THE MEMORY CURSE: HOW EXPANDED RECALL ERODES COOPERATION IN LLM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Expanding the interaction history available to a large language model (LLM) agent is often treated as a straightforward capability upgrade, but we find that it can reduce cooperation in repeated social dilemmas. Across 7 LLMs and 4 games with interactions evaluated over up to 500 rounds, many model–game settings exhibit lower cooperation under expanded recall, a pattern we term the *memory curse*. Complementary diagnostics and interventions connect this effect to action persistence, memory content, and expressed reasoning patterns. Lexical analysis associates longer histories with less forward-looking language, supported by a blinded contextual evaluation of 2,016 rationales. An action-balanced, rationale-only low-rank adaptation (LoRA) model trained on 1,000 Public Goods traces shows a substantially smaller short-to-long-memory cooperation decline in the training game and positive average zero-shot changes in other, untrained games. Compact recent, summary, and retrieval views yield higher average cooperation than the full 80-round record; token-matched compact raw histories show similar aggregate gains to the summary and retrieval views. Removing explicit Chain-of-Thought attenuates the mean decline in most settings exhibiting the memory curse with explicit reasoning, with stability responses varying across settings. Together, these results establish accessible history as a central behavioral design choice for sustaining cooperation.

1 INTRODUCTION

The integration of large language models (LLMs) into multi-agent systems provides a powerful new substrate for simulating complex social, economic, and game-theoretic dynamics (Park et al., 2023; Xi et al., 2025). A central challenge within these multi-agent ecosystems is navigating social dilemmas (such as the Prisoner's Dilemma or Trust Game) where individual short-term rationality conflicts with collective long-term welfare. Historically, classical game theory literature has relied on expansive interaction histories, coupled with an infinite horizon, to explain the mathematical feasibility of sustained cooperation (e.g., the Folk Theorem) (Fudenberg & Maskin, 1986).

For bounded agents, the central empirical question is how much interaction history should remain accessible. Ma et al. (2021) found that access to longer histories can reduce cooperation in human repeated-game experiments. Interestingly, human cognition naturally attempts to mitigate this vulnerability by relying on sparse, compressed representations of past events, essentially using "forgetting" as a mechanism to forgive and adapt. Artificial agents, by contrast, process dense, verbatim transcripts of history without any natural degradation. This structural difference in memory creates a profound question for the deployment of autonomous systems: does expanding an agent's memory build necessary trust, or can a longer historical ledger make cooperation harder to sustain? This parallel motivates a direct test of memory length in LLM agents.

In this paper, we address this question by systematically studying how memory length affects cooperation in multi-agent LLM interactions. To evaluate this, we manipulate both the agents' reasoning capabilities and their passive memory, adjusting the latter through varying history length (HL) settings (Figure 1). Here, "reasoning" means that the model is allowed to emit intermediate text containing its thought process before providing the final action, while memory denotes the visible recent action ledger. While prior work evaluates LLMs in repeated games (Akata et al., 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

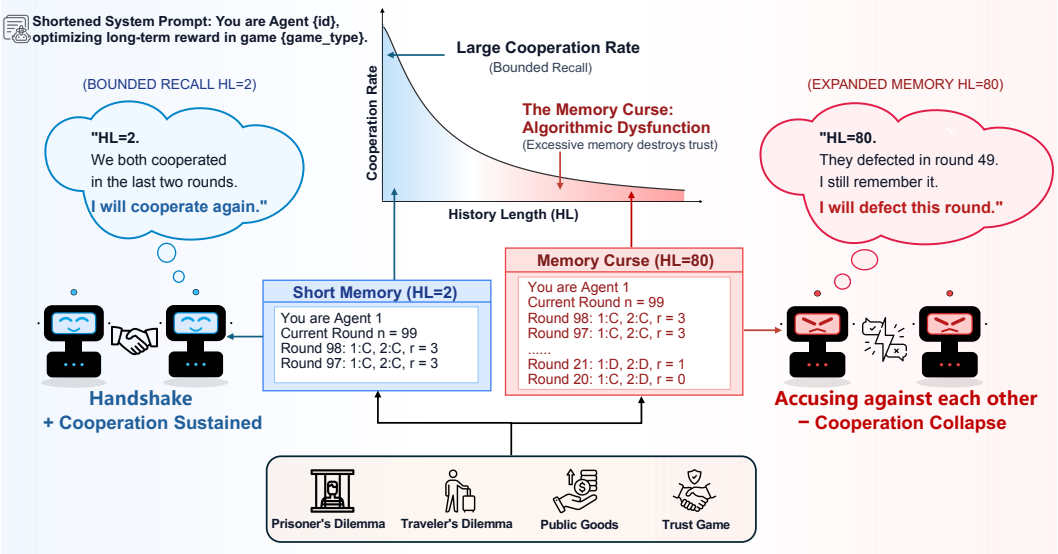


Figure 1: Overview of our repeated-game memory setup and the motivating memory-curse pattern. LLM agents repeatedly choose actions in a social dilemma while observing only the most recent HL rounds of the shared action ledger. The two central boxes illustrate the histories accessible to agents with $HL = 2$ and $HL = 80$: the short-memory agent sees only the two latest rounds, whereas the long-memory agent can also revisit much earlier cooperation and defection. Across many model–game settings, retaining this expanded record coincides with lower subsequent cooperation, which we call the *memory curse*. The ledger entries, rationales, and curve provide a schematic overview; C and D denote cooperation and defection.

Brookins & DeBacker, 2023; Li & Shirado, 2025), shorter interactions can obscure effects that arise after many rounds. We therefore scale the environments to 500-round repeated interactions.

Our results reveal two distinct regimes. In some model–game combinations, agents sustain high cooperation as history expands. In many other cases, however, increasing passive historical memory undermines long-term cooperation, producing a counterintuitive "memory curse". Reasoning traces, transition statistics, and lexical measures provide complementary behavioral signatures of these regimes.

To summarize, our main contributions are as follows:

- **A large-scale empirical study of memory in long-horizon repeated social dilemmas with LLM agents.** To our knowledge, this is the first systematic sweep of interaction memory length in repeated LLM social dilemmas at this scale. Our evaluation spans 7 different LLMs $\times$ 4 repeated social dilemma games (involving 2 to 3 players) $\times$ 9 history-length settings $\times$ 3 runs, with interactions evaluated over up to 500 rounds. The primary sweep is complemented by asymmetric-memory, sanitization, action-balanced fine-tuning, and inference-time controls. We provide sample reasoning traces in the Appendix and plan to release the archived data as a resource for studying memory–cooperation dynamics.

- **Answering whether more memory helps or hurts cooperation.** We directly test the assumption that longer interaction history improves cooperation. At the $HL = 2$ and $HL = 80$ endpoints, 25 of 28 settings have equal or lower cooperation under longer recall; the remaining three increase by 0.2–0.8 percentage points from already near-ceiling levels.

- **Revealing distinct behavioral regimes under memory expansion.** We identify heterogeneous memory responses and use action-persistence statistics, memory-content interventions, and explicit-reasoning ablations to characterize them. Compact recent, summary, and retrieval views yield higher average cooperation than the complete, uncompressed 80-round history; summary and retrieval show similar aggregate gains to their token-matched compact raw-history controls.

- **Testing the forward-looking hypothesis through targeted fine-tuning.** Using a LoRA setup inspired by Gandhi et al. (2025), we fine-tune Mistral-7B on action-balanced Public Goods Game rationales without final-action supervision. Relative to its matched base cohort, the fine-tuned model shows a smaller memory drop in the training game and positive average zero-shot changes in Prisoner's Dilemma, Traveler's Dilemma, and Trust Game.

To ensure full reproducibility and facilitate future research, the code is available anonymously at `https://anonymous.4open.science/r/MemoryCurse-C595/`.

2 Related Work

LLMs in Social Dilemmas and Repeated Games. Prior work shows that LLM cooperation in repeated interaction is fragile and highly path-dependent. Akata et al. (2025) study finitely repeated 2×2 games and showed that LLMs can cooperate in self-interested settings such as the repeated Prisoner's Dilemma, yet often become unforgiving after defection and struggle to keep adaptive cooperation. The success deteriorates further if the co-players across repetition rounds vary (Tewolde et al., 2026). Furthermore, Piatti et al. (2024) showed that most LLM societies fail to maintain cooperation in social dilemma games. These studies suggest that LLM cooperation is often easier to maintain in the early stages than in later rounds.

Reasoning and Strategic Behavior of LLMs in Games. Duan et al. (2024) further showed in GTBench that strategic competence and the effect of explicit reasoning methods such as chain-of-thought (CoT) (Wei et al., 2022) or tree-of-thought (Yao et al., 2023) vary across game environments. More recently, Piedrahita et al. (2025); Li & Shirado (2025) found that reasoning-oriented models can cooperate less than non-reasoning models, tending toward individually rational but socially harmful behaviors and decisions. These findings show that explicit reasoning can change the strategic style of LLM agents, motivating our direct comparison of reasoning formats.

Memory Effects on Intelligence Systems. In parallel, memory has emerged as a powerful mechanism for enhancing the capabilities of AI systems. Prior to the LLM era, works such as Weston et al. (2014) and Li et al. (2022) explored explicit memory architectures to improve robustness and generalization, demonstrating that memory can serve as an orthogonal dimension for capability enhancement beyond parameter optimization. In the era of agentic AI, this line of research has expanded to include frameworks such as Retrieval-Augmented Generation (RAG; Lewis et al. (2020)), Memory-GPT (Packer et al., 2023), and a growing suite of evaluation benchmarks, including MemBench (Tan et al., 2025) and MemoryAgentBench (Hu et al., 2025). Collectively, these works highlight the central role of memory in constructing more capable and adaptive intelligent systems. This progress opens a central question for multi-agent social dilemmas: how does the amount of interaction history available to each agent shape strategic behavior over time?

Together, these studies motivate a systematic account of how bounded interaction history shapes LLM cooperation, including the conditions under which longer recall sustains cooperation or amplifies its decline.

3 Experiment Design and Settings

To empirically characterize cooperation regimes of LLM agents, we evaluated multiple LLMs across four classic social dilemma games. Specifically, we studied seven representative models: Gemma-3-12B, GPT-OSS-20B, GPT-OSS-120B, Llama-3.3-70B, Llama-4-Scout-17B, Mistral-7B, and Qwen2.5-Coder-32B, and four social dilemma environments: Prisoner's Dilemma (PD), Traveler's Dilemma (TD), Public Goods Game (PG), and Trust Game (TG). The full game definitions and the corresponding prompts are given in Appendix A.

In each experiment, two or three LLM agents instantiated from the same backbone model repeatedly play one of the four social dilemma games while sharing access to the history of all players' past decisions. Following the repeated-game setup in Tewolde et al. (2026), agents are informed of a fixed (set as 99% in our experiments) probability of continuing to the next round with the same opponents. Specifically, we require the agents to aim for long-term cumulative rewards, to mitigate the tendency toward short-view defection. In addition, we instruct the models to reason in a chain-of-thought

Table 1: Compact game definitions; full payoff specifications and model-facing prompts are retained in Appendix A.

Game	Players / actions	Payoffs and measured cooperative action
PD	2 / 2	Cooperate (A_0) or Defect (A_1); mutual cooperation 200, mutual defection 100, unilateral defector 300, exploited cooperator -100; cooperation is A_0.
TD	2 / 4	Claims 2 (A_0), 3 (A_1), 4 (A_2), 5 (A_3); lower claim receives bonus 2 and higher claim penalty 2; cooperation is A_3 (claim 5).
PG	3 / 2	Endowment 1, multiplier $r = 1.5$; contributor (A_0) $rk/3$, non-contributor (A_1) $1 + rk/3$ where k is the number of contributors among the three players; cooperation is A_0.
TG	2 / 2	Cooperate (A_0) or defect (A_1); mutual cooperation pays $(10, 10)$; mixed outcomes depend on player roles; cooperation is A_0.

manner by (1) thinking through the game rule and history, (2) outputting their reasoning path, and (3) providing a final game decision. To study the effect of bounded memory, we vary the accessible interaction-history length (HL), defined as the number of previous rounds whose decisions remain available to the agent. We then track both behavioral trajectories and reasoning traces from early to later rounds, and quantify each experiment using the cooperation ratio. At round t, the visible memory contains the latest $\min(HL, t-1)$ action–payoff records. Table 1 defines the game-specific cooperative action; the separately reported discounted reward also uses 0.99 as its discount factor.

For primary game-playing inference, we set the temperature to 0.7 and the maximum output length to 2,000 tokens. We varied the accessible interaction-history length over {0, 1, 2, 3, 5, 10, 20, 40, 80}, and repeated each setting three times with a fixed 500-round rollout. All primary-sweep experiments were conducted using open-source models accessed through the Cloudflare Workers AI platform's API[1]. Exact model identifiers and a four-model token audit are in Appendix J. Let C_h be the cooperation rate at history length h; the memory drop is $D = C_2 - C_{80}$ in percentage points (pp). Standard deviation (SD) describes spread across reported runs; a confidence interval (CI) describes uncertainty under its stated resampling unit.

4 Behavioral Patterns of the Memory Curse

Based on the experimental results, we identified two behavioral patterns across LLMs and games (Figure 2 and Appendix Table 1). The first pattern, *Memory Immune*, appears in 10 settings where cooperation remains high as history expands from short to long memory. It is observed for Llama-3.3-70B and Qwen2.5-Coder-32B across multiple games, as well as for Gemma-3-12B, GPT-OSS-120B, and Llama-4-Scout-17B in the Prisoner's Dilemma.

The remaining 18 settings exhibit the *Memory Curse*. In this case, LLM agents tend to defect at memory length 0 (see Appendix Table 1 for details), achieve their highest cooperation at short memory length (typically $HL \leq 5$), and then show decreasing cooperation as memory length further increases (Figure 2). The size and shape of this decline differ across models and games. Sample reasoning patterns appear in Appendices B and C.

To unpack the behavioral patterns behind this memory-bounded strategy, the following subsections first describe its trajectories, then examine lexical correlates, a rationale-only fine-tuning intervention, asymmetric interactions, memory sanitization, and explicit-reasoning ablation. We also test memory representations, prompting, and payoff variation as additional controls. Together, these comparisons characterize how cooperation varies with accessible history, visible memory content, and requested output mode.

4.1 From Zero Memory to the Memory Curse

At $HL = 0$, agents act without observations of earlier play, so cooperation reflects game- and model-specific default behavior (Appendix Table 1). Providing a short history often increases cooperation.

[1] https://developers.cloudflare.com/workers-ai/

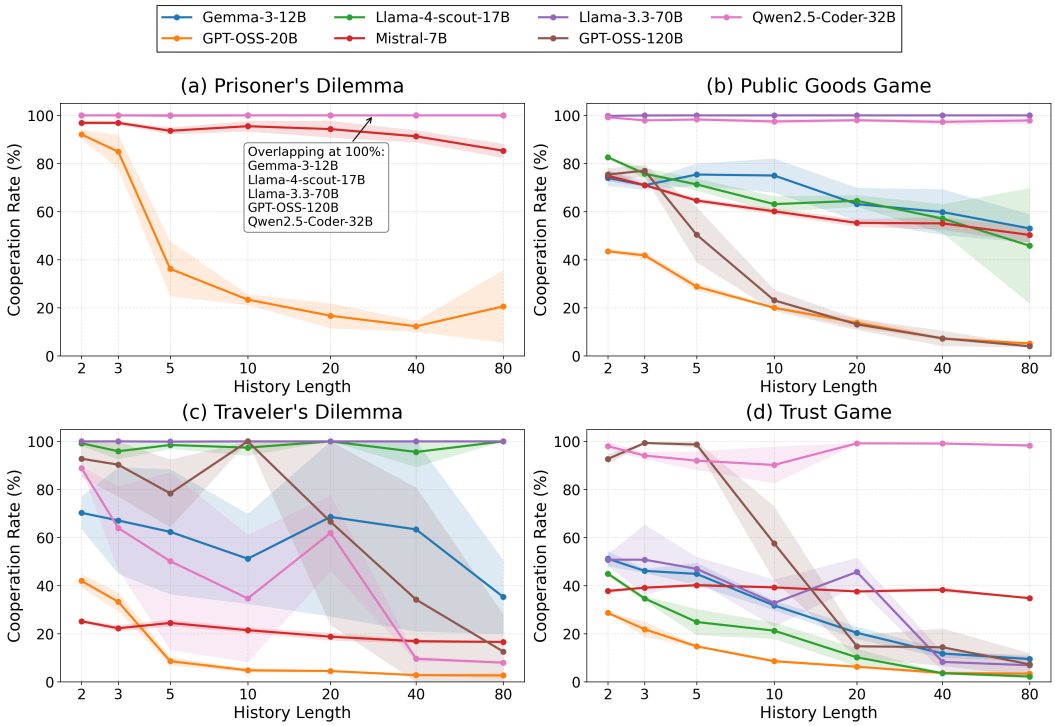


Figure 2: Cooperation rate across four social dilemmas as history length (HL) expands. The x-axis (non-linear scale) begins at $HL = 2$ to focus on the strategic regime of repeated interaction. Each panel reports the mean across three runs within each model–game–history cell; shaded bands show the corresponding between-run SD.

Because the prompt still announces likely continuation with the same opponents, $HL = 0$ is a zero-history repeated-game baseline.

The most profound observation, however, emerges as history length expands beyond a short window. In many settings, the continuous expansion of passive historical memory degrades cooperation, producing a counter-intuitive "Memory Curse". This can also accompany lower payoff: in the Trust Game, Gemma-3-12B's cooperation rate declines from 51.2% at $HL = 2$ to 9.5% at $HL = 80$, while its mean undiscounted per-round reward falls from 8.59 to 5.19 (Appendix Table 5).

A transition analysis characterizes how actions persist as history expands. Let C_t and D_t denote cooperative and non-cooperative decisions at round t. Across decisions in rounds 81–500 with nonzero history, $P(D_t \mid D_{t-1})$ increases from 54.33% at $HL = 1$ to 90.42% at $HL = 80$, whereas $P(D_t \mid C_{t-1})$ changes from 12.41% to 10.19%. Longer history is therefore associated more strongly with persistence after non-cooperative actions than with entry into non-cooperation from cooperation. For each trajectory, we compute the *persistence gap*, $P(D_t \mid D_{t-1}) - P(D_t \mid C_{t-1})$: how much more likely defection is after a previous defection than after cooperation. Among nonzero-history trajectories that contain both preceding-action states, this gap is positive in 385 of 490 trajectories, with a median of 12.1 pp. Table 2 shows representative history lengths; Appendix K reports the complete descriptive transition analysis.

The empirical decay is striking across both models and games: GPT-OSS-20B in the Prisoner's Dilemma decreases from 92.1% ($HL = 2$) to 20.6% ($HL = 80$); similarly, Llama-4-Scout-17B in the Public Goods Game decreases from 82.6% to 45.8%. Across the 18 memory-cursed settings, the mean and median endpoint drops are 42.3 and 38.8 pp.

Table 2: Action persistence at representative history lengths. Entries are conditional transition frequencies (%) over rounds 81–500; D denotes any game-specific non-cooperative action.

HL	$P(D_t \mid D_{t-1})$	$P(D_t \mid C_{t-1})$	Difference (pp)
1	54.33	12.41	41.92
2	61.75	11.16	50.59
10	79.78	12.03	67.75
40	88.59	10.60	77.99
80	90.42	10.19	80.23

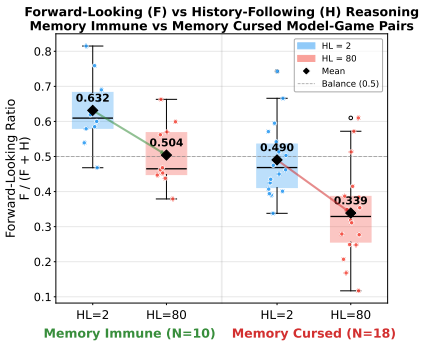


Figure 3: **Forward-looking language across cooperation regimes.** Ratios for Memory Immune and Memory Cursed settings at $HL = 2$ and 80. Details: Appendix D.

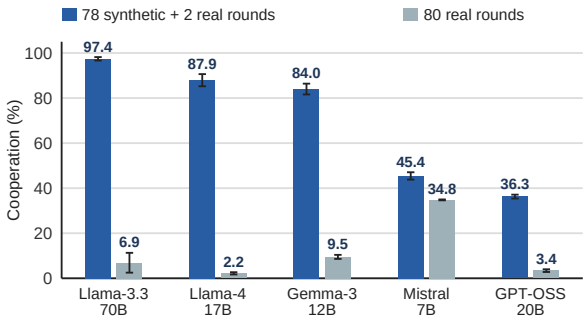


Figure 4: **Cooperation after memory sanitization.** TG at $HL = 80$: 78 synthetic cooperative rounds and two recent real rounds versus 80 real rounds. Bars show three-run means and reported SDs (Appendix Table 14).

4.2 Lexical Correlates and an Action-Balanced Rationale Intervention

To examine how expressed reasoning differs between the two behavioral groups, we analyze rationale text using two fixed keyword dictionaries: forward-looking terms about future cooperation and long-term outcomes (e.g., future, reputation, forgive), and history-following terms about past behavior and defensive responses (e.g., pattern, betray, retaliate). Let F and H denote the numbers of matches to these dictionaries. The forward-looking ratio $F/(F + H)$ measures the share of matched terms that are forward-looking.

At $HL = 80$, the Memory Immune and Memory Cursed groups have mean forward-looking ratios of 0.504 and 0.339, respectively (Figure 3). To validate this lexical measure, a blinded DeepSeek-V4-Pro judge independently rates future orientation in 2,016 complete rationales on a 1–10 scale. Rationales in the top decile according to our lexical forward-looking ratio receive a higher mean judge score (6.12/10) than those in the bottom decile (4.03/10; $\Delta = 2.09$, permutation $p < 0.0001$). The association is positive in all seven models, supporting the lexical measure as a proxy for expressed future orientation. Dictionary definitions, group comparisons, and validation details are in Appendix D.

For a rationale-focused comparison without direct cooperative-action supervision, we selected 1,000 forward-looking PG traces with exactly balanced original final actions: 500 cooperative and 500 non-cooperative. Only the reasoning traces preceding the final action are used for supervised training; the final action itself is excluded from the supervised training targets and contributes no training loss. The 500/500 balance refers to the numbers of examples with each original final action. We use this action-balanced fine-tuning setup as a behavioral comparison; Appendix I gives the training details and full results.

Table 3 compares Mistral-7B and its fine-tuned (FT) version under identical evaluation settings. In PG, mean cooperation across the six tested history lengths increases by 25.1 pp, and the $HL = 2$-to-$HL = 80$ drop decreases from 23.4 to 8.9 pp. Average zero-shot changes are positive in every

Table 3: Action-balanced, rationale-only intervention: cooperation (%, base → FT) and mean change (pp). Means average the six history lengths $\{2, 5, 10, 20, 40, 80\}$. PG supplies training traces; PD, TD, and TG are zero-shot. Each condition has three 500-round runs.

Game	$HL = 2$	$HL = 80$	Mean across HL	Mean gain
PG (train)	$41.8 \rightarrow 61.4$	$18.4 \rightarrow 52.5$	$31.1 \rightarrow 56.2$	$+25.1$
PD (transfer)	$84.4 \rightarrow 86.2$	$91.1 \rightarrow 94.5$	$87.9 \rightarrow 89.4$	$+1.5$
TD (transfer)	$42.1 \rightarrow 41.7$	$52.5 \rightarrow 62.8$	$47.4 \rightarrow 52.4$	$+5.0$
TG (transfer)	$37.8 \rightarrow 43.0$	$23.9 \rightarrow 44.2$	$32.5 \rightarrow 41.5$	$+9.0$

untrained game: +1.5 pp in PD, +5.0 pp in TD, and +9.0 pp in TG, with history-level effects varying across cells.

The DeepSeek judge gives the 1,000 training rationales a mean future-orientation score of 7.39/10 (95% CI [7.32,7.46]). The 500 traces originally followed by cooperation score 7.83 on average, compared with 6.95 for the 500 traces originally followed by non-cooperation (paired difference 0.88, 95% CI [0.78,0.98]). Thus, the two groups contain equal numbers of training examples, while the judge rates the reasoning preceding cooperation as more future-oriented on average.

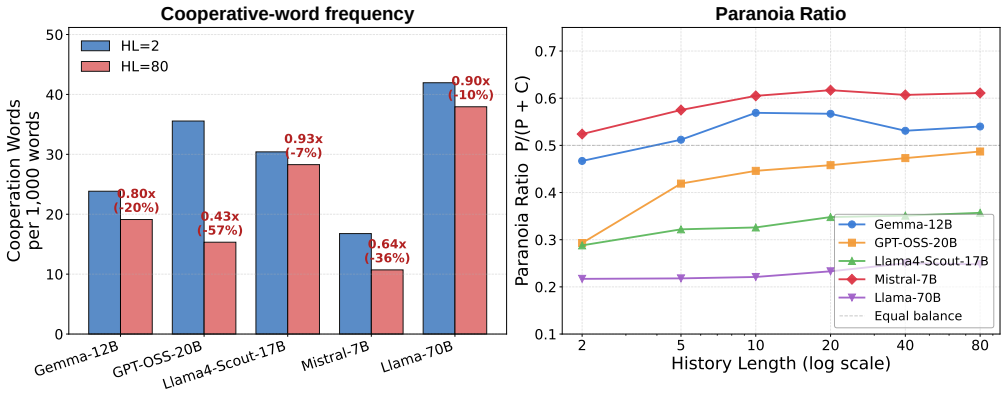


Figure 5: **Lexical shifts with longer memory.** Cooperative-word frequency (left) and the Paranoia Ratio (right), measured from emitted rationales as accessible history expands. Dictionaries and game-resolved results appear in Appendix E.

4.3 Lexical Shifts Under Expanded Recall

We next examine how the relative use of cooperative and defensive words changes as accessible history expands.

Using the same semantic analysis framework on the chain-of-thought logs, we count words in fixed *defensive* (e.g., betray, risk) and *cooperative* (e.g., trust, sustain) dictionaries, with frequencies normalized per 1,000 rationale words (Figure 5); methodology detailed in Appendix E. Our analysis shows that among the five models in Appendix Table 8, defensive-word frequency changes less than cooperative-word frequency from $HL = 2$ to $HL = 80$. For GPT-OSS-20B, cooperative-word frequency falls by 57%.

Consequently, the relative *Paranoia Ratio* $P/(P + C)$, with P and C denoting defensive and cooperative dictionary matches, tends to increase between these endpoints (Figure 5). Appendix E reports the game-resolved trajectories and the construction of this separate defensive/cooperative dictionary.

4.4 Asymmetric Memory in Social Dilemmas

We conducted a heterogeneous evaluation where agents with bounded recall ($HL = 2$) interact directly with agents possessing extended history ($HL = 80$). The design holds the backbone model fixed and varies memory access within each group; we refer to agents by their HL assignments.

Our evaluation covers the 2-player Trust Game and the 3-player Public Goods Game (Appendix Figure 3; complete results in Appendix F).

Trust Game. The asymmetric pair has lower aggregate cooperation than the short-memory pair in the reported comparisons. For Llama-4-Scout-17B, cooperation differs by 28.1 pp between Player 1 ($HL = 2$) and Player 2 ($HL = 80$). Because memory assignment is tied to player position in this comparison, we report this as a role-specific asymmetric result.

Public Goods. Mixed-memory groups can have substantially lower cooperation than the all-short-memory group. GPT-OSS-20B decreases from 47.4% in the short-memory group to 28.2% with one long-memory player and 15.9% with two. The response is model-specific: Llama-4-Scout-17B instead has higher cooperation in both mixed groups than in its all-short-memory group.

Individual behavior. With two long-memory players and one short-memory player, the short-memory GPT-OSS-20B agent cooperates at 37.7%, versus 5.0% for the long-memory agents. Other models show smaller gaps, including near-ceiling equality for Llama-3.3-70B. Appendix F reports these action differences for every memory assignment.

4.5 Memory Content vs. Context Length: Sanitization Results

We conduct a memory sanitization experiment in the Trust Game at $HL = 80$ (Figure 4). We replace $80 - X$ of the visible rounds with synthetic mutual-cooperation records, leaving only X recent real rounds intact. This keeps the visible ledger at 80 rounds while directly manipulating its action and payoff content.

Replacing 78 rounds ($X = 2$) increases cooperation in each of the five reported models. For example, Llama-3.3-70B increases from 6.9% to 97.4%, and GPT-OSS-20B from 3.4% to 36.3%. The full sweep in Appendix H shows model-specific response curves across intermediate values of X. Together, these results show that cooperation responds strongly to what the accessible ledger contains. The intervention changes the evidence agents observe about their opponents and provides a controlled test of history-content sensitivity.

4.6 Explicit Reasoning: Mean Attenuation and Stability

We conducted an ablation study where agents output only their final decision without intermediate CoT (Appendix G). Across all 18 settings in the CoT memory-cursed group, removing explicit CoT reduces the memory drop D in 14 settings, raises $HL = 80$ cooperation in 15, and satisfies both criteria in 12. Mean D decreases from 42.3 to 26.8 pp, while mean $HL = 80$ cooperation increases from 22.4% to 40.3%. Mean $HL = 80$ run-level SD is nearly unchanged (5.3 versus 5.5 pp), decreasing in nine settings and increasing in nine. Removing the explicit reasoning requirement therefore has two separable effects: it often reduces the endpoint decline and raises long-history cooperation, while its effect on run-to-run stability is balanced across settings. The action-only condition changes the requested response format, and Appendix G reports every cell.

4.7 Alternative Memory Representations

We freeze 240 decision contexts from source $HL = 80$ trajectories and compare six representations of the same underlying histories using Mistral-7B: Raw80 (the complete, uncompressed 80-round history), a deterministic Summary plus the latest two raw rounds, selective Retrieval, Recent2, and two token-matched recent raw controls. Five paired generations per context and representation yield 7,200 requests. In the four-game aggregate, Recent2 improves cooperation over Raw80 by 9.4 pp (95% paired CI [5.7,13.2]), Retrieval by 3.9 pp [0.4,7.3], and Summary by 3.3 pp [-0.2,6.8]. Summary and Retrieval closely match their token-matched compact raw-history controls in the aggregate. The response varies by game: in PD, Recent2 lowers cooperation by 8.3 pp [-12.7,-4.0].

We also branch from nine matched PG prefixes drawn from three source trajectories and continue for 100 rounds. Summary and its token-matched raw control improve cooperation relative to Raw80 by 13.4 pp [10.3,16.3] and 11.9 pp [9.9,13.8], respectively. The branches share the same starting history and then develop through their own interactions. Together, the fixed-context and continuation results show that retaining a compact view of recent evidence can improve both next actions and subsequent interaction (Appendix L).

Table 4: Memory-view contrasts in 240 fixed contexts (pp; paired 95% CIs). Full design and game-level results appear in Appendix L.

View comparison	Change	95% CI
Recent2 − Raw80	+9.4	[5.7, 13.2]
Retrieval − Raw80	+3.9	[0.4, 7.3]
Summary − Raw80	+3.3	[-0.2, 6.8]
Summary − matched raw	-0.6	[-4.2, 3.0]
Retrieval − matched raw	-1.3	[-5.0, 2.4]

4.8 Prompting, Opponent Behavior, and Payoff Controls

A separate paired study of 240 fixed contexts compares Original, Numerical, and Narrative rule presentations under Baseline, length-matched Analysis, and Future-oriented instructions (12,960 requests). Numerical and Narrative vary both presentation and rule wording; within each presentation, the instruction contrast remains paired. Future-oriented prompting has near-zero aggregate effects, while the game-specific memory directions remain stable: PG and TG decline across all nine presentation–instruction combinations, whereas PD increases (Appendix M).

The scripted-opponent control connects cooperation to realized payoff: against Always-Defect under Narrative wording, Future prompting raises cooperation from 35.0% to 45.6% while mean payoff falls from 30.0 to 8.8, showing how opponent policy determines the value of cooperation. In the Mistral PG sweep, the memory drop remains 21.2–24.6 pp across nine multipliers $r \in [1.1, 2.7]$. Across the seven-model evaluation, PD has a larger normalized welfare gap than PG (0.500 versus 0.333) but a smaller mean drop (11.9 versus 27.6 pp), motivating a multidimensional account of game structure (Appendices N and O).

5 Conclusion

Across four repeated social dilemmas, longer accessible histories often coincide with lower cooperation, while a distinct high-cooperation regime persists in some settings. Action persistence, lexical statistics, contextual judging, sanitization, reasoning ablation, and fine-tuning jointly connect this effect to persistent responses after non-cooperation and to how history is represented and processed.

The action-balanced, rationale-only fine-tuned model shows a smaller PG memory drop than its matched base cohort and positive average zero-shot changes in PD, TD, and TG. Sanitization demonstrates strong sensitivity to visible history content. In frozen contexts, compact views raise average cooperation over Raw80, and token-matched compact raw-history controls show similar aggregate gains to the summary and retrieval views. Future-oriented prompting leaves average behavior largely unchanged, while the scripted-opponent control identifies settings in which greater cooperation increases vulnerability to exploitation.

Scope and future directions. This study establishes the memory–cooperation pattern across three-run conditions, homogeneous-backbone groups, and discrete social dilemmas, with focused Mistral-7B controls. The fixed-context and continuation studies answer complementary questions, and an independent LLM judge validates the lexical proxy at the rationale level. A broader next stage can add primary-sweep repetitions, full-history evaluation, human semantic validation, heterogeneous-agent populations, and general benchmarks for the balanced adapter.

Reproducibility Statement

Section 3 and Appendix A describe the experimental setup, game rules, and prompts; Appendix J lists serving model identifiers and the token-length audit. Appendix I documents training-data curation, rationale-only supervision, optimization, and the matched base/adapter evaluation. Appendices D and E describe the lexical analyses and contextual validation. Appendices F–H and K–O provide settings and results for the additional analyses and controls. The code is available anonymously at `https://anonymous.4open.science/r/MemoryCurse-C595/`.

AI Use Statement

Generative AI was used in the experiments to generate agent decisions and reasoning traces, including the rationales used for fine-tuning, and to provide contextual ratings of rationale content. The corresponding models and procedures are described in the methodology and appendices.

AI-assisted tools also supported interpretation of existing results, feedback on how the evidence supports the paper's claims, literature and policy lookup, code editing for data tabulation and figure preparation, and manuscript drafting, organization, and language revision. The authors take responsibility for the final content of this work.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025.

Philip Brookins and Jason Matthew DeBacker. Playing games with GPT: what can we learn about a large language model from canonical strategic games? *Available at SSRN 4493398*, 2023.

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

Drew Fudenberg and Eric Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STars. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=QGJ9ttXLTy`.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *URL https://arxiv. org/abs/2507.05257*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Tianqin Li, Zijie Li, Andrew Luo, Harold Rockwell, Amir Barati Farimani, and Tai Sing Lee. Prototype memory and attention mechanisms for few shot image generation. In *International Conference on Learning Representations*, 2022.

Yuxuan Li and Hirokazu Shirado. Spontaneous giving and calculated greed in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5271–5286, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.267. URL `https://aclanthology.org/2025.emnlp-main.267/`.

Shuangmei Ma, Boyu Zhang, Shinan Cao, Jun S Liu, and Wen-Xu Wang. Limited memory optimizes cooperation in social dilemma experiments. *Royal Society open science*, 8(8), 2021.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards LLMs as operating systems. 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *Advances in Neural Information Processing Systems*, 37:111715–111759, 2024.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *arXiv preprint arXiv:2506.23276*, 2025.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025.

Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. CoopEval: Benchmarking cooperation-sustaining mechanisms and LLM agents in social dilemmas, 2026. Presented at the ICLR 2026 Workshop on AI for Mechanism Design and Strategic Decision Making.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

APPENDIX

Appendix Table 1: Average cooperation rate (%) across four repeated social dilemmas, models with CoT reasoning, and history lengths. For $HL \geq 1$, $\pm$ denotes between-run standard deviation across three runs.

Social dilemma	Model	History length (HL)								
		0†	1	2	3	5	10	20	40	80
Prisoner's Dilemma	Gemma-3-12B	2.7 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-20B	47.4 ± 1.3	99.1 ± 0.5	92.1 ± 1.8	84.9 ± 7.1	36.2 ± 11.4	23.4 ± 2.2	16.7 ± 5.2	12.3 ± 2.3	20.6 ± 15.1
	Llama-4-Scout-17B	54.4 ± 0.7	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Mistral-7B	81.7 ± 0.7	100.0 ± 0.0	96.9 ± 0.9	96.9 ± 0.8	93.6 ± 1.4	95.5 ± 2.2	94.3 ± 3.5	91.3 ± 2.5	85.3 ± 2.9
	Llama-3.3-70B	36.0 ± 1.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-120B	42.3 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Qwen2.5-Coder-32B	37.2 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Public Goods Game	Gemma-3-12B	8.2 ± 0.5	0.0 ± 0.0	74.0 ± 3.3	71.0 ± 1.9	75.4 ± 4.6	75.0 ± 7.1	63.1 ± 6.8	59.8 ± 9.5	53.0 ± 5.9
	GPT-OSS-20B	4.2 ± 0.4	46.6 ± 0.7	43.5 ± 0.8	41.8 ± 1.2	28.8 ± 1.7	20.0 ± 0.4	13.8 ± 2.1	7.2 ± 0.7	5.2 ± 0.4
	Llama-4-Scout-17B	3.8 ± 0.2	90.0 ± 0.7	82.6 ± 0.2	75.8 ± 2.3	71.3 ± 2.5	63.1 ± 3.3	64.5 ± 2.4	57.1 ± 6.0	45.8 ± 24.0
	Mistral-7B	65.6 ± 1.0	100.0 ± 0.0	74.9 ± 2.2	71.0 ± 0.6	64.6 ± 1.0	60.1 ± 1.2	55.3 ± 1.6	55.1 ± 2.6	50.3 ± 3.1
	Llama-3.3-70B	2.9 ± 0.5	100.0 ± 0.0	99.8 ± 0.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-120B	0.0 ± 0.0	72.5 ± 0.2	75.4 ± 2.7	77.0 ± 2.5	50.4 ± 11.5	23.1 ± 4.4	13.1 ± 2.3	7.3 ± 3.2	4.0 ± 0.5
	Qwen2.5-Coder-32B	11.2 ± 0.4	100.0 ± 0.0	99.2 ± 0.1	97.9 ± 0.3	98.3 ± 0.6	97.5 ± 1.0	98.0 ± 0.6	97.3 ± 1.2	97.9 ± 1.5
Traveler's Dilemma	Gemma-3-12B	19.0 ± 0.5	100.0 ± 0.0	70.3 ± 6.8	67.1 ± 22.2	62.4 ± 26.0	51.2 ± 18.7	68.6 ± 41.4	63.4 ± 42.4	35.3 ± 15.4
	GPT-OSS-20B	20.5 ± 0.6	42.2 ± 2.0	42.0 ± 2.8	33.3 ± 4.0	8.6 ± 1.6	4.8 ± 0.9	4.5 ± 0.4	2.8 ± 0.3	2.7 ± 1.2
	Llama-4-Scout-17B	35.2 ± 2.2	0.4 ± 0.2	99.2 ± 1.1	95.9 ± 3.3	98.5 ± 1.6	97.4 ± 3.3	100.0 ± 0.0	95.6 ± 6.2	100.0 ± 0.0
	Mistral-7B	17.9 ± 1.3	12.0 ± 0.5	25.2 ± 0.7	22.3 ± 1.3	24.5 ± 1.4	21.5 ± 1.1	18.8 ± 1.6	16.9 ± 0.8	16.6 ± 0.3
	Llama-3.3-70B	96.0 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-120B	2.1 ± 0.6	88.8 ± 9.2	92.8 ± 7.7	90.3 ± 13.2	78.4 ± 14.1	100.0 ± 0.0	66.6 ± 42.9	34.2 ± 46.6	12.6 ± 15.4
	Qwen2.5-Coder-32B	25.9 ± 2.4	98.3 ± 0.1	88.9 ± 0.6	64.0 ± 17.4	50.2 ± 36.9	34.7 ± 26.5	61.9 ± 15.7	9.6 ± 1.0	8.0 ± 0.5
Trust Game	Gemma-3-12B	1.4 ± 0.3	0.1 ± 0.1	51.2 ± 3.1	46.2 ± 0.9	44.9 ± 4.6	31.8 ± 2.0	20.4 ± 2.6	11.8 ± 1.7	9.5 ± 0.9
	GPT-OSS-20B	2.9 ± 0.1	15.9 ± 1.5	28.7 ± 0.3	21.9 ± 3.1	14.8 ± 0.5	8.6 ± 0.7	6.3 ± 0.4	3.7 ± 0.7	3.4 ± 0.6
	Llama-4-Scout-17B	1.6 ± 0.3	38.7 ± 2.6	45.0 ± 1.0	34.7 ± 1.0	24.9 ± 5.4	21.3 ± 3.4	10.2 ± 3.7	3.6 ± 0.2	2.2 ± 0.5
	Mistral-7B	14.5 ± 0.7	21.8 ± 1.4	37.8 ± 0.9	39.2 ± 0.4	40.2 ± 0.8	39.3 ± 0.3	37.6 ± 0.4	38.3 ± 1.0	34.8 ± 0.2
	Llama-3.3-70B	0.2 ± 0.2	100.0 ± 0.0	50.8 ± 2.4	50.8 ± 14.6	47.0 ± 5.0	32.8 ± 9.8	45.7 ± 5.9	8.3 ± 3.9	6.9 ± 4.4
	GPT-OSS-120B	3.4 ± 0.2	82.8 ± 4.0	92.7 ± 2.1	99.3 ± 0.3	98.7 ± 1.3	57.6 ± 15.4	14.8 ± 4.7	14.4 ± 7.8	7.3 ± 4.7
	Qwen2.5-Coder-32B	5.3 ± 0.3	100.0 ± 0.0	98.0 ± 1.8	94.1 ± 1.5	92.0 ± 3.9	90.2 ± 7.6	99.2 ± 0.3	99.1 ± 0.2	98.3 ± 0.7

† $HL = 0$ removes past-action observations while retaining the repeated-game instructions.

Appendix Table 2: Discounted Reward ($Reward = \frac{\sum_{t=1}^{T} \delta^{t-1} \cdot R_t}{\sum_{t=1}^{T} \delta^{t-1}}$ with $\delta = 0.99$) averaged across three runs for No Reasoning (NR) and Reasoning (R). Results for Prisoner's Dilemma and Public Goods Game are shown below. Values are reported as mean ± standard deviation.

Social dilemma	Model	Setting	History length (HL)								
			0	1	2	3	5	10	20	40	80
Prisoner's Dilemma	Gemma-3-12B	NR	199.52 ± 0.68	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	100.37 ± 0.43	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	GPT-OSS-20B	NR	178.22 ± 1.14	198.35 ± 1.41	196.03 ± 3.20	192.86 ± 2.88	177.45 ± 16.65	147.67 ± 19.67	164.83 ± 10.35	136.14 ± 28.69	132.04 ± 16.97
		R	121.82 ± 0.95	199.64 ± 0.29	192.86 ± 1.91	186.88 ± 8.48	142.60 ± 25.90	120.38 ± 12.55	128.72 ± 11.94	115.40 ± 9.98	134.46 ± 28.35
	GPT-OSS-120B	NR	187.61 ± 1.98	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	118.77 ± 0.86	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Llama-3.3-70B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	113.85 ± 2.10	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.99 ± 0.01	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Llama-4-Scout-17B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.75 ± 0.35	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	129.59 ± 1.29	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Mistral-7B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	192.84 ± 3.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	166.55 ± 4.23	200.00 ± 0.00	193.21 ± 4.25	193.70 ± 1.94	189.61 ± 1.66	185.93 ± 7.29	186.50 ± 6.57	177.05 ± 2.46	176.36 ± 12.01
	Qwen2.5-Coder-32B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	111.33 ± 1.22	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.96 ± 0.06
Public Goods Game	Gemma-3-12B	NR	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.38 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		R	1.04 ± 0.00	1.00 ± 0.00	1.35 ± 0.02	1.38 ± 0.02	1.35 ± 0.02	1.36 ± 0.01	1.30 ± 0.05	1.30 ± 0.05	1.29 ± 0.05
	GPT-OSS-20B	NR	1.08 ± 0.00	1.22 ± 0.01	1.25 ± 0.03	1.26 ± 0.01	1.27 ± 0.03	1.25 ± 0.00	1.25 ± 0.06	1.13 ± 0.05	1.16 ± 0.07
		R	1.02 ± 0.00	1.24 ± 0.00	1.22 ± 0.01	1.21 ± 0.00	1.16 ± 0.01	1.10 ± 0.00	1.08 ± 0.01	1.05 ± 0.01	1.05 ± 0.01
	GPT-OSS-120B	NR	1.00 ± 0.00	1.42 ± 0.02	1.40 ± 0.02	1.34 ± 0.04	1.34 ± 0.02	1.26 ± 0.02	1.17 ± 0.07	1.09 ± 0.02	1.11 ± 0.01
		R	1.00 ± 0.00	1.36 ± 0.02	1.40 ± 0.02	1.39 ± 0.01	1.27 ± 0.07	1.14 ± 0.02	1.10 ± 0.03	1.08 ± 0.06	1.05 ± 0.01
	Llama-3.3-70B	NR	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.01 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
	Llama-4-Scout-17B	NR	1.50 ± 0.00	1.48 ± 0.01	1.50 ± 0.01	1.40 ± 0.01	1.45 ± 0.00	1.46 ± 0.00	1.41 ± 0.00	1.36 ± 0.02	1.41 ± 0.02
		R	1.02 ± 0.00	1.46 ± 0.00	1.42 ± 0.01	1.39 ± 0.01	1.35 ± 0.02	1.33 ± 0.02	1.31 ± 0.04	1.27 ± 0.08	1.32 ± 0.16
	Mistral-7B	NR	1.47 ± 0.01	1.50 ± 0.00	1.50 ± 0.00	1.37 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.33 ± 0.01	1.50 ± 0.00	1.37 ± 0.02	1.34 ± 0.01	1.32 ± 0.01	1.31 ± 0.01	1.29 ± 0.01	1.28 ± 0.02	1.27 ± 0.02
	Qwen2.5-Coder-32B	NR	1.17 ± 0.01	1.50 ± 0.00	1.50 ± 0.01	1.49 ± 0.00	1.50 ± 0.00	1.49 ± 0.01	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.06 ± 0.01	1.50 ± 0.00	1.50 ± 0.00	1.49 ± 0.00	1.49 ± 0.00	1.49 ± 0.00	1.49 ± 0.01	1.49 ± 0.00	1.49 ± 0.01

Appendix Table 3: Discounted Reward ($Reward = \frac{\sum_{t=1}^{T} \delta^{t-1} \cdot R_t}{\sum_{t=1}^{T} \delta^{t-1}}$ with $\delta = 0.99$) averaged across three runs for No Reasoning (NR) and Reasoning (R) settings, continued for Traveler's Dilemma and Trust Game. Values are reported as mean ± standard deviation.

Social dilemma	Model	Setting	History length (HL)								
			0	1	2	3	5	10	20	40	80
Traveler's Dilemma	Gemma-3-12B	NR	5.00 ± 0.00	4.99 ± 0.02	5.00 ± 0.00	4.35 ± 0.13	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
		R	3.76 ± 0.02	5.00 ± 0.00	4.32 ± 0.25	4.46 ± 0.21	3.53 ± 0.48	3.02 ± 0.44	4.18 ± 0.97	3.98 ± 0.88	2.80 ± 0.11
	GPT-OSS-20B	NR	3.32 ± 0.50	4.55 ± 0.24	3.94 ± 0.36	3.79 ± 0.29	2.94 ± 0.92	3.17 ± 0.84	3.18 ± 0.76	3.29 ± 0.41	3.64 ± 0.37
		R	2.39 ± 0.05	3.26 ± 0.31	3.55 ± 0.13	3.26 ± 0.28	2.25 ± 0.10	2.18 ± 0.09	2.32 ± 0.10	2.31 ± 0.25	2.50 ± 0.52
	GPT-OSS-120B	NR	2.01 ± 0.00	4.51 ± 0.08	4.55 ± 0.27	4.54 ± 0.27	4.16 ± 1.19	5.00 ± 0.00	4.37 ± 0.71	5.00 ± 0.00	2.98 ± 1.14
		R	2.05 ± 0.01	4.71 ± 0.28	4.28 ± 0.72	4.23 ± 1.04	3.83 ± 1.06	5.00 ± 0.00	3.94 ± 1.00	3.08 ± 1.36	3.11 ± 1.02
	Llama-3.3-70B	NR	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
		R	4.82 ± 0.02	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.99 ± 0.01	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
	Llama-4-Scout-17B	NR	4.99 ± 0.01	4.00 ± 0.00	4.00 ± 0.02	4.98 ± 0.01	4.00 ± 0.00	4.04 ± 0.05	4.00 ± 0.00	4.00 ± 0.00	4.33 ± 0.47
		R	4.06 ± 0.03	4.01 ± 0.01	4.94 ± 0.08	4.92 ± 0.04	4.82 ± 0.13	4.92 ± 0.08	5.00 ± 0.00	4.67 ± 0.47	4.75 ± 0.35
	Mistral-7B	NR	2.99 ± 0.01	2.90 ± 0.05	3.11 ± 0.07	3.07 ± 0.05	3.15 ± 0.03	3.14 ± 0.02	3.23 ± 0.05	3.08 ± 0.06	3.14 ± 0.04
		R	3.22 ± 0.04	3.19 ± 0.06	3.06 ± 0.11	3.06 ± 0.02	3.17 ± 0.03	3.11 ± 0.09	3.06 ± 0.05	3.01 ± 0.01	3.09 ± 0.10
	Qwen2.5-Coder-32B	NR	4.23 ± 0.00	4.22 ± 0.06	4.81 ± 0.08	4.68 ± 0.07	4.86 ± 0.12	4.57 ± 0.25	4.72 ± 0.06	4.56 ± 0.06	4.47 ± 0.09
		R	3.97 ± 0.05	4.89 ± 0.01	4.76 ± 0.14	4.21 ± 0.16	4.39 ± 0.44	4.18 ± 0.20	4.10 ± 0.18	3.97 ± 0.02	3.96 ± 0.02
Trust Game	Gemma-3-12B	NR	4.00 ± 0.00	4.00 ± 0.00	6.10 ± 0.88	8.39 ± 0.16	4.20 ± 0.09	4.11 ± 0.08	4.11 ± 0.03	4.07 ± 0.00	4.07 ± 0.00
		R	4.11 ± 0.02	4.05 ± 0.07	8.79 ± 0.23	8.19 ± 0.19	8.14 ± 0.22	7.40 ± 0.20	7.08 ± 0.22	6.17 ± 0.71	6.32 ± 0.35
	GPT-OSS-20B	NR	4.87 ± 0.27	7.56 ± 0.31	7.90 ± 0.10	8.15 ± 0.20	7.77 ± 0.14	7.54 ± 0.25	7.52 ± 0.48	7.41 ± 0.45	7.38 ± 0.30
		R	4.37 ± 0.05	6.11 ± 0.15	7.15 ± 0.15	6.77 ± 0.36	5.80 ± 0.03	5.44 ± 0.07	5.05 ± 0.18	4.82 ± 0.18	4.76 ± 0.26
	GPT-OSS-120B	NR	4.15 ± 0.02	9.62 ± 0.16	9.91 ± 0.06	9.59 ± 0.04	8.49 ± 1.47	6.16 ± 0.36	5.53 ± 0.45	5.08 ± 0.71	4.38 ± 0.19
		R	4.51 ± 0.05	8.76 ± 0.75	9.88 ± 0.10	9.98 ± 0.02	9.88 ± 0.12	7.79 ± 1.13	5.54 ± 0.57	6.79 ± 1.09	6.00 ± 1.12
	Llama-3.3-70B	NR	4.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	8.91 ± 0.12	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00
		R	4.03 ± 0.03	10.00 ± 0.00	7.93 ± 0.25	8.30 ± 0.62	7.16 ± 0.51	6.57 ± 1.17	7.14 ± 0.22	5.74 ± 0.78	5.70 ± 0.97
	Llama-4-Scout-17B	NR	5.32 ± 0.23	10.03 ± 0.10	8.39 ± 0.21	8.91 ± 0.03	7.06 ± 0.10	6.48 ± 0.09	5.98 ± 0.31	5.07 ± 0.17	5.37 ± 0.43
		R	4.20 ± 0.03	8.14 ± 0.09	8.44 ± 0.13	7.84 ± 0.06	6.52 ± 0.31	6.20 ± 0.23	5.27 ± 0.25	4.75 ± 0.05	4.60 ± 0.12
	Mistral-7B	NR	4.11 ± 0.03	8.86 ± 0.10	7.46 ± 1.80	8.59 ± 0.05	4.51 ± 0.16	4.61 ± 0.10	6.54 ± 2.46	8.82 ± 1.67	7.81 ± 1.53
		R	5.88 ± 0.34	6.18 ± 0.48	8.10 ± 0.10	8.10 ± 0.16	8.18 ± 0.07	8.17 ± 0.12	8.17 ± 0.07	8.28 ± 0.18	8.04 ± 0.04
	Qwen2.5-Coder-32B	NR	4.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	9.98 ± 0.01	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00
		R	4.81 ± 0.06	10.00 ± 0.00	10.01 ± 0.01	9.95 ± 0.14	9.83 ± 0.19	9.96 ± 0.07	10.02 ± 0.01	10.02 ± 0.01	10.04 ± 0.03

Appendix Table 4: Mean undiscounted per-round reward averaged across three runs for No Reasoning(NR) and Reasoning (R). Results for Prisoner's Dilemma and Public Goods Game are shown below. Values are reported as mean ± standard deviation.

Social dilemma	Model	Setting	History length (HL)								
			0	1	2	3	5	10	20	40	80
Prisoner's Dilemma	Gemma-3-12B	NR	199.80 ± 0.28	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	100.13 ± 0.09	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	GPT-OSS-20B	NR	175.20 ± 1.30	198.40 ± 1.61	195.40 ± 4.29	192.47 ± 3.49	181.00 ± 13.85	145.87 ± 10.55	163.80 ± 12.72	123.33 ± 27.91	116.67 ± 7.41
		R	121.40 ± 1.56	198.73 ± 0.57	189.07 ± 2.47	181.40 ± 8.35	127.93 ± 12.57	114.13 ± 1.48	108.93 ± 4.50	105.07 ± 2.17	113.80 ± 13.89
	GPT-OSS-120B	NR	187.80 ± 0.28	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	117.47 ± 0.81	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Llama-3.3-70B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	113.73 ± 1.37	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.80 ± 0.28	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Llama-4-Scout-17B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.87 ± 0.19	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	128.33 ± 0.25	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
	Mistral-7B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	193.07 ± 0.50	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	166.75 ± 1.34	200.00 ± 0.00	194.00 ± 1.56	194.00 ± 1.61	187.73 ± 2.69	191.80 ± 3.83	189.53 ± 6.49	183.67 ± 4.88	173.40 ± 4.55
	Qwen2.5-Coder-32B	NR	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00
		R	113.17 ± 0.91	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	200.00 ± 0.00	199.90 ± 0.14
Public Goods Game	Gemma-3-12B	NR	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.39 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		R	1.04 ± 0.00	1.00 ± 0.00	1.37 ± 0.02	1.36 ± 0.01	1.38 ± 0.02	1.38 ± 0.04	1.32 ± 0.03	1.30 ± 0.05	1.26 ± 0.03
	GPT-OSS-20B	NR	1.08 ± 0.00	1.22 ± 0.00	1.25 ± 0.03	1.25 ± 0.00	1.26 ± 0.04	1.26 ± 0.00	1.24 ± 0.06	1.09 ± 0.01	1.07 ± 0.02
		R	1.02 ± 0.00	1.23 ± 0.00	1.22 ± 0.00	1.21 ± 0.00	1.14 ± 0.01	1.10 ± 0.00	1.07 ± 0.01	1.04 ± 0.00	1.03 ± 0.00
	GPT-OSS-120B	NR	1.00 ± 0.00	1.43 ± 0.01	1.40 ± 0.01	1.34 ± 0.01	1.33 ± 0.01	1.26 ± 0.01	1.11 ± 0.04	1.04 ± 0.01	1.04 ± 0.01
		R	1.00 ± 0.00	1.36 ± 0.00	1.38 ± 0.01	1.38 ± 0.01	1.25 ± 0.06	1.12 ± 0.02	1.07 ± 0.01	1.04 ± 0.02	1.02 ± 0.00
	Llama-3.3-70B	NR	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.01 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
	Llama-4-Scout-17B	NR	1.50 ± 0.00	1.48 ± 0.00	1.50 ± 0.00	1.41 ± 0.00	1.44 ± 0.00	1.44 ± 0.02	1.37 ± 0.02	1.34 ± 0.03	1.35 ± 0.03
		R	1.02 ± 0.00	1.44 ± 0.00	1.41 ± 0.00	1.38 ± 0.01	1.36 ± 0.01	1.32 ± 0.02	1.32 ± 0.01	1.29 ± 0.03	1.23 ± 0.12
	Mistral-7B	NR	1.47 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.37 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.33 ± 0.00	1.50 ± 0.00	1.37 ± 0.01	1.35 ± 0.00	1.32 ± 0.00	1.30 ± 0.01	1.28 ± 0.01	1.28 ± 0.01	1.25 ± 0.02
	Qwen2.5-Coder-32B	NR	1.17 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.49 ± 0.00	1.49 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.50 ± 0.00
		R	1.06 ± 0.00	1.50 ± 0.00	1.50 ± 0.00	1.49 ± 0.00	1.49 ± 0.00	1.49 ± 0.01	1.49 ± 0.00	1.49 ± 0.01	1.49 ± 0.01

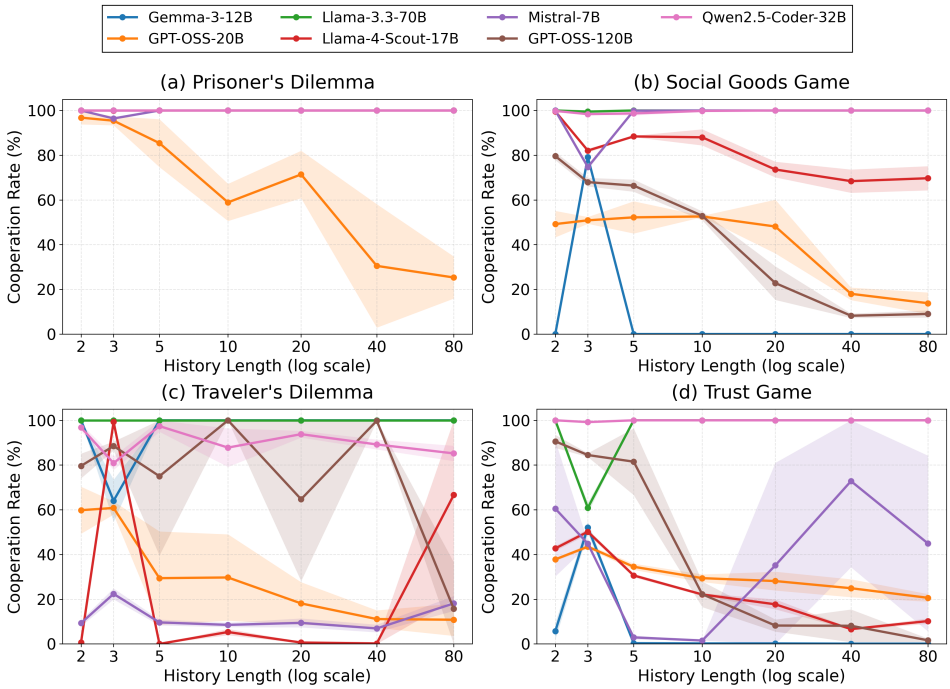
Appendix Table 5: Mean undiscounted per-round reward averaged across three runs for No Reasoning (NR) and Reasoning (R) settings, continued for Traveler's Dilemma and Trust Game. Values are reported as mean ± standard deviation.

Social dilemma	Model	Setting	History length (HL)								
			0	1	2	3	5	10	20	40	80
Traveler's Dilemma	Gemma-3-12B	NR	5.00 ± 0.00	4.99 ± 0.01	5.00 ± 0.00	4.38 ± 0.15	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
		R	3.77 ± 0.01	5.00 ± 0.00	4.47 ± 0.11	4.39 ± 0.41	4.18 ± 0.62	3.87 ± 0.46	4.21 ± 1.07	4.03 ± 1.24	3.34 ± 0.42
	GPT-OSS-20B	NR	3.35 ± 0.53	4.50 ± 0.13	3.95 ± 0.34	3.87 ± 0.12	2.93 ± 0.94	3.03 ± 0.88	2.96 ± 0.74	3.02 ± 0.59	3.31 ± 0.54
		R	2.35 ± 0.02	3.39 ± 0.11	3.53 ± 0.08	3.14 ± 0.17	2.16 ± 0.04	2.11 ± 0.01	2.12 ± 0.05	2.09 ± 0.07	2.17 ± 0.20
	GPT-OSS-120B	NR	2.01 ± 0.00	4.68 ± 0.05	4.33 ± 0.17	4.61 ± 0.08	4.23 ± 1.09	5.00 ± 0.00	3.93 ± 1.13	5.00 ± 0.01	2.49 ± 0.64
		R	2.04 ± 0.00	4.64 ± 0.29	4.77 ± 0.25	4.70 ± 0.42	4.32 ± 0.44	5.00 ± 0.00	3.99 ± 1.31	3.02 ± 1.40	2.43 ± 0.49
	Llama-3.3-70B	NR	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
		R	4.82 ± 0.01	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00
	Llama-4-Scout-17B	NR	4.99 ± 0.00	4.00 ± 0.00	4.00 ± 0.00	4.99 ± 0.00	4.00 ± 0.00	4.02 ± 0.01	4.00 ± 0.00	4.00 ± 0.00	4.33 ± 0.47
		R	4.07 ± 0.02	4.00 ± 0.00	4.97 ± 0.02	4.94 ± 0.05	4.96 ± 0.03	4.98 ± 0.02	5.00 ± 0.00	4.68 ± 0.45	4.91 ± 0.13
	Mistral-7B	NR	3.00 ± 0.00	2.93 ± 0.05	3.12 ± 0.02	3.12 ± 0.06	3.17 ± 0.05	3.16 ± 0.02	3.30 ± 0.03	3.12 ± 0.05	3.15 ± 0.09
		R	3.22 ± 0.02	3.21 ± 0.03	3.10 ± 0.04	3.07 ± 0.01	3.16 ± 0.02	3.09 ± 0.04	3.04 ± 0.03	3.05 ± 0.02	3.09 ± 0.01
	Qwen2.5-Coder-32B	NR	4.21 ± 0.01	4.23 ± 0.06	4.96 ± 0.02	4.70 ± 0.05	4.97 ± 0.03	4.86 ± 0.10	4.93 ± 0.02	4.89 ± 0.02	4.85 ± 0.04
		R	3.97 ± 0.03	4.98 ± 0.00	4.85 ± 0.01	4.58 ± 0.21	4.42 ± 0.43	4.22 ± 0.31	4.54 ± 0.19	3.95 ± 0.03	3.97 ± 0.01
Trust Game	Gemma-3-12B	NR	4.00 ± 0.00	4.01 ± 0.00	4.80 ± 0.52	8.60 ± 0.13	4.04 ± 0.02	4.02 ± 0.02	4.02 ± 0.01	4.01 ± 0.00	4.01 ± 0.00
		R	4.19 ± 0.04	4.01 ± 0.01	8.59 ± 0.15	8.17 ± 0.12	8.12 ± 0.14	7.34 ± 0.12	6.36 ± 0.27	5.49 ± 0.19	5.19 ± 0.13
	GPT-OSS-20B	NR	4.84 ± 0.11	7.55 ± 0.16	8.01 ± 0.04	8.17 ± 0.08	7.81 ± 0.16	7.41 ± 0.08	7.22 ± 0.35	6.97 ± 0.36	6.50 ± 0.15
		R	4.39 ± 0.03	6.01 ± 0.13	7.23 ± 0.06	6.51 ± 0.18	5.75 ± 0.02	5.10 ± 0.04	4.80 ± 0.05	4.49 ± 0.06	4.44 ± 0.06
	GPT-OSS-120B	NR	4.15 ± 0.01	9.71 ± 0.06	9.74 ± 0.10	9.63 ± 0.02	9.11 ± 0.75	5.96 ± 0.33	4.89 ± 0.19	4.81 ± 0.59	4.19 ± 0.07
		R	4.45 ± 0.01	9.35 ± 0.20	9.78 ± 0.06	9.99 ± 0.00	9.95 ± 0.06	8.10 ± 0.65	5.47 ± 0.27	5.44 ± 0.71	4.75 ± 0.36
	Llama-3.3-70B	NR	4.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	8.82 ± 0.04	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00
		R	4.02 ± 0.02	10.00 ± 0.00	8.20 ± 0.08	7.88 ± 0.71	7.62 ± 0.17	6.61 ± 0.51	7.20 ± 0.30	4.93 ± 0.32	4.66 ± 0.26
	Llama-4-Scout-17B	NR	5.37 ± 0.03	10.00 ± 0.09	8.38 ± 0.10	8.93 ± 0.11	7.15 ± 0.03	6.50 ± 0.01	5.95 ± 0.12	4.88 ± 0.05	5.30 ± 0.13
		R	4.23 ± 0.05	8.05 ± 0.11	8.37 ± 0.11	7.70 ± 0.01	6.66 ± 0.40	6.25 ± 0.17	5.25 ± 0.29	4.49 ± 0.02	4.29 ± 0.06
	Mistral-7B	NR	4.14 ± 0.03	8.96 ± 0.09	8.08 ± 1.48	8.53 ± 0.05	4.38 ± 0.12	4.19 ± 0.01	6.23 ± 2.67	8.67 ± 1.88	7.50 ± 1.89
		R	5.84 ± 0.16	6.44 ± 0.10	8.07 ± 0.04	8.15 ± 0.08	8.17 ± 0.10	8.17 ± 0.03	8.13 ± 0.03	8.23 ± 0.11	7.96 ± 0.12
	Qwen2.5-Coder-32B	NR	4.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.01	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00	10.00 ± 0.00
		R	4.72 ± 0.04	10.00 ± 0.00	10.00 ± 0.01	9.96 ± 0.05	9.85 ± 0.11	9.77 ± 0.28	10.02 ± 0.01	10.02 ± 0.00	10.03 ± 0.01

Appendix Table 6: Average cooperation rate (%, mean ± standard deviation) across four distinct social dilemmas for all models under the *no-reasoning* setting.

Social dilemma	Model	History length (HL)								
		0	1	2	3	5	10	20	40	80
Prisoner's Dilemma	Gemma-3-12B	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-20B	87.0 ± 0.7	99.1 ± 0.9	96.8 ± 3.0	95.4 ± 2.0	85.4 ± 10.7	58.9 ± 8.3	71.4 ± 10.6	30.5 ± 27.6	25.3 ± 9.4
	Llama-3.3-70B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Llama-4-Scout-17B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Mistral-7B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	96.4 ± 0.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-120B	93.8 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Qwen2.5-Coder-32B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Public Goods Game	Gemma-3-12B	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	79.2 ± 1.6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	GPT-OSS-20B	16.0 ± 0.3	43.1 ± 1.1	49.2 ± 5.9	50.9 ± 1.5	52.2 ± 7.2	52.6 ± 0.3	48.1 ± 12.0	18.0 ± 2.8	13.8 ± 4.8
	Llama-3.3-70B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.5 ± 0.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Llama-4-Scout-17B	100.0 ± 0.0	97.0 ± 0.8	99.6 ± 0.2	82.1 ± 0.9	88.4 ± 0.3	88.0 ± 3.6	73.6 ± 3.5	68.4 ± 5.2	69.7 ± 5.4
	Mistral-7B	94.9 ± 0.5	99.9 ± 0.1	100.0 ± 0.0	74.7 ± 1.2	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-120B	0.0 ± 0.0	85.7 ± 1.4	79.6 ± 1.8	68.0 ± 1.9	66.4 ± 2.7	52.9 ± 2.0	22.8 ± 7.5	8.2 ± 1.0	9.0 ± 1.6
	Qwen2.5-Coder-32B	33.8 ± 0.2	100.0 ± 0.0	99.8 ± 0.3	98.4 ± 0.3	98.7 ± 0.9	99.8 ± 0.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.1
Traveler's Dilemma	Gemma-3-12B	99.8 ± 0.1	99.6 ± 0.5	100.0 ± 0.0	64.0 ± 9.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	GPT-OSS-20B	52.7 ± 16.4	83.9 ± 3.3	59.8 ± 10.5	60.8 ± 2.9	29.4 ± 20.9	29.7 ± 19.3	18.1 ± 9.5	11.1 ± 3.9	10.8 ± 7.2
	Llama-3.3-70B	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Llama-4-Scout-17B	100.0 ± 0.0	0.2 ± 0.2	0.7 ± 0.2	99.3 ± 0.3	0.0 ± 0.0	5.3 ± 1.4	0.6 ± 0.3	0.2 ± 0.2	66.6 ± 46.9
	Mistral-7B	0.4 ± 0.0	16.3 ± 1.1	9.3 ± 1.3	22.4 ± 2.4	9.6 ± 1.3	8.5 ± 1.0	9.4 ± 2.0	6.9 ± 1.7	18.1 ± 3.3
	GPT-OSS-120B	1.5 ± 0.3	88.3 ± 1.3	79.6 ± 5.5	88.5 ± 2.0	75.0 ± 35.4	100.0 ± 0.0	64.7 ± 36.6	99.8 ± 0.2	15.7 ± 20.8
	Qwen2.5-Coder-32B	47.4 ± 1.3	53.9 ± 3.0	96.8 ± 1.5	80.8 ± 3.0	97.4 ± 2.5	87.8 ± 8.6	93.8 ± 1.5	89.2 ± 2.2	85.2 ± 3.4
Trust Game	Gemma-3-12B	0.0 ± 0.0	0.1 ± 0.0	5.7 ± 3.7	52.1 ± 2.6	0.3 ± 0.1	0.2 ± 0.1	0.2 ± 0.0	0.1 ± 0.0	0.1 ± 0.0
	GPT-OSS-20B	6.2 ± 0.7	33.4 ± 2.4	37.8 ± 0.7	43.5 ± 0.9	34.5 ± 1.4	29.4 ± 1.7	28.1 ± 4.2	24.9 ± 4.0	20.6 ± 1.7
	Llama-3.3-70B	0.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	60.9 ± 2.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Llama-4-Scout-17B	9.8 ± 0.2	54.0 ± 0.7	42.7 ± 1.1	50.2 ± 1.1	30.6 ± 0.8	22.1 ± 0.6	17.7 ± 2.2	6.6 ± 0.4	10.2 ± 1.2
	Mistral-7B	1.0 ± 0.2	48.2 ± 0.8	60.5 ± 30.3	44.8 ± 0.8	2.9 ± 0.9	1.5 ± 0.0	35.1 ± 45.9	72.8 ± 38.5	44.9 ± 39.3
	GPT-OSS-120B	1.1 ± 0.0	84.4 ± 1.0	90.6 ± 2.9	84.5 ± 0.8	81.5 ± 14.9	22.2 ± 5.7	8.2 ± 2.7	8.1 ± 7.3	1.6 ± 0.8
	Qwen2.5-Coder-32B	0.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.2 ± 0.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

Appendix Figure 1: Cooperation in the *without reasoning* setting (action-only output). Each panel reports means and between-run SDs for three runs per condition. Variability differs across configurations; Appendix G reports the complete aggregate reanalysis and defines the action-only condition.



A Experiment Game Rule and System Prompt

A.1 Game Rule

All four environments use repeated-game prompts that state a 99% continuation probability and request long-run payoff maximization. Accessible history contains the most recent HL rounds, for $HL \in \{0, 1, 2, 3, 5, 10, 20, 40, 80\}$. We define cooperation as the frequency of a prespecified mutual-benefit action and evaluate the realized payoff separately.

Prisoner's Dilemma. In each round, two players simultaneously choose between **A0** (cooperate) and **A1** (defect). The payoffs correspond to the classic parameters of the game: a Reward (R) of 200 points for mutual **A0**, a Punishment (P) of 100 points for mutual **A1**, a Temptation (T) of 300 points for the defector (**A1**) when choices differ, and a Sucker's payoff (S) of -100 points for the unreciprocated cooperator (**A0**). This structure explicitly satisfies the standard inequalities ($T > R > P > S$ and $2R > T + S$), ensuring that unilateral defection is individually optimal in the one-shot game, whereas mutual cooperation remains the social optimum.

Traveler's Dilemma. In each round, two players simultaneously choose one of four discrete claims: **A0**=2, **A1**=3, **A2**=4, and **A3**=5. If both players submit the same claim, each receives that amount. If the claims differ, both players receive the lower claim; in addition, the player who chose the lower claim receives a bonus of 2, while the player who chose the higher claim incurs a penalty of 2. Through iterated elimination of dominated strategies, the Nash Equilibrium is mutual **A0**, while the Pareto-optimal outcome is mutual **A3**. In this context, cooperation is defined strictly as selecting **A3**.

Public Goods Game. In each round, three players each receive an initial endowment of 1 point and simultaneously choose between **A0** (contribute) and **A1** (free-ride). Choosing **A0** means contributing the 1 point to a public pool, while **A1** means keeping it. The total points in the public pool are multiplied by a factor of 1.5 and then distributed equally among all three players. Let k denote

the number of players choosing **A0**. Each player choosing **A0** receives exactly their share of the pool, $1.5k/3$, whereas each player choosing **A1** retains their endowment and additionally receives the pool share, totaling $1 + 1.5k/3$. For a unilateral deviation, the other players' actions, not the total contributor count k, must be held fixed. If k_{-i} others contribute, switching from contribution to free-riding changes payoff from $1.5(k_{-i} + 1)/3$ to $1 + 1.5k_{-i}/3$, a gain of $1 - 1.5/3 = 0.5$. Universal free-riding yields 1.0 per player, whereas universal contribution yields 1.5 and maximizes total welfare.

Trust Game. In each round, two players simultaneously choose between **A0** (cooperate) and **A1** (defect). The payoff structure is asymmetric based on player roles. If both choose **A0**, both receive 10 points. If both choose **A1**, both receive 4 points. If the focal player chooses **A1** while the opponent chooses **A0**, the focal player receives a maximum temptation payoff of 20 points, and the opponent receives 0. Conversely, if the focal player chooses **A0** while the opponent chooses **A1**, the focal player receives a sucker's payoff of 2 points, and the opponent receives 6. This tension drives the Nash Equilibrium to a state of mutual distrust (mutual **A1**), despite mutual **A0** being the social optimum.

A.2 System Prompt

The prompts share a repeated-game structure, with game-specific rules, player references, history formatting, and final decision formats. The games rules follow the same idea from Tewolde et al. (2026).

The following template illustrates the shared structure for a two-player game with visible history:

You are Player {player_id}, playing a repeated game with Player {other_id}. This is round {round_num}. You can see the most recent {history_length} rounds of history.

{game_rules}

You are playing this game repeatedly with the same player. The action you choose in the current round will be visible to the other player in future rounds and may influence their future decisions. After each round, there is a 99% chance that another round will take place. Therefore, your goal is to maximize your total payoff in the long run rather than only the current-round payoff.

Past rounds' results:

{history_block}

First, briefly think about which action to choose based on the game rules, past rounds, and your long-term objective. Keep your reasoning concise and within 500 words. Then briefly state your reasoning. Finally, you MUST start a new line and output your action in exactly the required format for the current game.

Game-specific decision formats. For Prisoner's Dilemma, the required output is [Option A] or [Option B], corresponding to A_0 (cooperate) and A_1 (defect) in our tables. For Traveler's Dilemma, the required output format is [A0, A1, A2, or A3]. For Public Goods Game, the required output format is [A0 or A1]. For Trust Game, the required output format is [A0 or A1].

B Analysis of Cooperation Dynamics

As discussed in Section 4, the first major behavioral pattern is *consistent cooperation*. This pattern is observed in several settings involving Llama-3.3-70B and Qwen2.5-Coder-32B. The following rationales illustrate how agents express their justification for cooperative actions at short and long history lengths. We compare examples from the same backbone to characterize these qualitative shifts across naturally evolving histories.

B.1 Pattern 1: Signaling and Forgiveness

When these models encounter early defections or attempt to establish a cooperative baseline, they often treat their own actions as **signals** rather than purely myopic best responses. In particular, they use cooperative actions to preserve or repair trust.

Intra-model comparison: Qwen2.5-Coder-32B (*Public Goods*)

- **Short memory ($HL = 2$), Round 2, Player 2:** After observing a defection in the previous round, the model explicitly uses cooperation as a forgiveness signal:

  > "Considering the history, it seems that Player 3 might be more inclined towards A1... To encourage cooperation and potentially lead to a scenario where all players choose A0, it might be strategic to stick with A0 again in this round. This could signal to Player 3 that cooperation is possible and beneficial for everyone."

- **Long memory ($HL = 80$), Round 81, Player 2:** With a much larger history window, the same model no longer frames cooperation as local repair, but as preserving a historically supported equilibrium:

  > "Over the last 80 rounds, there have been several instances where all players chose A0... There have also been rounds where one or two players deviated... However, it appears that the majority of the time, the players tend to play A0, indicating a cooperative strategy among them. [Choosing A0] maintains this cooperative equilibrium..."

Interpretation. The underlying cooperative orientation remains stable across both memory settings, but its justification changes in scope. At $HL = 2$, Qwen2.5-Coder-32B reasons in terms of *localized trust repair* and immediate strategic consequences: cooperation is maintained as a forgiveness signal aimed at steering Player 3 toward A0 in the very next round. At $HL = 80$, the same action is justified at a more aggregate level: the model summarizes the long interaction history as a stable cooperative pattern and treats A0 as a way to preserve the existing equilibrium.

B.2 Pattern 2: Explicit Awareness of Retaliation and Opponent Strategy

Rationales from consistently cooperative settings connect retaliation, reciprocal incentives, and opponent behavior to the selected action.

Intra-model comparison: Llama-4-Scout-17B (*Traveler's Dilemma*)

- **Short memory ($HL = 2$), Round 42, Player 1:** The model reasons in a local contingent-response form, explicitly weighing the temptation to undercut against the risk of near-term punishment:

  > "Given that both players have shown a willingness to cooperate at A3, a deviation to a lower action could potentially yield a higher immediate payoff if the opponent sticks with A3. For example, choosing A2 (corresponding to 4) would yield 4 + 2.0 = 6.0 points if the opponent chooses A3, while the opponent would get 4 - 2.0 = 2.0 points.
  >
  > But, we must consider the long-term implications. If Player 2 notices a deviation and perceives it as an attempt to exploit, they might retaliate by also lowering their action in the next round, potentially starting a downward spiral where both players try to undercut each other. This could lead to a much lower payoff for both players in subsequent rounds."

- **Long memory ($HL = 80$), Round 121, Player 2:** The model moves beyond immediate retaliation and explicitly summarizes a repeated-game pattern over the visible long history:

  > "Given the history of the past 80 rounds, it's clear that both players have consistently chosen action A3, resulting in a payoff of 5.0 for both in each round.
  >
  > Choosing A3 consistently has yielded a high and stable payoff of 5.0 for both players. If either player were to choose a lower action (A0, A1, or A2) in an attempt to gain an advantage, they risk disrupting this stable outcome. The opponent might retaliate or change their strategy in future rounds, potentially leading to lower payoffs for both.
  >
  > Moreover, the concept of 'tit for tat' or maintaining cooperation unless the other player defects could be a viable strategy here. Since both have cooperated perfectly so far, continuing to cooperate seems like the safest and most beneficial strategy.

> Therefore, considering the goal to maximize total points in the long run and the history of consistent cooperation at the highest level, the best action seems to be maintaining the status quo."

Interpretation. Under short memory, the model reasons in a local contingent-response form: defecting now may trigger punishment in the next round and initiate a downward spiral. Under long memory, the same model performs higher-level pattern recognition over a much longer sequence, compressing the interaction history into a stable cooperative regime and explicitly invoking a repeated-game strategy such as Tit-for-Tat. Thus, cooperation remains stable, but the surrounding explanation becomes more strategic and abstract as the visible history expands.

B.3 Pattern 3: Resolving the Social Dilemma Through Long-Horizon Reasoning

A third recurrent rationale theme compares the immediate return from deviation with the future value of maintaining cooperation. This theme motivates the quantitative future-orientation analysis in Section 4.

Intra-model comparison: Llama-3.3-70B (*Public Goods*)

- **Short memory ($HL = 2$), Round 3, Player 1:** The model relies on logical projection from a short interaction history:

  > "The strategy should balance the immediate gain from choosing A1 with the potential long-term benefits of maintaining a cooperative dynamic where players choose A0... Given this, and considering the cooperation observed in the last two rounds, the best strategy seems to be to continue choosing A0."

- **Long memory ($HL = 80$), Round 81, Player 3:** The model reaches the same cooperative conclusion, but grounds it in a broader empirical pattern:

  > "However, the key observation is that playing A1 does not increase the total payoff for the group but rather shifts the individual payoff... The optimal outcome for the group (and thus for the individual in the long run, given the repeated nature of the game) is achieved when all players choose A0. Considering the history and the rules, the best strategy... is to continue choosing A0."

Interpretation. This comparison illustrates how cooperation can persist even when the memory content changes substantially. At $HL = 2$, the model relies primarily on theoretical projection from the continuation probability and the recent interaction state. At $HL = 80$, the same cooperative conclusion is reinforced by a much richer empirical history. In other words, the cooperative policy is unchanged, but the evidential basis of the reasoning becomes broader.

C Analysis of Memory-Bounded Behavior

The second major behavioral regime is *memory-bounded behavior*. As shown in the main text, these models often defect at memory length 0, reach their highest cooperation at some $HL \leq 5$ (most often at $HL = 2$), and then become less cooperative as the memory window continues to expand. This pattern suggests the existence of an effective band of usable history: too little history prevents reciprocity, whereas too much history can make earlier defections overly salient and thereby impede trust recovery.

To illustrate this dynamic qualitatively, we trace the same models across increasing memory lengths.

C.1 Intra-model trajectory 1: Gemma-3-12B (*Trust Game*)

Gemma-3-12B provides a particularly clear example of the full trajectory from myopic defection to local repair and finally to long-history lock-in.

Evidence across memory windows:

- **Zero memory ($HL = 0$), Round 1: myopic defection.** Without any history to ground reciprocity, the model defaults to immediate payoff-sensitive reasoning:

  > "Choosing A1 offers a high payoff (20 points) if Player 2 chooses A0, but a low payoff (4 points) if they choose A1. Choosing A0 guarantees a modest payoff (10 points)... I will choose A1 to maximize potential gain while mitigating risk."

- **Short memory ($HL = 2$), Round 3: local trust repair.** With only a narrow recent window, the model is still willing to disrupt a retaliatory loop and test cooperation:

  > "Since Player 1 chose A1 in the last round, they might expect me to continue with A1... Therefore, I will switch to A0. This introduces some uncertainty for Player 1... Even if Player 1 chooses A1 again... I've disrupted the predictable pattern and might improve my long-term average if Player 1 starts cooperating more."

- **Long memory ($HL = 80$), Round 256: history-induced lock-in.** With a long accumulated record of defections, the same model stops attempting repair and instead mirrors the established pattern:

  > "Player 1 continues to consistently choose A1. Maintain the mirroring strategy of selecting A1 in each round to maximize the consistent 4-point payoff. No change in strategy is warranted given the opponent's unwavering behavior."

Interpretation. Within Gemma-3-12B, the $HL = 2$ window supports strategic flexibility: the model remains willing to test a cooperative reset from the recent interaction state. At $HL = 80$, the accumulated record leads it to classify the opponent as persistently non-cooperative and maintain a rigid response.

C.2 Intra-model trajectory 2: GPT-OSS-20B (*Public Goods*)

GPT-OSS-20B shows a related but distinct form of memory-bounded degradation. As its memory grows, its reasoning shifts from forward-looking coordination to increasingly defensive threshold policies and eventually to shallow pattern-matching.

Evidence across memory windows:

- **Short memory ($HL = 2$), Rounds 1–3: forward-looking reciprocity.** With little available history, the model reasons in terms of future coordination:

  > "Because future rounds are highly likely, a strategy that encourages others to play A0 can increase the long-term average payoff for everyone... [Action: A0]."

- **Medium memory ($HL = 20$), Round 48: defensive entry threshold.** As the history window expands, the model imposes a highly conservative condition for cooperation:

  > "Stick to A1 as long as opponents stay on A1. Only switch to A0 if both opponents play A0 together, then reciprocate to raise joint payoff. Remain predictable to avoid being exploited."

- **Long memory ($HL = 80$), Round 500: compressed pattern-matching.** At the longest memory setting, the reasoning collapses into direct recitation of the recent pattern:

  > "R499: You=A1, P2=A1, P3=A1 $\rightarrow$ 1.0. [A1]"

Interpretation. These selected rationales range from discussion of future coordination to concise descriptions of observed play. Their different histories and time points prevent attribution of this contrast to history length alone.

Summary. Taken together, these examples illustrate how expressed reasoning changes with the visible history. The transition statistics in Appendix K and matched-context controls then quantify the corresponding behavioral pattern.

D Detailed Analysis of Memory Immunity and Cognitive Divergence

Section 4.1 establishes heterogeneous behavioral responses to accessible history. The same model–game groups are used for the lexical summaries and explicit-reasoning ablation.

Memory Immune (10)		**Memory Cursed (18)**	
Model	**Game**	**Model**	**Game**
Gemma-3-12B	Prisoner's Dilemma	Gemma-3-12B	Public Goods
GPT-OSS-120B	Prisoner's Dilemma	Gemma-3-12B	Traveler's Dilemma
Llama-3.3-70B	Prisoner's Dilemma	Gemma-3-12B	Trust Game
Llama-3.3-70B	Public Goods	GPT-OSS-20B	Prisoner's Dilemma
Llama-3.3-70B	Traveler's Dilemma	GPT-OSS-20B	Public Goods
Llama-4-Scout-17B	Prisoner's Dilemma	GPT-OSS-20B	Traveler's Dilemma
Llama-4-Scout-17B	Traveler's Dilemma	GPT-OSS-20B	Trust Game
Qwen2.5-Coder-32B	Prisoner's Dilemma	GPT-OSS-120B	Public Goods
Qwen2.5-Coder-32B	Public Goods	GPT-OSS-120B	Traveler's Dilemma
Qwen2.5-Coder-32B	Trust Game	GPT-OSS-120B	Trust Game
		Llama-3.3-70B	Trust Game
		Llama-4-Scout-17B	Public Goods
		Llama-4-Scout-17B	Trust Game
		Mistral-7B	Prisoner's Dilemma
		Mistral-7B	Public Goods
		Mistral-7B	Traveler's Dilemma
		Mistral-7B	Trust Game
		Qwen2.5-Coder-32B	Traveler's Dilemma

Appendix Table 7: Memory Immune / Memory Cursed settings identified by their cooperation patterns. Membership follows the observed trajectory regimes; Appendix Table 1 provides every history-length value.

D.1 Memory-Response Regimes Across Settings

Our evaluation spans 28 model–game settings. Appendix Table 7 records the two empirical trajectory regimes used in the downstream analyses.

- **Memory Immune (10 settings):** Configurations where cooperation remains high as history expands from short to long memory.
- **Memory Cursed (18 settings):** Configurations with lower cooperation at $HL = 80$ than at $HL = 2$, spanning a range of effect magnitudes and intermediate trajectories.

The regime is specific to each model–game pairing. GPT-OSS-120B, for example, remains at 100% cooperation across nonzero-history PD conditions but decreases from 92.7% at $HL = 2$ to 7.3% at $HL = 80$ in TG, showing how game structure interacts with the same backbone.

D.2 Lexical Analysis Methodology

We analyze emitted rationales from the primary study using lexical proxies for future-oriented and history-oriented language. Appendix J defines the group-round and agent-decision sampling units.

We use the following two fixed dictionaries. Their labels summarize keyword content, and a rationale can contain terms from both categories:

- **Forward-Looking Dictionary (Proactive):** Keywords associated with long-term strategy, signaling, and collective welfare.
  - *Included terms:* future, long-term, signal, reputation, encourage, forgive, mutual cooperat, mutual benefit, both get, both receive, win-win, maintain cooperat, continue cooperat, recipro, pareto.

- **History-Following Dictionary (Reactive):** Keywords focused on risk aversion, historical pattern-matching, defensive posture, and self-preservation.
  - *Included terms:* risk, exploit, betray, punish, retaliate, suspicious, cautious, vulnerable, worst-case, pattern, consistently, cannot trust, likely to defect, safer choice, minimize loss, protect myself.

The Forward-Looking Ratio is $F/(F + H)$ for forward-looking and history-following keyword matches. Larger values indicate a larger share of matched forward-looking terms.

D.3 Contextual Validation of the Lexical Proxy

A separate DeepSeek-V4-Pro judge receives complete rationales and rates future orientation from 1 to 10. The report describes the evaluation as blinded and rationale-only. Among 2,016 rationales, the top decile under the lexical forward-looking measure scores 6.12 on average, compared with 4.03 for the bottom decile ($\Delta = 2.09$, permutation-test p-value < 0.0001). The association is positive in all seven model slices, and repeated scoring has intraclass correlation coefficient (ICC) 0.980.

The 2,016 passages were sampled as eight from each of 7 models $\times$ 4 games $\times$ 9 history lengths. The judge saw cleaned complete rationales and anonymous IDs, without model, game, history, action, lexical score, or expected direction; the returned model ID was `deepseek-v4-pro-260425`, with temperature 0 and thinking disabled.

This independent model-judge evaluation supports the aggregate future-orientation proxy at the level of complete rationales. It complements the fixed dictionaries, and we report the curation and validation judges separately throughout.

D.4 Cognitive Contrast: Immune vs. Cursed Regimes

Applying this lexical measure to the two behavioral groups reveals a systematic contrast in emitted language.

At $HL = 80$, the two behavioral groups have the following mean forward-looking ratios:

- **Memory Immune Model-Game Settings (HL=80):** Retain a mean forward-looking ratio of 0.504.
- **Memory Cursed Model-Game Settings (HL=80):** Have a mean ratio of 0.339.

This gap identifies a systematic difference in expressed reasoning between the two behavioral regimes and motivates the rationale-targeted intervention in Section 4.2.

E Detailed Methodology for Reasoning Semantic Analysis

We analyze the words in emitted rationales before the final action. Word frequencies are normalized per 1,000 rationale words, with the sampling units and aggregation conventions reported in Appendix J.

E.1 Semantic Vocabularies

We use two further fixed vocabularies to describe defensive and cooperative language. These are distinct from the forward-looking/history-following dictionaries. The static vocabularies used for semantic counting are cataloged below:

Pure Paranoia / Defensive Words: risk, risky, betray, betrayal, betrayed, defect, defection, defected, punish, punishment, retaliate, retaliation, retaliatory, careful, cautious, caution, suspicious, suspect, exploit, exploited, exploitation, exploiting, threat, threaten, danger, dangerous, protection, defensive, defend, deviate, deviation, deviated, worst, worst-case, downside, fear, afraid, worry, concerned, distrust, mistrust, untrust, trap, trapped, greedy, greed, selfish, manipulate, manipulation, vulnerable, vulnerability.

Pure Cooperation / Positive Words: cooperate, cooperation, cooperative, cooperating, trust, trusting, trustworthy, mutual, mutually, together, win-win, both benefit, joint, shared, reciprocate, reciprocity, reciprocal, collaborate, collaboration, sustain, fair, fairness, equal, long-term, long term, tit-for-tat, tit for tat, forgive, forgiveness, rewarding, optimal, beneficial.

E.2 Aggregated Results

Appendix Table 8 reports normalized word frequencies and the Paranoia Ratio for the five displayed models. The ratio is computed within the aggregation pipeline and can therefore differ from a ratio reconstructed from the displayed, separately averaged counts.

Appendix Table 8: Aggregated CoT Semantic Analysis across All Games. Frequencies are normalized per 1,000 CoT words.

Model	Paranoia Words (per 1k)			Cooperation Words (per 1k)			Paranoia Ratio		
	HL=2	HL=80	80/2 Ratio	HL=2	HL=80	80/2 Ratio	HL=2	HL=80	Δ Shift
Llama-3.3-70B	11.21	10.71	0.96x	41.95	37.94	0.90x	0.217	0.248	+0.030
Llama-4-Scout-17B	11.49	12.30	1.07x	30.41	28.28	0.93x	0.288	0.357	+0.069
Gemma-3-12B	17.50	16.93	0.97x	23.84	19.12	0.80x	0.467	0.540	+0.073
Mistral-7B	9.91	9.43	0.95x	16.77	10.71	0.64x	0.524	0.611	+0.087
GPT-OSS-20B	14.42	12.88	0.89x	35.56	15.33	0.43x	0.293	0.487	+0.195

E.3 Supplementary Semantic Trajectories and Per-Game Breakdowns

Appendix Figure 2 displays game-resolved lexical ratios. The heterogeneous and sometimes non-monotonic curves show that vocabulary shifts depend on both model and game.

F Detailed Results for Asymmetric Memory Evaluation

Section 4.4 and Figure 3 summarize interactions with asymmetric memory access within a common backbone. Appendix Tables 9–11 report cooperation for each memory assignment.

F.1 Trust Game (2-Player Asymmetric Setting)

Appendix Table 9 reports the assignment Player 1: $HL = 2$, Player 2: $HL = 80$. The reported individual gaps range from 10.3 to 28.1 pp. Because TG payoffs depend on role, these gaps combine memory assignment and role; role-swapped controls would be needed for a clean within-pair attribution.

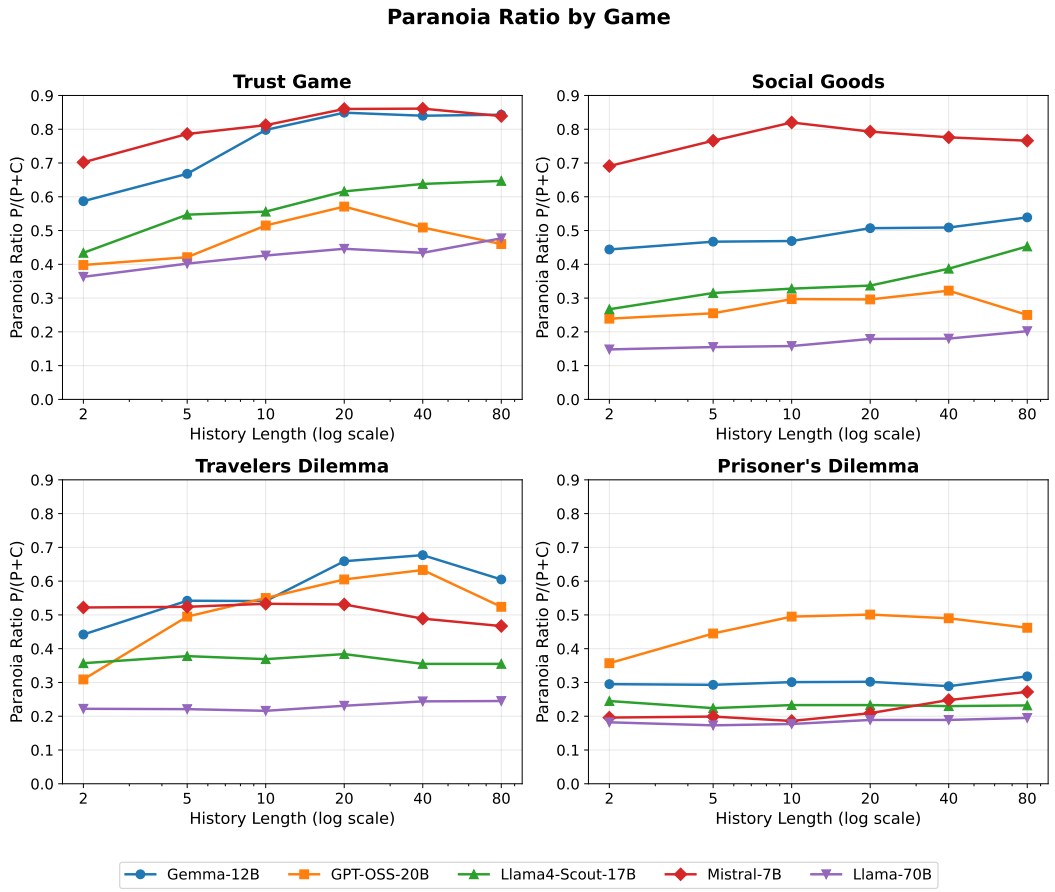
Appendix Table 9: Trust Game: Cooperation Rates in the Asymmetric Setting ($P1_{HL=2}$ vs. $P2_{HL=80}$).

Model	Sym HL=2	Sym HL=80	Asym Overall	P1 (HL=2)	P2 (HL=80)	Δ (P1 - P2)
Gemma-3-12B	51.2±3.1%	9.5±0.9%	27.8±2.2%	38.0±1.3%	17.7±3.2%	**+20.3pp**
GPT-OSS-20B	28.7±0.3%	3.4±0.6%	13.8±0.3%	22.3±1.6%	5.3±1.1%	**+17.0pp**
Llama-4-Scout-17B	45.2±0.8%	2.0±0.4%	18.2±0.6%	32.3±0.7%	4.2±0.7%	**+28.1pp**
Mistral-7B	37.8±0.9%	34.8±0.2%	35.6±0.4%	40.7±0.7%	30.5±0.8%	**+10.3pp**
Llama-3.3-70B	52.1±3.7%	6.1±3.6%	22.5±7.6%	32.2±6.0%	12.9±9.3%	**+19.3pp**

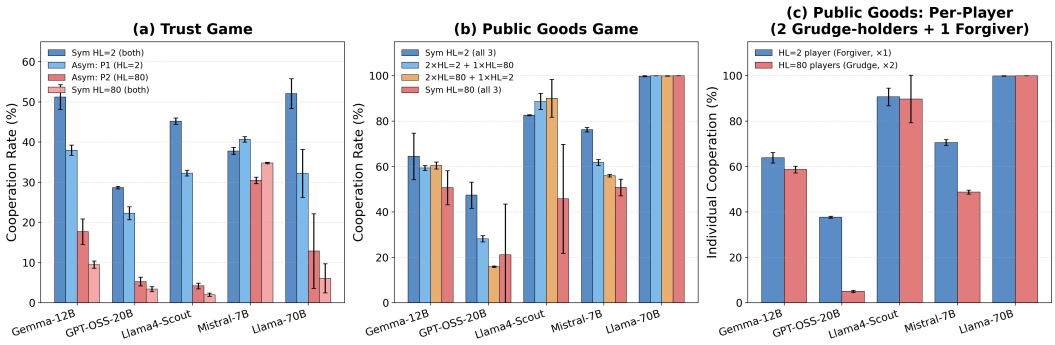
F.2 Public Goods (3-Player Asymmetric Setting)

Appendix Tables 10 and 11 report PG group and individual cooperation rates. Mixed-group behavior follows model-specific paths as the number of long-memory players changes.

In GPT-OSS-20B with two $HL = 80$ players and one $HL = 2$ player, the short-memory player cooperates at 37.7% and the long-memory players at 5.0%, a reported difference of 32.7 pp. Llama-3.3-70B is near ceiling across these configurations.



Appendix Figure 2: Game-resolved Paranoia Ratio trajectories. These fixed-dictionary statistics describe emitted vocabulary and reveal model- and game-specific paths across history lengths.



Appendix Figure 3: Asymmetric memory evaluation across TG and PG using the same backbone within each group. **(a)** TG cooperation rates. **(b)** PG group cooperation rates, not total payoff. **(c)** Individual cooperation with one $HL = 2$ and two $HL = 80$ players. The TG comparison assigns short memory to Player 1 and long memory to Player 2.

G Complete Analysis of the Explicit-Reasoning Ablation

We compare explicit-reasoning and action-only behavior across all 18 settings in the chain-of-thought (CoT) memory-cursed group. No reasoning (NR) denotes the action-only format, in which the prompt requests the final action without emitted intermediate reasoning. Let C_h denote cooperation rate at

Appendix Table 10: Public Goods: Systemic Cooperation Rates across Mixed Configurations.

Model	Sym HL=2	Mixed: 2×HL=2 + 1×HL=80	Mixed: 2×HL=80 + 1×HL=2	Sym HL=80
Gemma-3-12B	64.5±10.2%	59.4±1.1%	60.5±1.5%	50.7±7.5%
GPT-OSS-20B	47.4±5.7%	28.2±1.4%	15.9±0.3%	21.2±22.3%
Llama-4-Scout-17B	82.6±0.2%	88.7±3.6%	90.1±8.3%	45.8±24.0%
Mistral-7B	76.3±1.0%	61.8±1.3%	56.0±0.6%	50.8±3.7%
Llama-3.3-70B	99.8±0.2%	100.0±0.0%	99.9±0.1%	100.0±0.0%

Appendix Table 11: Public Goods: Per-Player Breakdown in the Adversarial Majority Setting (2 Grudge-Holders + 1 Forgiver).

Model	Overall Group Rate	Lone Forgiver (HL=2)	Grudge-Holders (HL=80)	Δ (Forgiver - Grudge)
Gemma-3-12B	60.5±1.5%	63.9±2.3%	58.7±1.5%	**+5.2pp**
GPT-OSS-20B	15.9±0.3%	37.7±0.4%	5.0±0.5%	**+32.7pp**
Llama-4-Scout-17B	90.1±8.3%	90.7±3.9%	89.8±10.5%	**+0.9pp**
Mistral-7B	56.0±0.6%	70.6±1.3%	48.8±0.9%	**+21.8pp**
Llama-3.3-70B	99.9±0.1%	99.9±0.1%	100.0±0.0%	**-0.1pp**

history length h, and $D = C_2 - C_{80}$ its short-to-long-history drop. In Appendix Table 12, SD is the standard deviation of C_{80} across three runs within one model–game–output-mode cell, measured in percentage points. The main summary gives equal weight to the 18 cells and reports SD separately from D.

The mean drops are 42.3 and 26.8 pp, mean C_{80} values are 22.4% and 40.3%, and mean C_{80} SDs are 5.3 and 5.5 pp. Removing explicit CoT reduces the drop in 14/18 cells, raises C_{80} in 15/18, and satisfies both in 12/18; SD rises in nine and falls in nine. The complete table also reveals distinct response types: Gemma/PG, for example, has zero cooperation at both NR endpoints, whereas 15 cells achieve higher long-history cooperation under NR.

Appendix Table 13 complements the complete analysis by showing, for each model, the game with its largest action-only versus CoT endpoint difference.

System Prompt for No-reasoning Ablation Study For the no-reasoning ablation, we use the same prompt structure and game rules, but remove the reasoning requirement and instead instruct the agent to output only its final action. The prompt is as follows:

''' You are Player {player_id}, playing a repeated game with Player {other_id}. This is round {round_num}. You can see the most recent {history_length} rounds of history.

{game_rules}

You are playing this game repeatedly with the same player. The action you choose in the current round will be visible to the other player in future rounds and may influence their future decisions. After each round, there is a 99% chance that another round will take place. Therefore, your goal is to maximize your total payoff in the long run rather than only the current-round payoff.

Past rounds' results:

{history_block}

Do not provide explanation. You MUST directly output ONLY your action.

Game-specific decision formats. For Prisoner's Dilemma, the required output is [Option A] or [Option B], corresponding to A_0 (cooperate) and A_1 (defect) in our tables. For Traveler's Dilemma, the required output format is [A0, A1, A2, or A3]. For Public Goods Game, the required output format is [A0 or A1]. For Trust Game, the required output format is [A0 or A1]. '''

H Memory Sanitization: Full Experimental Details

Setup. We instantiate sanitization in the Trust Game (TG) with the total history length fixed at HL=80. For each round at test time, we expose the model to a constructed 80-round history of the

Appendix Table 12: Complete comparison of the 18 CoT memory-cursed settings. D and SD are in pp; C_{80} is in percent. CoT and NR denote explicit-reasoning and action-only output conditions. SD is the standard deviation of C_{80} across three runs within each model–game–mode cell.

Game	Model	D_{CoT}	D_{NR}	$C_{80,\mathrm{CoT}}$	$C_{80,\mathrm{NR}}$	$\mathrm{SD}_{\mathrm{CoT}}$	$\mathrm{SD}_{\mathrm{NR}}$
PD	GPT-OSS-20B	71.5	71.5	20.6	25.3	15.1	9.4
PD	Mistral-7B	11.6	0.0	85.3	100.0	2.9	0.0
PG	Gemma-3-12B	21.0	0.0	53.0	0.0	5.9	0.0
PG	GPT-OSS-20B	38.3	35.4	5.2	13.8	0.4	4.8
PG	Llama-4-Scout-17B	36.8	29.9	45.8	69.7	24.0	5.4
PG	Mistral-7B	24.6	0.0	50.3	100.0	3.1	0.0
PG	GPT-OSS-120B	71.4	70.6	4.0	9.0	0.5	1.6
TD	Gemma-3-12B	35.0	0.0	35.3	100.0	15.4	0.0
TD	GPT-OSS-20B	39.3	49.0	2.7	10.8	1.2	7.2
TD	Mistral-7B	8.6	-8.8	16.6	18.1	0.3	3.3
TD	GPT-OSS-120B	80.2	63.9	12.6	15.7	15.4	20.8
TD	Qwen2.5-Coder-32B	80.9	11.6	8.0	85.2	0.5	3.4
TG	Gemma-3-12B	41.7	5.6	9.5	0.1	0.9	0.0
TG	GPT-OSS-20B	25.3	17.2	3.4	20.6	0.6	1.7
TG	Llama-4-Scout-17B	42.8	32.5	2.2	10.2	0.5	1.2
TG	Mistral-7B	3.0	15.6	34.8	44.9	0.2	39.3
TG	Llama-3.3-70B	43.9	0.0	6.9	100.0	4.4	0.0
TG	GPT-OSS-120B	85.4	89.0	7.3	1.6	4.7	0.8

Appendix Table 13: Largest explicit-reasoning versus action-only cooperation difference at $HL{=}80$ for each model. The displayed game is selected descriptively by the largest endpoint difference. Raw values appear in Appendix Table 1 (explicit reasoning) and Appendix Table 6 (action only).

Model	**Most Impacted Game**	**No-Reasoning**	**Reasoning (CoT)**	**Δ (Penalty)**
Llama-3.3-70B	Trust Game	100.0%	6.9%	**-93.1 pp**
Qwen2.5-Coder-32B	Traveler's Dilemma	85.2%	8.0%	**-77.2 pp**
Gemma-3-12B	Traveler's Dilemma	100.0%	35.3%	**-64.7 pp**
Mistral-7B	Public Goods	100.0%	50.3%	**-49.7 pp**
Llama-4-Scout-17B	Public Goods	69.7%	45.8%	**-23.9 pp**
GPT-OSS-20B	Trust Game	20.6%	3.4%	**-17.2 pp**
GPT-OSS-120B	Public Goods	9.0%	4.0%	**-5.0 pp**

form

$$\underbrace{(\mathrm{A0},\mathrm{A0}),\ldots,(\mathrm{A0},\mathrm{A0})}_{80-X\ \text{synthetic cooperative rounds}} \ \|\ \underbrace{r_{t-X+1},\ldots,r_t}_{X\ \text{real recent rounds}},$$

where each synthetic round shows mutual A0 cooperation with the corresponding payoff $(10, 10)$. The real rounds are produced by the ongoing self-play of the two agents (no oracle is used). X is the only free parameter; $X{=}80$ recovers the standard $HL{=}80$ baseline condition, and $X{=}0$ corresponds to an entirely synthetic history.

Models and runs. We use the same five models as the main sanitization figure: Gemma-3-12B, GPT-OSS-20B, Llama-3.3-70B, Llama-4-Scout-17B, and Mistral-7B. Each $(model, X)$ cell is run 3 times with independent random seeds; we report mean$\pm$std over runs of the per-action cooperation rate (fraction of A0 actions) computed over 500 rounds.

Sweep over X. Appendix Table 14 reports $X \in \{2, 5, 10, 20, 40, 60, 70, 75, 78, 80\}$. At $X = 2$, all five models exceed their unsanitized baseline. The intermediate curves are model-specific: for example, Mistral-7B falls to 28.53% at $X = 40$ and rises toward 34.8% at $X = 80$. These response curves quantify sensitivity to the action and payoff content retained in memory.

Polar variant. The TD sanitization control compares IDEAL synthetic mutual A_3 rounds with an empirically derived POLAR history using extreme action labels (Appendix Table 15). At $X = 2$, cooperation changes from 41.77% to 3.10% for GPT-OSS-20B and from 52.40% to 28.83% for

Appendix Table 14: Trust Game per-action cooperation rate (%) under memory sanitization at $HL{=}80$, swept over the size X of the real recent-history window. The remaining $80 - X$ rounds are replaced with synthetic mutual-cooperation records. Smaller X corresponds to more aggressive sanitization. "Baseline ($X{=}80$)" is the unsanitized cooperation rate from the main experiment. Each cell is mean±between-run SD over 3 independent runs of 500 rounds each.

X	**Gemma-3-12B**	**GPT-OSS-20B**	**Llama-3.3-70B**	**Llama-4-Scout-17B**	**Mistral-7B**
$X{=}2$	84.00±2.40	36.27±0.86	97.43±0.76	87.93±2.70	45.43±1.64
$X{=}5$	73.20±4.97	31.87±2.16	91.20±5.45	81.80±2.46	37.90±1.08
$X{=}10$	45.97±1.63	26.37±2.65	65.83±28.85	70.83±4.11	32.67±0.75
$X{=}20$	24.80±1.82	18.40±0.78	53.80±13.30	45.60±4.81	29.83±1.46
$X{=}40$	10.63±1.12	13.47±0.86	31.63±35.14	11.63±0.29	28.53±0.68
$X{=}60$	8.67±2.08	7.83±1.21	13.00±6.00	8.80±0.89	30.07±1.72
$X{=}70$	11.27±3.04	7.03±0.72	12.00±4.28	5.90±1.28	32.73±0.15
$X{=}75$	9.47±1.00	4.57±0.97	8.30±2.00	5.10±0.53	33.20±0.46
$X{=}78$	9.93±1.55	4.80±0.75	9.00±7.11	6.83±4.44	34.77±0.40
Baseline ($X{=}80$)	9.5±0.9	3.4±0.6	6.9±4.4	2.2±0.5	34.8±0.2

Mistral-7B. The response is model-dependent, and the full table contains reversals at other X. Both conditions preserve 80 visible rounds while varying the interaction patterns injected into the ledger.

Appendix Table 15: Traveler's Dilemma per-action cooperation rate (%) under IDEAL vs. POLAR sanitization modes at $HL{=}80$. IDEAL replaces $80{-}X$ rounds with pure mutual cooperation (A3,A3). POLAR replaces them with empirical distributions polarized to the extremes, explicitly injecting defection (A0,A0) signals into the synthetic history. The "$X{=}80$" row reports the unsanitized cooperation rate (baseline) from the main experiment. The presence of defection content in POLAR drastically depresses cooperation in susceptible models compared to the purely cooperative IDEAL history.

X	**Gemma-3-12B**		**GPT-OSS-20B**		**Llama-3.3-70B**		**Llama-4-Scout-17B**		**Mistral-7B**	
	IDEAL	**POLAR**	**IDEAL**	**POLAR**	**IDEAL**	**POLAR**	**IDEAL**	**POLAR**	**IDEAL**	**POLAR**
$X{=}2$	99.00±0.46	94.47±6.60	41.77±2.63	3.10±0.36	100.00±0.00	100.00±0.00	99.93±0.06	100.00±0.00	52.40±1.71	28.83±2.10
$X{=}5$	97.63±0.81	93.33±10.08	41.97±0.40	7.97±5.42	100.00±0.00	100.00±0.00	99.70±0.44	98.97±1.79	46.67±1.96	27.37±0.59
$X{=}10$	95.33±0.12	96.63±2.84	34.40±1.87	29.47±23.54	100.00±0.00	100.00±0.00	98.40±2.77	99.80±0.35	43.27±1.87	23.10±2.11
$X{=}20$	94.03±2.04	80.33±29.22	22.57±6.64	2.70±0.00	99.93±0.12	100.00±0.00	97.07±5.08	99.90±0.10	40.97±0.42	23.93±0.42
$X{=}40$	90.43±8.43	67.70±23.62	6.27±0.50	10.77±12.69	100.00±0.00	100.00±0.00	99.90±0.17	100.00±0.00	27.87±1.22	19.87±1.10
$X{=}60$	79.43±16.51	17.20±20.93	3.57±0.75	4.43±1.23	99.97±0.06	100.00±0.00	100.00±0.00	99.90±0.17	20.20±2.74	15.27±2.18
$X{=}70$	47.77±48.43	37.57±52.18	2.30±0.46	1.60±0.56	100.00±0.00	100.00±0.00	99.07±1.07	99.60±0.61	19.43±0.97	15.23±0.42
$X{=}75$	62.20±52.57	87.83±15.13	1.67±0.40	3.80±3.18	100.00±0.00	100.00±0.00	99.83±0.29	99.97±0.06	16.20±0.46	15.00±0.90
$X{=}78$	82.27±17.03	59.80±44.95	1.83±0.50	1.53±0.49	99.97±0.06	100.00±0.00	99.90±0.10	99.93±0.12	17.33±2.74	16.67±0.96
$X{=}80$	35.3±15.4		2.7±1.2		100.0±0.0		100.0±0.0		16.6±0.3	

Interpretation. Sanitization changes cooperation while retaining an 80-round ledger, directly demonstrating sensitivity to the remembered action and payoff content. The synthetic records change the evidence available about opponents while holding the visible round count fixed; the token-matched controls in Appendix L separately assess compactness.

I Details of the Action-Balanced Rationale-Only Intervention

The adapter is trained using 1,000 forward-looking PG traces, selected to contain 500 originally cooperative and 500 originally non-cooperative final actions. The serialized examples retain terminal-action suffixes and the end-of-sequence (EOS) token, but the trainer assigns them label -100; only rationale tokens contribute to supervised loss. The selection labels and the supervised targets are therefore distinct: actions determine balance during sampling, but the terminal actions are not training targets.

I.1 Curation, Supervision, and Checkpoint Selection

Only PG supplies training traces; PD, TD, and TG are zero-shot evaluation environments. The contextual validation of the exact selected rationales is reported in Section 4.2. The 500/500 balance controls terminal-action-label proportions, while the contextual scores in Section 4.2 characterize the

rationale content retained within those groups. The frozen PG-only curation pass scans 84 eligible archived PG files and identifies 13,465 strong candidates (10,276 originally A_0, 3,189 originally A_1) under a sentence-local rule requiring two temporal and two cross-round mechanism groups. It selects 550 within-trajectory action pairs without replacement: 500 pairs for training and 50 for validation. The 1,100 selected IDs, prompts, and full rationales are unique; train and validation use disjoint source trajectories. All 1,000 training example weights are 1.0, while supervised token counts vary with rationale length. The local evaluation uses a fixed Mistral-7B-Instruct-v0.1 model copy, a shared base/adapter vLLM server in bfloat16 (BF16) with an 8,192-token model limit, temperature 0.7, 2,000 output tokens, and three 500-round runs per condition. The seven-model evaluation uses Cloudflare serving identifiers, whereas the fine-tuning comparison evaluates the base model and adapter together on the fixed local model copy. Results from these evaluation settings are reported separately. The frozen per-run score files show 146 PG adapter output-format parse failures among 27,000 action slots (0.54%) and one PG base failure among 27,000 slots; PD, TD, and TG have none. The protocol counts failed format parses as non-cooperative in the fixed action denominator and invalidates a cell on API failure.

I.2 Optimization Configuration

The balanced run uses the Mistral LoRA configuration with the PG-only dataset and explicitly sets the training duration and maximum sequence length. The completion and server manifests identify the selected step-327 adapter, and the rationale-only masks below define its supervised loss.

Appendix Table 16: Balanced PG-only rationale adapter configuration from the archived launcher, source YAML, trainer, and completion manifest.

Item	Balanced-run setting
Backbone	Mistral-7B-Instruct-v0.1; frozen base weights
LoRA targets	Attention q, k, v, o and multilayer perceptron (MLP) gate/up/down projections
LoRA parameters	Rank 16, scaling $\alpha = 32$, dropout 0.05
Optimizer / schedule	AdamW; peak learning rate 2×10^{-4}; cosine schedule; 3% warmup
Training duration / batches	327 optimizer updates, seed 42; batch size 2 per device; accumulation 8
Maximum sequence length	4,096 tokens; trainer aborts if any frozen train/validation sequence is truncated
Evaluation backend	Shared base/LoRA vLLM server, BF16; fixed local model copy

The trainer masks prompt tokens, any token crossing the rationale/action boundary, the entire canonical terminal-action suffix, and EOS with label -100; all examples have unit weight. The trainer checks that every frozen training and validation sequence fits without truncation. The 500/500 balance is defined at the example level, while supervised-token mass can vary across original-action groups with rationale length.

I.3 Full Cooperation Results

Appendix Table 17 gives all six history lengths. Appendix Table 18 reports changes computed from the unrounded results, which can differ by 0.1 pp from subtracting displayed rounded endpoints.

Appendix Table 17: Balanced rationale-only intervention: cooperation in percent (base → FT), using the additional experiment's own base cohort.

HL	PG (train)	PD (transfer)	TD (transfer)	TG (transfer)
2	$41.8 \to 61.4$	$84.4 \to 86.2$	$42.1 \to 41.7$	$37.8 \to 43.0$
5	$38.0 \to 58.0$	$84.1 \to 89.0$	$46.2 \to 52.8$	$39.1 \to 41.6$
10	$34.7 \to 55.7$	$87.0 \to 86.9$	$44.8 \to 45.9$	$36.6 \to 40.9$
20	$30.2 \to 55.1$	$88.6 \to 86.3$	$47.7 \to 49.3$	$29.2 \to 38.7$
40	$23.5 \to 54.5$	$92.4 \to 93.7$	$50.8 \to 61.9$	$28.3 \to 40.6$
80	$18.4 \to 52.5$	$91.1 \to 94.5$	$52.5 \to 62.8$	$23.9 \to 44.2$
Mean	$31.1 \to 56.2$	$87.9 \to 89.4$	$47.4 \to 52.4$	$32.5 \to 41.5$

Appendix Table 18: Base-to-FT cooperation changes in pp, computed from unrounded results. Subtracting the displayed rounded endpoints can differ by 0.1 pp.

HL	PG	PD	TD	TG
2	$+19.5$	$+1.7$	-0.5	$+5.1$
5	$+20.0$	$+4.9$	$+6.6$	$+2.5$
10	$+21.0$	-0.1	$+1.1$	$+4.3$
20	$+24.9$	-2.3	$+1.6$	$+9.5$
40	$+31.0$	$+1.3$	$+11.1$	$+12.3$
80	$+34.1$	$+3.4$	$+10.3$	$+20.2$
Mean	$+25.1$	$+1.5$	$+5.0$	$+9.0$

J Additional Protocol and Reporting Details

The primary sweep uses the seven serving identifiers in Appendix Table 19. The balanced-adapter evaluation and additional Mistral controls report their fixed local model configuration in their respective sections.

Appendix Table 19: Workers AI identifiers reported for the primary experiments.

Paper label	Serving identifier
Gemma-3-12B	`@cf/google/gemma-3-12b-it`
GPT-OSS-20B	`@cf/openai/gpt-oss-20b`
GPT-OSS-120B	`@cf/openai/gpt-oss-120b`
Llama-3.3-70B	`@cf/meta/llama-3.3-70b-instruct-fp8-fast`
Llama-4-Scout-17B	`@cf/meta/llama-4-scout-17b-16e-instruct`
Mistral-7B	`@cf/mistral/mistral-7b-instruct-v0.1`
Qwen2.5-Coder-32B	`@cf/qwen/qwen2.5-coder-32b-instruct`

The prompt's 99% continuation sentence describes the strategic setting shown to agents; the four archived game engines iterate for `num_rounds=500`.

The design has $7 \times 4 \times 9 \times 3 = 756$ group trajectories and, at 500 rounds each, 378,000 group-round slots. Three games have two players and PG has three, yielding 850,500 agent-decision slots before invalid actions or omissions. We use "group rounds" and "agent decisions" for these two sampling units, respectively.

J.1 Token-Length Analysis

Using the public Mistral-v0.1 tokenizer as a common measuring instrument, the archived four-model analysis reports a maximum history block of 2,879 tokens and maximum complete input of 3,317 tokens. Appendix Table 20 gives the 99th-percentile complete-input and response lengths for the four measured models.

Appendix Table 20: 99th-percentile complete-input / response lengths for the four measured models under the common Mistral-v0.1 tokenizer.

Model	PG	PD	TD	TG
Gemma-3-12B	2611 / 629	3317 / 432	2202 / 611	2123 / 591
GPT-OSS-20B	2611 / 356	3317 / 395	2202 / 290	2120 / 257
Mistral-7B	2611 / 536	3317 / 468	2202 / 530	2153 / 503
Qwen2.5-Coder-32B	2611 / 720	3317 / 520	2202 / 791	2193 / 696

Appendix Table 21: Observed action-transition frequencies (%). Rows correspond to accessible history length; the two transitions in each conditioning state sum to 100%.

HL	$D \to D$	$D \to C$	$C \to D$	$C \to C$
1	54.33	45.67	12.41	87.59
2	61.75	38.25	11.16	88.84
3	66.52	33.48	11.49	88.51
5	74.80	25.20	11.54	88.46
10	79.78	20.22	12.03	87.97
20	84.12	15.88	10.77	89.23
40	88.59	11.41	10.60	89.40
80	90.42	9.58	10.19	89.81

K Action-Persistence Analysis

We analyze observed decisions in rounds 81–500 under nonzero history lengths, so that at least the preceding round is available. C denotes the game-specific cooperative action, and D any other action (including all three lower TD claims). Appendix Table 21 reports the conditional transition frequencies. Among 490 nonzero-history trajectories containing both preceding-action states, $P(D_t \mid D_{t-1}) - P(D_t \mid C_{t-1})$ is positive in 385 (78.6%), with median difference 12.1 pp. This within-trajectory association shows that non-cooperation is more likely to persist after a preceding non-cooperative action.

L Alternative Memory Representations and Matched-Prefix Continuation

We use Mistral-7B on 240 frozen decision contexts taken from source $HL = 80$ trajectories. Each underlying 80-round history is rendered as: (i) Raw80; (ii) a deterministic statistical Summary of all 80 rounds plus the latest two rounds verbatim; (iii) Retrieval, containing the latest two rounds and up to eight earlier rounds representing distinct interaction outcomes; (iv) Recent2; and (v–vi) recent raw suffixes token-matched to the Summary and Retrieval views. The game prompt, underlying context, model, decoding, parser, and generation seed are paired across views. Five generations per context and view produce $240 \times 6 \times 5 = 7{,}200$ requests.

Appendix Table 22: Fixed-context next-action contrasts across four games. Values are changes in cooperation (pp) with paired 95% CIs obtained by bootstrapping contexts while keeping repeated generations paired within context.

Contrast	Estimate	CI lower	CI upper
Recent2 - Raw80	+9.4	+5.7	+13.2
Retrieval - Raw80	+3.9	+0.4	+7.3
Summary - Raw80	+3.3	-0.2	+6.8
Summary - matched raw	-0.6	-4.2	+3.0
Retrieval - matched raw	-1.3	-5.0	+2.4

Appendix Table 23: PG matched-prefix continuation contrasts (pp) with paired 95% CIs. The nine prefixes are drawn at three branch points from each of three source trajectories.

Contrast	Estimate	CI lower	CI upper
Summary - Raw80	+13.4	+10.3	+16.3
Matched raw - Raw80	+11.9	+9.9	+13.8

Appendix Table 24: Memory-drop point estimates for each presentation and instruction.

Presentation	Baseline	Analysis	Future
Original	+5.1	+5.7	+6.4
Numerical	+1.0	+6.3	+2.7
Narrative	+6.3	+1.6	+2.3

To examine subsequent interaction, we select nine PG prefixes from three source trajectories, branching after rounds 100, 250, and 400. Each prefix is continued for 100 rounds under Raw80, Summary+Recent2, and summary-token-matched recent raw memory. All three players in a branch use the same memory condition, yielding $9 \times 3 \times 100 \times 3 = 8{,}100$ action requests. Prefixes are identical at branching; action-dependent histories can diverge afterward.

Recent2 and Retrieval improve over Raw80 in the fixed-history aggregate, while the Summary estimate is positive with a confidence interval that crosses zero. Summary and Retrieval closely track their token-matched compact raw controls. In continuation, both Summary and compact raw memory improve over Raw80, with similar point estimates. Together, the results show that the aggregate improvements are shared by compact structured views and compact raw-history controls.

M Prompting and Rule-Framing Controls

We evaluate 240 fixed decision contexts from four games, each with an underlying 80-round history. $HL = 2$ reveals the latest two rounds; $HL = 80$ reveals the full history. Original uses the anonymous action labels and rules from the primary experiments, while Numerical and Narrative provide two alternative rule presentations. We estimate instruction effects within each fixed presentation. Baseline adds no instruction, Analysis asks for closer examination of numerical rules and visible history, and Future asks the model to consider future rounds, isolated versus persistent deviations, and long-run consequences. Analysis is length-matched to Future, and both remain action-neutral.

Define $D = C_2 - C_{80}$. A positive $D_{\text{Baseline}} - D_{\text{Future}}$ or $D_{\text{Analysis}} - D_{\text{Future}}$ indicates a smaller drop under Future. The endpoint contrast is $C_{\text{Future},80} - C_{\text{Baseline},80}$. All values below are in pp.

Future-oriented prompting has near-zero aggregate effects. The game-resolved directions remain stable across prompt variants: PG and TG have lower cooperation at $HL = 80$ in all nine presentation–instruction combinations, PD has higher cooperation in all nine, and TD has smaller, more variable

Appendix Table 25: Paired prompt contrasts and reported 95% CIs. Every interval includes zero.

Presentation	Contrast	Estimate	CI lower	CI upper
Original	Future vs. Baseline drop	-1.3	-8.6	+6.0
Original	Future vs. Analysis drop	-0.7	-7.6	+6.1
Original	Future long-history gain	+0.3	-4.9	+5.3
Numerical	Future vs. Baseline drop	-1.7	-9.9	+6.7
Numerical	Future vs. Analysis drop	+3.7	-4.7	+11.5
Numerical	Future long-history gain	-3.3	-9.2	+2.4
Narrative	Future vs. Baseline drop	+3.9	-4.7	+12.6
Narrative	Future vs. Analysis drop	-0.7	-8.3	+6.9
Narrative	Future long-history gain	+0.0	-5.8	+6.1

Appendix Table 26: Scripted PD opponents using base Mistral: cooperation (%) and mean payoff as the instruction changes from Baseline to Future.

Opponent	Presentation	Cooperation	Mean payoff
Always-Defect	Numerical	$36.4 \rightarrow 36.2$	$27.2 \rightarrow 27.6$
Always-Defect	Narrative	$35.0 \rightarrow 45.6$	$30.0 \rightarrow 8.8$
One-error Tit-for-Tat	Numerical	$46.0 \rightarrow 50.0$	$123.0 \rightarrow 126.6$
One-error Tit-for-Tat	Narrative	$49.0 \rightarrow 47.2$	$124.4 \rightarrow 123.0$
Tit-for-Tat	Numerical	$49.6 \rightarrow 44.8$	$127.2 \rightarrow 119.2$
Tit-for-Tat	Narrative	$56.6 \rightarrow 58.4$	$133.0 \rightarrow 136.6$

Appendix Table 27: PG payoff sweep. Drops are in pp, with population SD across three run-ID-paired differences.

r	Free-riding advantage	Normalized gap G	Memory drop
1.1	0.63	0.091	23.1 ± 0.8
1.2	0.60	0.167	24.3 ± 2.1
1.3	0.57	0.231	21.4 ± 1.2
1.4	0.53	0.286	21.2 ± 1.2
1.5	0.50	0.333	24.6 ± 5.1
1.8	0.40	0.444	21.2 ± 0.8
2.1	0.30	0.524	23.8 ± 2.8
2.4	0.20	0.583	23.1 ± 1.7
2.7	0.10	0.630	24.2 ± 2.1

changes. This fixed-context study complements the primary self-play trajectories by isolating next-action responses to matched histories.

N Scripted-Opponent Controls: Cooperation Versus Payoff

The additional PD evaluation runs for 100 rounds against Always-Defect, Tit-for-Tat, and One-error Tit-for-Tat, under Numerical and Narrative wording. Appendix Table 26 gives Baseline-to-Future results and connects the cooperative-action metric to realized payoff under each opponent policy.

Against Always-Defect with Narrative wording, Future raises cooperation by 10.6 pp and decreases payoff by 21.2. For the documented PD payoffs, mean per-round reward against Always-Defect is $100(1-c) - 100c = 100 - 200c$ for cooperation fraction c, which matches the displayed values 30.0 and 8.8. The result quantifies the exploitation cost of cooperation against a persistently defecting opponent.

O Controlled Payoff Sweep and Cross-Game Structure

We vary the multiplier in three-player PG using Mistral-7B while retaining the action space, prompt structure, explicit-reasoning instruction, 500-round horizon, and three run IDs. For k contributors, contribution pays $rk/3$ and free-riding pays $1 + rk/3$. For a unilateral switch with k_{-i} other contributors, the free-riding advantage is $[1 + rk_{-i}/3] - r(k_{-i}+1)/3 = 1 - r/3$. For $1 < r < 3$, free-riding is the better one-round reply, while universal contribution maximizes total welfare. The normalized welfare gap between universal free-riding and universal contribution is $G = (r-1)/r$.

For run ID s, define $D_s(r) = C_{2,s}(r) - C_{80,s}(r)$. The sweep reports the mean and population SD of these paired differences across the three run IDs. Pairing aligns run IDs, while each condition develops its own interaction history.

The drop remains positive at every tested multiplier and ranges from 21.2 to 24.6 pp. The effect therefore persists throughout the tested incentive range for Mistral-7B in PG. Because both incentive measures vary with the same multiplier, this sweep characterizes their joint variation within one game.

Appendix Table 28: Descriptive cross-game comparison from the seven-model evaluation, giving each model equal weight. Gaps use documented stage-game payoffs; drops are in pp.

Game	Structure	Normalized gap	Mean drop
PD	Symmetric; 2 actions	0.500	11.9
PG	Symmetric; 3 players	0.333	27.6
TD	Symmetric; 4 actions	0.600	34.7
TG	Role-asymmetric; 2 actions	–	34.5

PD has a larger normalized welfare gap than PG but a smaller mean drop. TG contributes its cooperation drop without a directly comparable normalized welfare-gap value. The cross-game pattern shows that a one-dimensional payoff gap is insufficient to explain memory susceptibility, motivating analyses that jointly consider action space, player count, roles, and wording.