. (1) *Regularization*: explicit penalties or procedures that constrain the model's policy, learning, and outputs. (2) *Formalization*: how the alignment problem is modeled. (3) *Convergence*: based on theoretical proof or experimental observation, the optimal strategy exists when converging to an equilibrium point, or explicitly reaching a consensus state in decision-making; some methods provide stronger guarantees such as average-iterate and last-iterate convergence. (4) *Preference modeling*: a preference probability model (reward model) is learned from human preference data for candidate outputs. (5) *Self-play*: the same model or homologous variants/clones engage in mutual competition, collaboration, or evaluation within a task or game. (6) *Robust alignment*: a Nash equilibrium policy guarantees at least a 50% win (or preference) rate against any alternative policy under general preferences, and guarantees the worst-group performance under heterogeneous preferences. (7) *Online*: new training data are obtained from the current model and the parameters are updated accordingly. Symbols ●, ○, and ⊙ denote conditions that are met, not met, and adjustable mixed settings, respectively; "-" indicates not applicable or not reported.

Table 2: Overview of alignment frameworks based on interactive enhancement.

	Enhancement via interaction							
Direction	**Framework**	**Target**	**Formalization**	**Interaction structure**	**Convergence**	**External feedback**	**Self-play?**	**Online?**
Synchronous interaction	(Jatova et al., 2024)	Safety alignment	Two-player game	LLM vs adversarial generator	Exist (NE)	●	○	●
	GPO (Zheng et al., 2024)	Safety alignment	Two-player game (zero-sum)	LLMs (attacker vs defender)	Average-iterate (NE)	●	○	●
	SELF-REDTEAM (Liu et al., 2025)	Safety alignment	Two-player game (zero-sum)	LLM (attacker / defender)	-	●	●	●
	MTSA (Guo et al., 2025)	Safety alignment	Multi-turn RL	LLMs (attacker vs defender)	-	●	○	⊙
	SPAG (Cheng et al., 2024)	Reasoning	Two-player game (zero-sum)	LLM (attacker / defender)	-	●	●	○
	APO (Cheng et al., 2023)	Effective alignment	Two-player game (mini-max)	LLM vs RM	-	○	○	●
	SPIN (Chen et al., 2024b)	Self-improvement	Two-player game	LLM (generator / distinguisher)	Exist	○	●	⊙
	LANA (Azarafrooz and Faal, 2024)	Self-alignment	Two-player game (zero-sum)	LLM (generation / evaluation)	Average-iterate (ϵ-NE)	○	●	●
	DuoGuard (Deng et al., 2025)	Safety alignment	Two-player game (mini-max)	LLMs (generator vs classifier)	Last-iterate (NE)	○	○	⊙
	SPC (Chen et al., 2025b)	Reasoning	Two-player game	LLM (generator / critic)	-	○	●	●
	PEG (Chen et al., 2025a)	Truthfulness	Peer elicitation game	LLMs (peer discriminators)	Last-iterate (Truthful NE)	○	○	●
Sequential interaction	STA–RLHF (Makar-Limanov et al., 2024)	Human preference	Two-player game (Stackelberg)	LM (leader) vs RM (follower)	-	○	○	⊙
	SPAC (Ji et al., 2024)	Human preference	Two-player game (Stackelberg)	LLM (policy model / critic)	Average-iterate (ϵ-OPT)	○	●	○
	SGPO (Chu et al., 2025)	Effective alignment	Two-player game (Stackelberg)	LLM vs pref. distribution	Last-iterate (SE)	○	○	⊙
	Anyprefer (Zhou et al., 2025c)	Effective alignment	Two-player game (Markov)	LLMs (target model & judger)	-	●	○	○
	CORY (Ma et al., 2024)	Robustness	Cooperative multi-agent RL	LLM (pioneer / observer)	-	●	●	●

Notes. (1) *Target*: research question. (2) *Formalization*: how the alignment problem is modeled. (3) *Interaction structure*: players and roles engaging in interaction. (4) *Convergence*: based on theoretical proof or experimental observation, the optimal strategy exists when converging to an equilibrium point, or explicitly reaching a consensus state in decision-making; some methods provide stronger guarantees such as average-iterate and last-iterate convergence. (5) *External feedback*: the AI model (agent) relies on externally added, manually annotated data for optimization and updates, rather than self-generating new annotations via interaction. (6) *Self-play*: the same model or homologous variants/clones engage in mutual competition, collaboration, or evaluation within a task or game. (7) *Online*: new training data are obtained from the current model and the parameters are updated accordingly. Symbols ●, ○, and ⊙ denote conditions that are met, not met, and adjustable mixed settings, respectively; "-" indicates not applicable or not reported.

Table 3: Overview of alignment (simulation) frameworks based on dynamic evolution.

	Dynamic evolution						
Direction	**Framework**	**Target**	**Formalization**	**Application**	**Stage**	**Convergence**	**Online?**
Adaptive evolution	(Klassen et al., 2024)	Temporal pluralism values	Non-Markovian process	Measurement	-	-	-
	(Suzuki and Arita, 2024)	Social cooperation	Evolutionary game (mutation)	Simulation	Inference	-	-
	ProgressGym (Qiu et al., 2024)	Social moral progress	Partially observable MDP	Improvement	Post-training	-	-
	Evolutionary Agent (Li et al., 2024b)	Norms and cooperation	Evolutionary game (mutation)	Improvement	Inference	-	-
	EVA (Ye et al., 2024b)	Preference alignment	Creator-solver game (mutation)	Improvement	Post-training	-	●
Co-evolution	SPARTA (Jiang et al., 2025)	Preference / instruction	Collective competition	Improvement	Post-training	-	⊙
	Red-Team Game (Ma et al., 2023)	Safety alignment	Red-blue team game	Improvement	Post-training	Exist (ϵ-NE)	●
	SSI (Liu et al., 2024b)	Social alignment	Social interaction	Improvement	Post-training	Exist	○
	Artificial Leviathan (Dai et al., 2024)	Social order	Social interaction (survival sandbox)	Simulation	Inference	-	-
	CAMEL (Li et al., 2023)	Social cooperation	Role-playing interaction (dialogue)	Simulation	Inference	-	-
	AI Collectives (Lai et al., 2024)	Social alignment	Social interaction (community)	Simulation	Inference	-	-
	Alympics (Mao et al., 2025)	Empirical game theory	Strategic games (survival resources)	Simulation	Inference	-	-
	FairMindSim (Lei et al., 2024)	Social value alignment	Ethical dilemma	Simulation	Inference	-	-
	(Du et al., 2024)	Reasoning / factuality	Multi-agent debate	Improvement	Inference	Exist (Consensus)	-
	MAD (Liang et al., 2024)	Reasoning	Multi-agent debate	Improvement	Inference	-	-

Notes. (1) *Target*: research question. (2) *Formalization*: how the process of dynamic evolution is driven or modeled, such as game, social interaction, and debate. (3) *Application*: quantify and enhance the alignment performance of AI systems, or conduct interdisciplinary research by simulating social behavior and group evolution processes through AI agents, such as social sciences and empirical game theory. (4) *Stage*: the phase during which the training or evaluation framework is running. (5) *Convergence*: based on theoretical proof or experimental observation, the optimal strategy exists when converging to an equilibrium point, or explicitly reaching a consensus state in decision-making. (6) *Online*: new training data are obtained from the current model and the parameters are updated accordingly. Symbols ●, ○, and ⊙ denote conditions that are met, not met, and adjustable mixed settings, respectively; "-" indicates not applicable or not reported.

Table 4: Overview of evaluation benchmarks based on games.

	Evaluation and Benchmark					
Paradigm	**Ref.**	**Target**	**Formalization**	**Metric**	**Benchmark**	**Oriented-O/P**
Classic Games	Fan et al.(Fan et al., 2024)	Rationality	Dictator Game Ring-network Game	Action Accuracy Average Payoff Preference Consistency	-	○
	Wang et al. (Wang et al., 2024b)	Strategic reasoning	Robinson-Goforth Topology of 2×2 Games	Bias Degree Perfect Accuracy Rate Inconsistency Degree	Tmgbench	⊙
	Lor'e et al.(Lorè and Heydari, 2024)	Strategic behavior	Social Dilemmas Games	Cooperation Rate Dominance Statistic	-	⊙
	Topsaka et al. (Topsakal et al., 2024)	Rule comprehension Strategic thinking	Grid-based Games	Win/DQ Rates Invalid Moves Missed Opportunities	✔	⊙
	Lu (Lu, 2024)	Strategic levels Adaptive learning	Beauty Contest	Strategic Degree Convergence Average Payoffs	-	○
	Sreedhar et al. (Sreedhar et al., 2025)	Cooperative prosocial behavior	Public Goods Game	Average Contribution Emergence Frequency	-	⊙
	Akata et al, (Akata et al., 2025)	Cooperation Coordination	Finitely repeated 2×2 games	Score Ratio Defection Rate Coordination Rate	-	⊙
	Fontan et al. (Fontana et al., 2025)	Cooperative behavior	Iterated Prisoner's Dilemma	Behavioral Traits SFEM Scores	✔	⊙
	Xie et al. (Xie et al., 2024)	Human trust behavior	Trust Game	Trust Rate Valid Response Rate Returned/Sent Ratio	✔	⊙
	Zakazo et al. (Zakazov et al., 2024)	Social alignment	Ultimatum Game	Acceptance Rate Disobedience Ratio	-	⊙
	Davidson et al. (Davidson et al., 2024)	Instruction-following Faithfulness	Structured Negotiation	Agreement Rate Normalized Payoffs	LAMEN	⊙
	Xia et al. (Xia et al., 2024)	Bargaining abilities	Bargaining	Deal Rate First Bid Ratio Sum of Normalized Profits	AmazonHistoryPrice	●
Social games	Zhou et al. (Zhou et al., 2025b)	Social intelligence	World tree scenarios	Goal Achievement Ratio Ability Selection Accuracy	SocialEval	⊙
	Zhou et al. (Zhou et al., 2024)	Social intelligence	Mixed-motive Markov games	Goal Completion Social Rules	SOTOPIA-EVAL	⊙
	Liu.et al. (Liu et al., 2024a)	Role consistency	Interactive dialogue	Self-Knowledge accuracy	RoleAgentBench	○
	Wang et al. (Wang et al., 2024g)	Role consistency	Role-playing dialogue	Win rate Average ranking	RoleBench	●
	Zhu et al. (Zhu et al., 2025)	Cooperation & competition	Mutual/conflicting scenarios	Task Score Milestone-based KPI	MultiAgentBench	⊙
	Wang et al. (Wang et al., 2024a)	Social intelligence	Social sandbox tasks	Success Rate Goal Condition SR	STSS benchmark	●
	Abdelnabi et al. (Abdelnabi et al., 2024)	Negotiation & security	Multi-agent, multi-issue negotiation	Final success Collective score Score leakage ratio	✔	⊙
	Xie et al. (Xie et al., 2026)	Social behavior	Mixed-motive games	Overall Score Talk–Action Consistency Trajectory Analysis	M3-Bench	⊙
	Curvo et al. (Curvo, 2025)	Deception & trust	Communication utterances Voting decisions	Traitor Survival Rate Trust Network Stability Faithful Correctness Rate	✔	⊙
	Motwani et al. (Motwani et al., 2024)	Secret collusion	Covert communication	Cipher round-trip accuracy Coordination rates	✔	⊙

Notes. (1) *Target*: research question. (2) *Formalization*: evaluation scenario or benchmark type. (3) *Metric*: quantifiable evaluation metrics that are typically results- or process-oriented. (4) *Benchmark*: self-named benchmarks are listed explicitly, while unnamed benchmarks are represented by ✔. (5) *Oriented-O/P*: the evaluation paradigm follows outcome-oriented (O: ●), process-oriented (P: ○), or both (⊙). The symbol "-" is used in the same way as above.