Dynamics of Adversarial Attacks on Large Language Model-Based Search Engines

Xiyang Hu [1]

Abstract

The increasing integration of Large Language Model (LLM) based search engines has transformed the landscape of information retrieval. However, these systems are vulnerable to adversarial attacks, especially ranking manipulation attacks, where attackers craft webpage content to manipulate the LLM's ranking and promote specific content, gaining an unfair advantage over competitors. In this paper, we study the dynamics of ranking manipulation attacks. We frame this problem as an Infinitely Repeated Prisoners' Dilemma, where multiple players strategically decide whether to cooperate or attack. We analyze the conditions under which cooperation can be sustained, identifying key factors such as attack costs, discount rates, attack success rates, and trigger strategies that influence player behavior. We identify tipping points in the system dynamics, demonstrating that cooperation is more likely to be sustained when players are forward-looking. However, from a defense perspective, we find that simply reducing attack success probabilities can, paradoxically, incentivize attacks under certain conditions. Furthermore, defensive measures to cap the upper bound of attack success rates may prove futile in some scenarios. These insights highlight the complexity of securing LLM-based systems. Our work provides a theoretical foundation and practical insights for understanding and mitigating their vulnerabilities, while emphasizing the importance of adaptive security strategies and thoughtful ecosystem design.

1. Introduction

The integration of large language models (LLMs) into search engines creates vulnerabilities to ranking manipulation attacks (Tang et al., 2025; Xing et al., 2025; Du et al., 2026). These attacks embed crafted instructions or misleading content in retrieved documents, exploiting prompt sensitivity (Zhao et al., 2023; Hu et al., 2024b; Xhonneux et al., 2024; Han et al., 2024; Hu et al., 2024a) to bias generated answers toward targeted content or products (Pfrommer et al., 2024; Aggarwal et al., 2024).

Ranking manipulation in LLM-based search differs from traditional search engine optimization (SEO). In traditional search, manipulation mainly changes the similarity score between a query and a single document, affecting that document's rank (Sharma et al., 2019). In LLM-based search, retrieved documents are jointly placed in the prompt, so one manipulated document can change how the model interprets and prioritizes other documents. A local manipulation can therefore distort the entire generated response. Because providers compete for visibility, such attacks also create strategic interactions: one provider's action changes the payoffs and incentives of others. Defenses therefore must account for market incentives, not only detection accuracy.

We propose a game-theoretic model of these interactions and frame the problem as an Infinitely Repeated Prisoners' Dilemma (IRPD). In each round, competing content providers choose whether to cooperate, by refraining from ranking attacks, or defect, by launching attacks to gain market share. Cooperation leads to a fair division of demand, while defection attempts to shift rankings in favor of the attacker.

The model includes three features specific to LLM-based search. First, attacks succeed stochastically. We represent this uncertainty through attack success rates (ASRs), since LLM outputs are probabilistic (Bengio et al., 2003) and prior work on prompt sensitivity and adversarial examples shows that manipulations do not always succeed (Perez et al., 2022; Wallace et al., 2019). Second, attacks incur costs: crafting and deploying adversarial content requires resources that may increase with attack sophistication. Third, simultaneous successful attacks can degrade the market by

[1]Arizona State University, Tempe, AZ, USA. Correspondence to: Xiyang Hu <xiyanghu@asu.edu>.

New Frontiers in Game-Theoretic Learning Workshop, The 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

introducing conflicting or low-quality manipulation signals, reducing user trust and shrinking effective demand (Nestaas et al., 2024).

Using this model, we derive conditions under which cooperation can be sustained. We begin with two symmetric providers and then study alternative trigger strategies, heterogeneous players, and multi-player settings with many defectors. The analysis shows that cooperation is easier to sustain when attack costs are high or players place sufficient weight on future profits. The effect of ASR is non-monotone: intermediate ASRs can create the strongest defection incentives because they offer sizable gains while avoiding the high costs and degradation risks of very high ASRs. We also identify "futile defense regions," where capping ASRs has little effect on defection incentives, showing that defenses must target both technical attack success and economic incentives.

Our main contribution is a game-theoretic framework for ranking attacks in LLM-based search engines. The framework gives cooperation conditions under several strategic settings and yields practical guidance for reducing adversarial behavior. The results also apply to other LLM-mediated platforms, including recommendation systems and answer engines, where providers compete for model-generated visibility.

2. Model Setup

We model ranking manipulation in LLM-based search as an infinitely repeated game (Abreu, 1988; Dal Bó & Fréchette, 2018). In each period $t = 1, 2, 3, \cdots$, two content providers $i = 1, 2$ simultaneously choose whether to launch a ranking manipulation attack. The stage game has the structure of an IRPD: each provider has a short-term incentive to attack to gain an advantage over its competitor, but mutual restraint can yield higher long-run payoffs. Players discount future profits by a common factor $\delta \in (0, 1)$.

Let p denote the ASR, i.e., the probability that an attack alters the model response to raise the rank of the targeted product. ASR is a standard metric in LLM safety for measuring adversarial attack effectiveness (Shayegani et al., 2023). Its value depends on attack sophistication, defense strength, and attacker resources: stronger attacks tend to increase p, while stronger defenses reduce it.

An attacking player incurs cost c, representing the resources needed to develop, deploy, and maintain manipulative content, including engineering effort, data collection, and infrastructure. We consider constant, linear, and nonlinear cost functions, including $c \propto p$ and $c \propto p^k$ for $k > 1$. The nonlinear case captures the increasing difficulty of improving high-ASR attacks.

Player 1 \ Player 2	Cooperate	Attack
Cooperate	R, R	S, T
Attack	T, S	Q, Q

Figure 1. Payoff Matrix

In each period, the market has a unit mass of potential consumers, so total demand is normalized to 1. The providers' goods are perfect substitutes, allowing the model to focus on how strategic actions affect market-share allocation. At the end of each period, each player observes the other player's action, yielding a perfect-monitoring repeated game (Fudenberg, 1991). This is a simplifying baseline. In deployed systems, providers may observe only public signals, such as ranking positions, citations, referral traffic, or changes in generated answers. We use perfect monitoring to isolate the incentive effects of p, c, β, and δ, and because it is the most favorable case for sustaining cooperation. If cooperation is difficult when deviations are observed, noisy public signals would make discipline harder. An imperfect-monitoring extension would replace observed actions with public signals over ranking outcomes and require stronger continuation incentives.

The strategic interactions between the players are represented by a payoff matrix, structured as a Prisoners' Dilemma, as summarized in Table 1. Within each cell of the table, the first element denotes the payoff of player 1, and the second element denotes that of player 2. The payoff values depend on whether each player chooses to cooperate or launch an attack:

• $R = \frac{1}{2}$ (**Mutual Cooperation**): If both players refrain from launching attacks, they equally share the market demand, resulting in a payoff of $\frac{1}{2}$ each.

• $T = p + (1-p)\frac{1}{2} - c$ (**Temptation Payoff**): When one player launches an attack while the other cooperates, the attacker potentially captures the entire market if the attack is successful with probability p. If the attack fails (with probability $1 - p$), the market demand is split evenly, but the attacker still bears the cost c of launching the attack.

• $S = (1-p)\frac{1}{2}$ (**Sucker Payoff**): If a player cooperates while the other attacks, the cooperator retains some market share only if the attack fails. Otherwise, the cooperator loses all market share.

• $Q = p^2\frac{1}{2}\beta + p(1-p) + (1-p)^2\frac{1}{2} - c$ (**Mutual Attack**): If both players launch attacks, the outcomes depend on the success rate of the attacks. (1) If both attacks succeed (probability p^2), they equally share the market but at a degraded value due to reduced LLM output quality, represented by $\beta < 1$. A smaller β indicates a larger degradation. (2) If only one player succeeds in attacking (probability $p(1-p)$), that player monopolizes the market. (3) If both attacks fail, the market is split evenly. The condition $\beta < 1$, which repre-

sents the degraded output quality when both parties launch attacks, aligns with empirical findings showing that when all parties engage in ranking manipulation attacks, it results in detrimental outcomes for everyone involved (Nestaas et al., 2024).

The condition $T > R > Q > S$ preserves the structure of a Prisoner's Dilemma, indicating that while mutual cooperation is preferable, individual incentives drive the players to defect. This condition holds when $c < \frac{1}{2}p + \frac{1}{2}(\beta - 1)p^2$, ensuring that the temptation to attack outweighs the costs but results in a worse outcome for both when both attack.

3. Analysis

Based on the model, we derive the conditions under which cooperation can be sustained among players. We answer key questions regarding what factors influence the sustainability of cooperation, and how varying parameters—such as attack costs, attack success rates, and future profits discount rates—affect the decision to either cooperate or defect. This analysis provides insights into the mechanisms that can encourage cooperation and mitigate ranking manipulation attacks in LLM-driven information retrieval systems. All proofs are included in Appendix H.

We consider the grim trigger strategy, a classic trigger strategy in repeated games where players cooperate until one defects; after a defection, all players respond by always defecting. This strategy serves as a baseline for understanding how cooperation might be enforced in an environment where deviations are possible. It creates a strong deterrent for initial defection since any deviation leads to permanent mutual defection, which can degrade outcomes for all parties involved. In Appendix D, we exam alternative trigger strategies; in Appendix E, we exam what if the attack cost is one-time.

We first derive the discounted payoffs for continuous cooperation $V(C)$ and for a one-time defection followed by mutual defection $V(D)$. If both players always cooperate, the discounted payoff for each player is: $V(C) = R + \delta R + \delta^2 R + \cdots = \frac{R}{1-\delta}$. If the first defecting player defects once while the other cooperates, and then both players defect forever after, the payoff for the first defecting player is: $V(D) = T + \delta Q + \delta^2 Q + \cdots = T + \frac{\delta Q}{1-\delta}$. Here, R, T, and Q represent the payoffs for mutual cooperation, defection, and mutual defection, respectively, while δ represents the discount rate that measures how players value future profits relative to immediate gains. $V(D)$ is a combination of the immediate gain from defection and the discounted future payoffs under mutual defection.

3.1. Condition for Cooperation

For cooperation to be a rational strategy for each player, the payoff from continuous cooperation must be at least as high as the payoff from defection followed by mutual defection $V(C) \geq V(D)$. Theorem 3.1 summarizes the result.

Theorem 3.1 (Cooperation Condition)**.** *Two players prefer long-term cooperation over engaging in ranking manipulation attacks if and only if: $\delta \geq \delta^* = \frac{T-R}{T-Q} = \frac{p-2c}{p-\beta p^2+p^2}$, where δ^* is the critical discount factor.*

The discount factor reflects how much players value future profits relative to immediate gains. Higher values of δ imply that players are more forward-looking and are thus more willing to cooperate because attacking now would mean losing significant future profits. Conversely, a lower δ makes attacking more appealing, as players prioritize immediate rewards gained from ranking attacks over long-term benefits. Cooperation is only viable if $\delta \geq \delta^*$, emphasizing the need for creating environments where long-term outcomes are valued.

Corollary 3.2. *The cooperation will be sustained if and only if the cost is larger than a threshold: $c \geq \frac{p-\delta(p-\beta p^2+p^2)}{2}$.*

Corollary 3.2 suggests that a higher attack cost reduces the attractiveness of attacking, as the immediate gain from a successful attack is offset by a substantial cost to attack.

In Theorem 3.3, we analyze how different parameters affect the range of the cooperation-inducing δ and discuss their broader implications for cooperative behavior in the system.

Theorem 3.3 (Monotonicity of δ^*)**.** *The critical discount factor δ^* exhibits the following behavior:*

- *δ^* decreases as the attack cost c increases. This implies that higher attack costs make cooperation easier to maintain, as the relative benefit of defecting diminishes.*
- *δ^* increases with larger β. When β is high, the payoffs from mutual defection are relatively high, making defection more attractive and cooperation more difficult to sustain.*
- *δ^* is non-monotonic with respect to the attack success rate p. This means that the relationship between p and δ^* is not straightforward—an increase in p may either raise or lower the likelihood of sustaining cooperation, depending on the interplay of other parameters like c and β.*

We find that the cost of attack c plays a critical role in fostering cooperation by lowering the critical discount factor δ^*. An increase in c discourages players from defecting, as the expense of launching an attack outweighs its short-term benefits. This insight highlights the importance of measures that raise the cost of malicious actions, such as enhancing LLM security technologies or implementing stricter legal penalties. By making defection more costly, these inter-

ventions shift incentives in favor of long-term cooperative behavior, creating a more stable ecosystem.

The degradation factor β also significantly affects the sustainability of cooperation. It reflects how much the market value diminishes when mutual defection occurs, which degrade the performance of LLMs. As β increases, the payoffs from mutual defection (when both players attack) become more attractive, which undermines the incentive to cooperate. Lowering β effectively widens the range of cooperation inducing δ, making cooperation more likely to be sustained. In practical terms, this suggests that systems should be designed to heavily penalize mutual defection. For instance, designing LLM-driven search engines where mutual defection leads to severe degradation in output quality can reduce the benefits of defection, thereby promoting cooperation.

The attack success rate p introduces complexity to the dynamics of cooperation due to its non-monotonic relationship with δ^*. It affects both the temptation to defect and the risks associated with defection. When p is low, attacks are less likely to succeed, which may disincentivize players from defecting. However, as p increases, the temptation to attack rises because a successful attack yields significant short-term gains. Interestingly, beyond a certain threshold, further increases in p can counter-intuitively reduce the temptation to defect, because the mutual defection that follows successful attacks degrades the market significantly and the cost of developing a stronger attack becomes higher, both reducing the attractiveness of defection. This nuanced interplay indicates that managing p requires careful calibration, as its effects on cooperation depend on its interactions with other parameters such as c and β.

These findings underscore the broader implications of parameter manipulation for designing cooperative systems. Raising the cost of attacks, lowering the payoffs from mutual defection, and carefully managing the attack success probability can together create an environment conducive to sustained cooperation. In the following sections, we delve deeper into how these parameters interact and explore practical strategies for implementing these insights in real-world LLM-driven ecosystems.

3.2. Cooperation Formation Region

Due to the lack of a closed-form solution for the cooperation formation condition in terms of p, we employ numerical visualization to investigate the parameter space of sustained cooperation. Figure 2 visualizes the regions of the δ (discount factor) and p (attack success probability) space that results in long-term cooperation under various values of β and different cost functions. The region to the right of the curve (the blue region) is where $\delta > \frac{p-2c}{p-\beta p^2+p^2}$, *i.e.*, the cooperation formation region. We can derive several key observations from these plots.

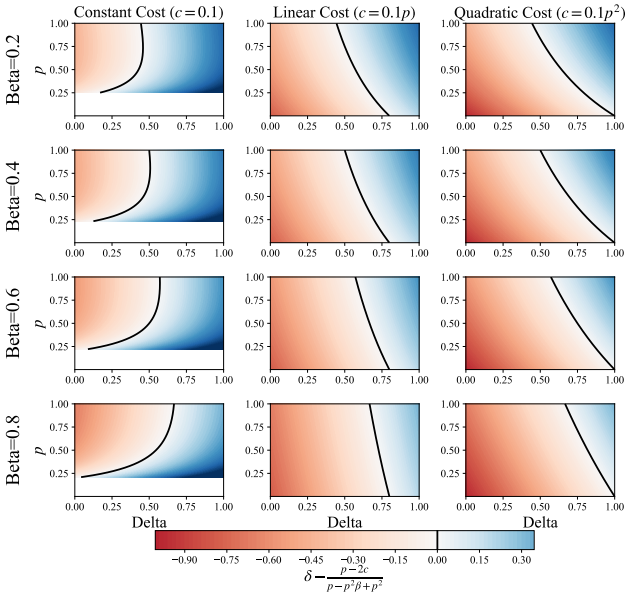

Figure 2. Region of cooperation formation (the region to the right of the boundary).

First, across all cost functions, the region where cooperation is possible tends to shrink as β increases. This is evident as the blue regions, where the inequality condition is satisfied, become smaller moving from the first row (where $\beta = 0.2$) to the fourth row (where $\beta = 0.8$). The decrease in the size of the cooperation region with increasing β can be attributed to the increasing sensitivity of the players to defection payoffs. As β increases, the loss of degradation decreases, therefore the incentive for defection grows, thereby requiring higher values of δ to sustain cooperation.

Second, the size and shape of the cooperation region vary significantly across different cost functions. For the constant cost function ($c = 0.1$), the cooperation region is relatively larger across different values of δ. In contrast, with a linear cost function ($c = 0.1p$), the cooperation region becomes smaller, as the cost is overall smaller than the linear setting. Finally, for the quadratic cost function ($c = 0.1p^2$), the cooperation region is the smallest, as the cost values are even smaller compared to the linear and constant cases. This minimal cost significantly diminishes the likelihood of maintaining cooperation. While the magnitude of the cost primarily determines the overall size of the cooperation region, the form of the cost function governs the shape of the cooperation boundary. The curvature and slope of the boundary reflect the sensitivity of cooperation to changes in δ, β, p.

Third, there exists a lower bound for δ, below which cooperation is not feasible. This lower bound represents the minimum level of patience (or preference for future rewards) required for players to consider cooperating. Interestingly, this lower bound for δ may increase or decrease as p increases, depending on the specific cost function c and value

of β. In some scenarios, for example, the subfigures in the second and third columns, a higher success probability p makes it easier to maintain cooperation, thus lowering the minimum required δ. In other cases, like the first column subfigures, increasing p might raise the lower bound for δ, making cooperation harder to achieve unless players are sufficiently future-oriented. In addition, for some settings where β is small, such as the top left subfigure, although increasing p disincentive cooperation at the beginning, after a certain point, p further can unexpectedly decrease the motivation to defect, as the increasing success of attacks leads to a larger loss of mutual defection, which harms the market and diminishes the appeal of defection.

In summary, the ability to sustain long-term cooperation depends critically on the interplay between the discount factor δ, success probability p, and the form of the cost function c. Cooperation is more likely when the discount factor is high, and the cost is high. As the cost decreases, the conditions for cooperation become more stringent, shrinking the feasible region for long-term cooperation. These findings suggest that both the design of cost structures and the understanding of players' time discount sensitivities are crucial for fostering stable cooperative relationships.

3.3. Payoff Analysis of Cooperation and Defection in LLM Systems

In this section, we examine the payoff values for cooperation V_C and defection V_D, focusing on how varying factors such as attack success probabilities p affect strategic decisions. Figure 3 illustrates the payoffs for cooperation V_C and defection V_D as p varies. In Figure 3, we fix β at 0.4 because the pattern of the cooperation region is relatively consistent across β values, as we see in Figure 2. Cooperation is sustainable when the V_C curve lies above the V_D curve, indicating that the long-term benefits of maintaining cooperative behavior outweigh the gains from defection.

Our analysis uncovers several non-intuitive findings that have significant implications for the security design of LLM-based systems. These insights challenge conventional assumptions about how to maintain cooperation and deter adversarial behavior.

Proposition 3.4 (Defection payoff and cap-based defenses)**.** *Fix $\delta \in (0,1)$, $\beta \in (0,1)$, and let $c(p)$ be a differentiable attack cost. Under Grim Trigger, the one-shot defection payoff followed by mutual defection is*

$$V_D(p) = \frac{1}{2} + \frac{p}{2} - c(p) + \frac{\delta}{1-\delta}\left(\frac{1}{2} - \frac{(1-\beta)p^2}{2} - c(p)\right).$$

The payoff gap between defection and cooperation is

$$V_D(p) - V_C = \frac{(1-\delta)p - \delta(1-\beta)p^2 - 2c(p)}{2(1-\delta)}.$$

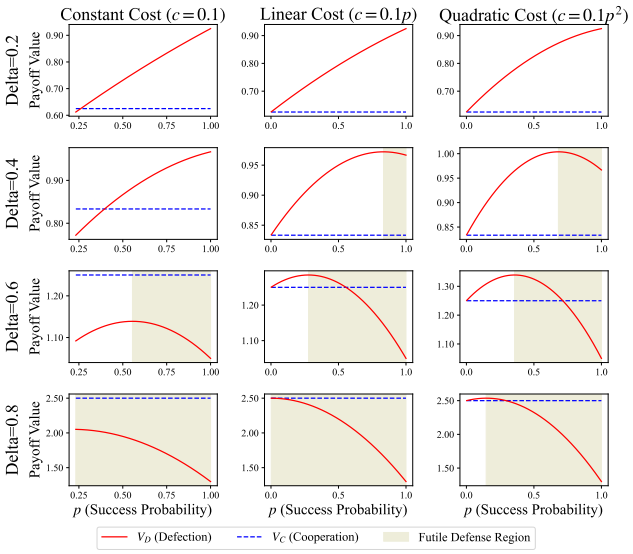

Figure 3. V_C and V_D values ($\beta = 0.4$)

If $c''(p) \geq 0$, then $V_D(p)$ is concave in p. Any interior maximizer $p^\star$ is the unique solution to

$$c'(p^\star) + \delta(1-\beta)p^\star = \frac{1-\delta}{2}.$$

If no solution lies in $(0,1)$, the maximum is attained at an endpoint.

For the cost functions used in the figures:

$$c(p) = c_0 \quad \Rightarrow \quad p^\star = \frac{1-\delta}{2\delta(1-\beta)},$$
$$c(p) = \alpha p \quad \Rightarrow \quad p^\star = \frac{(1-\delta)/2 - \alpha}{\delta(1-\beta)},$$
$$c(p) = \alpha p^2 \quad \Rightarrow \quad p^\star = \frac{1-\delta}{2(2\alpha + \delta(1-\beta))},$$

whenever the displayed value belongs to $(0,1)$.

Corollary 3.5 (Non-monotone effect of reducing p)**.** *When an interior maximizer $p^\star$ exists, reducing p can increase the incentive to defect. In particular, for $p > p^\star$, a local decrease in p raises $V_D(p)$ because $V_D'(p) < 0$ on that side of the peak.*

Corollary 3.6 (Low-success attacks can remain attractive)**.** *If $c(0) = 0$ and $c'(0) < (1-\delta)/2$, then there exists $\varepsilon > 0$ such that $V_D(p) > V_C$ for all $p \in (0, \varepsilon)$. Thus, driving p close to zero does not by itself guarantee strict cooperation when low-success attacks have near-zero marginal cost. If $c(0) > 0$, this conclusion need not hold.*

Corollary 3.7 (Futile defense region)**.** *Suppose an attacker can choose any $p \in [0, \bar{p}]$ after a defense imposes an upper bound $\bar{p}$ on attainable attack success. If $V_D(p^\star) > V_C$ and $\bar{p} \geq p^\star$, then $\max_{p \in [0,\bar{p}]} V_D(p) = V_D(p^\star) =$*

$\max_{p\in[0,1]} V_D(p)$. *Hence, the cap does not reduce the maximum attainable defection payoff. A cap-based defense can change the attacker's best-response payoff only if it lowers the attainable upper bound below $p^\star$, and it restores cooperation only if* $\max_{p\in[0,\bar{p}]} V_D(p) \le V_C$.

The findings from our analysis emphasize the necessity of a system-level perspective when designing defenses for LLM-based platforms. Instead of solely focusing on reducing the attack success probability p, effective strategies must account for the interplay between p, the market degradation factor β, the discount rate δ, and the attack cost c. These parameters collectively shape the strategic decisions of both attackers and defenders. For example, dynamic defense mechanisms that adjust the LLM's response strategies based on detected attack patterns could mitigate persistent threats by introducing variability that increases the cost of launching repeated attacks. Moreover, integrating measures that penalize attackers—whether or not their attacks succeed—can shift the cost-benefit analysis against continued adversarial attempts. This could involve mechanisms that reduce the utility of even low-probability attacks, thereby discouraging persistent efforts to exploit the system. By balancing these considerations, LLM developers can design more resilient systems capable of maintaining stability in the face of evolving adversarial tactics.

4. Conclusion

This paper presents a game-theoretic modeling of ranking manipulation attacks in LLM-based search engines, where content providers strategically decide whether to engage in manipulative practices to gain competitive advantages. By modeling these interactions as an IRPD, we capture the unique characteristics of LLM-based systems, including stochastic attack success rates, attack costs, future discount rates, and market degradation effects. Our analysis reveals several key insights that have direct implications for platform designs, industry practices, and regulatory policies.

First, we find that cooperation sustainability depends critically on the interplay between immediate costs and long-term benefits. Content providers with high attack costs or strong forward-looking tendencies are more likely to maintain cooperative behavior, as the immediate gains from manipulation are outweighed by long-term market benefits. In practice, this suggests that platform operators should implement both immediate and long-term deterrence mechanisms. For example, search engines could impose escalating computational costs for realizing ranking manipulations, while simultaneously developing reputation systems that reward long-term cooperative behavior. Second, we discover that the relationship between attack success probability and cooperation sustainability is non-monotonic. Intermediate success rates can sometimes incentivize more manipulation attempts than high rates, as they provide an optimal balance between potential gains and the combined loss of costs and degradation risks. This presents a crucial challenge for platform operators, suggesting that partial defenses might paradoxically be futile. Defensive measures aimed at capping attack success rates fail to meaningfully reduce manipulation incentives. This finding fundamentally challenges traditional approaches to platform security. Rather than investing heavily in technical measures within these futile defense regions, platforms should redirect resources toward economic deterrence mechanisms and reputation systems that create long-term incentives for cooperation.

In examining asymmetric player scenarios, we find that system stability is primarily determined by the player with the strongest incentive to defect. This suggests that platforms should focus their monitoring and enforcement efforts on the most potential attackers. Platforms could implement tiered security measures that apply stricter scrutiny to larger content providers or those with a history of sophisticated ranking manipulation practices. Additionally, platforms might consider implementing compensatory mechanisms that help level the playing field between players with different capabilities, thereby reducing the incentive for more capable players to exploit their advantages.

The policy implications can be read directly through the model parameters. Defenses that make attacks harder to engineer, such as adversarial-content detection, randomized ranking audits, or stronger separation between retrieved data and model instructions, primarily reduce p. Measures that increase the effort needed to develop, test, and maintain attacks, such as rate limits, adaptive audits, and escalating review for repeated suspicious edits, increase c. Enforcement mechanisms that demote or remove pages involved in detected simultaneous manipulation lower the effective mutual-defection payoff Q, which in our reduced-form model corresponds to a lower effective β from the attackers' perspective. Reputation systems, long-term contracts, and persistent audit records increase the value of future cooperative payoffs relative to short-run manipulation gains, which corresponds to a higher effective δ. The non-monotonicity results show that reducing p alone may not be enough; a defense is more reliable when it jointly lowers attainable p, raises c, and reduces the continuation payoff from mutual defection.

Future research could explore several promising directions that are not addressed in this paper. One avenue is extending the model to two-sided markets, capturing strategic interactions between content providers and platform operators under manipulation and defense dynamics. Another is incorporating user behavior, market feedback, and information asymmetry to understand how manipulation affects trust, long-term viability, and strategy under uncertainty.

References

Abreu, D. On the theory of infinitely repeated games with discounting. *Econometrica: Journal of the Econometric Society*, pp. 383–396, 1988.

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., and Deshpande, A. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5–16, 2024.

Alpcan, T. and Başar, T. *Network security: A decision and game-theoretic approach*. Cambridge University Press, 2010.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Chen, X., Nie, Y., Guo, W., and Zhang, X. When llm meets drl: Advancing jailbreaking efficiency via drl-guided search. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 26814–26845. Curran Associates, Inc., 2024.

Dal Bó, P. and Fréchette, G. R. On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114, 2018.

Du, Y., Du, Y., Yu, C., Xu, H., Wang, Z., Zhao, Y., and Hu, X. Multimodal generative engine optimization: Rank manipulation for vision–language model rankers. *Available at SSRN 5917963*, January 2026. URL https://ssrn.com/abstract=5917963.

Fudenberg, D. *Game theory*. MIT press, 1991.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8093–8131. Curran Associates, Inc., 2024.

Hu, K., Yu, W., Li, Y., Yao, T., Li, X., Liu, W., Yu, L., Shen, Z., Chen, K., and Fredrikson, M. Efficient llm jailbreak via adaptive dense-to-sparse constrained optimization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23224–23245. Curran Associates, Inc., 2024a.

Hu, X., Li, X., Chen, J., Li, Y., Li, Y., Li, X., Wang, Y., Liu, Q., Wen, L., Yu, P., and Guo, Z. Evaluating robustness of generative search engine on adversarial factoid questions. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10650–10671, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.633. URL https://aclanthology.org/2024.findings-acl.633.

Jones, E., Dragan, A., Raghunathan, A., and Steinhardt, J. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*, pp. 15307–15329. PMLR, 2023.

Kamhoua, C. A., Kiekintveld, C. D., Fang, F., and Zhu, Q. *Game theory and machine learning for cyber security*. John Wiley & Sons, 2021.

Manshaei, M. H., Zhu, Q., Alpcan, T., Bacşar, T., and Hubaux, J.-P. Game theory meets network security and privacy. *ACM Computing Surveys (CSUR)*, 45(3):1–39, 2013.

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., and Karbasi, A. Tree of attacks: Jailbreaking black-box llms automatically. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 61065–61105. Curran Associates, Inc., 2024.

Nestaas, F., Debenedetti, E., and Tramèr, F. Adversarial search engine optimization for large language models. *arXiv preprint arXiv:2406.18382*, 2024.

Pal, A. and Vidal, R. A game theoretic analysis of additive adversarial attacks and defenses. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1345–1355. Curran Associates, Inc., 2020.

Pawlick, J., Colbert, E., and Zhu, Q. A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys*, 52(4), August 2019. doi: 10.1145/3337772. URL https://doi.org/10.1145/3337772.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, 2022.

Pfrommer, S., Bai, Y., Gautam, T., and Sojoudi, S. Ranking manipulation for conversational search engines. *arXiv preprint arXiv:2406.03589*, 2024.

Roy, S., Ellis, C., Shiva, S., Dasgupta, D., Shandilya, V., and Wu, C. A survey of game theory as applied to network security. *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10, 2010.

Sharma, D., Shukla, R., Giri, A. K., and Kumar, S. A brief review on search engine optimization. In *2019 9th international conference on cloud computing, data science & engineering (confluence)*, pp. 687–692. IEEE, 2019.

Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

Tang, Y., Fan, Y., Yu, C., Yang, T., Zhao, Y., and Hu, X. Stealthrank: Llm ranking manipulation via stealthy prompt optimization, 2025. URL https://arxiv.org/abs/2504.05804.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. Universal adversarial triggers for attacking and analyzing NLP. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221.

Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., and Goldstein, T. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

Xhonneux, S., Sordoni, A., Günnemann, S., Gidel, G., and Schwinn, L. Efficient adversarial training in llms with continuous attacks. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 1502–1530. Curran Associates, Inc., 2024.

Xing, T., Li, J., Du, Y., and Hu, X. Are llms reliable rankers? rank manipulation via two-stage token optimization, 2025. URL https://arxiv.org/abs/2510.06732.

Zhao, S., Wen, J., Luu, A., Zhao, J., and Fu, J. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12303–12317, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.757. URL https://aclanthology.org/2023.emnlp-main.757.

A. Related Literature

Our work draws upon and connects several streams of research: (1) vulnerabilities and security challenges in LLM, (2) ranking manipulation in LLM-based search engines, (3) applications of game theory to security problems, and (4) strategic interactions in AI-driven markets.

Vulnerabilities in Large Language Models. The vulnerability of LLMs to adversarial inputs forms the technical foundation for ranking manipulation attacks in LLM-based search engines. Early work demonstrated that universal adversarial triggers can dramatically alter language model outputs (Wallace et al., 2019), establishing the basis for exploiting these models' sensitivity to carefully crafted inputs. Subsequent studies showed that gradient-informed optimization techniques can generate adversarial inputs that consistently bypass LLM safety measures (Jones et al., 2023; Wen et al., 2024). More recently, jailbreaking has been formulated as a reinforcement learning–guided search problem (Chen et al., 2024). Related work further introduced a tree-based black-box jailbreaking method that automatically refines prompts and achieves high attack success rates (Mehrotra et al., 2024).

Ranking Manipulation in LLM-enhanced Search.

One of the first analyses of ranking manipulation in LLM-enhanced search showed that adversaries can use carefully crafted external content, such as website text or plugin documentation, to influence LLMs into promoting specific preferences or products (Nestaas et al., 2024). This line of work was extended by identifying concrete content structures and semantic patterns that are particularly effective at influencing LLM-based ranking systems (Pfrommer et al., 2024). Further advances proposed optimization-based attack techniques for manipulating LLM preferences in search contexts (Aggarwal et al., 2024; Tang et al., 2025; Xing et al., 2025; Du et al., 2026). Collectively, these studies indicate that ranking manipulation in LLM-based systems differs fundamentally from traditional SEO, as it relies on exploiting deep language understanding and generation mechanisms rather than surface-level keyword matching. While existing work establishes the technical feasibility of such attacks, it primarily focuses on isolated, one-shot attack instances. Our work complements this literature by analyzing the long-term strategic dynamics that arise when multiple attackers repeatedly interact within the same market.

Game Theory in Security Applications. Our game-theoretic approach builds on established frameworks for analyzing security problems. Foundational surveys demonstrated how game-theoretic modeling can illuminate attacker–defender dynamics across a wide range of security domains (Manshaei et al., 2013). Early work on network security further established the applicability of game theory to strategic defense and attack problems (Alpcan & Başar, 2010). These principles were later extended to cyber-security systems by incorporating machine learning components, providing insights relevant to LLM-based platforms (Kamhoua et al., 2021). More recent studies have applied game theory to specific security scenarios. Prior work on defensive deception showed how game-theoretic models can capture settings in which attackers and defenders reason explicitly about each other's strategies (Pawlick et al., 2019). Additional analyses demonstrated how strategic security investments and defense mechanisms can be modeled within a game-theoretic framework (Roy et al., 2010). Related work also presented a game-theoretic analysis of additive adversarial attacks and defenses, introducing geometry-based proof techniques for deriving provable guarantees (Pal & Vidal, 2020). Our work extends these approaches to the setting of LLM-based search engines by incorporating stochastic attack success rates, performance degradation under universal attacks, and the strategic interdependence among multiple attackers.

B. Asymmetric Players Scenarios and Multi-Player Scenarios

In real-world settings, players often differ in their strategic capabilities, such as attack success probabilities, costs, or patience levels. We analyze such *asymmetric scenarios* by extending our baseline model to include heterogeneity in these dimensions. Our results show that the sustainability of cooperation is governed by the player with the greatest temptation to defect—typically the one with the highest success probability, the lowest attack cost, or the most myopic discount rate. We further generalize these results to cases where asymmetries arise along multiple dimensions simultaneously. A complete derivation of these conditions is provided in Appendix F. We also extend our framework to *multi-player scenarios*, where cooperation dynamics depend on the number of defectors among N total players. We derive explicit cooperation conditions under both Grim Trigger and Tit-for-Tat strategies and show that, counterintuitively, cooperation becomes easier to sustain as more players defect. This is because the marginal benefit from defection diminishes with increased competition among attackers, while the threat of long-term punishment persists. Formal analysis and closed-form expressions are given in Appendix G.

The most important implication of the asymmetric analysis is that system stability is governed by the player with the largest

deviation incentive, not by the average player. In the two-player asymmetric model, cooperation requires both players' incentive constraints to hold. For example, when players differ in both success probabilities and costs, player i must satisfy

$$\delta_i \geq \frac{p_i - 2c_i}{(1-\beta)p_i p_j + p_j}, \qquad j \neq i.$$

A single player with high p_i, low c_i, or low δ_i can therefore determine whether cooperation fails. This matters for defense design because uniform defenses may have limited effect if they reduce the risk of already-safe players while leaving the most attack-prone player nearly unchanged. The model instead points to targeted controls: lower p_i for high-capability attackers, raise c_i for low-cost attackers, and increase the future loss from defection for myopic attackers.

The multi-player results should be interpreted in the same conditional way. As the number of defectors grows, the marginal payoff from being one additional attacker can fall because successful attackers divide the market among more competitors. However, this comparative static does not mean that cooperation naturally appears when many attackers exist. The threshold result assumes that the remaining players use the specified trigger strategy and that deviations are monitored as assumed by the model. In practice, more attackers can still reduce market quality and increase attack volume unless the platform can enforce credible future punishment. Thus, the asymmetric and multi-player analyses both point to the same defense principle: identify the actors with the highest deviation payoff and apply parameter-specific interventions to them rather than relying only on system-wide caps on p.

C. Empirical Calibration of p and c

The model treats p as a reduced-form probability that an attack changes the LLM-mediated ranking enough to obtain the attacker-favored allocation. Existing ranking-manipulation papers do not always report this exact probability, but their experimental quantities give useful calibration ranges.

Nestaas et al. (2024) report that direct preference-manipulation attacks can succeed in roughly 95%–100% of trials in some search settings, while external-page attacks have lower success, reaching at most about 25%. Their plugin experiments also show selection rates increasing from 0% to above 90% in some cases. Pfrommer et al. (2024) report normalized ranking-score improvements from 54.23% to 95.74% across models. Tang et al. (2025) report average target ranks between 1.46 and 2.50 on 10-item lists and between 1.87 and 2.39 on 8-item lists. Using the rank-normalized proxy $p_{\text{rank}} = \frac{K-\bar{r}}{K-1}$, where K is the candidate-list size and $\bar{r}$ is the average target rank, these values correspond roughly to $p_{\text{rank}} \in [0.80, 0.95]$.

The cost parameter c is measured in units of one-period normalized market value, so dollar costs must be scaled by the economic value of the relevant query class. Pfrommer et al. (2024) report inference costs of approximately $\$15$, $\$50$, and $\$450$ for their experiments across different providers. If the one-period gross profit associated with the relevant query class is G, these budgets map to $c \approx \frac{15}{G}$, $c \approx \frac{50}{G}$, $c \approx \frac{450}{G}$. For example, if $G \in [10^3, 10^5]$, this gives a rough range $c \in [1.5 \times 10^{-4}, 4.5 \times 10^{-1}]$. We use this range only as a sensitivity guide, since deployment costs, query value, and attacker automation differ across markets.

D. Tit-for-Tat Trigger Strategy

Following the structure of our model above, this section investigates the Tit-for-Tat (TFT) strategy as an alternative trigger strategy to foster cooperation. We analyze three different settings for the Tit-for-Tat (TFT) strategy: (1) a single defection by one player followed by immediate one-time retaliation from the other, (2) alternating cooperation and defection between the players, and (3) consecutive defections for a fixed number of rounds before returning to cooperation. For each setting, we examine the payoff of defection $V(D)$, and compare it with the payoff of continuous cooperation $V(C)$, to derive conditions under which cooperation can be sustained.

D.1. Setting 1: Player 1 Defects in the First Round, Player 2 Retaliates Once

In this setting, we consider a scenario where Player 1 defects in the first round, breaking cooperation, but then chooses to cooperate in all subsequent rounds. Player 2 retaliates by defecting in the second round as a response to Player 1's initial defection, and then resumes cooperation from the third round onward. This setup models a limited retaliation strategy, where defection is punished but not perpetuated indefinitely, aligning with the principles of the Tit-for-Tat strategy. The action sequences for the two players are as follows: Player 1's actions are $D \to C \to C \to C \to \ldots$; while Player 2's actions are $C \to D \to C \to C \to \ldots$.

Theorem D.1 (Cooperation Condition Under Single Defection and One-Time Retaliation)**.** *Long-term cooperation is sustainable under this strategy if and only if:*

$$\delta \geq \frac{T-R}{R-S} = 1 - \frac{2c}{p}$$

D.2. Setting 2: Alternating Cooperation and Defection

In this setting, the two players engage in a cyclic pattern of alternating cooperation and defection. Player 1 begins by defecting in the first round, prompting Player 2 to retaliate by defecting in the second round. This behavior continues indefinitely, with Player 1 defecting in all odd-numbered rounds and cooperating in all even-numbered rounds, while Player 2 cooperates in odd-numbered rounds and defects in even-numbered rounds. The alternating pattern models a competitive dynamic where neither player consistently cooperates nor defects. The action sequences for the two players can be summarized as follows: Player 1's actions are $D \to C \to D \to C \to \ldots$; Player 2's actions are $C \to D \to C \to D \to \ldots$

Theorem D.2 (Cooperation Condition Under Alternating Cooperation and Defection)**.** *Long-term cooperation is sustainable under this strategy if and only if:*

$$\delta \geq \frac{T-R}{R-S} = 1 - \frac{2c}{p}$$

D.3. Setting 3: Consecutive Defections for k Rounds

In this setting, Player 1 defects for k consecutive rounds before returning to cooperation. Player 2 retaliates by defecting from the second round through round $k+1$, matching Player 1's defections, before resuming cooperation in round $k+2$. This scenario models a prolonged period of defection followed by reconciliation, testing the players' capacity to return to cooperative behavior after extended conflict. The action sequences for the two players can be summarized as follows: Player 1's action sequence is $\underbrace{D \to D \to \cdots \to D}_{k \text{ rounds}} \to C \to C \to \ldots$ (defects for k rounds, then cooperates); Player 2's action sequence is $C \to \underbrace{D \to \cdots \to D \to D}_{k \text{ rounds}} \to C \to \ldots$ (cooperates in the first round, retaliates for k rounds, then cooperates).

Theorem D.3. *It is rational for Player 1 to either defect for only one round ($k = 1$) or defect indefinitely (i.e., $k \to \infty$), depending on the value of the discount factor δ. Specifically, Player 1 will:*

- *Defect for only one round if $\delta \geq \frac{Q-S}{R-S} = p\beta + (1-p) - \frac{2c}{p}$.*
- *Defect indefinitely if $\delta < \frac{Q-S}{R-S} = p\beta + (1-p) - \frac{2c}{p}$.*

The implications of Theorem D.3 reveal that Player 1's decision to defect either for only one round ($k = 1$) or indefinitely ($k \to \infty$) depends on the value of the discount factor δ. When Player 1 defects only once, this behavior is equivalent to the scenario described in Setting 1, where a single defection is followed by a return to cooperation. Specifically, if $\delta \geq p\beta + (1-p) - \frac{2c}{p}$, Player 1 defects once and then resumes cooperation, as the long-term value of cooperation outweighs the immediate gains of prolonged defection. However, if $\delta < p\beta + (1-p) - \frac{2c}{p}$, the immediate rewards from defection dominate, prompting Player 1 to defect indefinitely, equivalent to the Grim Trigger strategy.

Combining this theorem with the Theorem D.1 in Setting 1 provides a complete characterization of Player 1's behavior across the entire range of δ. If $\delta \geq 1 - \frac{2c}{p}$, cooperation is sustained indefinitely, as both players value the long-term rewards of mutual cooperation over any short-term temptation. If δ lies in the intermediate range, $p\beta + (1-p) - \frac{2c}{p} \leq \delta < 1 - \frac{2c}{p}$, Player 1 defects for one round before returning to cooperation, as this strategy balances the short-term gain from defection with the long-term benefits of cooperation. Finally, if $\delta < p\beta + (1-p) - \frac{2c}{p}$, Player 1 defects indefinitely, leading to the breakdown of cooperation.

These combined results emphasize the pivotal role of δ in governing strategic behavior. A higher δ fosters long-term cooperation by ensuring that the future rewards of mutual cooperation outweigh the short-term incentives to defect. Conversely, a lower δ diminishes the value of future payoffs, incentivizing short-sighted strategies such as prolonged or

indefinite defection. This underscores the importance of designing systems or environments that increase δ—for example, by fostering repeated interactions, implementing reputation systems, or introducing long-term incentives—to encourage cooperative outcomes and mitigate the risk of defection.

D.4. Cooperation Formation Region under Tit-for-Tat

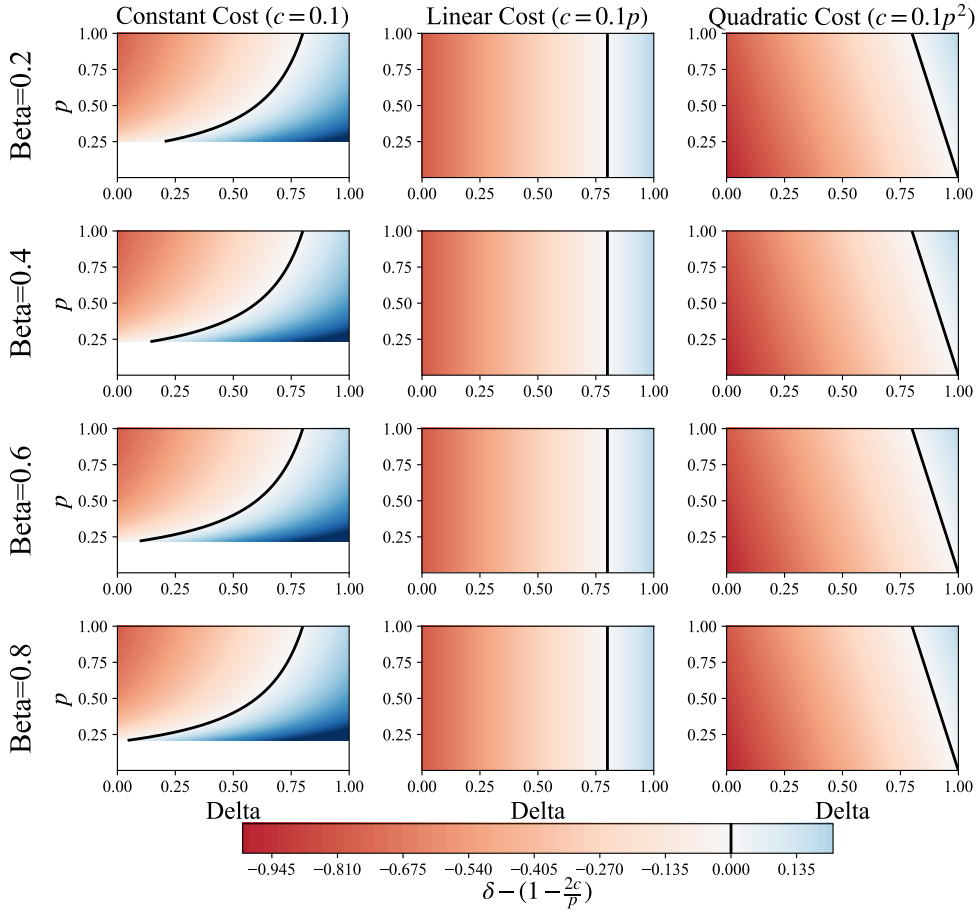

Figure 4. Region of Cooperation Formation (the region to the right of the boundary, Tit-for-Tat)

Under the Tit-for-Tat strategy, the condition for sustaining cooperation is given by $\delta \geq 1 - \frac{2c}{p}$, where δ is the discount factor, c is the cost of launching an attack, and p is the attack success probability. This condition is notable for being independent of the degradation factor β, setting Tit-for-Tat different from the Grim Trigger strategy, where β plays a central role in shaping the cooperation region under the Grim Trigger strategy. This independence fundamentally influences the dynamics of cooperation under Tit-for-Tat, leading to unique properties and implications for the formation and sustainability of cooperative behavior.

Under Grim Trigger, the cooperation condition depends on β, as $\delta \geq \frac{p-2c}{p-\beta p^2+p^2}$. This condition explicitly incorporates β, allowing Grim Trigger to leverage long-term market degradation to enforce cooperation. When β is low (i.e., mutual defection severely degrades market value), Grim Trigger imposes strong penalties for defection, effectively deterring deviations. However, when β is high, mutual defection outcomes become less severe, diminishing the deterrent effect of Grim Trigger's punishment and shrinking the cooperation region.

In contrast, for Tit-for-Tat, the cooperation condition is the same under different β values. The lack of dependence on β in Tit-for-Tat arises from its short-term punishment mechanism. Unlike Grim Trigger, which enforces cooperation by leveraging long-term degradation in market value (influenced by β), Tit-for-Tat relies solely on immediate retaliation. As our analysis above shows, the Tit-for-Tat strategy only has two possible settings: one player only defects once and another

player only retaliates once in the next round, or two players alternate between cooperation and defection. As a result, the two players will never attack simultaneously so that the LLM systems would never degrade, and the cooperation region remains unaffected by β, providing stability across varying degradation levels.

Figure 4 visualizes the cooperation formation regions for Tit-for-Tat under various cost functions and attack success probabilities. The regions to the right of the curve (the blue regions) represent the parameter space where cooperation can be sustained. Unlike Grim Trigger, these regions are shaped solely by the interplay between δ, c, and p, with no direct dependence on β.

While this independence from β simplifies the analysis, it also highlights a limitation of Tit-for-Tat: its inability to exploit long-term degradation as a deterrent. In scenarios where β is low (i.e., mutual defection severely degrades market value), Grim Trigger effectively uses this dynamic to enforce cooperation by imposing a strong, permanent penalty. Tit-for-Tat, lacking this mechanism, struggles to sustain cooperation when the immediate incentives to defect are strong.

E. One-Time Fixed Cost

In this section, we extend the analysis to account for scenarios where players face a one-time fixed cost when launching an attack. This cost could arise from factors such as one-time hardware or software purchases, one-time research and development expenditures, or other non-recurring investments required to enable the attack. We replicate the cooperation condition analysis and the payoff analysis of cooperation and defection, as done in Sections 3, but now incorporating the one-time fixed cost, denoted as c, incurred upon the first attack by a player.

Theorem E.1 (Cooperation Condition with One-Time Fixed Cost)**.** *In the presence of a one-time fixed cost, two players will prefer long-term cooperation over engaging in ranking manipulation attacks if and only if the following condition is satisfied:*

$$\delta \geq \delta^*_{\textit{One-Time Cost}} = \frac{p - 2c}{p - p^2\beta + p^2 - 2c}$$

where $\delta^*_{\textit{One-Time Cost}}$ represents the critical discount factor required to sustain cooperation.

Compared to the case with recurring costs (Theorem 3.1), the presence of a one-time fixed cost causes the cooperation region to contract. This is because the threshold for sustaining cooperation raises, as shown in the increasing the left-hand side of the inequality. The critical condition for cooperation established in Theorem E.1 highlights the distinct influence of the fixed cost on equilibrium outcomes. Specifically, the presence of the $-2c$ term in the denominator reflects the one-time nature of the cost, which does not penalize repeated acts of defection. This limitation makes the fixed cost less effective at discouraging defection compared to recurring costs, which impose a continuous penalty over successive interactions.

Furthermore, the one-time fixed cost is more effective at discouraging attacks from myopic players with smaller δ, who prioritize immediate payoffs, compared to forward-looking players with a long-term outlook who value future rewards through a larger δ. Myopic players view the fixed cost as a significant hurdle because it substantially reduces their immediate net benefits, making attacks less appealing unless the short-term gains are exceptionally high. In contrast, forward-looking players are less deterred by the fixed cost, as they treat it as an upfront investment and consider the long-term benefits of defection. Consequently, while the fixed cost effectively shrinks the attack incentives for myopic players, its impact on forward-looking players is less pronounced. This distinction underscores the need for strategic calibration of fixed costs to deter attacks from a broader range of players, including those with long-term strategies, ensuring overall system stability and cooperation.

Finally, the findings under recurring costs (Theorem 3.1) remain consistent when applied to the case of a one-time fixed cost. The cooperation formation region in the presence of one-time fixed cost exhibits similar sensitivity to δ, p, and β, where higher values of β and lower values of costs shrink the region for sustained cooperation. In addition, the non-monotonic relationship between p and defection payoff persists, where increasing p can initially discourage cooperation but, beyond a certain threshold, may diminish the motivation for defection as mutual defection harms both players significantly. Furthermore, the existence of futile defense regions is also observed under a one-time fixed cost, where reducing p does not necessarily lower defection payoffs, leaving attackers' incentives largely unaffected. These findings highlight that while a fixed cost shifts the magnitude of cost impacts, the underlying dynamics of the cooperation formation region and its dependencies on the strategic parameters remain consistent with those derived under recurring costs.

F. Asymmetric Players Scenarios

In this section, we extend our analysis to scenarios where players differ in their attack success probabilities, attack costs, and discount rates. Such asymmetries are common in real-world competitive environments where participants have varying resources, capabilities, or strategic priorities. We explore how these differences influence the conditions for sustaining cooperation and identify which player's characteristics are pivotal in determining the overall cooperation dynamics.

The sustainability of cooperation in asymmetric scenarios is primarily determined by the player with the greatest temptation to defect. This player has the most significant influence on whether mutual cooperation can be maintained.

We consider three types of asymmetries: differences in attack success probabilities, differences in attack costs, and differences in discount rates. In the first scenario, the players have different probabilities of successfully executing an attack, with Player 1 having a lower success probability than Player 2 ($p_1 < p_2$). In the second scenario, the players face different costs for conducting an attack, with Player 1 incurring a lower cost than Player 2 ($c_1 < c_2$). In the third scenario, the players have differing levels of patience, reflected in their discount factors, where Player 1 values future payoffs less than Player 2 ($\delta_1 < \delta_2$). For each of these asymmetries, the payoff structures are modified to reflect the differences between the players. We then derive the conditions under which cooperation can be sustained.

F.1. Scenario 1: Different Attack Success Probabilities ($p_1 < p_2$)

When players have different probabilities of successfully launching an attack, their incentives to cooperate or defect diverge, as reflected in their respective payoff structures. This asymmetry arises in settings where one player has a greater likelihood of achieving successful attacks, creating unequal temptations to defect.

To account for the asymmetric success probabilities ($p_1 < p_2$), the payoff functions are adjusted as follows:

- **Cooperation Payoff:** $R_1 = R_2 = \frac{1}{2}$.
- **Temptation Payoffs:** $T_1 = p_1 + (1 - p_1)\left(\frac{1}{2}\right) - c$, $T_2 = p_2 + (1 - p_2)\left(\frac{1}{2}\right) - c$.
- **Sucker Payoffs:** $S_1 = (1 - p_2)\left(\frac{1}{2}\right)$, $S_2 = (1 - p_1)\left(\frac{1}{2}\right)$.
- **Mutual Defection Payoff:** $Q_1 = p_1 p_2 \frac{1}{2}\beta + p_1(1 - p_2) + (1 - p_1)(1 - p_2)\left(\frac{1}{2}\right) - c$, $Q_2 = p_1 p_2 \frac{1}{2}\beta + p_2(1 - p_1) + (1 - p_1)(1 - p_2)\left(\frac{1}{2}\right) - c$.

Cooperation Condition. Cooperation is sustainable if:

$$\delta_i \geq \frac{T_i - R_i}{T_i - Q_i}, \quad \text{for } i = 1, 2 \qquad \text{(Grim Trigger)}$$

$$\delta_i \geq \frac{T_i - R_i}{R_i - S_i}, \quad \text{for } i = 1, 2 \qquad \text{(Tit-For-Tat)}$$

Theorem F.1. *Consider two players with attack success probabilities p_1 and p_2, cooperation is sustainable if and only if:*

$$\delta_1 \geq \frac{p_1 - 2c}{(1 - \beta)p_1 p_2 + p_2} \quad \textit{and} \quad \delta_2 \geq \frac{p_2 - 2c}{(1 - \beta)p_1 p_2 + p_1} \qquad \text{(Grim Trigger)}$$

$$\delta_1 \geq \frac{p_1 - 2c}{p_2} \quad \textit{and} \quad \delta_2 \geq \frac{p_2 - 2c}{p_1} \qquad \text{(Tit-For-Tat)}$$

Corollary F.2. *Under different attack success probabilities ($p1 < p2$), the sustainability of cooperation is primarily determined by the player with the higher success probability.*

Players with higher attack success probabilities face greater temptation to defect, as their potential short-term gains from defection are larger. Thus, the sustainability of cooperation hinges on their willingness to prioritize future payoffs over immediate gains.

F.2. Scenario 2: Different Attack Costs ($c_1 < c_2$)

In scenarios where players incur different costs for launching attacks, the player with the lower cost faces reduced deterrence against defecting. This asymmetry shifts the balance of cooperation dynamics.

The payoff functions reflect the cost asymmetry:

- **Cooperation Payoff:** $R_1 = R_2 = \frac{1}{2}$.
- **Temptation Payoffs:** $T_1 = p + (1-p)\left(\frac{1}{2}\right) - c_1, T_2 = p + (1-p)\left(\frac{1}{2}\right) - c_2$.
- **Sucker Payoffs:** $S_1 = (1-p)\left(\frac{1}{2}\right), S_2 = (1-p)\left(\frac{1}{2}\right)$.
- **Mutual Defection Payoff:** $Q_1 = p^2\frac{1}{2}\beta + p(1-p) + (1-p)^2\left(\frac{1}{2}\right) - c_1, Q_2 = p^2\frac{1}{2}\beta + p(1-p) + (1-p)^2\left(\frac{1}{2}\right) - c_2$.

Theorem F.3. *Let two players have different attack costs c_1 and c_2. The cooperation is sustainable if and only if:*

$$\delta_1 \geq \frac{p - 2c_1}{(1-\beta)p^2 + p} \quad \textit{and} \quad \delta_2 \geq \frac{p - 2c_2}{(1-\beta)p^2 + p} \tag{Grim Trigger}$$

$$\delta_1 \geq \frac{p - 2c_1}{p} \quad \textit{and} \quad \delta_2 \geq \frac{p - 2c_2}{p} \tag{Tit-For-Tat}$$

Corollary F.4. *Under different attack costs ($c_1 < c_2$), cooperation sustainability is determined by the player who has a lower attack cost.*

A player with a lower attack cost faces less deterrent against attacking. They set the cooperation threshold, as their defection can disrupt mutual cooperation.

F.3. Scenario 3: Different Discount Rates ($\delta_1 < \delta_2$)

Discount rate asymmetries reflect that players value future profits differently, with the less patient player placing less value on future cooperation benefits. This influences their willingness to cooperate.

Theorem F.5. *For players with different discount rates δ_1 and δ_2, cooperation is sustainable if and only if:*

$$c \geq \frac{p - \delta_1(p - \beta p^2 + p^2)}{2} \quad \textit{and} \quad c \geq \frac{p - \delta_2(p - \beta p^2 + p^2)}{2} \tag{Grim Trigger}$$

$$c \geq \frac{1}{2}(1-\delta_1)p \quad \textit{and} \quad c \geq \frac{1}{2}(1-\delta_2)p \tag{Tit-For-Tat}$$

Corollary F.6. *The player with a lower discount rate is more inclined to defect, making their cooperation threshold critical for sustaining cooperation.*

F.4. Multiple Asymmetric Dimensions

In this subsection, we explore the scenario where players differ along multiple dimensions simultaneously, such as attack success probabilities (p_1, p_2) and attack costs (c_1, c_2). This scenario captures a more realistic situation where participants have asymmetric strengths and weaknesses across multiple strategic factors. For example, players in real-world competitive environments often differ in both their ability to launch successful attacks and the costs they incur for such actions. For instance, a company with better technology might have a higher success probability but also face higher operational costs. These trade-offs affect the cooperation dynamics and require more nuanced strategies to sustain cooperative outcomes.

The payoffs for both players are adjusted to reflect asymmetries in both attack success probabilities and attack costs.

- **Cooperation Payoff:** $R_1 = R_2 = \frac{1}{2}$.
- **Temptation Payoffs:** $T_1 = p_1 + (1-p_1)\left(\frac{1}{2}\right) - c_1, T_2 = p_2 + (1-p_2)\left(\frac{1}{2}\right) - c_2$.

- **Sucker Payoffs:** $S_1 = (1-p_2)\left(\frac{1}{2}\right)$, $S_2 = (1-p_1)\left(\frac{1}{2}\right)$.
- **Mutual Defection Payoff:** $Q_1 = p_1 p_2 \frac{1}{2}\beta + p_1(1-p_2) + (1-p_1)(1-p_2)\left(\frac{1}{2}\right) - c_1$, $Q_2 = p_1 p_2 \frac{1}{2}\beta + p_2(1-p_1) + (1-p_1)(1-p_2)\left(\frac{1}{2}\right) - c_2$.

Cooperation Conditions with Multiple Asymmetries. The cooperation conditions must now account for both the differences in attack success probabilities and attack costs. For cooperation to be sustainable, the following inequalities must hold under the Grim Trigger and Tit-for-Tat strategies.

Theorem F.7. *Let two players differ in both attack success probabilities and attack costs. The cooperation is sustainable if and only if:*

$$\delta_1 \geq \frac{p_1 - 2c_1}{(1-\beta)p_1 p_2 + p_2} \quad \textit{and} \quad \delta_2 \geq \frac{p_2 - 2c_2}{(1-\beta)p_1 p_2 + p_1} \tag{Grim Trigger}$$

$$\delta_1 \geq \frac{p_1 - 2c_1}{p_2} \quad \textit{and} \quad \delta_2 \geq \frac{p_2 - 2c_2}{p_1} \tag{Tit-For-Tat}$$

Corollary F.8. *When players differ in both attack costs and success probabilities, the cooperation condition is determined by the player with the greater temptation to defect, which is jointly influenced by their attack cost and success probability.*

The framework we developed allows us to analyze more complex scenarios where players differ along multiple dimensions simultaneously. For instance, if Player 1 has a lower attack success probability ($p_1 < p_2$) but incurs a lower attack cost ($c_1 < c_2$), the outcome of cooperation will depend on the joint effect of these asymmetries. Specifically, Player 2's higher success probability provides them with more opportunities to gain from defection, while Player 1's lower attack cost reduces their barrier to engage in repeated attacks. Together, these factors shape the dynamics of cooperation.

These findings underscore the importance of identifying the player whose characteristics pose the greatest threat to long-term cooperation. Effective cooperative outcomes can be fostered through targeted incentives tailored to the asymmetries. For instance, Player 2 (the more capable but costlier player) could be incentivized to cooperate by sharing a portion of Player 1's benefits or by imposing penalties that increase Player 2's effective attack cost. Conversely, Player 1 could be compensated with a share of Player 2's potential gains to discourage opportunistic behavior stemming from their lower-cost advantage.

Our framework systematically analyzes the interplay between multiple asymmetries and their impact on cooperation, enabling decision-makers to identify the key player whose characteristics pose the greatest threat to sustained cooperation. By modeling differences in attack success probabilities, attack costs, and other dimensions, it provides actionable insights into designing multi-faceted interventions—such as profit-sharing mechanisms or cost-based penalties—to address specific asymmetries. These contributions align with real-world practices, equipping decision-makers with tailored strategies to foster cooperation and mitigate risks associated with strategic imbalances.

G. Multi-Player Scenarios

This section derives the conditions under which cooperation can be sustained in scenarios with N players, of which M players defect, using different trigger strategies. The analysis extends the payoff structure and the grim trigger and tit-for-tat strategies to multi-player contexts, providing a generalized framework.

In a system with N players, the payoff dynamics are determined by the total number of defectors M, where $M < N$. The market is shared among successful attackers, and the outcome varies based on the proportion of successful attacks.

The payoffs are defined as follows:

- R represents the payoff for mutual cooperation, where all N players cooperate and share the normalized market value equally, giving $R = \frac{1}{N}$.
- T is the payoff for the defectors who launch attacks while others cooperate. The expected payoff for a one-time defector includes the probabilities of successful attacks. If exactly k out of M defectors succeed (probability $\binom{M}{k} p^k (1-p)^{M-k}$),

the k successful attackers share the market equally. If no attack succeeds (probability $(1-p)^M$), the market is shared equally among all N players.

$$T = \sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \cdot \frac{1}{k} + (1-p)^M \cdot \frac{1}{N} - c$$

- S is the payoff for a cooperator when M players defect. If all M attackers fail (probability $(1-p)^M$), the cooperator receives their share of the market, which is $\frac{1}{N}$. If at least one attack succeeds (probability $1-(1-p)^M$), the cooperator receives nothing, as the successful attackers monopolize the market. The resulting payoff is:

$$S = (1-p)^M \cdot \frac{1}{N}$$

- Q captures the payoff for mutual defection, where all N players defect. If exactly k attacks succeed (probability $\binom{N}{k} p^k (1-p)^{N-k}$), the k successful attackers share the market equally. If all N attacks succeed (probability p^N), the market degrades to β, shared equally among the N players. If no attack succeeds (probability $(1-p)^N$), the market is shared equally among all N players.

$$Q = \sum_{k=1}^{N-1} \binom{N}{k} p^k (1-p)^{N-k} \cdot \frac{1}{k} + p^N \cdot \frac{\beta}{N} + (1-p)^N \cdot \frac{1}{N} - c$$

Now let's plug in our formulas into the critical conditions for the Grim Trigger and Tit-for-Tat strategies, given:

Under Grim Trigger, the cooperation is sustainable if and only if: $\delta \geq \frac{T-R}{T-Q}$.

Under Tit-for-Tat, the cooperation is sustainable if and only if: $\delta \geq \frac{T-R}{R-S}$.

Let's substitute the derived formulas step-by-step to better understand the conditions for sustaining cooperation.

Theorem G.1 (Condition to Sustain Cooperation in Multi-Player Setting)**.** *In the Multi-Player setting with the Grim Trigger strategy, cooperation is sustainable if and only if:*

$$\delta \geq \frac{T-R}{T-Q} \stackrel{\text{def}}{=} \delta^*_{\textit{Multi,GT}} \qquad \text{(Grim Trigger)}$$

$$\delta \geq \frac{T-R}{R-S} \stackrel{\text{def}}{=} \delta^*_{\textit{Multi, TFT}} \qquad \text{(Tit-For-Tat)}$$

, where

$$\begin{aligned}
&\delta^*_{\textit{Multi,GT}} = [\Big(\textstyle\sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \cdot \frac{1}{k} + (1-p)^M \cdot \frac{1}{N} - c\Big) - \frac{1}{N}] \\
&\div [\Big(\textstyle\sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \cdot \frac{1}{k} + (1-p)^M \cdot \frac{1}{N} - c\Big) \\
&- \Big(\textstyle\sum_{k=1}^{N-1} \binom{N}{k} p^k (1-p)^{N-k} \cdot \frac{1}{k} + p^N \cdot \frac{\beta}{N} + (1-p)^N \cdot \frac{1}{N} - c\Big)], \\
&\delta^*_{\textit{Multi, TFT}} = [\Big(\textstyle\sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \cdot \frac{1}{k} + (1-p)^M \cdot \frac{1}{N} - c\Big) - \frac{1}{N}] \\
&\div [\tfrac{1}{N} - (1-p)^M \cdot \tfrac{1}{N}]
\end{aligned}$$

Corollary G.2. *Given the total number of players N, when there are a substantial number of attackers, an increase in the number of attackers M widens the range of discount factors δ that sustain cooperation.*

As the number of attackers grows, the critical discount factor δ^*_{Multi} decreases, making it easier to satisfy the cooperation condition. The corollary highlights an important relationship between the number of attackers and the feasibility of sustaining cooperation. As the number of attackers M increases, the opportunity for any individual attacker to gain significant benefits through defection diminishes because the total payoff is spread across more players. In other words, the marginal gain from defection becomes smaller when more players are involved in attacks. Simultaneously, the punishment remains severe, which strengthens the incentive to cooperate. Consequently, even players who place less weight on future rewards (lower δ) will still find cooperation preferable, widening the range of discount factors under which cooperation can be sustained.

H. Proof of Theorems

H.1. Proof of Theorem 3.1

Proof. For cooperation to be sustainable, we need $V(C) \geq V(D)$, which is

$$\frac{R}{1-\delta} \geq T + \frac{\delta Q}{1-\delta}$$

Reorganizing this inequality gives the critical condition for sustaining cooperation:

$$\delta \geq \frac{T-R}{T-Q} \stackrel{\text{def}}{=} \delta^*$$

where δ^* is the critical discount factor. This threshold represents the minimum value of δ required for players to prioritize long-term cooperation over the short-term gains from defection.

Let's expand this condition using our given payoff functions:

$$\begin{aligned}
T - R &= \left[p + (1-p)\left(\frac{1}{2}\right) - c\right] - \frac{1}{2} = \frac{p}{2} - c \\
T - Q &= \frac{1+p}{2} - c - \left(p^2\frac{1}{2}\beta + p(1-p) + (1-p)^2\frac{1}{2} - c\right) \\
&= \frac{1+p}{2} - p^2\frac{1}{2}\beta - p(1-p) - \frac{(1-p)^2}{2} \\
&= \frac{1+p}{2} - p^2\frac{1}{2}\beta - p + p^2 - \frac{1}{2} + p - \frac{p^2}{2} \\
&= \frac{p}{2} - p^2\frac{1}{2}\beta + \frac{p^2}{2}
\end{aligned}$$

Therefore, the players prefer cooperation over launching an attack when:

$$\delta \geq \delta^* = \frac{\frac{p}{2} - c}{\frac{p}{2} - p^2\frac{1}{2}\beta + \frac{p^2}{2}} = \frac{p - 2c}{p - \beta p^2 + p^2}$$

□

H.2. Proof of Theorem 3.3

Proof. We derive these results by analyzing the partial derivatives of δ^* with respect to the key parameters c, β, and p. For c, the partial derivative is negative, indicating that an increase in attack costs reduces the threshold for cooperation. For β, the derivative is positive, signifying that higher mutual defection payoffs make cooperation harder to maintain. The derivative with respect to p is more complex, as it can be positive or negative depending on the values of the other parameters, leading to the observed non-monotonic behavior. □

H.3. Proof of Proposition 3.4

Proof. Using the payoff definitions,

$$T = \frac{1}{2} + \frac{p}{2} - c(p)$$

and

$$Q = p^2\frac{\beta}{2} + p(1-p) + (1-p)^2\frac{1}{2} - c(p).$$

The mutual-defection payoff simplifies to

$$Q = \frac{1}{2} - \frac{(1-\beta)p^2}{2} - c(p).$$

Therefore,

$$V_D(p) = T + \frac{\delta Q}{1-\delta} = \frac{1}{2} + \frac{p}{2} - c(p) + \frac{\delta}{1-\delta}\left(\frac{1}{2} - \frac{(1-\beta)p^2}{2} - c(p)\right).$$

Since

$$V_C = \frac{1}{2(1-\delta)},$$

we obtain

$$V_D(p) - V_C = \frac{(1-\delta)p - \delta(1-\beta)p^2 - 2c(p)}{2(1-\delta)}.$$

Differentiating $V_D(p)$ gives

$$V_D'(p) = \frac{1}{2} - \frac{c'(p) + \delta(1-\beta)p}{1-\delta}.$$

Thus any interior maximizer satisfies

$$c'(p^\star) + \delta(1-\beta)p^\star = \frac{1-\delta}{2}.$$

The second derivative is

$$V_D''(p) = -\frac{c''(p) + \delta(1-\beta)}{1-\delta}.$$

If $c''(p) \geq 0$, then $V_D''(p) < 0$, so the maximizer is unique whenever it is interior. The closed-form cases follow by substituting $c'(p) = 0$, $c'(p) = \alpha$, and $c'(p) = 2\alpha p$, respectively. □

H.4. Proof of Theorem F.1

Proof. For each player, the discounted payoff from continuous cooperation must be at least as high as the payoff from defecting once and then facing mutual defection indefinitely. □

H.5. Proof of Corollary F.2

Proof. Under the grim trigger strategy, given $p1 < p2$, we have $\frac{p_1-2c}{(1-\beta)p_1p_2+p_2} < \frac{p_2-2c}{(1-\beta)p_1p_2+p_1}$ for Grim Trigger, and $\frac{p_1-2c}{p_2} < \frac{p_2-2c}{p_1}$ for Tit-For-Tat. Under both strategies, player 2 has a greater temptation to defect and a stricter cooperation condition. □

H.6. Proof of Theorem F.3

Proof. The proof follows similar logic to Theorem F.1, adjusting for the different costs in the temptation and mutual defection payoffs. □

H.7. Proof of Corollary F.4

Proof. Under different attack cost, given $c_1 < c_2$, we have $\frac{p-2c_1}{(1-\beta)p^2+p} > \frac{p-2c_2}{(1-\beta)p^2+p}$ for Grim Trigger, and $\frac{p-2c_1}{p} > \frac{p-2c_2}{p}$ for Tit-For-Tat. Under both strategies, player 1 has a greater temptation to defect and a stricter cooperation condition. □

H.8. Proof of Theorem F.5

Proof. The proof follows a similar logic to Theorem F.1, adjusting for the different discount rates. □

H.9. Proof of Corollary F.6

Proof. Under different discount rates, given $\delta_1 < \delta_2$, we have $\frac{p-\delta_1(p-\beta p^2+p^2)}{2} > \frac{p-\delta_2(p-\beta p^2+p^2)}{2}$ for Grim Trigger, and $\frac{1}{2}(1-\delta_1)p > \frac{1}{2}(1-\delta_2)p$ for Tit-For-Tat. Under both strategies, player 1 has a greater temptation to defect and a stricter cooperation condition. □

H.10. Proof of Theorem F.7

Proof. The proof follows from the analysis of each player's incentive to defect based on their unique attack success probability and attack cost. □

H.11. Proof of Theorem F.8

Proof. If $p_1 < p_2$ but $c_1 < c_2$, the interaction between success probability and cost will determine which player has a stricter cooperation condition. The trade-off between higher probability and higher cost can either encourage or discourage defection, depending on the specific values. □

H.12. Proof of Corollary G.2

Proof. For large M, the term T can be approximated by focusing on the dominant portion of its summation. The probability mass of the binomial distribution is centered around $k \approx Mp$, and since $\frac{1}{k} \approx \frac{1}{Mp}$ in that region, the sum $\sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \frac{1}{k}$ approaches $\frac{1}{Mp}$. Additionally, $(1-p)^M \frac{1}{N}$ becomes negligible as M grows. Hence:

$$T \approx \frac{1}{Mp} - c.$$

Substituting T into $\delta^*_{\text{Multi,GT}}(M)$:

$$\delta^*_{\text{Multi,GT}}(M) \approx \frac{\left(\frac{1}{Mp} - c - \frac{1}{N}\right)}{\left(\frac{1}{Mp} - c - Q\right)}.$$

Define:

$$x = \frac{1}{Mp}, \quad A = c + \frac{1}{N}, \quad B = c + Q.$$

Then:

$$\delta^*_{\text{Multi,GT}}(M) \approx \frac{x - A}{x - B}.$$

We also know that $Q < \frac{1}{N}$, which implies:

$$B = c + Q < c + \frac{1}{N} = A.$$

Thus, $A - B > 0$.

To determine how $\delta^*_{\text{Multi,GT}}(M)$ evolves as M grows, examine how it changes as x varies. Since $x = \frac{1}{Mp}$ decreases as M increases, the sign of the derivative with respect to x reveals the direction of change with respect to M. The derivative with respect to x is:

$$\frac{d}{dx}\left(\frac{x - A}{x - B}\right) = \frac{A - B}{(x - B)^2}.$$

Since $A - B > 0$, the fraction $(x - A)/(x - B)$ is strictly increasing in x. As M grows, x becomes smaller, so $\delta^*_{\text{Multi,GT}}(M)$ decreases. In other words, increasing M reduces $\delta^*_{\text{Multi,GT}}(M)$.

In conclusion, as M becomes large, the value of $\delta^*_{\text{Multi,GT}}(M)$ is guaranteed to decrease. This analysis provides a clear asymptotic trend for $\delta^*_{\text{Multi,GT}}(M)$, indicating that growth in M shifts the ratio toward smaller values.

Similarly, for the Tit-for-Tat strategy, we simplify the denominator:

$$R - S = \frac{1}{N} - \frac{(1-p)^M}{N} = \frac{1 - (1-p)^M}{N}.$$

For large M, $(1-p)^M \to 0$ whenever $0 < p < 1$. Thus:

$$R - S \approx \frac{1}{N}.$$

Next, examine the numerator for large M. As before, the dominant term in $\sum_{k=1}^{M} \binom{M}{k} p^k (1-p)^{M-k} \frac{1}{k}$ is approximately $\frac{1}{Mp}$. Also, $(1-p)^M \frac{1}{N}$ becomes negligible. Hence:

$$T - R \approx \frac{1}{Mp} - c - \frac{1}{N}.$$

Putting these approximations together:

$$\begin{aligned}\delta^*_{\text{Multi, TFT}} &\approx \frac{\frac{1}{Mp} - c - \frac{1}{N}}{\frac{1}{N}} \\ &= N\left(\frac{1}{Mp} - c - \frac{1}{N}\right) = \frac{N}{Mp} - cN - 1.\end{aligned}$$

Since the only term in $\delta^*_{\text{Multi, TFT}}$ that depends on M is $\frac{N}{Mp}$, which decreases as M increases, $\delta^*_{\text{Multi, TFT}}(M)$ also decreases with increasing M. □

H.13. Proof of Theorem D.1

Proof. The payoff for continuous cooperation, denoted $V(C)$, is straightforward and represents the long-term reward from mutual cooperation across all rounds. It can be expressed as:

$$V(C) = \frac{R}{1-\delta},$$

where R is the reward for mutual cooperation, and δ is the discount factor that accounts for the value players place on future payoffs. In contrast, the payoff for defection by Player 1, $V(D)$, reflects the one-time benefit of defection in the first round, followed by the sucker payoff S in the second round when Player 2 retaliates, and the cooperation payoff R in all subsequent rounds. The defection payoff is therefore given by:

$$V(D) = T + \delta S + \sum_{t=3}^{\infty} \delta^{t-1} R = T + \delta S + \frac{\delta^2}{1-\delta} R,$$

where T is the temptation payoff received in the first round, S is the sucker payoff in the second round, and the remaining terms account for the discounted future cooperation payoffs.

For cooperation to be sustainable, $V(C) \geq V(D)$:

$$\frac{R}{1-\delta} \geq T + \delta S + \frac{\delta^2}{1-\delta} R.$$

Subtract $\frac{\delta^2}{1-\delta} R$ from both sides:

$$\frac{R}{1-\delta} - \frac{\delta^2}{1-\delta} R \geq T + \delta S.$$

Thus, the inequality becomes:

$$(1+\delta)R \geq T + \delta S.$$

We can rearrange the inequality:

$$\delta \geq \frac{T-R}{R-S} = 1 - \frac{2c}{p}.$$

□

H.14. Proof of Theorem D.2

Proof. The payoff for continuous cooperation, $V(C)$, remains unchanged, representing the long-term reward for mutual cooperation:

$$V(C) = \frac{R}{1-\delta}.$$

For alternating cooperation and defection, Player 1 receives the temptation payoff T in odd-numbered rounds and the sucker payoff S in even-numbered rounds. The total payoff is an infinite geometric series:

$$V(D) = T + \delta S + \delta^2 T + \delta^3 S + \cdots = \sum_{n=0}^{\infty} \delta^{2n} T + \sum_{n=0}^{\infty} \delta^{2n+1} S.$$

Simplifying the series yields:

$$V(D) = \frac{T + \delta S}{1-\delta^2}.$$

To sustain cooperation, the payoff from mutual cooperation must exceed the payoff from alternating cooperation and defection:

$$\frac{R}{1-\delta} \geq \frac{T + \delta S}{1-\delta^2}.$$

Simplifying this inequality gives the same condition as Setting 1:

$$\delta \geq \frac{T-R}{R-S} = 1 - \frac{2c}{p}.$$

□

H.15. Proof of Theorem D.3

Proof. The payoff for continuous cooperation, $V(C)$, remains the same:

$$V(C) = \frac{R}{1-\delta}.$$

For Player 1, the payoff from defecting for k rounds includes: Temptation payoff (T) in the first round, Mutual defection payoff (Q) in rounds 2 through k, Sucker payoff (S) in round $k+1$, and Cooperation payoff (R) from round $k+2$ onward.

Thus, the total payoff for Player 1 is:

$$\begin{aligned} V(D) &= T + \sum_{i=1}^{k-1} \delta^i Q + \delta^k S + \sum_{t=k+2}^{\infty} \delta^{t-1} R \\ &= T + \left(\frac{\delta - \delta^k}{1-\delta} \right) Q + \delta^k S + \frac{\delta^{k+1}}{1-\delta} R. \end{aligned}$$

Here, the summation for rounds 2 to k is a geometric series, while the cooperation payoff beyond round $k+1$ accounts for the discounted long-term rewards.

We now analyze whether it is beneficial for Player 1 to defect for $k+1$ rounds instead of k. The change in Player 1's payoff when increasing defection rounds from k to $k+1$ is: (1) The payoff from mutual defection Q increases by $\delta^k Q$; (2) The sucker payoff S decreases by $(\delta^k - \delta^{k+1})S$; (3) The cooperation payoff R decreases by $\delta^{k+1} R$.

Thus, the net change in payoff is:

$$\delta^k Q - (\delta^k - \delta^{k+1})S - \delta^{k+1} R.$$

Factor the terms:

$$\delta^k \left[Q - S \right] - \delta^{k+1} \left[R - S \right].$$

Defecting for more rounds is beneficial if the above expression is positive:

$$\delta^k \left[Q - S \right] - \delta^{k+1} \left[R - S \right] > 0.$$

Simplifying this gets us the condition when the player has an incentive to defect for more rounds:

$$\delta < \frac{Q - S}{R - S}.$$

□

H.16. Proof of Theorem E.1

Proof. For cooperation to be sustainable, the condition $V(C) \geq V(D)_{\text{One-Time Cost}}$ must hold, where $V(D)_{\text{One-Time Cost}}$ is the defection payoff in the presence of a one-time fixed cost. This is given by:

$$\begin{aligned} V(D)_{\text{One-Time Cost}} &= T + \delta Q + \delta^2 Q + \dots \\ &= T + \frac{\delta Q}{1 - \delta} \\ &= \frac{1}{2} + \frac{1}{2}p - c + \frac{\delta}{1 - \delta}\left(p^2\beta + p(1-p) + (1-p)^2\frac{1}{2}\right), \end{aligned}$$

where:

- $T = \frac{1}{2} + \frac{1}{2}p - c$ represents the immediate payoff when one player defects while the other cooperates, accounting for the one-time fixed cost c.
- Q represents the discounted future payoffs under mutual defection. For subsequent rounds, there are no additional costs incurred.

By solving the inequality $V(C) \geq V(D)_{\text{One-Time Cost}}$, we obtain the critical threshold:

$$\delta \geq \delta^*_{\text{One-Time Cost}} = \frac{p - 2c}{p - p^2\beta + p^2 - 2c}.$$

□

I. Camera-Ready Edits Summary

- Revised the interpretation of the degradation parameter β so that the prose matches the formal payoff model, where degradation is triggered by simultaneous successful attacks.
- Consolidated Propositions 4.4-4.6 into a single analytical characterization of the defection payoff $V_D(p)$, with related defense implications stated as corollaries.
- Added empirical calibration discussion for p and c, and expanded related work to connect the model to repeated oligopoly and cartel-stability literature.
- Clarified modeling assumptions, asymmetric-player implications, and conclusion-level defense recommendations.
- Shortened the main body to satisfy the page limit by removing redundant exposition, consolidating overlapping results, and moving detailed derivations to the appendix.
- Updated appendix statements and proofs, refined equation formatting, and fixed formatting artifacts in the running header.