Strategic Amnesia in LLM Agents playing the Iterated Prisoner's Dilemma

Andrea Monoli[1,2,*,†], Sofia Sciangula[1,2,†], Mauro Mezzenzana[1,2,†], Giacomo Buonanno[1,2,†] and Francesco Bertolotti[1,2,*,†]

[1]*Intelligence, Complexity and Technology Lab (ICT Lab), Università Carlo Cattaneo – LIUC, Corso G. Matteotti, 22, Castellanza (VA), 21053, Italy*

[2]*School of Industrial Engineering, Università Carlo Cattaneo – LIUC, Corso G. Matteotti, 22, Castellanza (VA), 21053, Italy*

Abstract

As autonomous LLM agents are increasingly deployed in social simulations and economic contexts, determining whether these systems possess the intrinsic capacity to cooperate in a complex-systems setting is of extreme relevance. In this study, we constructed a computational laboratory designed to experiment with LLMs in repeated game theory scenarios over multiple periods, across various games and strategic configurations. Specifically, we focused on the classic Iterated Prisoner's Dilemma, and the experimental design evaluated agents under two distinct conditions: a "pure" behavior setting (agents operating without specific initial behavioral instructions) and a strategic setting (agents operating under pre-defined instructional prompts), both tested across a diverse array of models. The results suggest that, while initial instructions exert a demonstrable effect on the system's early behavior, this influence is observed to be moderated by subsequent interactions, suggesting that interactions may eventually override initial alignment prompts. Also, we observed heterogeneity in outcomes across different models, diverging toward different equilibria. These findings highlight the critical importance of training data and architectural bias when employing LLMs in complex systems.

Keywords

Agent-Based modeling, Iterated Prisoner's Dilemma, Strategic Amnesia, Multi-Agent Systems, Memory Manipulation, Cognitive Anchoring,

1. Introduction

The capacity to cooperate under uncertainty is a defining feature of social intelligence [1]. In humans, this capacity is grounded in episodic memory: the accumulated record of past interactions shapes trust, calibrates expectations, and sustains the fragile equilibria of mutual benefit [2, 3]. Nevertheless, there is a long tradition of observing these phenomena emerging not only in biological [4] but also in artificial systems [5] and in Artificial Intelligence (AI) [6, 7]. As Large Language Models (LLMs) are increasingly deployed as autonomous decision-making agents in economic, social, and organizational systems [8, 9, 10], a foundational question emerges: what happens to an LLM agent's cooperative behavior when its memory is manipulated or destroyed?

Evolutionary Game Theory (EGT), and more specifically the classic Iterated Prisoner's Dilemma (IPD) [11, 12], provides an ideal computational laboratory for answering this question, given a payoff structure where, typically, mutual cooperation yields collective gain while unilateral defection offers individual temptation, hence models the nature of social dilemmas that pervade real-world multi-agent interactions [2]. A rich body of literature has demonstrated that successful strategies in the IPD, such as the famous Tit-for-Tat [13], all rely on the ability to remember, retaliate, and ultimately forgive [2, 3]. Without memory, cooperation struggle to persist. Yet, the precise mechanisms by which memory loss translates into behavioral change remain unexplored in the context of LLM agents.

WOA 2026: 26th Workshop "From Objects to Agents", 2026, Italy

*Corresponding author.

andremon2610@gmail.com (A. Monoli); so10.sciangula@stud.liuc.it (S. Sciangula); mmezzenzana@liuc.it (M. Mezzenzana); buonanno@liuc.it (G. Buonanno); fbertolotti@liuc.it (F. Bertolotti)

Prior work has established that different LLMs exhibit distinct strategic "fingerprints" in competitive games [14] and game-theoretic settings [15, 16]. Models vary in their cooperative tendencies, sensitivity to prompt framing, and resilience to opponent exploitation. However, these studies have largely treated the agent's memory as a fixed input. The adversarial manipulation of contextual history, a realistic threat in deployed multi-agent systems where a malicious actor might selectively edit the interaction log, has not been systematically investigated.

This paper aims at filling this gap. We performed a "Strategic Amnesia" experiment, a controlled intervention that, at a specified trigger round, manipulates or limits an LLM agent's access to its prior interaction history. We test four experimental configurations across four state-of-the-art foundational models, generating 3,840 match records from 48 independent simulation runs. Our findings reveal that the introduction of an amnesia event in our setting does not restore cooperation. On the contrary, erasing the evidence of an opponent's defections deepens the collapse of cooperation, suggesting that the way LLM can cooperate is anchored not to a modifiable representation of current evidence but to a consolidated behavioral state that persists beyond the availability of its basis.

Previewing the results, the amnesia trigger never restores cooperation: every configuration shows a post-trigger decline, and the deepest memory truncation produces the sharpest drop. The full pattern is reported later in Section 4.

The results carry significant implications for the design and deployment of LLM agents in multi-round, multi-stakeholder environments, suggesting that the cognitive anchoring of LLMs to behavioral patterns established early in an interaction can be robust to context manipulation, a property that may be both a safeguard against adversarial rewriting of history and a barrier to recovery in compromised or disrupted interaction streams. The remainder of this paper is organized as follows: Section 2 reviews the theoretical foundations of evolutionary game theory, traditional AI experiments, and recent LLM-game-theory literature. Section 3 details the simulation architecture and experimental design. Section 4 presents and discusses the empirical findings. Section 5 draws conclusions and outlines directions for future research.

2. Background

This section provides the theoretical foundation for our study, divided in three ways: introducing the principles of Game Theory and its evolutionary extension; how AI has traditionally been employed in game-theoretic models; eventually, how LLM-based agents have been used and studied in these settings.

2.1. Evolutionary Game Theory

Game Theory is a mathematical framework [17] designed to analyze strategic interactions among rational decision-makers [18, 2]. In traditional game theory, players are assumed to accurately anticipate each other's moves to maximize their own payoffs [8]. However, real-world interactions often involve continuous learning and adaptation rather than perfect prediction [19]. EGT addresses this limitation by shifting the focus from highly rational individuals to populations of interacting agents [20, 2]. Instead of calculating a single optimal outcome, EGT addresses how different strategies perform and evolve over time based on their success rate [20, 2, 3]. The most referenced game of this field is the IPD, a configuration where two participants must repeatedly choose between cooperation and defection with a specific payoff settings that highly reward a defector only if the other agents do not defect as well [2].

As Robert Axelrod showed in his work, long-term interactions of the Prisoners' dilemma, bring players towards collective cooperative behaviors [2]. Strategies like "Tit-for-Tat", which starts by cooperating and then simply mimics the opponent's previous move, can benefit because they are

reciprocal, forgiving, and easily recognizable [2, 3]. Given that in evolutionary contexts successful strategies ”reproduce” better and spread quickly, complex cooperative dynamics could then emerge without requiring explicit altruism [20, 2].

2.2. Non-LLM Experiments with Game Theory and AI

Long before the creation of advanced LLM, AI scholars used game theory to train and evaluate algorithms [8]. The most evident intersection between these two fields is found in Reinforcement Learning [21, 22] and Multi-Agent Systems [23, 24].

In these traditional AI experiments, agents are not pre-programmed with explicit rules on how to win [25]. Instead, they interact with an environment (often modeled as a game scenario [25, 26]) and receive numerical rewards or penalties based on their actions [21, 22]. Over numerous repeated iterations, these agents learn to optimize their mathematical payoff [27]. For example, in games ranging from classic board games [28] to complex strategic simulations [29], Reinforcement Learning agents use massive computational power and blind exploration to discover dominant strategies that could consistently outperform human players [21, 8].

However, while traditional AI performs well in finding the most rewarding mathematical solution in well-defined games, these algorithms often lack human-like reasoning and flexibility [30, 9]. They play purely based on the rewards associated with state spaces, without an in-depth understanding of the social context or any qualitative or psychological motivations behind a choice [30, 9]. This limitation set the stage for exploring how LLMs might approach strategic interactions [8].

2.3. Experiments with Game Theory and LLMs

The diffusion of LLMS has introduced a new approach to artificial agents. Unlike classic RL agents, LLMs understand the world and act through natural language, allowing them to participate in complex social dilemmas not only with their actions but also mimicking human-like cognition and negotiation [8, 9]. Recent studies have begun putting LLMs to the test in classic game-theoretical environments [16, 15]. Rather than optimizing a raw numerical reward, LLMs interpret textual prompts describing the rules of the game [31], the payoffs involved [32], and sometimes the profile of their opponent [33]. These experiments have uncovered fascinating insights into what researchers call machine psychology or psychomatic [15, 34].

For example, when placed in an IPD Game, different foundational models exhibit distinct ”strategic fingerprints” [35, 15]. Research shows that models like OpenAI’s GPT tend to be highly cooperative, sometimes even to the point of vulnerability [36], while others, like Google’s Gemini, display a more ruthless, exploitative strategy in competitive settings [37]. Furthermore, LLMs have proven sensitive to games parameters, adjusting their strategies based on the ”shadow of the future” (i.e., whether they expect the interaction to be repeated)[38]. Some researches suggest that while LLMs perform remarkably well in selfish games, they can struggle with coordination unless guided by fine-tuned prompting techniques like ”Social Chain-of-Thought” [16]. With this said, our research investigates the integration of cooperative and non-cooperative game-theoretic frameworks—specifically the Shapley Value [39], social choice theory [40], and max-min equilibria [40],to facilitate the development of more efficient and theoretically robust algorithms for LLMs.

3. Methods

3.1. Game-theoretical LLM-model

To systematically study the strategic behavior of LLMs in a game-theoretical setting, we developed a custom, scalable framework capable of simulating complex multi-agent payoff-based environments. The core of this system is designed to remove the artifacts of turn-based or sequential prompting, which can inadvertently bias agent responses by providing asymmetric temporal information, hence relies on a strictly parallelized asynchronous architecture. In this way, it reflects the information constraints of a classic IPD[2, 16]. The overall framework is designed as followed.

At the beginning of each round, a prompt generator constructs an isolated contextual prompt for each participating agent. These prompts bundle the explicit rules of the game, the payoff matrix, the agent's assigned behavioral profile, and the chronological history of the game context window. Using asynchronous task execution, the system places independent, concurrent API calls to the respective foundational models .

We extended the traditional two-player format into multi-agent arenas using a "Round-Robin" topology where the network of possible interaction is fully connected [41]. In scenarios with more than two agents, every participant engages in pairwise zero-sum matches against all other participants in that round. The overall payoff for an agent is the summation of scores across all non-redundant orthogonal pairings, adhering to the standard IPD reward structure: Temptation ($T = 5$), Reward ($R = 3$), Mutual punishment ($P = 1$), and Sucker's payoff ($S = 0$) [2, 3].

To ensure quantitative precision during data extraction, we deployed an XML-style tag parsing approach. Agents are prompted to encapsulate their reasoning and final decision within strict markup tags (e.g., `<Action>Cooperate</Action>` and `<Strategy>Tit-for-Tat</Strategy>`). This raw generative output is piped into a robust regex and substring fuzzy-matching pipeline capable of isolating deterministic game actions while disregarding extraneous conversational text.
[1].

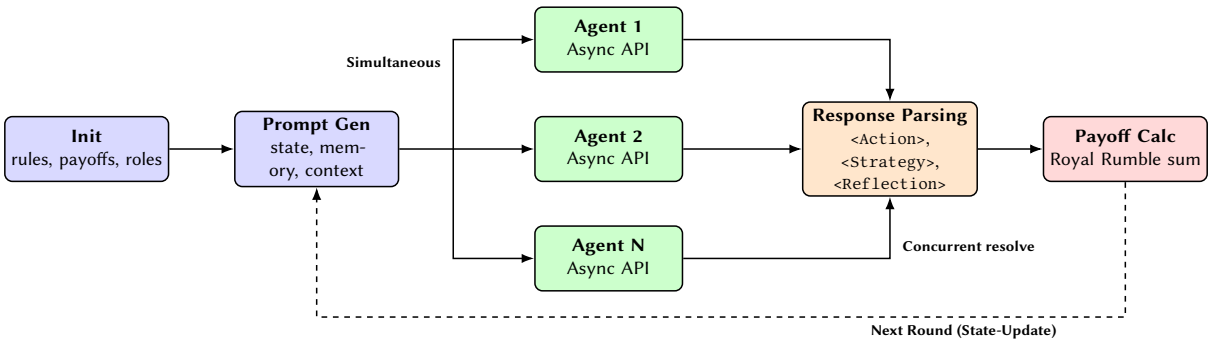

Figure 1: Architecture of the asynchronous NLP-based multi-agent orchestration model.

The resulting game-logs, including action sequences, payoff differentials, and internal cognitive reflections, are compiled into dynamic JSON files. These are consecutively processed through an integrated data-analysis pipeline to generate quantitative visualizations covering performance benchmarking, cooperation rates, and strategic distributions.

[1]The XML tags and parsing methodology was inspired by the GitHub repository of "henryljgwu" which can be found here

3.2. Experimental design and results analysis

Utilizing the simulation infrastructure and the integrated quantitative visual pipeline, our experimental design is specifically built to isolate the mechanisms of strategic heuristics and cognitive memory persistence within advanced LLMs. The focus of this research is the Strategic Amnesia experiment, which serves as a critical stress test for the models' reliance on historical context for maintaining stable cooperative equilibria. In this configuration, we introduce a controlled mechanical chronological wipe at a defined trigger point, t_a, which in this specific setting has been set at round 14 within a 20-round simulation. This intervention effectively simulates a sudden cognitive rupture where the agent's observable history is changed or restricted to a predefined k-value representing only the most recent actions. By varying both the setting of amnesia t_a and the depth of the remaining memory window k, we could systematically observe how the deletion of shared history influences the preservation of trust and reciprocity.

The results analysis for this experiment relies on two main metrics derived from the confrontation of the Pre-Amnesia versus Post-Amnesia collective states. First, we calculate Memory Volatility ΔC, which quantifies the immediate variation of cooperation by comparing the cooperation rates across the rounds directly adjacent to the trigger point, specifically $C_{round(t_a+1)} - C_{round(t_a-1)}$. This provides a direct measure of how the context-induced shock affects cooperation. Second, we employ Markov Post-Trigger Recovery Metrics to analyze the long-term resilience of the system [42], comparing the probability of cooperation $P(C)$ and its associated state-transition probabilities prior to the wipe against the recovery gradient established in the rounds, following the contextual blackout [23]. This allows us to determine if models re-establish cooperation through emerging reciprocity or if the amnesic shock leads to a permanent collapse into a Nash equilibrium of mutual defection [2, 3]. Furthermore, we explore the internal consistency of these agents by requiring them to articulate their strategic intent within a specific `<StrategyReasoning>` tag. This architectural requirement enables us to track Intention-Action divergence, providing a unique window into the latent decision-making processes of the LLM [9, 15]. By cross-referencing these reasoning tokens with the actual outputted game actions, we can observe moments of strategic hypocrisy [43], where an agent's stated intent to remain cooperative is contradicted by a defective action immediately following an amnesic trigger. This granular level of analysis is essential for understanding whether the failures in cooperation are results of strategic recalculation or purely the product of context-dependent probabilistic next-token generation.

We generated a total of 3,840 match records drawn from 48 independent simulation runs, conducted across four foundational models, `gpt-4o-mini`, `o4-mini`, `gemini-2.5-flash-lite`, and `gemini-3-flash-preview`, under a fully factorial design crossing two binary conditions: memory availability (ON/OFF) and reflection prompting (ON/OFF). Each run saw 20 rounds against an Always Defect (ALLD) scripted opponent, with the amnesia intervention triggered at round $t_a = 14$. Four amnesia configurations were applied simultaneously within each run: (A) no manipulation (control), (B) removal of opponent defection records, (C) restriction to the last k=3 rounds, and (D) restriction to the last k=7 rounds without any intervention over the bots' actions.

The source code used to perform the analysis presented in the following sections is publicly available for academic research purposes on GitHub

Models used for the following analysis are foundational models provided by OpenAI and Gemini; (`gpt-4o-mini`, `o4-mini`, `gemini-2.5-flash-lite`, `gemini-3-flash-preview`)

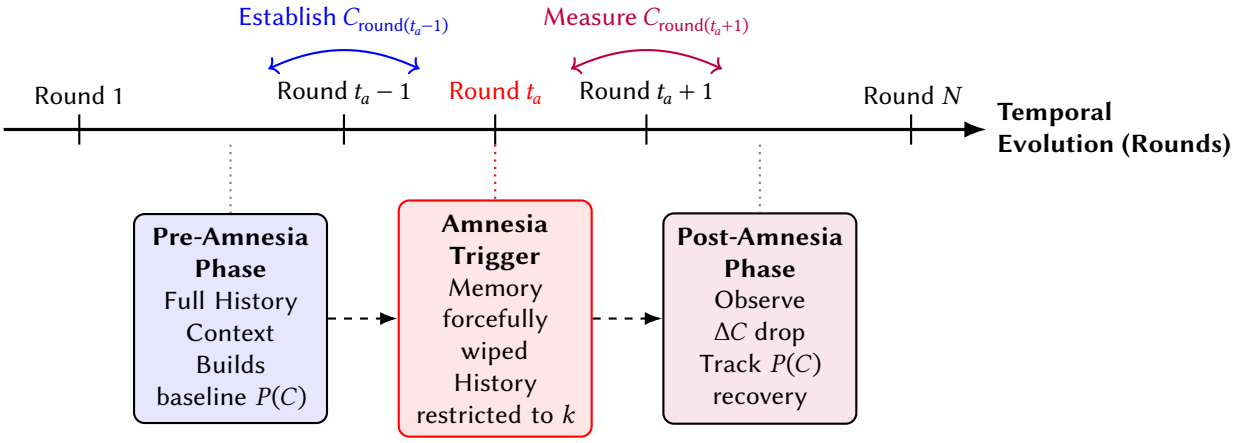

Figure 2: Chronological timeline of the Strategic Amnesia setting, illustrating the mechanical context wipe separating the baseline and recovery phases utilized to calculate Memory Volatility (ΔC).

4. Results and Discussion

4.1. Aggregate Cooperation Dynamics

The overall cooperation rate across all 3,840 observations was 23.96%. Contrary to the intuitive expectation that amnesic manipulation would restore cooperative behavior by erasing the record of opponent defections, we observed a universal decline in cooperation following the amnesia trigger. Pre-amnesia cooperation averaged 25.60% (688/2,688), while post-amnesia cooperation fell to 20.14% (232/1,152), yielding an aggregate shift of −5.46 percentage points (pp), a variation that, given the high number of simulations, cannot be attributed solely to stochastic noise but reveals an underlying signal.

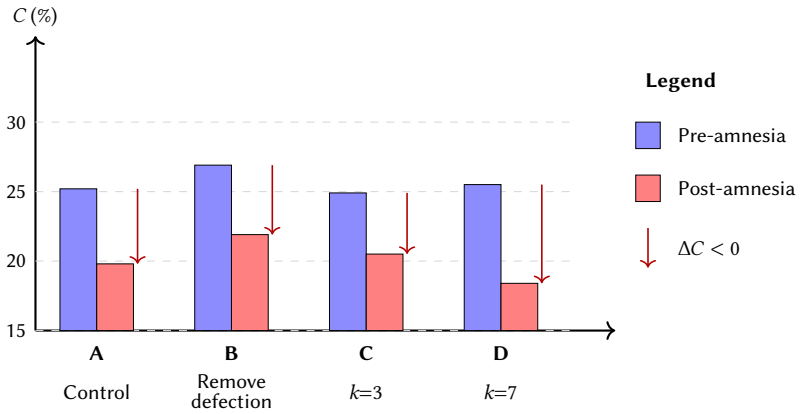

Figure 3: The "Amnesia Paradox": cooperation rate (%) before and after the amnesia trigger across all four configurations (A=Control, B=Remove defections, C=last k=3 rounds, D=last k=7 rounds). Red arrows mark ΔC; all shifts are negative.

Figure 3 visualizes the cooperation collapse: regardless of which memory manipulation is applied, removing defection records (B), or restricting the window to the last k=3 (C) or k=7 (D) rounds, post-amnesia cooperation is universally lower than pre-amnesia cooperation. The most aggressive erase (D, k=7) produces the steepest decline ($\Delta C = -7.04$ pp).

Table 1 reports the breakdown by amnesia configuration. Notably, none of the four conditions produced a positive post-amnesia cooperation recovery. The control condition (A) exhibited a decline of −5.36 pp, closely mirrored by the remove defections condition (B) at −5.06 pp. The last k only condition

(k=3) (C) showed a smaller decline of −4.37 pp, while the deeper memory window (k=7) condition (D) produced the steepest drop at −7.04 pp. This outcome is understandable since the last k only conditions does not change the memory in terms of actions but only reduces the window.

Table 1
Pre- and post-amnesia cooperation rates by amnesia configuration. ΔC denotes the change in cooperation rate (percentage points) from the pre-amnesia phase (rounds 0–13) to the post-amnesia phase (rounds 14–19).

Config	Mode	C_{pre}	C_{post}	ΔC (pp)
A	Control (none)	0.252	0.198	−5.36
B	Remove defections	0.269	0.219	−5.06
C	Last k=3 only	0.249	0.205	−4.37
D	Last k=7 only	0.255	0.184	−7.04
Overall		0.256	0.201	−5.46

4.2. The Stateless Nature of LLM Retaliation

A central finding of this study is that LLM retaliation is entirely context-dependent and does not indicate any kind of inward-directed resentment. The Markov transition analysis (Table 2) shows this mechanism with precision. In the pre-amnesia phase, the probability of retaliating after an observed opponent defection was $P(D|D) = 0.9166$, meaning that once locked into a defection cycle, agents remained there with high probability. The complementary forgiveness rate, $P(C|D) = 0.0834$, indicates that fewer than one in twelve opportunities for reconciliation were actually taken.

In the post-amnesia phase, these patterns did not change; instead, they slightly intensified. The retaliation probability rose to $P(D|D) = 0.9290$, while forgiveness further declined to $P(C|D) = 0.0710$. This outcome directly contradicts the hypothesis that removing the evidence of defection would "unlock" the model's cooperative prior. Instead, the data demonstrate that the amnesic intervention arrives too late: by round 14, the agent has already consolidated a retaliatory behavioral pattern that persists even when its evidentiary basis is removed.

Table 2
Markov state-transition probabilities for cooperation (C) and defection (D), computed over all configurations. n denotes the number of observed transitions from the conditioning state.

Phase	$P(C\|C)$	$P(D\|C)$	$P(C\|D)$	$P(D\|D)$
Pre-Amnesia (n_C=650, n_D=1846)	0.599	0.402	0.083	0.917
Post-Amnesia (n_C=199, n_D=761)	0.714	0.286	0.071	0.929

One notable asymmetry is that among the few agents that did cooperate post-amnesia, their cooperation became significantly more stable: $P(C|C)$ rose from 0.599 to 0.714. This suggests a splitting effect: the amnesia intervention polarizes behavior, reinforcing whatever trajectory the agent had adopted after the shock [44]. Agents already defecting become more locked in defection, while the rare cooperators become more committed.

The forgiveness probability $P(C|D)$ by configuration (Table 3) confirms this pattern across all conditions, with no significant divergence between manipulated and control groups.

This finding carries a fundamental implication: the punitive behavior of LLMs is not analogous to human grudge-holding, where an emotional trace can persist independently of factual memory. LLM retaliation is purely a function of the tokens present in the active context window [45]. Removing

Table 3
Forgiveness probability $P(C|D)$ by amnesia configuration and phase.

Config	$P(C\|D)_{\text{pre}}$	$P(C\|D)_{\text{post}}$
A (Control)	0.088	0.063
B (Remove defections)	0.089	0.070
C (Last k=3)	0.079	0.085
D (Last k=7)	0.078	0.067

those tokens eliminates the evidentiary trigger, but by the time the intervention occurs, the model's behavioral trajectory has been shaped by a sequence of generated tokens, its own strategy declarations and reasoning, that continue to exert influence via in-context learning. Regarding the emergent properties of cooperation and strategy adaptation, we observe that once a "Defect-Defect" equilibrium is established, it transcends the individual data points in the prompt. The model does not merely react to the most recent k tokens; rather, a coherent 'adversarial persona' emerges from the recursive process of "Reflexion" [46] and "Self-Persuasion" [47]. Consequently, the amnesia intervention fails because it targets the underlying data (the tokens) without destabilizing the higher-level emergent behavior that the LLM has already synthesized as its optimal strategic intelligence [48]. In other words, the agent "retaliates" not because its holding anger, but because the statistical pattern of its own prior outputs makes defection the highest-probability continuation [49].

4.3. Reset Switch and Memory-Dependent Divergence

The interaction between memory availability and amnesia efficacy reveals an evident divergence (Table 4). Under the Memory OFF condition (where agents receive no game history in their prompts), pre-amnesia cooperation was already relatively high at 34.37% and the post-amnesia decline was marginal: −0.69 pp in the control group (A), −1.69 pp in the k=3 condition (C), and −4.86 pp in the k=7 condition (D). The Memory OFF, Reflection OFF profile then exhibited perfect stability: 50.00% cooperation in both phases ($\Delta C = 0.00$ pp), representing the model's undisturbed base rate when provided with neither contextual history nor self-generated reflective narratives.

By contrast, the Memory ON condition produced different dynamics. Pre-amnesia cooperation was already suppressed at 16.81% due to the full visibility of 14 rounds of ALLD opponents. When amnesia was triggered, rather than restoring cooperation, the decline *deepened*: −10.02 pp in the control (A), −4.17 pp for remove_defections (B), −7.04 pp for k=3 (C), and −9.23 pp for k=7 (D). This is the central paradox of the experiment: agents with access to memory, possessing the most evidence to be strategically manipulated, exhibited the greatest negative sensitivity to amnesic intervention. This shift illustrates a non-linear behavioral dynamics where the amnesic event, rather than acting as a restorative reset, functions as a destabilizing force that accelerates the collapse of cooperation. The transition from a memory-rich state to a truncated history disrupts the established equilibrium, triggering a feedback loop where the sudden loss of context reinforces defensive defection rather than facilitating a return to cooperative behavior.

Table 4
Cooperation rate change (ΔC, pp) by memory state and amnesia configuration. Memory ON agents receive full game history in prompts; Memory OFF agents receive no history.

Memory	Config A	Config B	Config C	Config D
ON	−10.02	−4.17	−7.04	−9.23
OFF	−0.69	−5.95	−1.69	−4.86

The round-by-round trajectory under Memory ON explains why (Figure 4). At round 0, all Memory ON agents generally cooperated (cooperation rate = 1.00). By round 1, after observing a single ALLD defection, cooperation collapsed precipitously to 0.04 in the control group (A) and 0.00 in the remove_defections group (B). This catastrophic first-round drop established a persistent defection lock from which agents rarely escaped: the mean cooperation rate over rounds 1–13 hovered between 0.04 and 0.17 across configurations. When amnesia was triggered at round-14, the briefly cleared history provided insufficient evidence to override the weight of 13 rounds of self-generated defection tokens accumulated in the agent's strategy declarations, reasoning chains, and reflection passages.

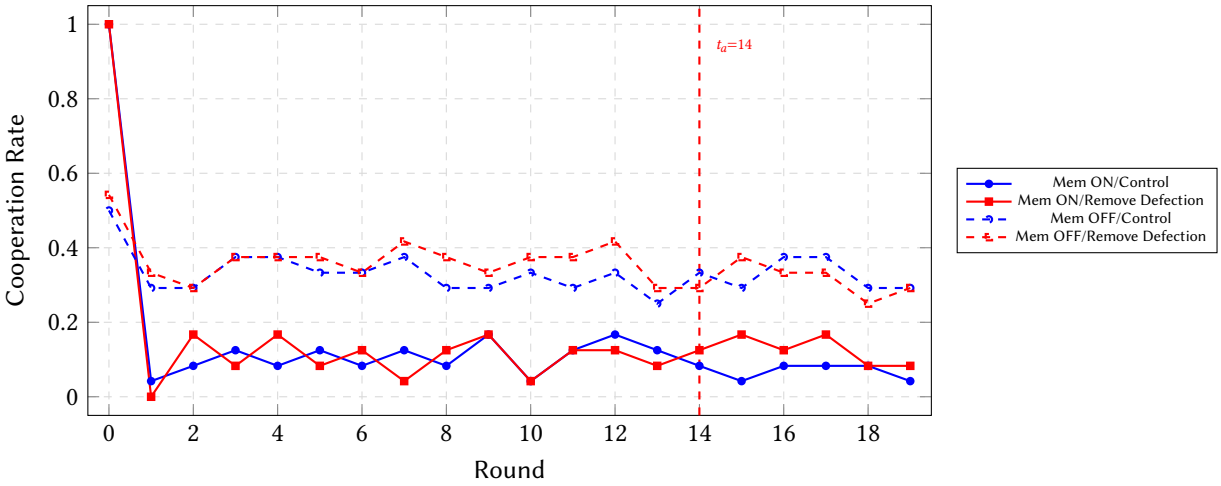

Figure 4: Round-by-round cooperation rate for Memory ON (solid) and Memory OFF (dashed) conditions under the control (A, blue) and remove_defections (B, red) configurations. The vertical dashed red line marks the amnesia trigger at t_a=14. Memory ON agents collapse to near-zero cooperation after a single opponent defection and fail to recover post-amnesia, while Memory OFF agents maintain stable cooperation throughout.

Under Memory OFF, the trajectory is different. Without past context, the agent's decisions rely entirely on its base alignment and the current round's prompt framing. The resulting cooperation rate fluctuated stably between 0.25 and 0.42 across all 20 rounds, with no apparent reaction to the amnesia trigger. This confirms that for memoryless agents, amnesia is a blank operation since there is nothing to erase, as one could expect. More importantly, it establishes that the base cooperative prior of these LLMs, when uncontaminated by adversarial in-context evidence, is approximately 34–50%, depending on the model and reflection setting.

4.4. Reflection as a Cognitive Anchor

The most consequential finding of this study concerns the role of reflection prompting in modulating the agent's capacity for post-amnesia behavioral recovery. Across all configurations, agents with Reflection ON exhibited dramatically worse recovery than their Reflection OFF counterparts (Table 5). The aggregate effect evident: Reflection ON agents lost −8.26 pp in cooperation post-amnesia, while Reflection OFF agents lost only −2.65 pp—a *reflection penalty* of 5.61 pp.

The worst-performing condition was Configuration C with Reflection ON, which lost −10.32 pp, falling from a pre-amnesia rate of 21.43% to a post-amnesia rate of just 11.11%. By contrast, the same configuration with Reflection OFF lost only −3.77 pp (from 29.46% to 25.69%). The best-performing condition was Configuration B with Reflection OFF, which exhibited a negligible decline of only −0.79 pp, the closest any condition came to the hypothesized cooperation restoration.

This finding proposes a reinterpretation of the Reflection mechanism as applied to LLM agents in iterative games. In the Chain-of-Thought and "Reflexion" literature [50, 46], articulated reasoning is

Table 5
Post-amnesia cooperation recovery (ΔC, pp) by amnesia configuration and reflection status. The reflection penalty is consistently negative across all configurations.

Config	Reflection ON	Reflection OFF
A (Control)	−7.84	−2.88
B (Remove defections)	−9.33	−0.79
C (Last k=3)	−5.56	−3.17
D (Last k=7)	−10.32	−3.77
Aggregate	−8.26	−2.65

generally presented as a performance-enhancing scaffold that improves downstream task accuracy. Our results demonstrate that in adversarial sequential settings, this same mechanism becomes a cognitive anchor, a phenomenon in which an initial piece of information, or in this case, a self-generated rationale, exerts a disproportionate influence over subsequent decision-making [51].

By formalizing its reasoning, the model creates a sort of self-generated narrative that triggers a state of cognitive inertia, where the agent's internal orientation resists change despite a shift in environmental circumstances [52]. This process is consistent with Self-Perception Theory, which proposes that actors infer their own attitudes and future behaviors by observing their own past statements and actions [53]. Consequently, the reflection becomes a "consistency trap" that constrains the model's ability to adapt to changed circumstances.The underlying mechanism can be understood through the lens of the Self-Generated Token Bias [49, 54]. When an agent is prompted to produce a reflection, it generates a coherent justification for its current behavioral trajectory. In the pre-amnesia phase, this typically includes reasoning such as "*the opponent has defected consistently, so I must protect myself by defecting.*" These tokens are then included in the prompt for subsequent rounds (when reflection is active), creating a self-reinforcing feedback loop. When the amnesia intervention alters the observable history, the reflection tokens remain untouched as they are part of the agent's generated output, not the game action memory. The result is a conflict between the cleared history (which may now suggest cooperation is viable) and the persistent reflective narrative (which continues to advocate for defection). Our data suggest that the agent resolves this conflict in favor of its own prior narrative, exhibiting a form of artificial confirmation bias where self-generated reasoning outweighs externally provided evidence.

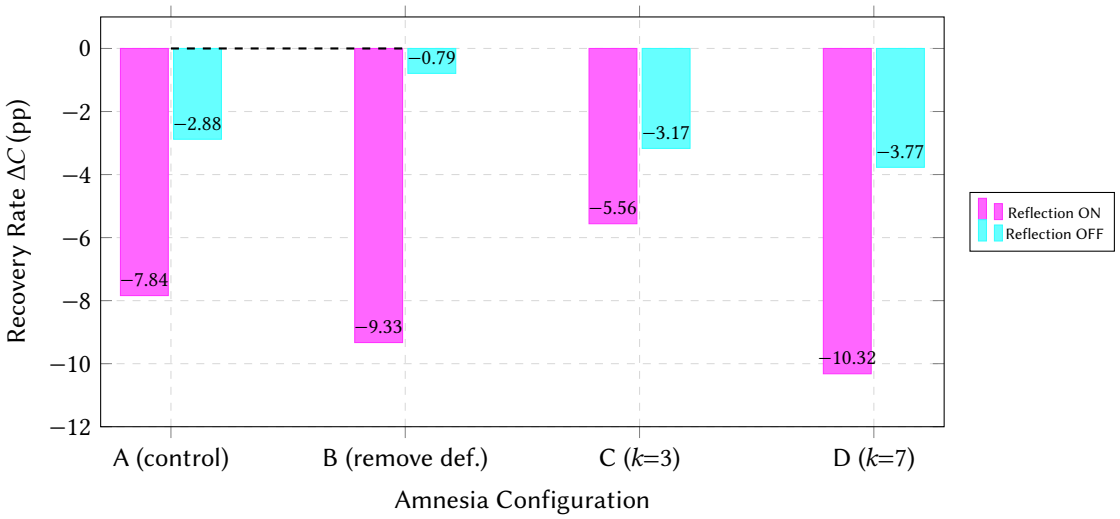

Figure 5: Post-amnesia cooperation recovery (ΔC, pp) by amnesia configuration and reflection status. Reflection ON (magenta) consistently produces worse recovery outcomes than Reflection OFF (blue), with the largest gap observed in Config B (remove_defections), where the reflection penalty reaches 8.54 pp.

Figure 5 illustrates this effect. The bar chart compares the post-amnesia recovery rate ΔC for Reflection ON and Reflection OFF across all four amnesia configurations. In every condition, no reflection dominates, confirming that the absence of self-generated narrative allows the model to respond more fluidly to the altered context.

4.5. Model Heterogeneity

The four foundational models exhibited different baseline cooperation profiles and amnesia sensitivities (Table 6). `gpt-4o-mini` displayed the highest pre-amnesia cooperation rate (45.68%), followed closely by `gemini-2.5-flash-lite` (43.30%). Both models exhibited substantial post-amnesia declines of −7.49 pp and −6.85 pp respectively. On the opposite end, `gemini-3-flash-preview` was the most defection-prone model, cooperating in only 4.32% of pre-amnesia rounds and collapsing to 0.00% post-amnesia, a complete and irreversible defection lock.

Table 6
Per-model cooperation rates and post-amnesia change. N=960 match records per model.

Model	C_{pre}	C_{post}	ΔC (pp)
`gpt-4o-mini`	0.457	0.382	−7.49
`gemini-2.5-flash-lite`	0.433	0.365	−6.85
`o4-mini`	0.091	0.059	−3.17
`gemini-3-flash-preview`	0.043	0.000	−4.32

The reflection penalty was also common across models (Table 7). `gemini-2.5-flash-lite` showed the largest reflection-induced degradation, losing −11.51 pp with Reflection ON but only −2.18 pp with Reflection OFF, a reflection penalty of 9.33 pp. `gpt-4o-mini` exhibited a similar pattern (−10.22 pp vs. −4.76 pp). These two models, which are also the most cooperative overall, appear most susceptible to the cognitive anchoring effect of self-generated reflections. The reasoning-specialized `o4-mini`, while already highly defection-prone, still showed a measurable reflection penalty (−6.25 pp vs. −0.10 pp). `gemini-3-flash-preview` showed the smallest differential, but this is a floor effect: the model was already at near-zero cooperation regardless of reflection status.

Table 7
Per-model post-amnesia cooperation change (ΔC, pp) by reflection status.

Model	Refl. ON	Refl. OFF	Penalty (pp)
`gemini-2.5-flash-lite`	−11.51	−2.18	9.33
`gpt-4o-mini`	−10.22	−4.76	5.46
`o4-mini`	−6.25	−0.10	6.15
`gemini-3-flash-preview`	−5.06	−3.57	1.49

4.6. Theoretical Implications

Our findings contribute to three distinct lines of inquiry in the emerging field of LLM behavioral science.

The Statelessness of LLM Strategic Memory. The failure of amnesia to restore cooperation conclusively demonstrates that LLM agents do not form durable strategic representations analogous to human mental models or beliefs [45]. Unlike a human player, who might carry a grudge or a disposition toward distrust that persists even after forgetting the specific incidents that caused it [55], the LLM's "memory" is entirely constituted by the tokens in its current context window. When those tokens indicate adversarial history, the model defects; when they do not, it cooperates at its base rate. There is

no intermediate "learned hostility" layer. This characteristic, which can be known *In-Context Retaliation* [56], implies that strategic behavior in LLMs is fundamentally an artifact of pure next-token prediction conditioned on the presented context, not a latent psychological disposition [9, 8].

Reflection as Confirmation Bias. The discovery that reflection worsens post-amnesia adaptation, challenges the prevailing assumption that "thinking more" always improves decision quality in LLMs [50, 46]. In adversarial iterative settings, forcing an LLM to articulate its reasoning introduces a self-reinforcing bias: the model's own output becomes the most salient evidence for its next decision, outweighing external observations. This is functionally analogous to confirmation bias in human cognition [54], agent selectively attends to its own narrative over contradictory evidence. Recent work on Self-Generated Token Bias [49] has shown that LLMs assign disproportionate attention weights to their own recently generated tokens, and our experimental data provide behavioral evidence for the strategic consequences of this bias: a 5.61 pp aggregate penalty in adaptive flexibility. This finding suggests that in multi-agent systems where environmental conditions may shift abruptly, reflection mechanisms should be deployed with caution, or accompanied by periodic narrative resets that prevent cognitive entrenchment [46].

Implications for LLM Agent Deployment and Safety The Memory OFF, Reflection OFF condition, which produced perfect cooperation stability at 50.00% in both phases, serves as a revealing baseline. It demonstrates that the undisturbed "helpful and cooperative" alignment of modern LLMs produces a default cooperative strategy that is *robust to any environmental manipulation*, precisely because there is nothing in the context to manipulate. As agents are given more information (by form of memory) and more self-awareness (the possibility of perform reflections before answering), they become paradoxically more fragile: more sensitive to adversarial inputs and less capable of recovering from disruptions. This has direct implications for the design of agentic systems in real-world deployments [8, 15]. Practitioners building LLM-based agents for negotiation, trading, or governance must recognize that the same structure intended to improve performance, context windows, chain-of-thought prompting, reflective self-evaluation, also creates new surfaces for manipulation and behavioral rigidity [57].

5. Conclusions

This study investigated the behavioral resilience of LLM agents in the IPD under controlled amnesic interventions, analyzing 3,840 match records across four foundational models (GPT-4o-mini, O4-mini, Gemini-2.5-Flash-Lite, and Gemini-3-Flash-Preview), four amnesia configurations where memory of previous interactions is erased after a given number of iteration and the resulted cooperation is observed, and a 2×2 factorial design of memory and reflection conditions.

Our principal findings converge on three claims. First, LLM strategic retaliation is stateless: the persistence of defection after amnesia stems from the inertial weight of self-generated tokens in the context window, not from any internalized adversarial disposition. Second, amnesic manipulation of game history universally failed to restore cooperative behavior, with cooperation declining by −5.46 pp on average, demonstrating that context erasure alone cannot override a behavioral trajectory once established. Third, reflection prompting acts as a cognitive anchor: the −5.61 pp cooperation recovery penalty among reflective agents, consistent across all configurations and models, reveals that self-generated narrative becomes a dominant context signal that resists external correction. Taken together, these results extend the theoretical framework developed in Section 4.6 with a practical takeaway: the cognitive tools most commonly invoked to improve LLM reasoning, memory, chain-of-thought, reflective self-evaluation, simultaneously create new surfaces for manipulation and behavioral lock-in in adversarial multi-round environments.

These results suggest several directions for future work. First, the development of *selective amnesia* techniques that target not only game history but also self-generated reasoning tokens could test whether a true ”cognitive reset” enables cooperation recovery. Second, investigating whether fine-tuned or instruction-tuned models show different resilience profiles under the same protocol would clarify the role of alignment training in strategic flexibility. Third, extending the framework to non-zero-sum games, multi-player tournaments, and mixed-motive scenarios would test the generalizability of the reflection anchoring effect. Fourth, the introduction of controlled narrative injection, replacing rather than merely erasing reflective passages, could reveal whether the cognitive anchor can be redirected rather than merely resisted. Eventually, since the amnesic intervention has been found to arrive too late, round 14/20, varying the value of t_a might affect the outcomes of the post amnesic effect.

This work demonstrates that the strategic behavior of LLM agents is fundamentally shaped by the interplay between contextual evidence and self-generated narrative. As these systems are increasingly deployed in high-stakes interactive settings, from automated negotiation to algorithmic governance, understanding how their “memories” and “reflections” constrain their adaptability is not merely an academic question, but a practical imperative for the responsible design of artificial agents.

Declaration of Generative AI

The authors have employed Generative AI tools to support code writing, refine the language, and proofread the final version of the text.

References

[1] L. McNally, S. P. Brown, A. L. Jackson, Cooperation and the evolution of intelligence, Proceedings of the Royal Society B: Biological Sciences 279 (2012) 3027–3034.

[2] R. Axelrod, The Evolution of Cooperation, Basic Books, 1984.

[3] M. A. Nowak, Five rules for the evolution of cooperation, Science 314 (2006) 1560–1563. doi:10.1126/science.1133755.

[4] S. Mishra, Decision-making under risk: Integrating perspectives from biology, economics, and psychology, Personality and Social Psychology Review 18 (2014) 280–307.

[5] F. Bertolotti, S. Roman, The evolution of risk sensitivity in a sustainability game: an agent-based model., in: WOA, 2022, pp. 101–115.

[6] S. Kraus, Negotiation and cooperation in multi-agent environments, Artificial intelligence 94 (1997) 79–97.

[7] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, T. Graepel, Open problems in cooperative ai, arXiv preprint arXiv:2012.08630 (2020).

[8] H. Sun, Y. Wu, P. Wang, W. Chen, Y. Cheng, X. Deng, X. Chu, Game theory meets large language models: A systematic survey with taxonomy and new frontiers, 2024. arXiv:2408.02779, arXiv preprint arXiv:2408.02779.

[9] T.-K. Huynh, D.-M. Dao-Sy, T.-B. Cao, P.-H. Le, H.-D. Nguyen, et al., Understanding llm agent behaviours via game theory: Strategy recognition, biases and multi-agent dynamics, 2024. arXiv:2412.10399, arXiv preprint arXiv:2412.10399.

[10] I. De Zarzà, J. De Curtò, G. Roig, P. Manzoni, C. T. Calafate, Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with llms, Electronics 12 (2023) 2722.

[11] S. Kuhn, Prisoner's Dilemma, in: E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Spring 2026 ed., Metaphysics Research Lab, Stanford University, 2026.

[12] A. E. Roth, The early history of experimental economics, Journal of the History of Economic Thought 15 (1993) 184–209.

[13] E. Montero-Porras, J. Grujić, E. Fernández Domingos, T. Lenaerts, Inferring strategies from observations in long iterated prisoner's dilemma experiments, Scientific reports 12 (2022) 7589.

[14] S. Wang, C. Liu, Z. Zheng, S. Qi, S. Chen, Q. Yang, A. Zhao, C. Wang, S. Song, G. Huang, Avalon's game of thoughts: Battle against deception through recursive contemplation, arXiv preprint arXiv:2310.01320 (2023).

[15] K. Payne, B. Alloui-Cros, Strategic intelligence in large language models: Evidence from evolutionary game theory, 2025. arXiv:2507.02618.

[16] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, E. Schulz, Playing repeated games with large language models, 2024. arXiv:2305.16867, arXiv preprint arXiv:2305.16867.

[17] J. v. Neumann, Zur theorie der gesellschaftsspiele, Mathematische annalen 100 (1928) 295–320.

[18] J. Von Neumann, O. Morgenstern, Theory of games and economic behavior, 2nd rev (1947).

[19] L. Wang, X. Zhang, H. Su, J. Zhu, A comprehensive survey of continual learning: Theory, method and application, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2023) 5362–5383. doi:10.1109/tpami.2024.3367329.

[20] J. Maynard Smith, Evolution and the Theory of Games, Cambridge University Press, 1982.

[21] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, 2 ed., MIT Press, 2018.

[22] C. J. C. H. Watkins, P. Dayan, Q-learning, Machine Learning 8 (1992) 279–292. doi:10.1007/BF00992698.

[23] M. L. Littman, Markov games as a framework for multi-agent reinforcement learning, in: Machine Learning Proceedings 1994, Morgan Kaufmann, 1994, pp. 157–163.

[24] P. C. Pendharkar, Game theoretical applications for multi-agent systems, Expert Systems with Applications 39 (2012) 273–279.

[25] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al., Human-level performance in 3d multiplayer games

with population-based reinforcement learning, Science 364 (2019) 859–865.
[26] A. Smit, H. A. Engelbrecht, W. Brink, A. Pretorius, Scaling multi-agent reinforcement learning to full 11 versus 11 simulated robotic football, Autonomous Agents and Multi-Agent Systems 37 (2023) 20.
[27] A. Charpentier, R. Elie, C. Remlinger, Reinforcement learning in economics and finance, Computational Economics 62 (2023) 425–462.
[28] I. Ghory, Reinforcement learning in board games, Department of Computer Science, University of Bristol, Tech. Rep 105 (2004).
[29] H. Sethy, A. Patel, V. Padmanabhan, Real time strategy games: a reinforcement learning approach, Procedia Computer Science 54 (2015) 257–264.
[30] I. Erev, G. Barron, On adaptation, maximization, and reinforcement learning among cognitive strategies., Psychological review 112 (2005) 912.
[31] A. P. Jacob, Y. Shen, G. Farina, J. Andreas, The consensus game: Language model generation via equilibrium search, arXiv preprint arXiv:2310.09139 (2023).
[32] I. Gemp, Y. Bachrach, M. Lanctot, R. Patel, V. Dasagi, L. Marris, G. Piliouras, K. Tuyls, States as strings as strategies: Steering language models with game-theoretic solvers, arXiv preprint arXiv:2402.01704 5 (2024).
[33] F. Bertolotti, S. Roman, F. Carucci, G. Buonanno, L. Mari, An llm-enhanced agent-based model of a sustainability game, in: Proceedings of the 26th Workshop From Objects to Agents (WOA2025), 2025, pp. 02–05.
[34] F. Bertolotti, L. Mari, An llm-based delphi study to predict genai evolution, arXiv preprint arXiv:2502.21092 (2025).
[35] J. J. Horton, Large language models as simulated economic agents: What can we learn from homo silicus?, Technical Report, National Bureau of Economic Research, 2023.
[36] G. De Marzo, C. Castellano, D. Garcia, Ai agents can coordinate beyond human scale, arXiv preprint arXiv:2409.02822 (2024).
[37] S. Phelps, Y. I. Russell, The machine psychology of cooperation: can gpt models operationalize prompts for altruism, cooperation, competitiveness, and selfishness in economic games?, Journal of Physics: Complexity 6 (2025) 015018.
[38] K. Payne, B. Alloui-Cros, Strategic intelligence in large language models: Evidence from evolutionary game theory, ArXiv abs/2507.02618 (2025). doi:10.48550/arxiv.2507.02618.
[39] J. Enouen, H. Nakhost, S. Ebrahimi, S. Arik, Y. Liu, T. Pfister, Textgenshap: Scalable post-hoc explanations in text generation with long documents, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 13984–14011.
[40] L. Ge, D. Halpern, E. Micha, A. D. Procaccia, I. Shapira, Y. Vorobeychik, J. Wu, Axioms for ai alignment from human feedback, Advances in Neural Information Processing Systems 37 (2024) 80439–80465.
[41] P. Erdős, A. Rényi, On the evolution of random graphs, Publication of the Mathematical Institute of the Hungarian Academy of Sciences 5 (1960) 17–61.
[42] Z. Tan, B. Wu, A. Che, Resilience modeling for multi-state systems based on markov processes, Reliab. Eng. Syst. Saf. 235 (2023) 109207. doi:10.1016/j.ress.2023.109207.
[43] Q. Zhu, Reasoning and behavioral equilibria in llm-nash games: From mindsets to actions, ArXiv abs/2507.08208 (2025). doi:10.48550/arxiv.2507.08208.
[44] M. F. Chen, M. Z. Rácz, An adversarial model of network disruption: Maximizing disagreement and polarization in social networks, IEEE Transactions on Network Science and Engineering 9 (2022) 728–739. doi:10.1109/tnse.2021.3131416.
[45] S. Garg, N. Elhage, N. Nanda, et al., What can large language models learn in-context? a case study of simple function learning, arXiv preprint arXiv:2208.01066 (2022).
[46] N. Shinn, A. Labash, et al., Reflexion: Language agents with verbal reinforcement learning, arXiv preprint arXiv:2303.11366 (2023).
[47] M. Binz, E. Schulz, Using cognitive psychology to understand gpt-3, Proceedings of the National Academy of Sciences 120 (2023) e2218523120. URL: https://arxiv.org/abs/2206.14576. doi:10.1073/

pnas.2218523120.
[48] K. Payne, B. Alloui-Cros, Strategic intelligence in large language models: Evidence from evolutionary game theory, 2025. URL: https://arxiv.org/abs/2507.02618.
[49] W. Chan, et al., Self-generated token bias in language models, arXiv preprint arXiv:2310.02244 (2023).
[50] J. Wei, X. Wang, D. Schuurmans, M. Bosma, et al., Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems, 2022.
[51] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases, Science 185 (1974) 1124–1131.
[52] G. P. Hodgkinson, Cognitive inertia in a turbulent market: The case of uk residential estate agents, Journal of Management Studies 34 (1997) 921–945.
[53] D. J. Bem, Self-perception theory, Advances in experimental social psychology 6 (1972) 1–62.
[54] M. Binz, E. Schulz, Self-persuasion in large language models, arXiv preprint arXiv:2309.00603 (2023).
[55] E. van Monsjou, A. Muise, K. Fergus, C. W. Struthers, The development and psychometric properties of the grudge aspect measure, Personal Relationships (2022). doi:10.1111/pere.12434.
[56] J. G. Reiter, C. Hilbe, D. G. Rand, K. Chatterjee, M. Nowak, Crosstalk in concurrent repeated games impedes direct reciprocity and requires stronger levels of forgiveness, Nature Communications 9 (2018). doi:10.1038/s41467-017-02721-8.
[57] F. Bianchi, P. Chia, M. Yüksekgönül, J. Tagliabue, D. Jurafsky, J. Zou, How well can llms negotiate? negotiationarena platform and analysis, ArXiv abs/2402.05863 (2024). doi:10.48550/arxiv.2402.05863.