IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024

Theory-Grounded LLM Societies for Emergent Coordination

J. DE CURTÒ[1,2], (Member, IEEE) and I. DE ZARZÀ[3], (Member, IEEE)

[1]Department of Computer Applications in Science & Engineering, BARCELONA Supercomputing Center, 08034 Barcelona, Spain
[2]Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas, 28015 Madrid, Spain
[3]Department of Human centered AI, Data & Software, LUXEMBOURG Institute of Science and Technology, 4362 Esch-sur-Alzette, Luxembourg

Corresponding author: J. de Curtò (e-mail: jdecurto@icai.comillas.edu).

ABSTRACT Recent generative-agent frameworks demonstrate that large language models (LLMs) can produce socially plausible behavior in multi-agent simulations, yet their strategy dynamics remain ungrounded: agents update beliefs through narrative reasoning alone, without the evolutionary and game-theoretic mechanisms known to govern cooperation in structured populations. We introduce Coevolutionary Generative Societies (CGS), an architecture that fuses LLM deliberation with evolutionary cooperation theory (EC-theory) to produce stable, resilient cooperative equilibria on interaction networks. Each agent maintains episodic memory and updates its cooperation propensity through a three-component rule combining Fermi-function social learning from neighbors, sigmoid individual reinforcement normalized by network degree, and a β-weighted advisory signal from a Solver–Critic–Aggregator deliberation pipeline, where the solver proposes an action, a critic evaluates strategic soundness, and a deterministic aggregator fuses both signals, flipping the proposed action on strong disagreement. We evaluate CGS on Watts–Strogatz, Barabási–Albert, and Erdös–Rényi networks ($n = 40$, $T = 50$ steps) under a standard Prisoner's Dilemma with an adversarial shock at $t = 30$ that forces 15% of agents to defect. CGS achieves peak cooperation rates of 0.80 and recovers from adversarial shocks within 8 time steps, whereas a narrative-only ablation (pure LLM-driven updates, no EC dynamics) plateaus at 0.55 cooperation and exhibits no dynamic response to perturbation. A no-critic ablation confirms that the full deliberation pipeline contributes measurable gains over solver-only inference. Cross-topology analysis reveals that small-world networks yield the highest cooperation and lowest payoff inequality (Gini $= 0.236$), while scale-free networks amplify inequality through hub concentration (Gini $= 0.420$). Ten-seed replication establishes that the qualitative distinction between CGS and a narrative-only ablation is statistically significant after Bonferroni correction (Cohen's $d = 4.59$ on cooperation responsiveness, $p_{\text{Bonf}} < 0.001$; CGS-EC remains dynamically adaptive in 9 of 10 seeds whereas the narrative-only baseline converges to a static equilibrium in 10 of 10). Population scaling to n = 80, a four-model sweep, alternative game substrates (Snowdrift, Stag Hunt), and five non-LLM learning and imitation baselines confirm that the dynamics are preserved across initialisations, scales, LLM families, and game classes. Comparison against GNN structural baselines, which replace LLM deliberation with graph message-passing, quantifies the LLM advisory premium at $+20.2\%$ cooperation over the best structural baseline, establishing the respective contributions of evolutionary structure and language-model advisory. An OAT sensitivity analysis confirms that the fixed architectural weights are robust (NSI < 0.1 for κ and w_ℓ), and temperature robustness experiments across $T_s \in \{0.0, 0.2, 0.5, 0.8\}$ confirm insensitivity to LLM sampling stochasticity (CV $= 0.6\%$ on final cooperation). Our results establish that grounding LLM-agent societies in evolutionary game-theoretic update rules is essential for emergent coordination *within the tested regime* ($n \leq 80$, $T = 50$, PD/Snowdrift/Stag Hunt, single shocks up to 15%), and that the topology of the interaction network significantly modulates both the level and the equity of cooperation.

INDEX TERMS Multi-agent systems, large language models, evolutionary game theory, cooperation emergence, Prisoner's Dilemma, network topology, generative agents

I. INTRODUCTION

The emergence and stability of cooperation among self-interested agents is one of the oldest open problems at the intersection of biology, economics, and computer science. Evolutionary game theory has established that network structure, imitation dynamics, and population heterogeneity are decisive for whether cooperation can invade and persist in populations playing social dilemmas [1]–[3]. On graphs, the celebrated result of Ohtsuki et al. [4] shows that natural selection favors cooperators when the benefit-to-cost ratio exceeds the average degree, while Santos and Pacheco [5] demonstrate that scale-free topologies can amplify cooperation through hub-mediated reinforcement. The Fermi imitation rule [6], [7], which governs the probability that an agent adopts a neighbor's strategy as a sigmoid function of payoff difference, has become the standard microscopic update mechanism in this literature [8].

In parallel, the rapid maturation of large language models (LLMs) [9], [10] has opened a fundamentally different approach to multi-agent simulation. Park et al. [11] introduced *generative agents*, LLM-driven characters that maintain memory streams, reflect on experience, and plan actions in natural language, demonstrating emergent social behavior in a sandbox environment. Subsequent work has extended this paradigm to multi-agent debate [12], [13], role-playing societies [14], and surveys of LLM-based agent architectures more broadly [15], [16]. A growing body of research has also probed LLM behavior in canonical strategic games: Akata et al. [17] show that LLMs exhibit cooperation biases in repeated Prisoner's Dilemma, Fontana et al. [18] find that language models tend to cooperate more than humans, and Fan et al. [19] systematically evaluate rationality across game-theoretic benchmarks. Piatti et al. [20] study cooperation emergence in LLM societies but without formal evolutionary grounding.

Despite this progress, a critical gap persists: *generative-agent frameworks and evolutionary cooperation theory have developed largely in isolation.* LLM-based agents update strategies through narrative reasoning, "I should cooperate because my neighbor seems trustworthy", without any formal mechanism ensuring that these updates respect the payoff gradients, imitation pressures, and frequency-dependent selection that evolutionary theory identifies as essential for stable cooperation [21], [22]. Conversely, evolutionary game-theoretic models employ fixed behavioral rules (Fermi imitation, replicator dynamics) that cannot capture the contextual, memory-dependent, and linguistically mediated decision-making that LLMs provide [23]. Prior work by de Curtò and de Zarzà [24], based on the initial theoretical formulation proposed in [25], demonstrated that LLM-generated narratives can influence cooperation rates in multi-agent networks and introduced preliminary evolutionary update rules, but the update dynamics lacked the full three-component decomposition proposed here, and the deliberation architecture relied on call inference without internal step verification. Relative to that work, the present paper contributes three specific advances: (1) a three-component decomposition of the strategy update into Fermi social learning, degree-normalised reinforcement, and trust-weighted advisory, replacing the single coupled rule of [24]; (2) a Solver–Critic–Aggregator deliberation pipeline with a deterministic aggregator and an action-flip rule on strong disagreement, replacing one-call inference; and (3) a systematic evaluation battery (three topologies, two ablations, alternative games, non-LLM baselines, 10-seed statistics, cost and failure-mode analysis) absent from prior formulations. We therefore frame the contribution as the first systematic decomposition and evaluation of a theory-grounded LLM society, not as the first coupling of LLMs with evolutionary dynamics.

This paper bridges the two paradigms. We introduce Co-evolutionary Generative Societies (CGS), an architecture in which each agent's strategy update is governed by a principled EC-theory rule, combining Fermi-function social learning, degree-normalized individual reinforcement, and a heterogeneous trust-weighted LLM advisory signal, while the LLM component itself is structured as a Solver–Critic–Aggregator deliberation pipeline that provides verified, confidence-calibrated recommendations to the evolutionary update. The key insight is that grounding LLM advice within a formal evolutionary update rule yields qualitatively different dynamics than either mechanism alone: the evolutionary component prevents the LLM from inducing erratic or payoff-incoherent strategy shifts, while the LLM enriches the evolutionary dynamics with context-sensitive, memory-informed reasoning that fixed rules cannot provide.

We evaluate CGS on three canonical network topologies, Watts–Strogatz small-world [26], Barabási–Albert scale-free [27], and Erdös–Rényi random [28], under a standard Prisoner's Dilemma with an adversarial shock that forces a fraction of agents to defect mid-simulation. Two ablation studies isolate the contribution of each architectural component: a *narrative-only* baseline removes EC-theory grounding entirely (pure LLM-driven updates), and a *no-critic* ablation bypasses the critic stage of the deliberation pipeline.

Our contributions are as follows:

1) **A hybrid architecture** that formally integrates LLM deliberation into an evolutionary cooperation-theoretic update rule, combining Fermi social learning, sigmoid reinforcement, and trust-weighted LLM advisory into a three-component strategy update.
2) **A Solver–Critic–Aggregator deliberation pipeline** in which a solver proposes actions with calibrated confidence, a critic evaluates strategic soundness, and a deterministic aggregator fuses both signals, including action flips on strong disagreement, providing verified recommendations to the evolutionary update.
3) **Ablation evidence** demonstrating that EC-theory grounding is essential, not optional: the narrative-only baseline achieves 0.55 peak cooperation versus 0.80 for the full CGS system, exhibits no dynamic recovery from adversarial shocks, and converges to a static equilibrium in *ten of ten* independent replication seeds while

CGS-EC remains adaptive in *nine of ten* (Cohen's $d = 4.59$ on cooperation responsiveness, $p_{\text{Bonf}} < 0.001$).

4) **Cross-topology analysis** showing that network structure significantly modulates both cooperation level and payoff equity, with small-world networks producing the highest cooperation and lowest inequality (Gini $= 0.236$), while scale-free networks amplify disparities through hub concentration (Gini $= 0.420$).
5) **Population, model, structural, *game-substrate*, and *statistical* robustness.** Population scaling at $n \in \{40, 60, 80\}$ confirms that pre-shock cooperation and payoff equity are scale-invariant; a multi-model sweep across four LLMs reveals a two-cluster structure with advisory signal quality, not confidence, as the operative performance factor; ten-seed Bonferroni-corrected testing confirms the core dynamical contrast is statistically significant; and comparison against three GNN structural baselines without any LLM component quantifies the LLM advisory premium at $\Delta\rho_C = +0.100$ ($+20.2\%$ over the best structural baseline), decomposing the CGS advantage into structural and advisory subcontributions that ablation studies alone cannot resolve.
6) **Failure-mode characterisation and empirical equilibrium map.** A shock-magnitude sweep locates a resilience boundary between 15% and 30% shocked agents (and shows repeated shocks are more damaging than a single large one), while a no-shock initial-condition sweep reveals a critical-mass bifurcation near $\mu_p \approx 0.35$ separating a low- from a high-cooperation attractor.
7) **Public release** of the CGS evaluation framework and complete experimental results to facilitate reproducible assessment of cooperation dynamics in LLM-agent societies.[1]

The remainder of this paper is organized as follows. Section II reviews related work at the intersection of evolutionary game theory and LLM-based multi-agent systems. Section III presents the CGS architecture, including the EC-theory update rule, the Solver–Critic–Aggregator pipeline, and the memory-informed prompt construction. Section IV describes the experimental setup and ablation design. Section V reports quantitative results across topologies, ablations, and thirteen additional experiments (multi-seed replication, population scaling, multi-model sweep, GNN structural baselines, OAT parameter sensitivity together with temperature robustness, alternative games, non-LLM baselines, cost, failure modes, and equilibrium). Section VI discusses implications and limitations. Section VII concludes.

II. RELATED WORK

CGS draws on and contributes to three research streams: evolutionary cooperation on networks, LLM-based multi-agent systems, and the nascent intersection where language models participate in game-theoretic interactions.

[1]Code and data available at: https://github.com/drdezarza/cgs

A. EVOLUTIONARY COOPERATION ON NETWORKS

The study of cooperation in structured populations originates with Axelrod's tournament experiments [1] and has since been formalized through evolutionary game theory on graphs [3], [21]. Nowak [2] identified five mechanisms, kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection, through which natural selection can favor cooperators. Of these, *network reciprocity* is most relevant to our setting: Ohtsuki et al. [4] proved that cooperation is favored on graphs when $b/c > k$ (benefit-to-cost ratio exceeds average degree), and Santos and Pacheco [5] showed that the heterogeneous degree distributions of scale-free networks amplify cooperation by concentrating interactions around high-degree hubs.

The microscopic update rule governing strategy revision is critical. The Fermi imitation rule, introduced by Szabó and Töke [6] and further analyzed by Traulsen et al. [7], defines the probability that agent o adopts agent z's strategy as a sigmoid function of their payoff difference: $p_{oz} = 1/(1 + \exp[-\kappa(\pi_z - \pi_o)])$, where κ controls selection intensity. This rule interpolates between neutral drift ($\kappa \to 0$) and deterministic best-response ($\kappa \to \infty$), and has become the standard in spatial evolutionary games [8]. Coevolutionary extensions allow the network itself or behavioral parameters to co-adapt with strategies [22], [29], demonstrating that dynamic rewiring can further promote cooperation. Hauert and Doebeli [30] showed that the effect of spatial structure depends on the specific dilemma: while it promotes cooperation in the Prisoner's Dilemma, it can inhibit cooperation in the Snowdrift game, underscoring the importance of studying specific game–topology combinations.

CGS inherits the Fermi imitation mechanism and network-structured interactions from this tradition, but augments the update rule with two additional components, individual reinforcement and LLM advisory, that have no analogue in classical evolutionary models.

A complementary line of work studies game–environment feedback and behavioural adaptation in social dilemmas, including reinforcement-learning-driven adaptive decision rules in three-strategy evolutionary games [31], feedback-coupled behavioural vaccination dynamics [32], the evolutionary basis of prosocial mask-wearing [33], and replicator-based analyses of the human vaccine dilemma [34]. These establish that coupling behavioural updating to environmental or payoff feedback can qualitatively change equilibrium selection, a principle the CGS reinforcement component operationalises.

B. LLM-BASED MULTI-AGENT SYSTEMS

The generative agents framework of Park et al. [11] demonstrated that LLM-driven agents equipped with memory streams, reflection, and planning can produce emergent social phenomena, information diffusion, relationship formation, coordinated activities, in a simulated town. This work catalyzed rapid development of LLM-based multi-agent architectures. CAMEL [14] introduced role-playing communica-

tive agents with inception prompting for autonomous cooperation. Multi-agent debate frameworks [12], [13] showed that inter-agent deliberation improves factual accuracy and reasoning quality, motivating the use of adversarial verification (solver–critic patterns) within individual agent pipelines. Xi et al. [15] and Guo et al. [16] provide comprehensive surveys of the rapidly growing landscape of LLM agent architectures, identifying memory, planning, and tool use as the core capabilities.

A key limitation of these systems is that agent behavior emerges entirely from narrative reasoning: strategy updates are implicit in the LLM's text generation, with no formal connection to the payoff structures or population dynamics that would constrain rational or evolutionary behavior. When agents in these frameworks ''cooperate'' or ''defect,'' they do so because the language model's next-token distribution happens to favor cooperative language, not because cooperation is payoff-improving given the local strategic context. CGS addresses this by making the LLM a *contributor* to a formally grounded update, rather than the sole determinant of strategy revision.

C. LLMS IN STRATEGIC AND GAME-THEORETIC SETTINGS

A growing literature evaluates LLM behavior in canonical games. Brookins and DeBacker [35] found that GPT-based models exhibit mixed rationality in strategic games, sometimes deviating substantially from NASH equilibrium predictions. Akata et al. [17] demonstrated that LLMs playing repeated Prisoner's Dilemma exhibit cooperation biases and sensitivity to framing, suggesting that their strategic behavior is mediated by language-level heuristics rather than payoff computation. Fontana et al. [18] confirmed this finding at scale, showing that multiple LLM families cooperate at rates significantly above human baselines. Fan et al. [19] provided the most systematic evaluation to date, testing LLMs across dictator games, ultimatum games, and public goods games, concluding that while LLMs approximate human-like behavior, they lack consistent strategic reasoning.

Piatti et al. [20] represent the closest precedent to our work, studying cooperation emergence in LLM societies under social dilemmas. However, their agents update strategies through LLM-generated reflections without evolutionary grounding, and the analysis does not include network structure or formal imitation dynamics. De Curtò and de Zarzà [24] demonstrated that LLM-generated influence narratives can shift cooperation rates in networked multi-agent populations and introduced preliminary EC-theory update rules coupling LLM output with evolutionary dynamics, establishing the feasibility of theory-grounded behavioral steering. However, the strategy update lacked the full three-component decomposition (Fermi social learning, degree-normalized reinforcement, trust-weighted advisory), and the deliberation architecture relied on a unique call LLM inference without internal verification.

D. POSITIONING OF CGS

Table 1 summarizes the positioning of CGS relative to the most relevant prior work. The distinguishing features of our approach are: (1) a formally grounded EC-theory update rule that combines Fermi social learning with LLM advisory rather than relying on either mechanism alone; (2) a structured Solver–Critic–Aggregator deliberation pipeline with deterministic fusion, rather than a one-call inference or unstructured multi-agent debate; (3) episodic memory that feeds both the evolutionary update (through payoff and cooperation history) and the LLM prompt (through natural-language summaries); and (4) systematic evaluation across network topologies with ablation studies isolating each architectural component.

III. METHOD

We present the CGS architecture in four parts: the interaction network and game substrate (Section III-A), the agent model with episodic memory (Section III-B), the Solver–Critic–Aggregator deliberation pipeline (Section III-C), and the EC-theory strategy update rule (Section III-D). Figure 1 provides an overview.

A. INTERACTION NETWORK AND GAME SUBSTRATE

Agents are placed on an undirected graph $G = (V, E)$ with $|V| = n$. We consider three canonical topologies: (1) Watts–Strogatz small-world [26] with mean degree $k = 6$ and rewiring probability $p = 0.12$, exhibiting high clustering and short path lengths; (2) Barabási–Albert scale-free [27] with preferential attachment parameter $m = 3$, producing a power-law degree distribution; and (3) Erdös–Rényi random graphs [28] with connection probability $p = 0.12$, serving as a null-model baseline. All graphs enforce connectivity: if the generated graph contains multiple components, bridge edges are added between a random node in the largest component and a random node in each smaller component.

The game substrate is the standard one-shot Prisoner's Dilemma (PD) with payoff matrix:

$$\begin{pmatrix} R & S \\ T & P \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 5 & 1 \end{pmatrix}, \tag{1}$$

where R is mutual cooperation reward, T is temptation to defect, S is the sucker's payoff, and P is mutual defection punishment, satisfying $T > R > P > S$ and $2R > T + S$. At each time step t, every agent simultaneously chooses an action $a_o(t) \in \{C, D\}$, and receives a payoff accumulated over all interactions with its neighbors:

$$\pi_o(t) = \sum_{z \in \mathcal{N}(o)} u\big(a_o(t),\, a_z(t)\big), \tag{2}$$

where $\mathcal{N}(o)$ denotes the neighbors of agent o in G and $u(\cdot, \cdot)$ returns the appropriate entry from (1).

B. AGENT MODEL AND EPISODIC MEMORY

Each agent o is characterized by a *strategy* $\sigma_o(t) \in [0, 1]$ representing its cooperation propensity, and a set of heterogeneous parameters drawn independently at initialization:

TABLE 1: Comparison of CGS with related approaches. EC = Evolutionary Cooperation theory; SCA = Solver–Critic–Aggregator.

Approach	LLM agents	EC-theory update	Deliberation	Episodic memory	Network topology	Ablations
Park et al. [11]	✓	—	—	✓	—	—
Akata et al. [17]	✓	—	—	Partial	—	—
Piatti et al. [20]	✓	—	—	✓	—	Limited
Fan et al. [19]	✓	—	—	—	—	—
de Curtò & de Zarzà [24]	✓	Preliminary	One-call	—	✓	—
Santos & Pacheco [5]	—	✓	—	—	✓	—
Szabó & Fáth [3]	—	✓	—	—	✓	—
CGS (ours)	✓	✓	SCA pipeline	✓	✓	Full

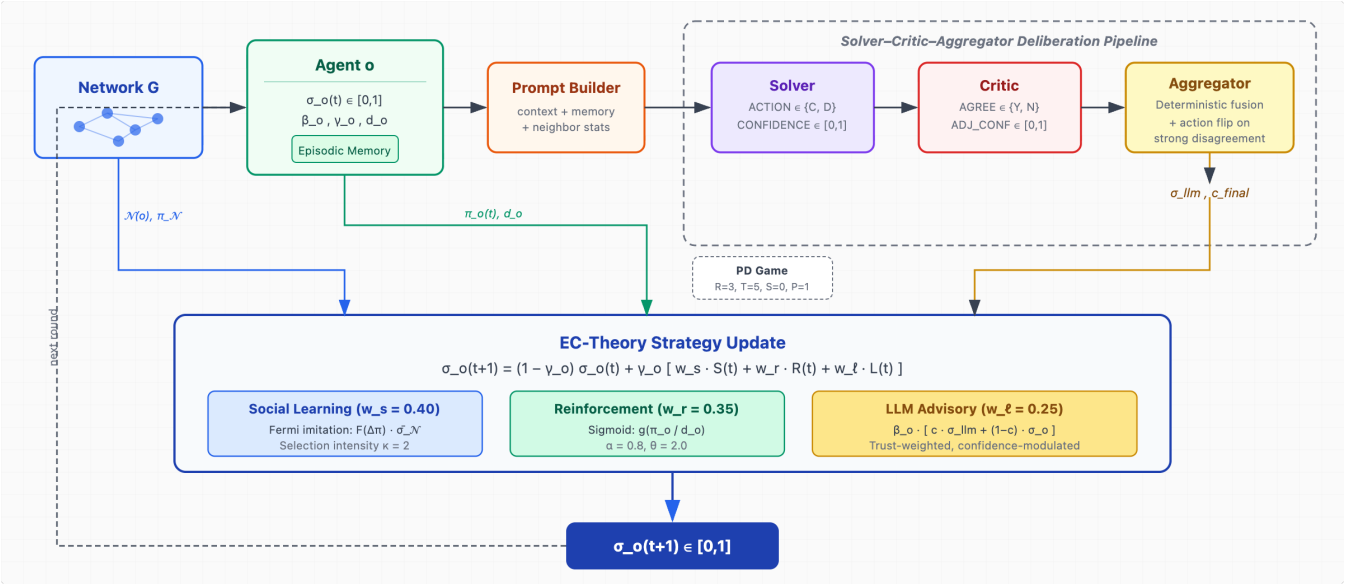

FIGURE 1: High-level architecture of the CGS framework. Each agent's strategy update combines three components: Fermi-function social learning from neighbors, sigmoid individual reinforcement, and a β-weighted LLM advisory signal produced by the Solver–Critic–Aggregator deliberation pipeline. The LLM receives a memory-informed prompt summarizing the agent's interaction history, network position, and neighbor cooperation statistics.

- **Learning rate** $\gamma_o \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$, clipped to $[0.01, 0.5]$, controlling the speed of strategy adaptation.
- **LLM trust** $\beta_o \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$, clipped to $[0.01, 0.99]$, weighting the influence of the LLM advisory signal relative to the agent's own propensity.
- **Prosocial propensity** $\sigma_o(0) \sim \mathcal{N}(\mu_p, \sigma_p^2)$, clipped to $[0.01, 0.99]$, serving as the initial strategy.

Each agent maintains an *episodic memory buffer* $\mathcal{M}_o$ of bounded horizon H, storing the H most recent interaction records. Each record $m_o^{(t)} = (t, a_o, \pi_o, r_{\mathcal{N}})$ captures the time step, action taken, payoff received, and fraction of cooperating neighbors. The memory is summarized into a natural-language string for the LLM prompt through a deterministic function that computes aggregate statistics:

$$\text{summary}(\mathcal{M}_o) = \big(\text{depth},\ \bar{a}_o^C,\ \bar{\pi}_o,\ \bar{r}_{\mathcal{N}}^{\text{recent}},\ \text{last-3 episodes}\big), \tag{3}$$

where $\bar{a}_o^C$ is the agent's historical cooperation rate, $\bar{\pi}_o$ its average payoff, and $\bar{r}_{\mathcal{N}}^{\text{recent}}$ the mean neighbor cooperation over the five most recent steps. The last three individual episode records are included verbatim to provide temporally grounded detail.

Action selection. At each time step, agent o's action is determined stochastically from a mixed cooperation probability that blends the agent's current propensity with the LLM's recommendation:

$$p_C^{(o)} = (1 - \beta_o)\,\sigma_o(t) + \beta_o\big[c \cdot \sigma_{\text{llm}} + (1 - c)\,\sigma_o(t)\big], \tag{4}$$

where $c \in [0, 1]$ is the confidence returned by the deliberation pipeline, $\sigma_{\text{llm}} \in \{0, 1\}$ encodes the recommended action (C = Cooperate $= 1$, D = Defect $= 0$), and $a_o(t) = C$ with probability $p_C^{(o)}$, D otherwise. This formulation ensures that agents with low LLM trust ($\beta_o \to 0$) rely primarily on their propensity, while those with high trust are more responsive to the LLM signal, modulated by the pipeline's own confidence.

C. SOLVER–CRITIC–AGGREGATOR DELIBERATION

The LLM advisory signal is produced by a three-stage deliberation pipeline operating on an underlying language model

(Llama-3.3-70B-Instruct [10] served via the Nebius AI Studio API):

Stage 1: Solver. The solver receives a memory-informed prompt encoding the agent's full decision context, time step, network topology, degree, current cooperation propensity, neighbor cooperation statistics from the previous round, and the episodic memory summary (3). It is instructed to output a structured response: ACTION $\in \{C, D\}$, CONFIDENCE $\in [0, 1]$, and a brief reason. The system prompt constrains the solver to reason about long-run payoff maximization in the repeated PD.

Stage 2: Critic. The critic receives the solver's proposal (action and confidence) together with the full agent context, and is instructed to evaluate strategic soundness. It outputs: AGREE $\in \{\text{YES}, \text{NO}\}$, ADJUSTED_CONF $\in [0, 1]$, and a brief note. The critic's role is adversarial verification: it may lower confidence when the solver's reasoning is inconsistent with the agent's history or network position, or reject the proposal entirely.

Stage 3: Aggregator. A deterministic (non-learned) aggregator fuses the solver and critic signals:

$$c_{\text{final}} = \begin{cases} 0.7\, c_s + 0.3\, c_k & \text{if AGREE} = \text{YES}, \\ 0.4\, c_s + 0.6\, c_k & \text{if AGREE} = \text{NO}, \end{cases} \tag{5}$$

where c_s and c_k are the solver and critic confidences, respectively. On *strong disagreement*, defined as the critic disagreeing with $c_k < 0.3$ while the solver has $c_s > 0.7$, the aggregator *flips* the proposed action and sets $c_{\text{final}} = c_k$. This mechanism prevents high-confidence but strategically unsound recommendations from propagating to the evolutionary update.

The asymmetric weighting reflects the critic's adversarial role: on disagreement, the critic carries the majority signal to prevent high-confidence but strategically unsound proposals from propagating; formal decision-theoretic grounding of this rule remains an avenue for future work.

All LLM responses are cached via SHA-1 hashing of the concatenated system prompt and user prompt, with exponential-backoff retry (up to 5 attempts) for API resilience. The structured output format (key–value pairs) is parsed via regular expressions with fallback defaults (ACTION $= C$, CONFIDENCE $= 0.5$) to ensure robustness against malformed LLM outputs.

D. EC-THEORY STRATEGY UPDATE

After payoffs are computed and actions recorded in memory, each agent updates its strategy through a three-component rule grounded in evolutionary cooperation theory [2]:

$$\sigma_o(t+1) = (1-\gamma_o)\,\sigma_o(t) + \gamma_o \Big[\underbrace{w_s \cdot S_o(t)}_{\text{social}} + \underbrace{w_r \cdot R_o(t)}_{\text{reinforcement}} + \underbrace{w_\ell \cdot L_o(t)}_{\text{LLM advisory}}\Big], \tag{6}$$

where γ_o is the agent's learning rate, $w_s = 0.40$, $w_r = 0.35$, and $w_\ell = 0.25$ are the component weights, and each component is defined as follows.

Social learning (Fermi imitation). *Intuition: if my neighbors are doing better than me, I shift toward their strategies; if I am doing better, I stay the course.* The social component implements the Fermi imitation rule [6], [7]. Let $\bar{\pi}_{\mathcal{N}(o)}$ and $\bar{\sigma}_{\mathcal{N}(o)}$ denote the mean payoff and mean strategy of o's neighbors. The imitation probability is:

$$F(\Delta\pi_o) = \frac{1}{1 + \exp\big(-\kappa\,(\bar{\pi}_{\mathcal{N}(o)} - \pi_o)\big)}, \tag{7}$$

with selection intensity $\kappa = 2$. The social target interpolates between the neighbor mean and the agent's own strategy, weighted by the imitation probability:

$$S_o(t) = F(\Delta\pi_o)\,\bar{\sigma}_{\mathcal{N}(o)} + \big(1 - F(\Delta\pi_o)\big)\,\sigma_o(t). \tag{8}$$

When neighbors outperform agent o, F is large and the agent shifts toward the neighbor consensus; when o outperforms its neighbors, it retains its own strategy.

Individual reinforcement. *Intuition: if my payoff per interaction is high, I reinforce cooperation; if it is low, I pull back, regardless of what my neighbors are doing.* The reinforcement component maps the agent's degree-normalized payoff to a cooperation target through a sigmoid:

$$R_o(t) = \frac{1}{1 + \exp\big(-\alpha\,(\pi_o/d_o - \theta)\big)}, \tag{9}$$

where $d_o = \deg(o)$ is the agent's degree, $\alpha = 0.8$ controls sensitivity, and $\theta = 2.0$ is the inflection point. Normalization by degree ensures that well-connected agents (who accumulate higher absolute payoffs) are not spuriously driven toward full cooperation; instead, the relevant signal is payoff *per interaction*.

LLM advisory. *Intuition: the deliberation pipeline's recommendation shifts my strategy in proportion to how much I trust it and how confident the pipeline is; low trust or low confidence leaves me unchanged.* The LLM component integrates the deliberation output, modulated by the agent's heterogeneous trust parameter β_o:

$$L_o(t) = \beta_o \big[c \cdot \sigma_{\text{llm}} + (1-c)\,\sigma_o(t)\big] + (1-\beta_o)\,\sigma_o(t), \tag{10}$$

where c is the final aggregated confidence and $\sigma_{\text{llm}} \in \{0, 1\}$ encodes the recommended action. When confidence is low, the LLM signal degrades gracefully to the agent's own strategy; when trust is low, the entire component reduces to $\sigma_o(t)$.

[2]The CGS social-learning component is a pairwise Fermi imitation rule, not a replicator equation; we use "evolutionary cooperation theory" to denote the Fermi-imitation-plus-reinforcement family and reserve "replicator" for the explicit replicator-only baseline of Section V-K.

TABLE 2: Notation. Population-level metrics are reported as means over the final 10 steps unless stated otherwise.

Symbol	Meaning
$\sigma_o(t)$	cooperation propensity of agent o
γ_o	learning rate (adaptation speed)
β_o	LLM trust (advisory weight)
μ_p	mean initial prosocial propensity
κ	Fermi selection intensity
α	reinforcement sigmoid sensitivity
θ	reinforcement inflection point
w_s, w_r, w_ℓ	social / reinforcement / advisory weights
S_o, R_o, L_o	social, reinforcement, advisory targets
c	aggregated deliberation confidence
σ_{llm}	LLM recommended action (C=1, D=0)
$\pi_o(t)$	accumulated payoff of agent o
$\rho_C(t)$	population cooperation rate
$\mathcal{G}(t)$	payoff Gini coefficient
$\bar{c}(t)$	mean deliberation confidence
H	episodic memory horizon
f_{shock}	shocked fraction of agents
n, T	population size, horizon

The final strategy is clipped to $[0, 1]$ after the update. The cumulative payoff $\Pi_o(t) = \sum_{\tau=0}^{t} \pi_o(\tau)$ is also maintained for analysis. See Table 2 for notation.

E. NARRATIVE-ONLY BASELINE

To isolate the contribution of EC-theory grounding, we define a *narrative-only* baseline that replaces the three-component update (6) with a direct application of the LLM signal:

$$\sigma_o(t+1) = c \cdot \sigma_{\text{llm}} + (1 - c)\, \sigma_o(t), \tag{11}$$

where c and σ_{llm} are as above. This removes social learning (no Fermi imitation), individual reinforcement (no payoff feedback), and trust heterogeneity (no β_o modulation), leaving the LLM as the sole driver of strategy evolution. The deliberation pipeline remains active, ensuring that any performance difference is attributable to the update rule rather than the quality of LLM advice.

F. ADVERSARIAL SHOCK

To test resilience and recovery capacity, we introduce an exogenous adversarial shock at a specified time step t_{shock}. At this step, a fraction f_{shock} of agents is selected uniformly at random and their strategies are set to $\sigma_o = 0$ (pure defection). The population must then recover through the endogenous dynamics of the update rule. Recovery time is measured as the number of steps after t_{shock} until the cooperation rate returns to within one standard deviation of its pre-shock mean.

G. METRICS

We track the following quantities at each time step:

- **Cooperation rate**: $\rho_C(t) = \frac{1}{n} \sum_{o=1}^{n} \mathbb{1}[a_o(t) = C]$.
- **Average payoff**: $\bar{\pi}(t) = \frac{1}{n} \sum_{o=1}^{n} \pi_o(t)$.
- **Payoff inequality**: the Gini coefficient $\mathcal{G}(t)$ of the payoff distribution $\{\pi_o(t)\}_{o=1}^{n}$.
- **Strategy distribution**: mean, standard deviation, minimum, and maximum of $\{\sigma_o(t)\}_{o=1}^{n}$.

TABLE 3: Agent parameter distributions. All values are clipped to the specified range after sampling.

Parameter	Symbol	Mean	Std	Range
Learning rate	γ_o	0.15	0.07	$[0.01, 0.50]$
LLM trust	β_o	0.35	0.15	$[0.01, 0.99]$
Prosocial propensity	$\sigma_o(0)$	0.55	0.15	$[0.01, 0.99]$

- **LLM confidence**: mean aggregated confidence $\bar{c}(t)$ across agents.
- **Rolling stability**: the standard deviation of ρ_C over a sliding window of width $w = 8$, with convergence declared when this falls below a threshold $\epsilon = 0.02$.

IV. EXPERIMENTAL SETUP

All experiments are executed from a jupyter notebook using the Nebius AI Studio API for open-weight model serving, with `meta-llama/Llama-3.3-70B-Instruct` as the primary underlying language model; the multi-model sweep of Section V-G additionally evaluates `DeepSeek-V3.2`, `Qwen3-235B-A22B-Instruct-2507`, and `Hermes-4-70B` under identical conditions.

A. AGENT POPULATION

Each experiment uses $n = 40$ agents with heterogeneous parameters drawn from truncated Gaussians (Table 3). All agents share a common memory horizon of $H = 20$ steps and are initialized on the same random seed ($s = 0$) across conditions, ensuring identical initial populations for fair comparison.

B. NETWORK CONFIGURATIONS

Three topologies are instantiated with $n = 40$ nodes and shared seed $s = 0$:

- **Watts–Strogatz small-world** [26]: each node initially connected to $k = 6$ nearest neighbors, with each edge rewired with probability $p = 0.12$. This produces high clustering and short average path lengths.
- **Barabási–Albert scale-free** [27]: grown by preferential attachment with $m = 3$ new edges per incoming node. The resulting power-law degree distribution concentrates interactions around a few high-degree hubs.
- **Erdös–Rényi random** [28]: each edge exists independently with probability $p = 0.12$. This yields a homogeneous degree distribution and serves as a null-model baseline.

All generated graphs are checked for connectivity; if isolated components exist, bridge edges are added between the largest component and each smaller component to ensure a connected network. Graph statistics (mean degree, clustering coefficient, average path length, degree standard deviation) are recorded for each topology.

C. SIMULATION PROTOCOL

Each experiment runs for $T = 50$ time steps. At every step, the following sequence is executed:

TABLE 4: Experimental conditions. All runs use $n = 40$, $T = 50$, seed $= 0$, shock at $t = 30$ with $f_{\text{shock}} = 0.15$.

Label	Topology	EC update	Critic	Purpose
Experiment 1: Multi-topology sweep				
CGS-EC-SW	small_world	✓	✓	Main result
CGS-EC-SF	scale_free	✓	✓	Topology effect
CGS-EC-ER	erdos_renyi	✓	✓	Structural baseline
Experiment 2: Narrative-only ablation				
Narr-only	small_world	—	✓	EC contribution
Experiment 3: No-critic ablation				
EC-no-critic	small_world	✓	—	Critic contribution

1) **Shock check**: if $t = t_{\text{shock}} = 30$, a fraction $f_{\text{shock}} = 0.15$ of agents (6 out of 40) is selected uniformly at random and their strategies are set to $\sigma_o = 0$.
2) **Decision phase**: each agent receives a memory-informed prompt and produces an action via the Solver–Critic–Aggregator pipeline (or solver-only, depending on the condition). Actions are recorded.
3) **Interaction phase**: all edges are evaluated under the PD payoff matrix (1), and payoffs are accumulated per agent.
4) **Neighbor statistics**: for each agent, the mean neighbor strategy $\bar{\sigma}_{\mathcal{N}(o)}$, mean neighbor payoff $\bar{\pi}_{\mathcal{N}(o)}$, and neighbor cooperation rate are computed.
5) **Memory update**: each agent stores its interaction record in the episodic memory buffer.
6) **Strategy update**: agents update $\sigma_o(t+1)$ via the EC-theory rule (6) (or the narrative-only rule (11), depending on the condition).
7) **Logging**: population-level metrics (cooperation rate, payoff statistics, Gini coefficient, strategy distribution, mean LLM confidence) are recorded; strategy snapshots are saved every 5 steps.

D. EXPERIMENTAL CONDITIONS

The full evaluation comprises five experimental runs organized into three groups (Table 4):

Experiment 1 (multi-topology sweep) evaluates the full CGS architecture, EC-theory update with Solver–Critic–Aggregator deliberation, on all three topologies. This group addresses the question of how network structure modulates cooperation emergence, payoff levels, and inequality under theory-grounded LLM societies.

Experiment 2 (narrative-only ablation) removes EC-theory grounding entirely. The strategy update is replaced by the narrative-only rule (11), which applies the LLM's recommendation directly without Fermi social learning, individual reinforcement, or trust heterogeneity. The deliberation pipeline (including the critic) remains active, ensuring that any observed difference is attributable solely to the update rule. This ablation is run on the small-world topology to enable direct comparison with CGS-EC-SW. It constitutes the paper's core claim: that EC grounding is essential, not optional.

Experiment 3 (no-critic ablation) retains the full EC-theory update but bypasses the critic stage: the solver's action and confidence are passed directly to the aggregator without adversarial verification. This isolates the contribution of the deliberation pipeline's depth: does internal verification (solver + critic) produce meaningfully better advisory signals than one-call inference (solver only)? This condition is also run on small-world for controlled comparison.

E. EVALUATION CRITERIA

We evaluate performance along five dimensions:

1) **Cooperation level**: the final cooperation rate ρ_C, averaged over the last 10 steps to smooth fluctuations, and the peak cooperation rate achieved during the run.
2) **Payoff efficiency**: the average payoff $\bar{\pi}$ in the final window, indicating how effectively the population converts cooperation into collective welfare.
3) **Equity**: the Gini coefficient $\mathcal{G}$ of the payoff distribution, with lower values indicating more equitable outcomes.
4) **Resilience**: the number of steps required to recover from the adversarial shock at $t = 30$, defined as the first post-shock step where $\rho_C(t) \geq \rho_C(t_{\text{shock}} - 1) - 0.05$.
5) **Stability**: whether the cooperation rate converges, measured by the rolling standard deviation falling below $\epsilon = 0.02$ over the final 10 steps, and the overall volatility pattern visible in the rolling stability time series.

F. REPRODUCIBILITY AND ADDITIONAL CONDITIONS

All experiments use a fixed random seed ($s = 0$) controlling both network generation and agent parameter initialisation. LLM responses are deterministically cached via content hashing, making the full pipeline reproducible across runs given the same API model version. Each run produces: (1) a per-timestep CSV with all population-level metrics, (2) a per-agent CSV with degree, β_o, γ_o, initial prosociality, final strategy, and cumulative payoff, (3) compressed strategy snapshots every 5 steps, and (4) a JSON configuration file recording all hyperparameters, graph statistics, and LLM API usage.

Beyond the five primary conditions, we conduct nine additional experiments to characterise the generality of the reported phenomena.

Multi-seed replication. Experiment 1 (CGS-EC vs. Narrative-only on the Watts–Strogatz topology) is repeated across *ten* independent random seeds ($s \in \{0, \ldots, 9\}$), each producing a fresh graph realisation and fresh agent-parameter draws from Table 3. Bootstrap 95 % confidence intervals (5 000 resamples) and pairwise Mann–Whitney U tests with Bonferroni correction ($m = 5$ metrics) assess whether the rank ordering between conditions is preserved.

Population scaling. CGS-EC is run on the small-world topology at $n \in \{40, 60, 80\}$ (seed $s = 0, T = 50$) to examine how cooperation dynamics change with population size.

Multi-model sweep. CGS-EC (small-world, $n = 40$, $s = 0$, $T = 50$) is re-executed with four LLMs served via the Nebius AI Studio API: `Llama-3.3-70B-Instruct` (Llama-70B, primary model), `DeepSeek-V3.2` (DeepSeek-V3), `Qwen3-235B-A22B-Instruct-2507` (Qwen3-235B), and `Hermes-4-70B` (Hermes-4-70B), to characterise the dependence of cooperation dynamics on the underlying language model.

GNN structural baselines. Three graph-neural-network-inspired update rules that replace LLM calls entirely with analytically defined message-passing aggregations establish a structural performance baseline: *GNN-Mean* (DeGroot-style mean consensus, $\sigma_i \leftarrow (1-\gamma_i)\sigma_i + \gamma_i \bar{\sigma}_{\mathcal{N}(i)}$), *GNN-PayW* (Fermi-weighted neighbour average, where each neighbour's strategy is weighted by $F(\pi_j - \pi_i)$ from (7)), and *GNN-2L* (a two-layer readout that mixes the direct-neighbour mean with the two-hop neighbourhood mean in a $0.6\,/\,0.4$ ratio). All three share the PD payoff substrate, adversarial shock, and agent initialisation of the primary conditions.

OAT sensitivity analysis. Six hyperparameters are swept one-at-a-time over five levels each (30 total runs, small-world, $n = 40$, seed 0): Fermi selection intensity $\kappa \in \{0.5, 1.0, 2.0, 3.0, 4.0\}$, reinforcement sensitivity $\alpha \in \{0.2, 0.5, 0.8, 1.2, 1.8\}$, mean LLM trust $\bar{\beta} \in \{0.10, 0.20, 0.35, 0.50, 0.65\}$, LLM advisory weight $w_\ell \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$, mean learning rate $\bar{\gamma} \in \{0.05, 0.10, 0.15, 0.25, 0.40\}$, and initial prosocial mean $\mu_p \in \{0.20, 0.35, 0.55, 0.70, 0.85\}$. All other parameters are held at their baseline values. The normalised sensitivity index (NSI) $= \max_k |\Delta y/y_0|/|\Delta x/x_0|$ quantifies the output elasticity per parameter. Temperature robustness is evaluated by re-running CGS-EC at $T_s \in \{0.0, 0.2, 0.5, 0.8\}$ (small-world, $n = 40$, seed 0).

Alternative game substrates. CGS-EC is re-run on the small-world topology with two further 2×2 games beyond the PD: Snowdrift ($T > R > S > P$; R=3, T=4, S=1, P=0) and Stag Hunt ($R > T > P > S$; R=4, T=3, P=2, S=0), under identical agent initialisation and shock protocol.

Non-LLM learning and imitation baselines. Five baselines that replace the LLM advisory entirely are evaluated on the primary PD/small-world setting: tabular ϵ-greedy Q-learning over a binned neighbour-cooperation state (α=0.2, γ=0.9, ϵ=0.1), aspiration learning (satisficing switch on normalised payoff), pure Fermi imitation, a discrete graph replicator, and a uniform random control.

Computational cost. Wall-clock time, API-call counts and cache statistics are recorded for CGS-EC at $n \in \{20, 40, 60\}$, and for the no-critic variant at $n = 40$, to isolate the cost of the critic stage.

Failure-mode and equilibrium probes. The shock fraction is swept over $f_{\text{shock}} \in \{0.05, 0.15, 0.30, 0.50, 0.75\}$ (single shock) and a persistent regime (five 15% shocks at $t \in \{20, 25, 30, 35, 40\}$). A separate no-shock sweep over $\mu_p \in \{0.10, \ldots, 0.85\}$ at $T = 80$ characterises the empirical attractor structure.

Complete solver and critic system prompts, a worked agent-decision prompt, the aggregator rule set, all hyperparameter values and the parameter-selection methodology are provided verbatim in the article GitHub repository (file `prompts_supplement.txt`).

V. RESULTS

We present results in thirteen parts: the main CGS dynamics on individual topologies (Section V-A), agent-level heterogeneity analysis (Section V-B), cross-topology comparison (Section V-C), ablation studies (Section V-D), multi-seed replication (Section V-E), population scaling (Section V-F), multi-model analysis (Section V-G), GNN structural baselines (Section V-H), one-at-a-time parameter sensitivity and temperature robustness (Section V-I), generalisation to other social dilemmas (Section V-J), non-LLM learning and imitation baselines (Section V-K), empirical equilibrium structure (Section V-L), and failure modes and the resilience boundary (Section V-M). All quantitative summaries report means over the last 10 time steps unless otherwise noted.

A. CGS DYNAMICS BY TOPOLOGY

1) Small-World Network

Figure 2 presents the four-panel dynamics for the CGS system on the Watts–Strogatz small-world network ($n = 40$, $k = 6$, $p = 0.12$; clustering coefficient $= 0.475$, average path length $= 2.61$). Cooperation rises rapidly from the initial mean of 0.55 to a peak of 0.80 at $t = 15$, driven by the combined effect of Fermi social learning, which amplifies cooperative strategies when they yield higher payoffs, and the LLM advisory signal, which consistently recommends cooperation with increasing confidence over the first 10 steps (Figure 3b, rising from 0.53 to 0.75). The pre-shock phase ($t = 0$–29) sustains cooperation in the 0.60–0.80 range with an average of 0.68 over $t = 25$–29.

The adversarial shock at $t = 30$ forces 15% of agents to pure defection, producing an immediate drop to 0.60 and a subsequent trough of 0.45 at $t = 34$ as defection propagates through the network. Recovery occurs within 8 steps (cooperation returns to 0.63 at $t = 38$), demonstrating the resilience of the EC-grounded update: Fermi imitation allows remaining cooperators to pull shocked agents back toward cooperative strategies when their payoffs recover. The post-shock final cooperation rate stabilizes at 0.595 ± 0.084.

Average payoff tracks cooperation closely, peaking at 16.73 and settling at 14.48 in the final window. Payoff inequality remains moderate (Gini $= 0.236$, range $[0.173, 0.319]$), reflecting the homogeneous degree distribution of the small-world topology (degree std $= 0.87$). The strategy distribution panel (Figure 2d) shows the population-level propensity converging to a mean of 0.53 with a moderate spread (±1 std band visible), indicating sustained heterogeneity rather than unanimous convergence.

The rolling stability analysis (Figure 3a) reveals that the cooperation rate does not formally converge, the rolling standard deviation remains above the $\epsilon = 0.02$ threshold throughout, but exhibits a characteristic pattern: periods of decreasing

volatility ($t = 10$–25) interrupted by the shock, followed by a gradual return to moderate fluctuation. Deliberation confidence (Figure 3b) stabilizes around 0.70–0.75 after an initial ramp-up phase, consistent with the LLM developing increasingly confident recommendations as episodic memory accumulates.

2) Scale-Free Network

Figure 4 shows the dynamics on the Barabási–Albert scale-free network ($m = 3$; clustering coefficient $= 0.221$, average path length $= 2.17$, degree std $= 3.81$). The pre-shock cooperation trajectory is qualitatively similar to the small-world case, reaching the same peak of 0.80 at $t = 15$ and maintaining 0.685 over $t = 25$–29. However, the post-shock behavior diverges sharply: cooperation drops more severely, reaching a minimum of 0.35 (versus 0.45 on small-world), and the final cooperation rate is substantially lower at 0.460 ± 0.080. Although the formal recovery metric indicates 1 step (cooperation briefly touches 0.65 at $t = 31$ before declining further), the system does not sustain recovery, instead settling into a lower-cooperation regime.

The most striking difference is in payoff inequality. The Gini coefficient on the scale-free topology averages 0.420 in the final window, nearly double the small-world value of 0.236, and ranges from 0.308 to 0.525. This reflects the power-law degree distribution: the highest-degree hub ($d = 17$) accumulates a cumulative payoff of 2,051 over 50 steps, while peripheral agents ($d \leq 3$) average only 351, a $3.89\times$ disparity. Average payoff is correspondingly lower (11.79 versus 14.48), as defection among peripheral agents reduces the collective surplus.

The strategy distribution (Figure 4d) shows wider post-shock spread compared to small-world, with the ± 1 std envelope expanding after $t = 30$. This reflects the amplified heterogeneity: hub agents and peripheral agents respond differently to the shock, with hubs recovering more slowly (final strategy 0.454) than periphery agents (0.480), likely because hubs interact with more defecting neighbors post-shock.

Deliberation confidence on the scale-free network drops more pronouncedly after the shock (Figure 5b), declining from 0.72 to 0.59 before partially recovering, indicating that the LLM recognizes increased strategic uncertainty in the disrupted network. The rolling stability analysis (Figure 5a) shows the highest post-shock volatility of any topology, peaking at 0.13 around $t = 35$, nearly double the small-world peak of 0.085.

3) Erdös–Rényi Random Network

Figure 6 presents the dynamics on the Erdös–Rényi random graph ($p = 0.12$; clustering coefficient $= 0.205$, average path length $= 2.41$, degree std $= 2.31$). The Erdös–Rényi topology occupies an intermediate position between the small-world and scale-free networks on most metrics. Peak cooperation reaches 0.775 at $t = 28$, slightly below the 0.80 achieved by both other topologies, and the pre-shock average (0.68 over $t = 25$–29) is comparable. The post-shock minimum of 0.425 is intermediate between small-world (0.45) and scale-free (0.35), but recovery is the slowest of the three topologies at 13 steps, with cooperation returning to within tolerance at $t = 43$. The final cooperation rate stabilizes at 0.548 ± 0.086.

Notably, the Erdös–Rényi network is the only topology where the system approaches formal convergence: the rolling standard deviation of cooperation rate drops below the $\epsilon = 0.02$ threshold (reaching 0.009) during $t = 20$–27 (Figure 7a), before the shock disrupts the near-equilibrium. This is consistent with the lack of structural heterogeneity: without hubs (scale-free) or high clustering (small-world), the random graph provides a more uniform selective environment where the Fermi imitation dynamics can settle into a quasi-stationary state.

Payoff inequality (Gini $= 0.359$, range $[0.260, 0.424]$) falls between the small-world (0.236) and scale-free (0.420) values, driven by a moderate degree range of $[1, 11]$ (degree std $= 2.31$, between small-world's 0.87 and scale-free's 3.81). The degree–payoff correlation remains very strong ($r = 0.979$, $p < 10^{-4}$), with cumulative payoffs spanning $[133, 1{,}471]$, a $11.1\times$ disparity.

Figure 8 presents the agent-level analysis for the Erdös–Rényi network. The degree–payoff relationship (panel b) is near-linear ($r = 0.979$), similar to scale-free but with a narrower degree range. Initial prosociality shows a weak positive trend with final strategy ($r = 0.251$, $p = 0.119$), consistent with the small-world pattern of partial persistence.

B. AGENT-LEVEL HETEROGENEITY

Figure 9 and Figure 10 present the agent-level analysis for small-world and scale-free topologies, respectively. Three scatter plots decompose the relationships between agent parameters, network position, and outcomes.

Trust vs. final strategy (panel a). On the small-world network, LLM trust β_o shows no significant correlation with final strategy ($r = 0.102$, $p = 0.533$), suggesting that the evolutionary components (Fermi imitation and reinforcement) dominate long-run strategy determination, with the LLM advisory acting as a perturbation rather than a primary driver. The scale-free network exhibits a similar pattern: trust does not predict final cooperation propensity, consistent with the $w_\ell = 0.25$ weight assigned to the LLM component in the update rule (6).

Network position vs. payoff (panel b). This is the panel where topology effects are most pronounced. On the small-world network, degree and cumulative payoff are strongly correlated ($r = 0.806$, $p < 10^{-4}$), but the narrow degree range ($[4, 8]$) limits the absolute disparity: the highest-earning agent accumulates 1,016 versus 472 for the lowest. On the scale-free network, the correlation is near-perfect ($r = 0.992$, $p < 10^{-4}$) and the degree range spans $[1, 17]$, producing a cumulative payoff range of $[110, 2{,}051]$, an $18.6\times$ disparity. This confirms the classical result from evolutionary game theory [5] that scale-free topologies amplify payoff concentration at hubs, and shows that this effect persists even when agents have access to LLM deliberation.

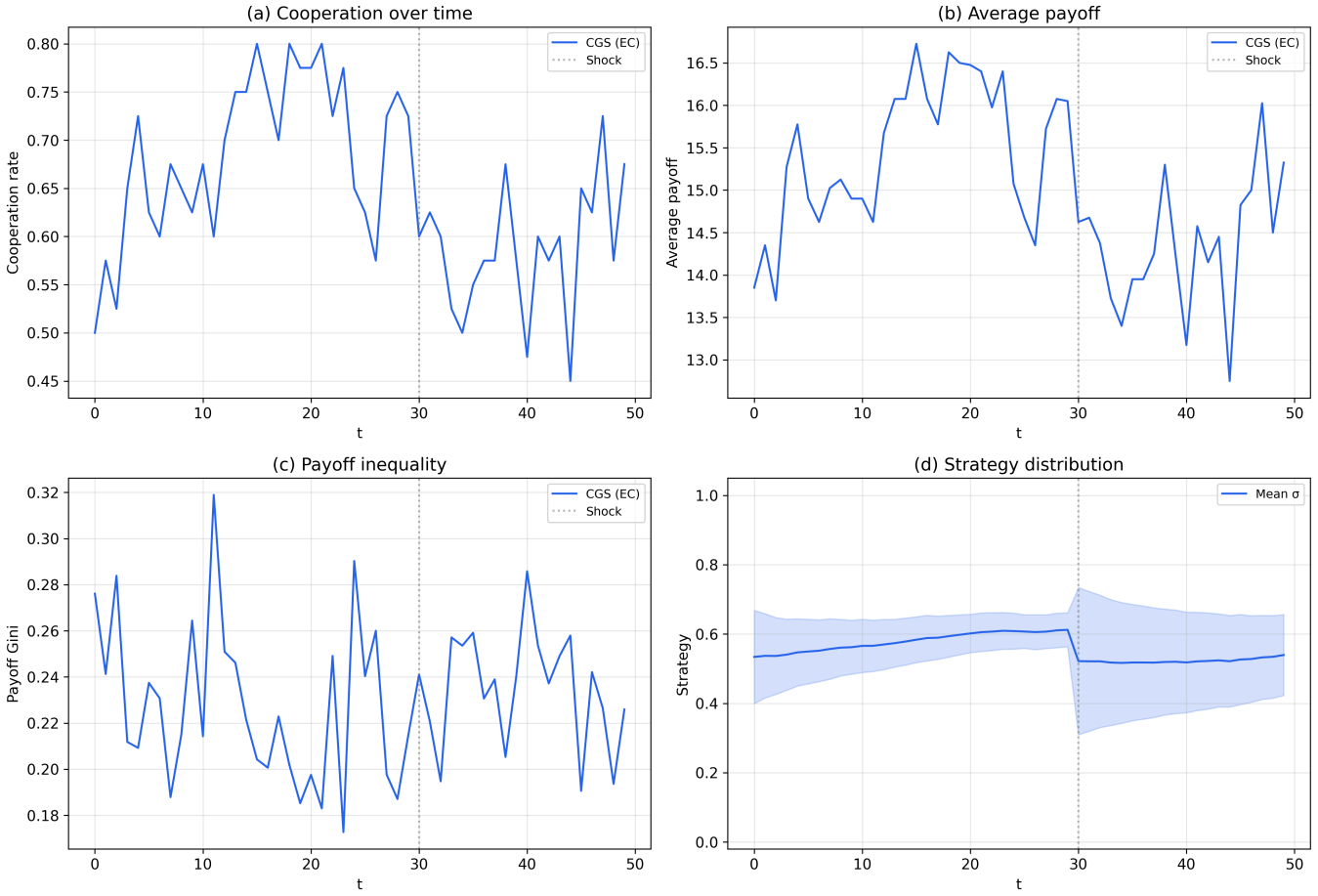

FIGURE 2: CGS dynamics on the small-world network. (a) Cooperation rate over time, showing peak cooperation of 0.80 at $t = 15$ and recovery from the adversarial shock (vertical dashed line at $t = 30$) within 8 steps. (b) Average payoff, peaking at 16.73. (c) Payoff inequality (Gini coefficient), remaining moderate (0.18–0.32). (d) Strategy distribution with ± 1 std envelope, showing sustained heterogeneity.

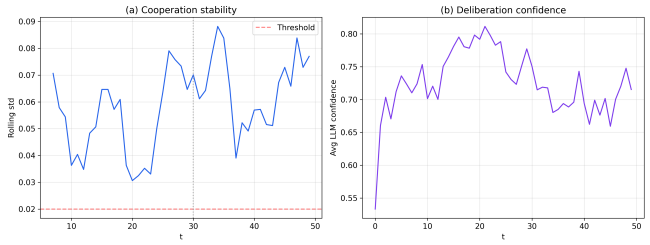

FIGURE 3: Stability analysis for the small-world topology. (a) Rolling standard deviation of cooperation rate (window = 8); the system does not formally converge but exhibits structured fluctuation. (b) Mean LLM deliberation confidence, rising from 0.53 to a plateau around 0.73.

Initial vs. final strategy (panel c). On the small-world network, initial prosociality shows a marginally significant positive correlation with final strategy ($r = 0.296$, $p = 0.064$), indicating partial persistence of initial dispositions through the evolutionary process. On the scale-free network this correlation vanishes ($r = 0.040$, $p = 0.808$), suggesting that the more heterogeneous interaction structure overwrites initial conditions more thoroughly, an effect consistent with the stronger selection pressures in scale-free networks where hubs mediate disproportionate strategy diffusion.

C. CROSS-TOPOLOGY COMPARISON

Table 5 consolidates the key metrics across all five experimental conditions. The topology sweep reveals a clear ordering on most dimensions.

Cooperation. Small-world networks produce the highest sustained cooperation ($\rho_C^{\text{final}} = 0.595$), followed by Erdös–Rényi (0.548) and scale-free (0.460). All three topologies reach comparable peak cooperation (0.775–0.800), indicating that the pre-shock dynamics are topology-insensitive; the divergence emerges primarily in post-shock recovery and long-run stability. The high clustering of small-world networks (0.475 vs. 0.205–0.221 for the others) appears to provide cooperative clusters that resist invasion by defection, consistent with the network reciprocity mechanism identified in evolutionary theory [4].

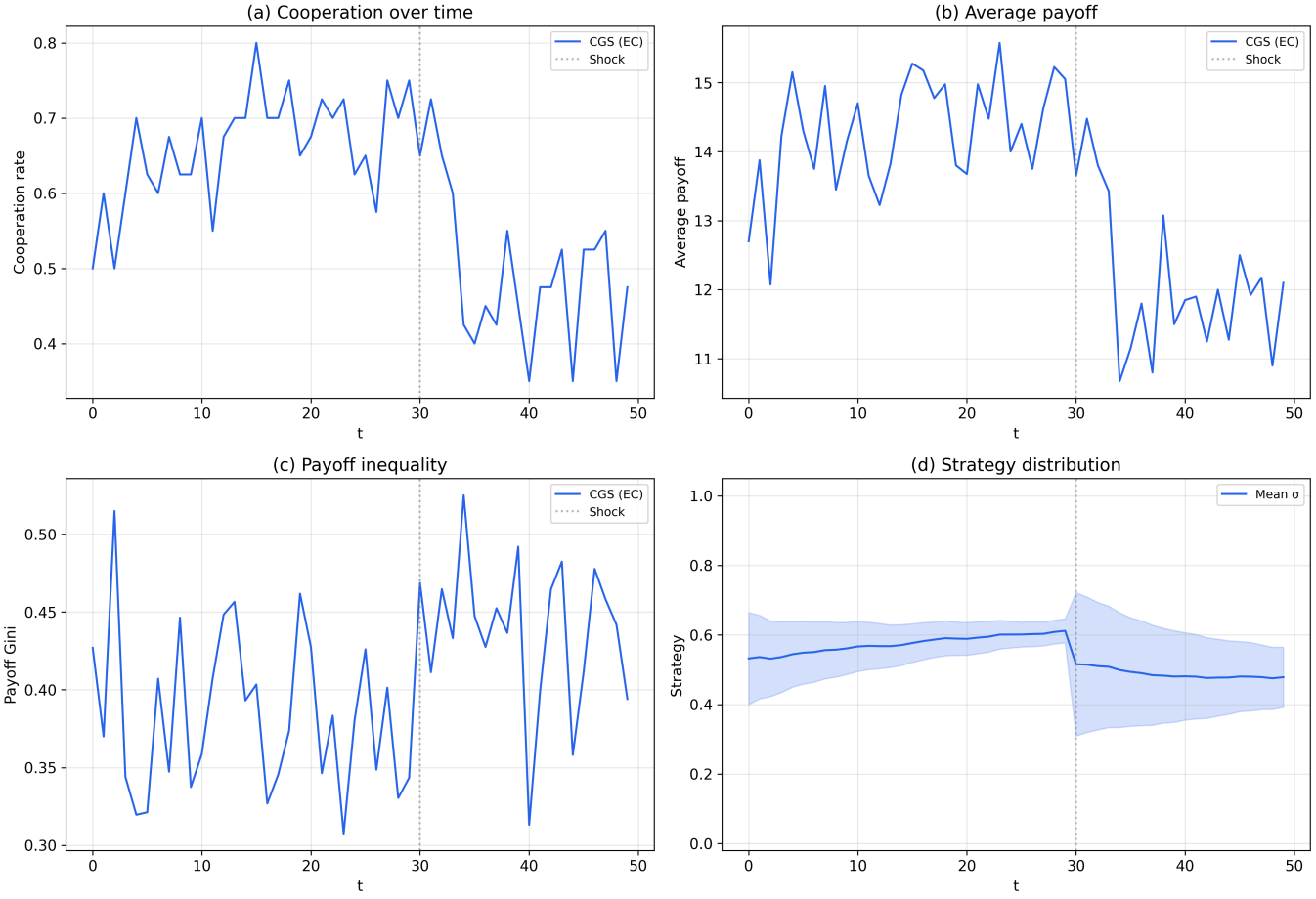

FIGURE 4: CGS dynamics on the scale-free network. (a) Cooperation reaches the same peak of 0.80 but suffers a deeper post-shock decline (minimum 0.35). (b) Average payoff is lower overall (11.79 final) due to hub–periphery asymmetry. (c) Payoff inequality is substantially higher (Gini ≈ 0.42, peaking at 0.53). (d) Strategy distribution shows wider post-shock spread.

TABLE 5: Summary of results across all experimental conditions. ρ_C^{final}: mean cooperation over last 10 steps. ρ_C^{peak}: maximum cooperation. $\bar{\pi}^{\text{final}}$: mean payoff over last 10 steps. $\mathcal{G}^{\text{final}}$: mean Gini over last 10 steps. Recovery: steps to return within 0.05 of pre-shock cooperation.

Condition	**Topology**	**EC**	**Critic**	ρ_C^{peak}	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	**Recovery**
CGS-EC-SW	small_world	✓	✓	0.800	0.595 ± 0.084	14.48	0.236	8 steps
CGS-EC-SF	scale_free	✓	✓	0.800	0.460 ± 0.080	11.79	0.420	1 step†
CGS-EC-ER	erdos_renyi	✓	✓	0.775	0.548 ± 0.086	12.36	0.359	13 steps
Narrative-only	small_world	—	✓	0.550	0.525 ± 0.000	13.62	0.248	0‡
EC-no-critic	small_world	✓	—	0.825	0.568 ± 0.076	14.16	0.239	8 steps

†Brief touch at $t = 31$; cooperation subsequently declines further. ‡No dynamic response to shock.

Payoff and inequality. Payoff efficiency follows the cooperation ordering: small-world (14.48) > Erdös–Rényi (12.36) > scale-free (11.79). However, inequality exhibits a different pattern driven by degree heterogeneity: scale-free ($\mathcal{G} = 0.420$) $\gg$ Erdös–Rényi (0.359) > small-world (0.236). The degree standard deviation is the strongest predictor of inequality (3.81, 2.31, and 0.87 respectively), confirming that structural heterogeneity translates directly into payoff disparity through the accumulated-neighbor-payoff mechanism (2).

Resilience. Recovery time does not follow the cooperation ordering: Erdös–Rényi is slowest (13 steps), small-world intermediate (8 steps), and scale-free nominally fastest (1 step). However, scale-free's rapid "recovery" is misleading, cooperation briefly touches the threshold at $t = 31$ before declining further, reflecting the volatility of hub-mediated dynamics rather than genuine resilience. Small-world's 8-step recovery with sustained post-recovery cooperation represents the most robust resilience pattern.

Figure 11 provides a direct visual overlay of the three

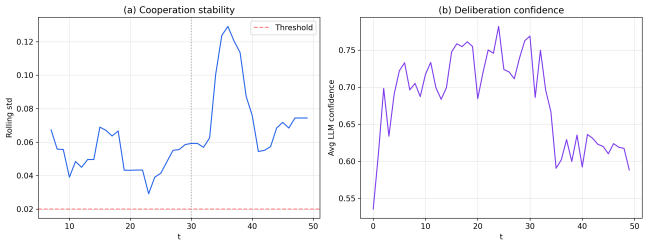

FIGURE 5: Stability analysis for the scale-free topology. (a) Rolling std peaks at 0.13 post-shock, indicating substantially higher volatility than small-world. (b) LLM confidence drops more sharply after the shock, reflecting increased strategic uncertainty.

topologies across cooperation, payoff, and inequality, making the ordering and divergence points clearly visible. The cooperation trajectories (panel a) are nearly indistinguishable before $t = 10$ but progressively separate, with the gap widening sharply after the shock. The inequality panel (panel c) shows the most consistent separation: scale-free is persistently highest, small-world lowest, throughout the entire simulation.

D. ABLATION STUDIES

1) Narrative-Only Ablation: EC Grounding Is Essential

The narrative-only baseline provides the paper's central ablation result. Removing the EC-theory update (6) and replacing it with the pure LLM-driven rule (11) produces qualitatively different dynamics across every metric. Figure 12 overlays the CGS and narrative-only trajectories on the small-world network, making the contrast immediately visible across all four panels.

Cooperation plateaus immediately. The narrative-only condition reaches a peak of 0.550 within the first 5 steps and remains locked at 0.525–0.550 for the entire simulation. The final cooperation rate of 0.525 ± 0.000 (zero variance over the last 10 steps) reflects complete convergence to a static mixed equilibrium. This represents a 0.275 absolute reduction from the CGS peak of 0.800 (34% relative), and a 0.070 reduction from the CGS final rate (12% relative). Critically, the narrative-only system achieves formal convergence, the rolling standard deviation reaches exactly zero, but at the cost of all dynamic responsiveness.

No shock response. The adversarial shock at $t = 30$ produces no discernible effect in the narrative-only condition. Cooperation drops briefly from 0.550 to 0.500 and returns to 0.525 by $t = 32$, but this near-zero response reflects the absence of the evolutionary feedback loop: without Fermi social learning, the shock-induced defection does not propagate through the network via payoff-mediated imitation, and without individual reinforcement, agents do not adjust their propensities in response to changed payoff environments. The LLM advisory alone lacks the mechanism to either transmit or recover from exogenous perturbations.

Strategy polarization. Despite the population-level stasis, the narrative-only condition produces extreme agent-level heterogeneity: final strategies span the full $[0.0, 1.0]$ range with a standard deviation of 0.506. Agents converge to either full cooperation ($\sigma_o = 1.0$) or full defection ($\sigma_o = 0.0$), producing a bimodal distribution. This is a direct consequence of the narrative-only update (11): without the moderating influence of Fermi imitation (which pulls agents toward the neighbor average) and individual reinforcement (which maps payoffs to intermediate propensities via a sigmoid), the LLM's binary action recommendation (*C* or *D*) drives strategies monotonically toward the extremes over repeated application.

2) No-Critic Ablation: Deliberation Quality Matters

The no-critic ablation retains the full EC-theory update but bypasses the critic stage, passing the solver's action and confidence directly to the aggregator. This isolates the contribution of adversarial verification within the deliberation pipeline.

Performance is close but distinguishable. The no-critic condition achieves a peak cooperation of 0.825, marginally higher than the full CGS system's 0.800, and a final rate of 0.568 ± 0.076, slightly below CGS's 0.595 ± 0.084. Payoff (14.16 vs. 14.48) and Gini (0.239 vs. 0.236) are nearly identical, and shock recovery time is the same (8 steps). The differences, while consistent in direction, are modest.

The critic's role is confidence calibration. The most informative difference is in LLM confidence: the no-critic condition produces higher average confidence (0.778 vs. 0.698 in the final window), reflecting the absence of the critic's downward pressure on overconfident proposals. Without the critic, the solver's confidence is taken at face value, which occasionally produces high-confidence but strategically suboptimal recommendations. The aggregator's action-flip mechanism (triggered when $c_k < 0.3$ and $c_s > 0.7$) is necessarily inactive in this condition, removing the self-correction capability that prevents the most egregious advisory errors.

Interpretation. The modest performance gap between CGS and the no-critic condition suggests that the EC-theory grounding is the primary contributor to emergent cooperation, with the deliberation pipeline playing a secondary but non-negligible role. The critic provides value primarily through confidence calibration and error correction rather than through fundamentally different action recommendations, consistent with the observation that the solver already has access to the full agent context and generally makes reasonable proposals.

Figure 13 provides a direct three-way comparison of the ablation conditions. The cooperation panel (a) visually confirms the hierarchy: both EC-grounded conditions (CGS and EC-no-critic) track each other closely and far exceed the narrative-only flat line, while the payoff panel (b) shows the corresponding pattern in collective welfare. The near-overlap of the CGS and EC-no-critic traces throughout most of the simulation, with the primary divergence in post-shock recovery amplitude, reinforces the conclusion that EC grounding is the dominant factor and critic verification is a secondary refinement.

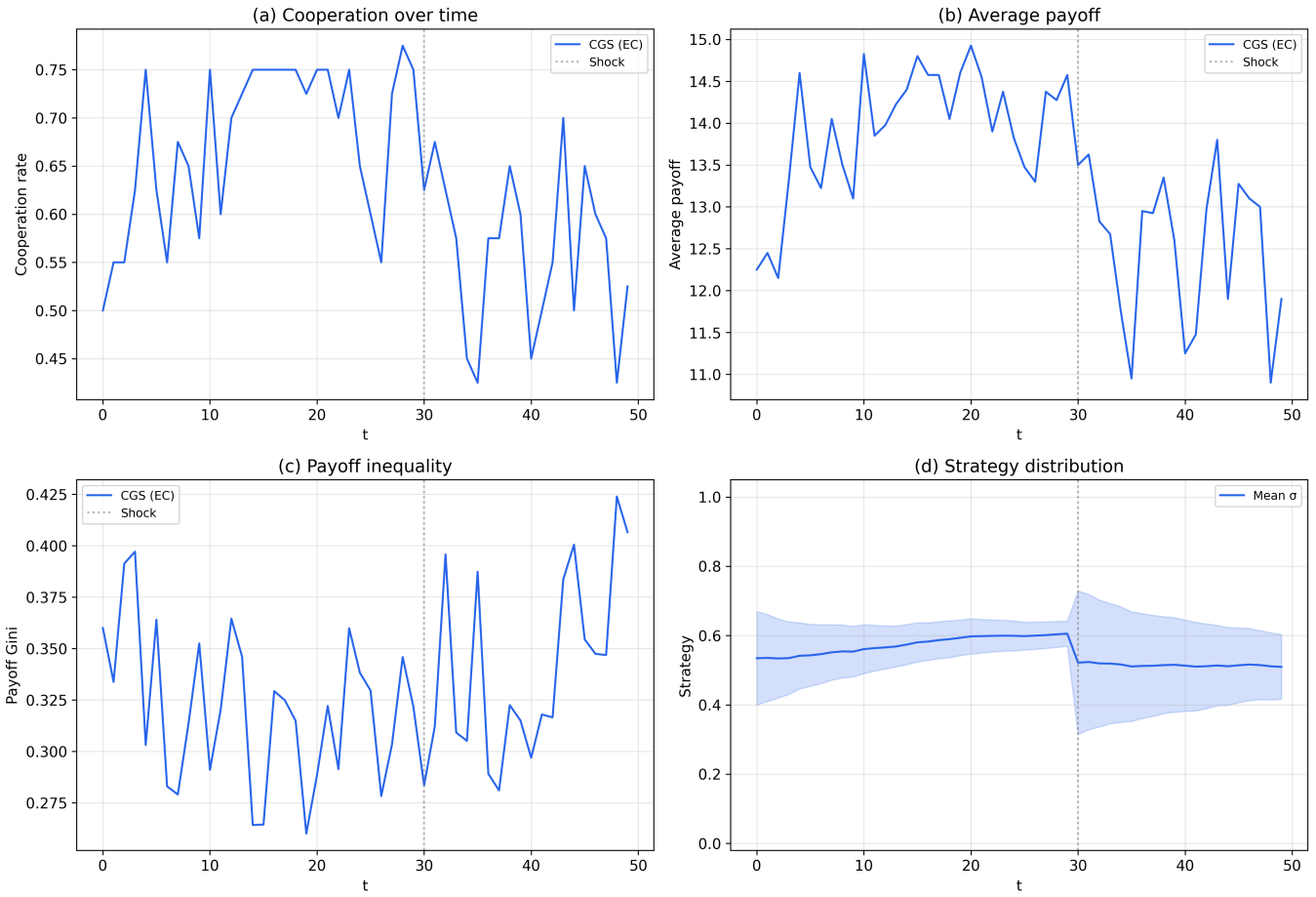

FIGURE 6: CGS dynamics on the Erdös–Rényi random network. (a) Cooperation peaks at 0.775 and recovers from the shock in 13 steps, the slowest recovery. (b) Average payoff (12.36 final). (c) Payoff inequality intermediate between small-world and scale-free (Gini ≈ 0.36). (d) Strategy distribution with post-shock widening.

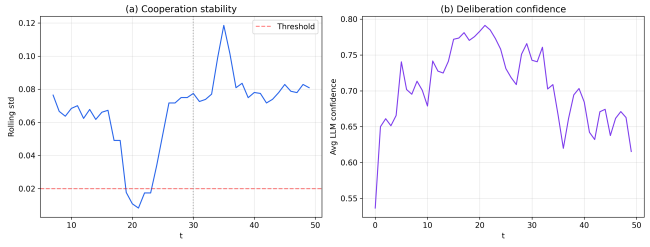

FIGURE 7: Stability analysis for the Erdös–Rényi topology. (a) The rolling std drops below 0.02 during $t = 20$–27, the only topology achieving near-convergence before the shock. (b) LLM confidence shows a post-shock decline from 0.78 to 0.62.

E. MULTI-SEED REPLICATION (TEN SEEDS)

We replicate the CGS-EC vs. narrative-only contrast across ten independent seeds ($s \in \{0, \ldots, 9\}$), doubling the Round-1 sample. Table 6 reports bootstrap 95 % confidence intervals and Table 7 the Bonferroni-corrected Mann–Whitney tests; Figure 14 visualises both.

The core distinction is now statistically significant. The cooperation standard deviation, a direct proxy for dynamic responsiveness, separates the two conditions completely (Mann–Whitney $U = 100$, the maximum value, $p_{\text{Bonf}} < 0.001$, Cohen's $d = 4.59$). Peak cooperation is also significant after correction ($p_{\text{Bonf}} = 0.015$, $d = 1.69$). Final cooperation, payoff and Gini do not differ significantly, which is the expected and correct picture: the claim has always concerned dynamic adaptivity, not the raw final level, and the two conditions reach similar mean final cooperation (0.608 vs. 0.596).

Perfect qualitative split. CGS-EC remains dynamically adaptive (fails to formally converge) in nine of ten seeds, whereas the narrative-only baseline converges to a static equilibrium in ten of ten seeds. This is a clean replication of the central mechanism: EC grounding sustains adaptive dynamics while pure LLM-driven updates collapse to fixed points, now established with statistical significance rather than as a qualitative trend.

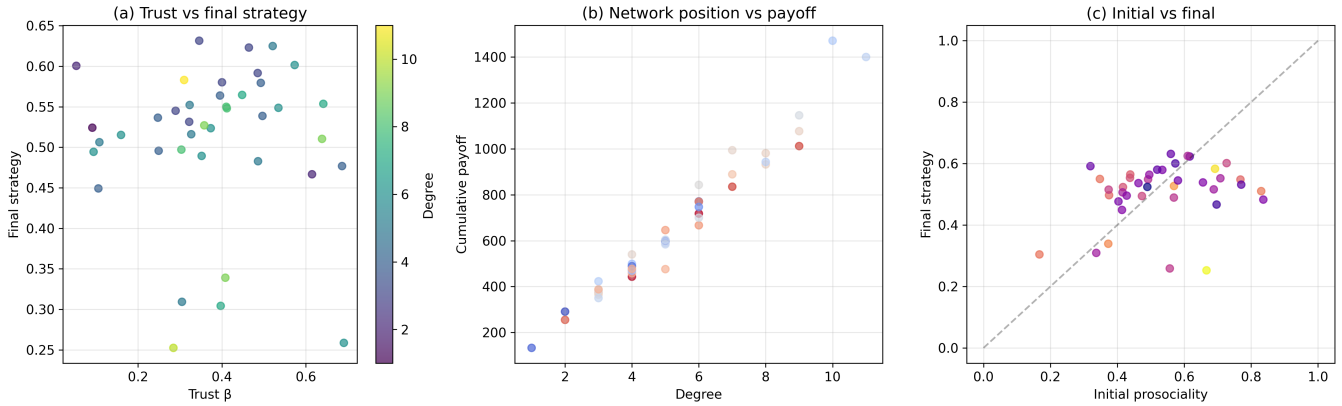

FIGURE 8: Agent-level analysis on the Erdös–Rényi network. (a) Trust β vs. final strategy: no clear pattern. (b) Degree vs. cumulative payoff: strong correlation ($r = 0.98$) spanning a $11.1\times$ range. (c) Initial vs. final strategy: weak positive trend ($r = 0.25$, $p = 0.119$).

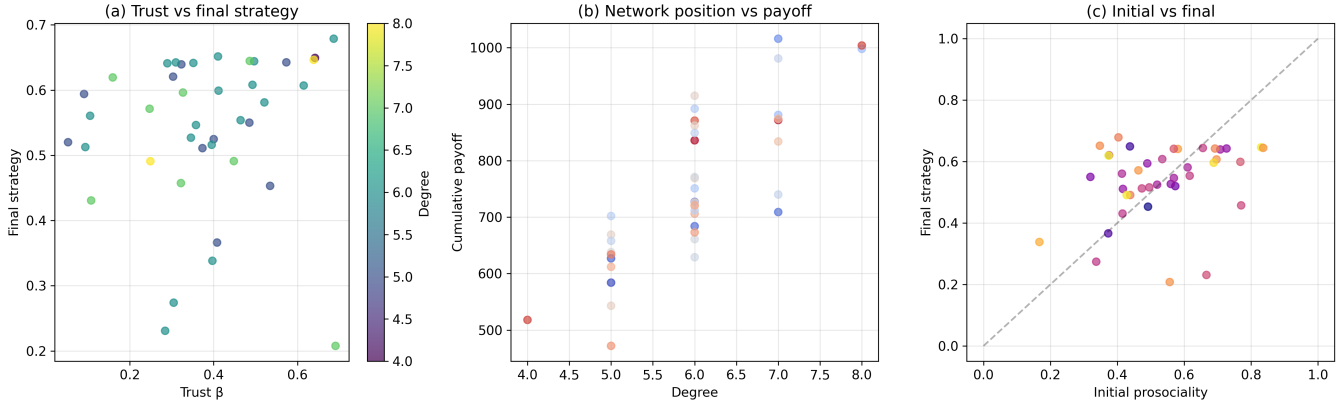

FIGURE 9: Agent-level analysis on the small-world network. (a) Trust β vs. final strategy (colored by degree): no significant correlation. (b) Degree vs. cumulative payoff: strong positive correlation ($r = 0.81$) within a narrow degree range. (c) Initial prosociality vs. final strategy: marginal persistence ($r = 0.30$, $p = 0.064$).

TABLE 6: Ten-seed bootstrap 95 % CIs (5 000 resamples), small-world, $n = 40$, $T = 50$, shock at $t = 30$.

Metric	CGS-EC	Narrative-only
ρ_C^{final}	0.608 [0.571, 0.647]	0.596 [0.549, 0.643]
$\text{std}(\rho_C)$	0.061 [0.051, 0.070]	0.006 [0.003, 0.009]
$\bar{\pi}^{\text{final}}$	14.62 [14.24, 15.01]	14.35 [13.81, 14.93]
$\mathcal{G}^{\text{final}}$	0.237 [0.229, 0.244]	0.222 [0.210, 0.235]
ρ_C^{peak}	0.820 [0.795, 0.845]	0.698 [0.643, 0.750]

TABLE 7: Ten-seed Mann–Whitney U tests: CGS-EC vs. Narrative-only (two-sided, Bonferroni corrected, $m = 5$). d = Cohen's d.

Metric	U	p_{raw}	p_{Bonf}	d	Sig.
Final cooperation	55.0	0.733	1.000	0.18	ns
Cooperation std	100.0	0.0002	0.0009	4.59	***
Final payoff	62.0	0.385	1.000	0.33	ns
Payoff Gini	72.0	0.104	0.521	0.83	ns
Peak cooperation	89.0	0.0031	0.015	1.69	*

F. POPULATION SCALING

Figure 15 presents cooperation, payoff, and Gini trajectories for CGS-EC at $n \in \{40, 60, 80\}$; Table 8 consolidates the summary metrics.

Pre-shock dynamics are scale-invariant. The pre-shock cooperation trajectory is essentially identical across all three population sizes, with mean pre-shock cooperation of 0.683, 0.677, and 0.669 for $n = 40$, 60, 80 respectively. Peak cooperation reaches 0.800 at $n \in \{40, 60\}$ and 0.775 at $n = 80$, and payoff efficiency is consistent across scales ($\bar{\pi}^{\text{final}} \in [14.07, 14.73]$). These results confirm that the Fermi imitation and LLM advisory dynamics produce comparable collective outcomes within this population range.

Post-shock recovery slows with scale. Recovery time increases from 1 step ($n = 40$) to 5 steps ($n = 60$), and the $n = 80$ simulation does not return to within 0.05 of the pre-shock level within $T = 50$ steps, reaching a trough of 0.425 at $t = 42$. This pattern is mechanistically expected: at $f_{\text{shock}} = 0.15$, the shock perturbs 6, 9, and 12 agents at

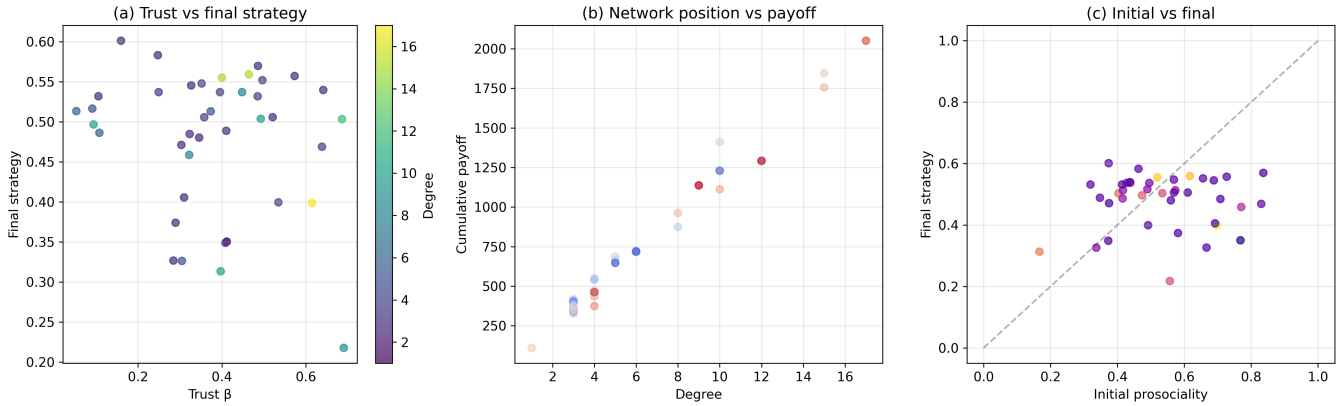

FIGURE 10: Agent-level analysis on the scale-free network. (a) Trust β vs. final strategy: no correlation. (b) Degree vs. cumulative payoff: near-perfect correlation ($r = 0.99$) spanning a $18.6\times$ disparity between the highest-degree hub and the lowest-degree periphery. (c) Initial vs. final strategy: correlation vanishes ($r = 0.04$), indicating that network structure dominates initial conditions.

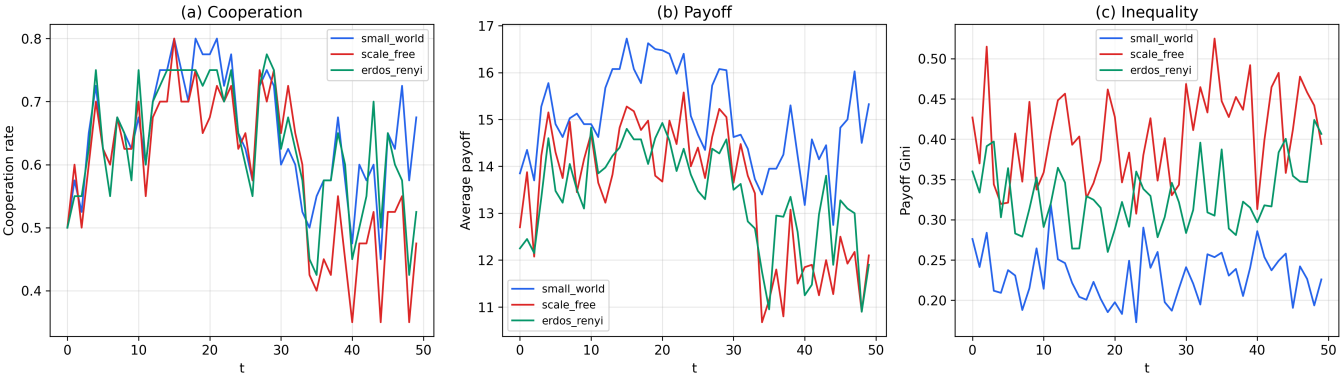

FIGURE 11: Cross-topology comparison of CGS dynamics. (a) Cooperation rate: trajectories diverge after the shock, with small-world sustaining the highest levels. (b) Average payoff: small-world consistently outperforms. (c) Payoff inequality (Gini): scale-free is persistently highest, reflecting hub-driven disparity; small-world is lowest throughout.

TABLE 8: Population scaling: CGS-EC on small-world (seed 0, $T = 50$). ρ_C^{pre}: mean cooperation over $t \in [0, 29]$. Recovery: steps to return within 0.05 of pre-shock cooperation; "—" indicates no full recovery within $T = 50$.

n	ρ_C^{pre}	ρ_C^{peak}	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	**Recovery**
40	0.683	0.800	0.598 ± 0.079	14.51	0.236	1 step
60	0.677	0.800	0.617 ± 0.070	14.73	0.236	5 steps
80	0.669	0.775	0.566 ± 0.074	14.07	0.254	—

$n = 40, 60$, and 80 respectively, so the absolute magnitude of the perturbation scales with n. Payoff Gini rises only modestly ($0.236 \to 0.236 \to 0.254$), remaining well below the scale-free values observed in Section V-A2.

These results establish that the core CGS cooperation phenomena extend beyond the $n = 40$ regime of the primary experiments. The main consequence of larger populations is a proportionally longer post-shock recovery horizon, suggesting that future experiments at $n > 80$ should extend T accordingly.

G. MULTI-MODEL ANALYSIS

Table 9 and Figure 16 report CGS-EC performance across four LLMs.

Two-cluster structure. The four models partition into two behavioural clusters. The *high-cooperation cluster* (Llama-70B, Hermes-4-70B) achieves final cooperation of 0.603 and 0.573 respectively, peak cooperation of 0.800 in both cases, and payoff efficiency consistent with the primary results ($\bar{\pi} \in [14.27, 14.57]$, $\mathcal{G} \in [0.235, 0.244]$). The *low-cooperation cluster* (DeepSeek-V3, Qwen3-235B) achieves identical final cooperation of 0.338 with substantially lower payoffs ($\bar{\pi} = 11.38$) and higher inequality ($\mathcal{G} \in [0.273, 0.278]$).

Cross-model coefficient of variation (CV) is 31.3% for final cooperation but only 9.7% for peak cooperation and 8.2% for the Gini coefficient, indicating that while the level of sustained cooperation depends on the underlying model, the structural properties of EC-grounded dynamics, the capacity to reach cooperative peaks and the low-inequality profile characteristic of small-world interactions, are substantially

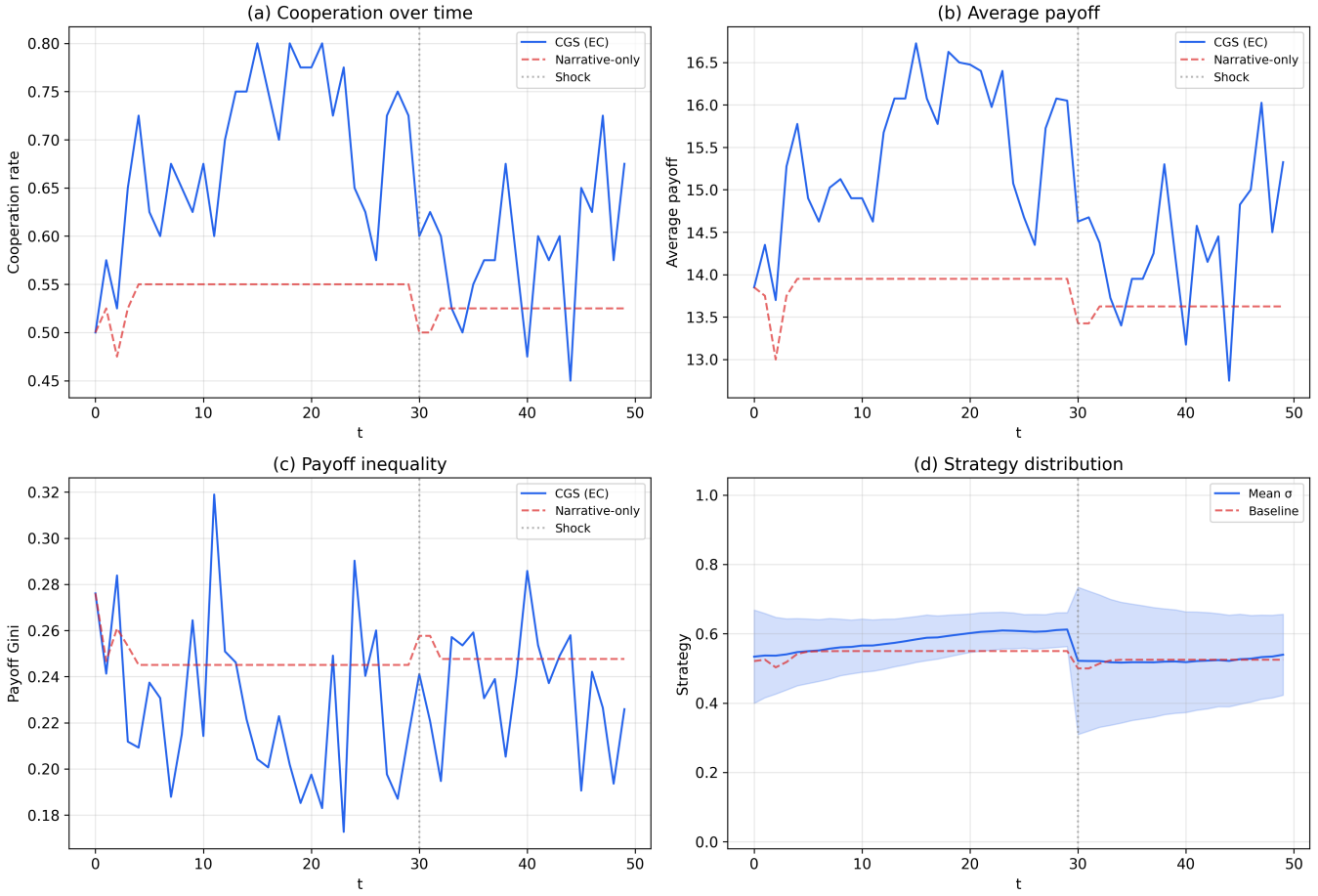

FIGURE 12: CGS (EC + Deliberation) vs. narrative-only baseline on the small-world network. (a) Cooperation: CGS reaches 0.80 and recovers from the shock; the narrative-only baseline plateaus at 0.55 with no dynamic response. (b) Payoff: CGS consistently outperforms. (c) Inequality: CGS achieves lower Gini during cooperative peaks. (d) Strategy distribution: CGS maintains moderate heterogeneity (shaded band) while the baseline is static.

TABLE 9: Multi-model sweep: CGS-EC on small-world ($n = 40$, seed 0, $T = 50$, shock at $t = 30$). $\bar{c}^{\text{final}}$: mean aggregated LLM confidence over the last 10 steps. Recovery: steps to return within 0.05 of pre-shock cooperation; 0 steps indicates brief threshold touch without sustained recovery.

Label	**Model**	ρ_C^{peak}	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	$\bar{c}^{\text{final}}$	**Recovery**
Llama-70B	Llama-3.3-70B-Instruct	0.800	0.603 ± 0.076	14.57	0.235	0.700	8 steps
DeepSeek-V3	DeepSeek-V3.2	0.650	0.338 ± 0.093	11.38	0.273	0.716	0 steps
Qwen3-235B	Qwen3-235B-A22B-2507	0.725	0.338 ± 0.108	11.38	0.278	0.787	16 steps
Hermes-4-70B	Hermes-4-70B	0.800	0.573 ± 0.090	14.27	0.244	0.728	15 steps

preserved.

Confidence does not predict cooperation. Deliberation confidence is similar across all four models (pre-shock range: 0.711–0.746; post-shock: 0.703–0.741), and is highest for the low-cooperation model (Qwen3-235B: $\bar{c} = 0.787$). This confirms that advisory signal quality, its alignment with the local payoff gradient, rather than its confidence magnitude determines downstream cooperation outcomes. The EC-theory update rule (6) provides partial insulation against low-quality advisory signals through the social learning and reinforcement components, which together carry $w_s + w_r = 0.75$ weight independently of LLM output quality.

Recovery behaviour. DeepSeek-V3 formally recovers in 0 steps but, in a pattern analogous to the scale-free topology results of Section V-A2, cooperation subsequently declines rather than being sustained. Qwen3-235B achieves the longest recovery (16 steps) with a sustained post-shock level of 0.338, while Hermes-4-70B recovers in 15 steps with a final cooperation of 0.573, outperforming DeepSeek-V3 at comparable model scale.

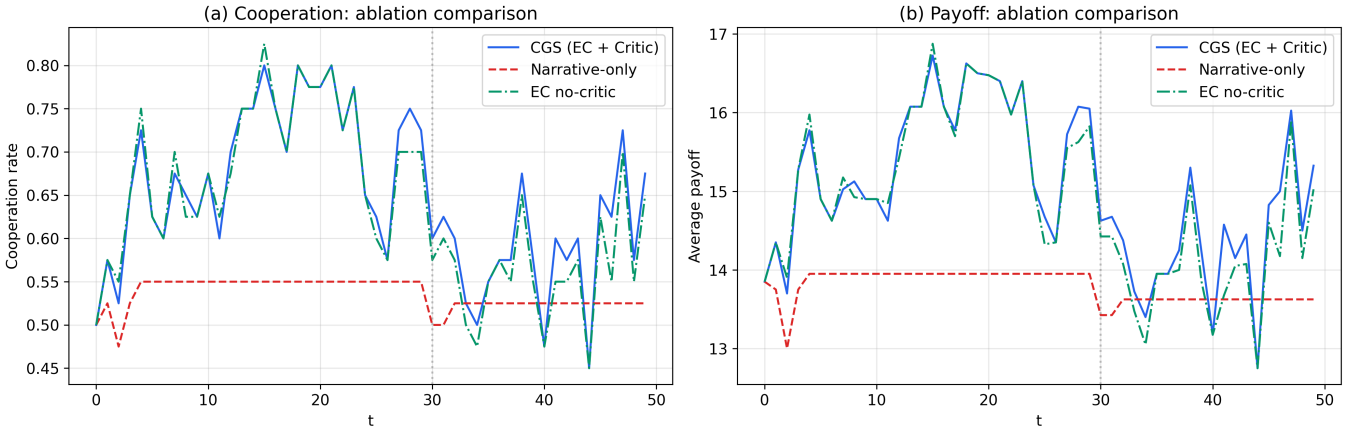

FIGURE 13: Three-way ablation comparison on the small-world network. (a) Cooperation rate: the narrative-only baseline (red dashed) plateaus at 0.55 while both EC-grounded conditions exhibit dynamic cooperation with peaks of 0.80–0.83. (b) Payoff: the narrative-only baseline maintains a flat $\approx$ 13.6 while EC-grounded conditions fluctuate between 13–17, achieving higher peaks.

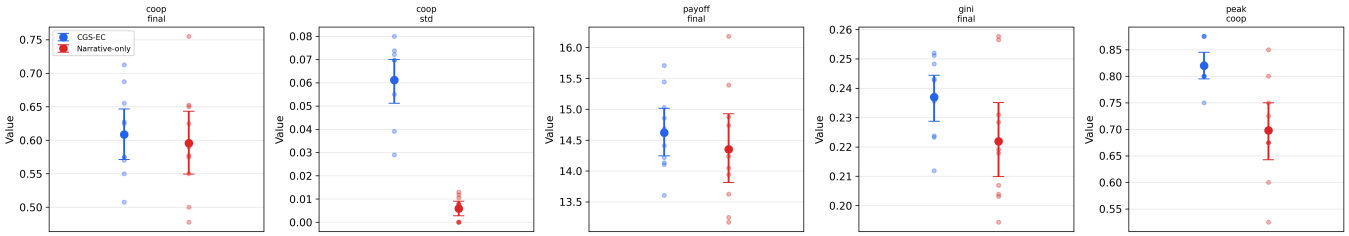

FIGURE 14: Ten-seed bootstrap 95 % CIs for five summary metrics. Large markers = bootstrap mean; small dots = individual seeds. The cooperation-std and peak-cooperation panels show fully separated intervals (significant after Bonferroni correction); final cooperation, payoff and Gini overlap, consistent with the claim being about dynamic adaptivity rather than final level.

H. GNN STRUCTURAL BASELINES

Table 10 and Figure 17 compare CGS-EC against three structural baselines that perform graph message-passing without any LLM component.

CGS-EC outperforms all structural baselines. CGS-EC achieves final cooperation of 0.595, exceeding GNN-Mean and GNN-2L (both 0.495) by $\Delta\rho_C = +0.100$ (+20.2% relative) and surpassing GNN-PayW (0.408) by +0.188 (+46.1%). Peak cooperation for CGS-EC (0.800) exceeds GNN-Mean and GNN-2L (0.775) and substantially exceeds GNN-PayW (0.725). Payoff efficiency follows the same ordering: CGS-EC (14.48) > GNN-Mean (13.44) $\approx$ GNN-2L (13.45) > GNN-PayW (12.36).

Payoff weighting without degree normalisation is counterproductive. GNN-PayW's Fermi-weighted aggregation most closely resembles the social learning component of the CGS EC-theory update, yet achieves the lowest cooperation of any evaluated method. This counterintuitive result is mechanistically informative: without the degree-normalised sigmoid of component (9), payoff-weighted imitation amplifies successful defectors during the post-shock period, driving peripheral agents further toward permanent defection rather than facilitating recovery. The +0.088 gap between GNN-PayW and GNN-Mean thus quantifies the benefit of degree normalisation within the EC update rule. The further +0.100 gap between GNN-Mean and CGS-EC quantifies the additional cooperation contribution of the LLM advisory signal after structural dynamics are fully accounted for.

Inequality. CGS-EC maintains the lowest payoff inequality ($\mathcal{G}^{\text{final}} = 0.236$), while all GNN baselines produce Gini values 13–16% higher (0.267–0.273). Without context-sensitive LLM advisory modulation, peripheral agents exposed to defection post-shock are more readily locked into permanent defection, amplifying the structural degree-driven payoff disparities.

Formal recovery and sustained cooperation. All three GNN baselines formally recover within 1 step, compared with 8 steps for CGS-EC. As with the scale-free results of Section V-A2, rapid formal recovery in the GNN baselines is followed by continued cooperation decline rather than sustained post-recovery levels. The 8-step recovery of CGS-EC is accompanied by sustained cooperation near $\rho_C \approx 0.60$, and recovery speed alone is therefore an insufficient resilience metric.

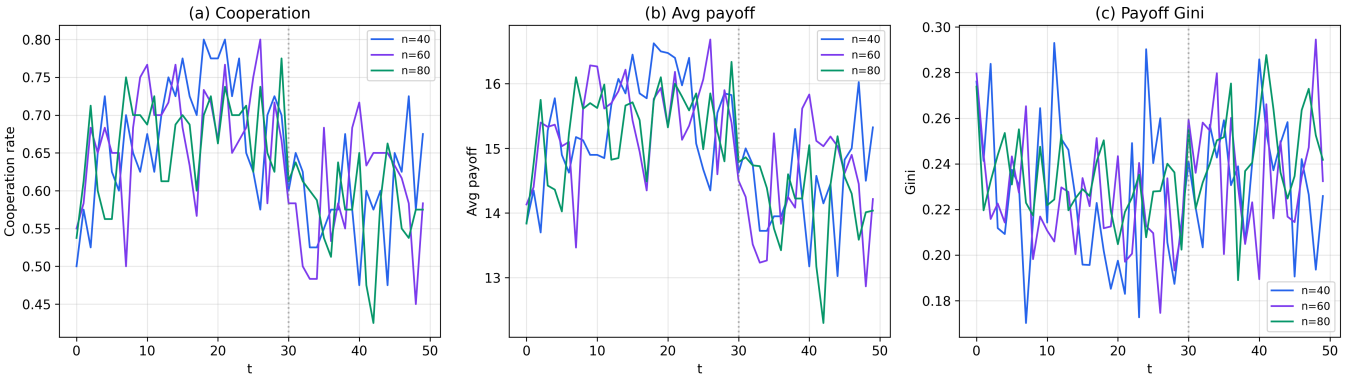

FIGURE 15: Population scaling: CGS-EC on small-world at $n \in \{40, 60, 80\}$ (seed 0, $T = 50$). (a) Pre-shock cooperation trajectories are nearly identical across population sizes; post-shock recovery slows with increasing n. (b) Average payoff is consistent across scales. (c) Payoff Gini remains below 0.255 at all sizes. Vertical dotted line marks the adversarial shock at $t = 30$.

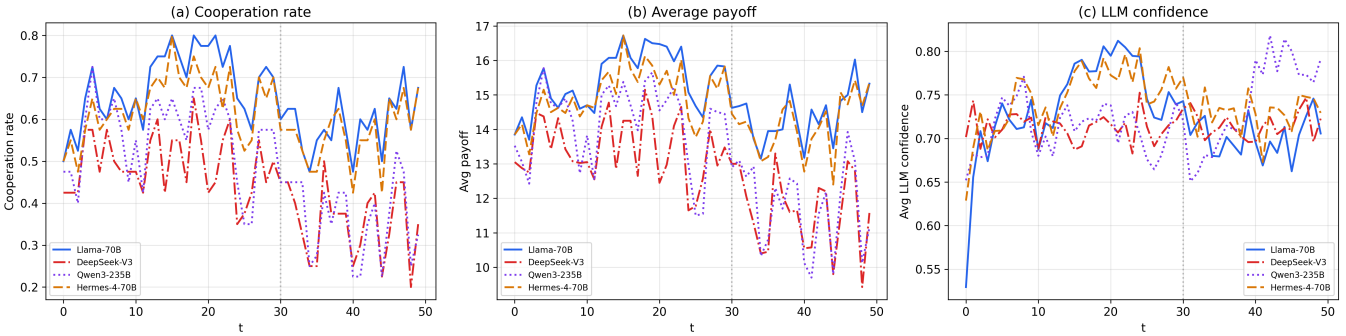

FIGURE 16: Multi-model sweep: cooperation rate (a), average payoff (b), and LLM deliberation confidence (c) across four LLMs. Llama-70B and Hermes-4-70B sustain high cooperation ($\rho_C^{\text{final}} > 0.57$), consistent with the primary results. DeepSeek-V3 and Qwen3-235B converge to a lower-cooperation regime ($\rho_C^{\text{final}} = 0.338$) despite comparable confidence levels. Vertical dotted line marks the adversarial shock at $t = 30$.

TABLE 10: CGS-EC vs. GNN structural baselines (small-world, $n = 40$, seed 0, $T = 50$, shock at $t = 30$). Recovery as defined in Section IV-E; 0 steps indicates brief threshold touch without sustained recovery.

Method	**Update rule**	ρ_C^{peak}	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	**Recovery**
CGS-EC	EC-theory + SCA deliberation	0.800	0.595 ± 0.084	14.48	0.236	8 steps
GNN-Mean	DeGroot mean consensus	0.775	0.495 ± 0.077	13.44	0.267	1 step
GNN-PayW	Fermi-weighted mean aggregation	0.725	0.408 ± 0.081	12.36	0.273	1 step
GNN-2L	Two-layer 0.6 / 0.4 readout	0.775	0.495 ± 0.076	13.45	0.267	1 step

I. OAT PARAMETER SENSITIVITY AND TEMPERATURE ROBUSTNESS

Table 11 reports normalised sensitivity indices (NSI) for four output metrics across six parameters. Figure 18 shows the full response curves, making the shape of each relationship visible beyond the scalar summary, while Figure 19 provides a compact cross-metric overview.

Architectural parameters are robust. The Fermi selection intensity κ has negligible influence on all outputs (NSI $= 0.017$, cooperation span of 0.7% across the full sweep from 0.5 to 4.0), confirming that the system is not finely tuned to the sigmoid steepness of the imitation rule. The LLM advisory weight w_ℓ is similarly insensitive (NSI $= 0.087$, cooperation span 9.5%), validating the fixed weighting scheme of (6): the output is robust to the precise allocation between structural and advisory components within the tested range. The reinforcement sensitivity α is moderate (NSI $= 0.137$), with the baseline $\alpha = 0.8$ sitting in the middle of the response curve.

Initial prosocial disposition is the dominant parameter. μ_p has the largest NSI (1.222) by a substantial margin, indicating that output sensitivity exceeds input variation. The response is strongly asymmetric: reducing μ_p from the baseline 0.55 to 0.20 collapses final cooperation to 0.255 (peak 0.375), while increasing it to 0.70–0.85 plateaus cooperation at 0.660–0.663 with diminishing returns. This asymmetry

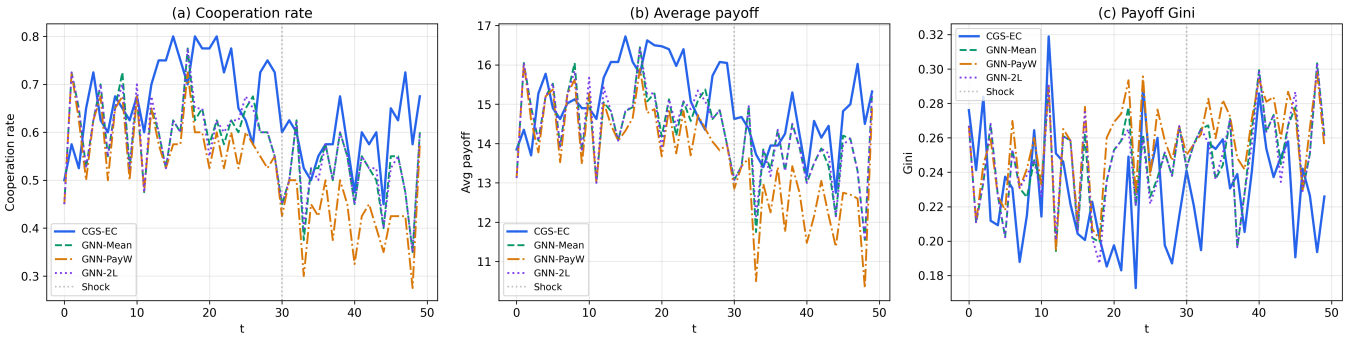

FIGURE 17: CGS-EC vs. GNN structural baselines (small-world, $n = 40$, seed 0, $T = 50$). (a) Cooperation rate: CGS-EC sustains the highest cooperation throughout; GNN baselines settle into lower post-shock regimes. (b) Average payoff: mirrors the cooperation ordering. (c) Payoff Gini: CGS-EC maintains the lowest inequality ($\mathcal{G} = 0.236$); all GNN baselines produce higher Gini (0.267–0.273). Vertical dotted line marks the adversarial shock at $t = 30$.

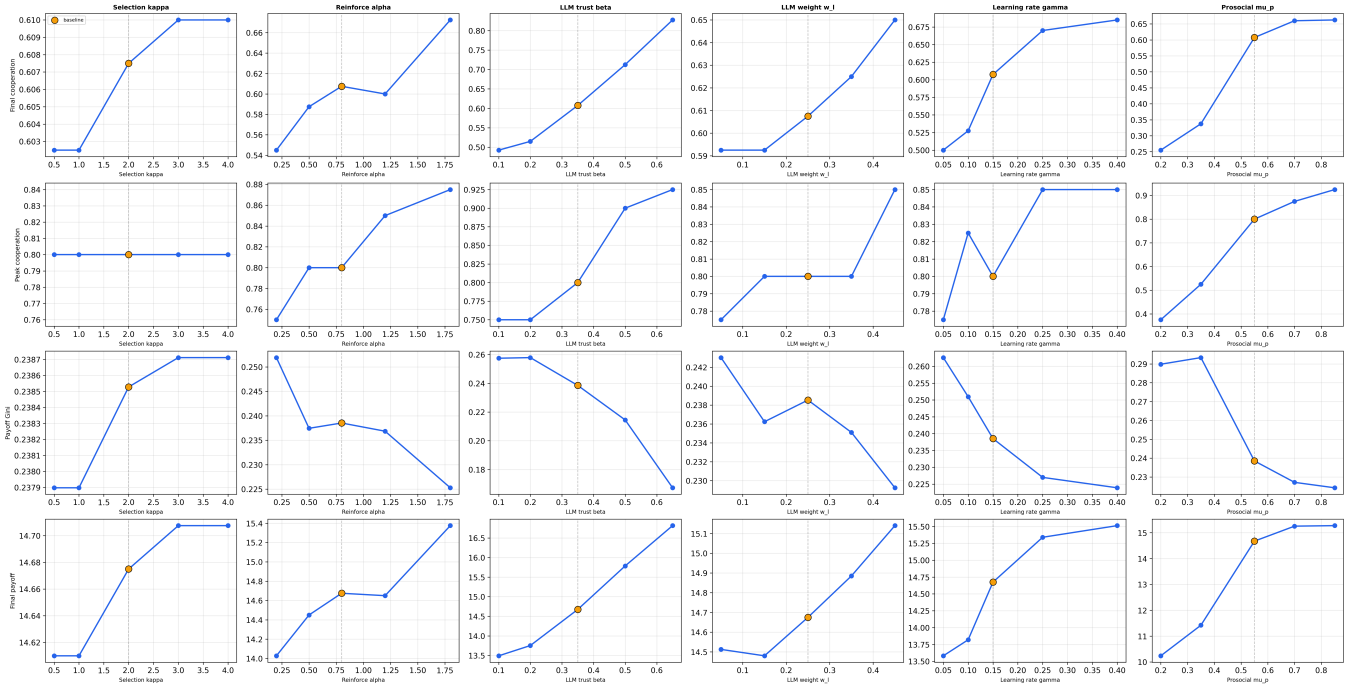

FIGURE 18: OAT sensitivity analysis: response curves for four output metrics (rows) across six parameters (columns). Orange dot = baseline value. Fermi selection intensity κ is essentially flat across all metrics (cooperation span $< 1\%$), while initial prosocial disposition μ_p exhibits a strong asymmetry with a critical-mass threshold near the baseline value. LLM trust $\bar{\beta}$ is the only parameter for which cooperation and Gini improve simultaneously across the full sweep.

reflects a critical-mass effect: populations initialised below a cooperative threshold cannot sustain sufficient payoff gradients for Fermi imitation to propagate cooperation, whereas populations above the threshold converge to similar long-run outcomes regardless of exact initialisation.

LLM trust monotonically improves both cooperation and equity. $\bar{\beta}$ (mean LLM trust) exhibits the second-highest NSI (0.423) with a monotone response: increasing $\bar{\beta}$ from 0.10 to 0.65 raises final cooperation from 0.493 to 0.828 and simultaneously reduces Gini from 0.257 to 0.167. This is the only parameter for which cooperation and equity improve jointly across the full sweep, suggesting that populations willing to weight the LLM advisory more heavily benefit from both higher collective welfare and more equitable payoff distributions. The baseline $\bar{\beta} = 0.35$ is therefore a conservative choice; practitioners with well-calibrated models can realise substantial gains by increasing LLM trust.

Temperature robustness.

Table 12 and Figure 20 report CGS-EC performance across four LLM sampling temperatures ($T_s \in \{0.0, 0.2, 0.5, 0.8\}$) on the small-world topology ($n = 40$, seed 0, shock at $t = 30$).

CGS dynamics are insensitive to sampling temperature. Final cooperation varies by only 0.0075 across the

TABLE 11: Normalised sensitivity indices (NSI) from the OAT sweep (small-world, $n = 40$, seed 0, $T = 50$). NSI > 0.3 indicates meaningful sensitivity; NSI < 0.1 indicates robustness to that parameter.

Parameter	ρ_C^{final}	ρ_C^{peak}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$
μ_p (prosocial mean)	**1.222**	**0.945**	**0.610**	**0.632**
$\bar{\beta}$ (LLM trust)	**0.423**	0.292	0.177	**0.349**
$\bar{\gamma}$ (learning rate)	**0.395**	0.094	0.175	0.157
α (reinforce sens.)	0.137	0.125	0.059	0.075
w_ℓ (LLM weight)	0.087	0.078	0.040	0.049
κ (Fermi intensity)	0.017	0.000	0.009	0.005

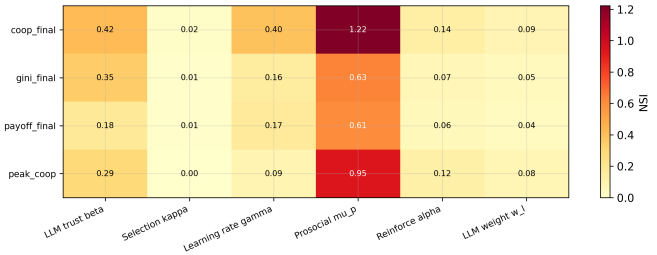

FIGURE 19: NSI heatmap: darker cells indicate greater output sensitivity to a given parameter. The architectural parameters w_ℓ and κ are robustly insensitive across all metrics; initial prosocial disposition μ_p and LLM trust $\bar{\beta}$ dominate.

full $[0.0, 0.8]$ range (CV $= 0.6\%$), payoff Gini varies by 0.0019 (CV $= 0.3\%$), and all four conditions reach peak cooperation of 0.80–0.83. This robustness follows from the architecture: at temperature $T_s = 0.0$ the LLM produces deterministic outputs, while at $T_s = 0.8$ outputs are stochastic, yet the EC-theory update rule absorbs the resulting advisory variance through the $w_\ell = 0.25$ weighting and the agent-level trust heterogeneity β_o. The baseline temperature of 0.2 used throughout the primary experiments therefore requires no special justification; any value in the tested range would yield equivalent results.

J. GENERALISATION TO OTHER SOCIAL DILEMMAS

The Prisoner's Dilemma is only one class of social dilemma. To assess whether the architecture transfers, we re-run CGS-EC on the small-world topology with two further canonical 2×2 games, Snowdrift (SD, $T > R > S > P$, anti-coordination) and Stag Hunt (SH, $R > T > P > S$, a coordination game with two pure equilibria), under identical agent initialisation and adversarial shock. Table 13 and Figure 21 report the outcomes.

The dynamics are qualitatively identical across all three games: final cooperation lies in $[0.600, 0.640]$, peak cooperation in $[0.800, 0.850]$, and *all three recover from the shock in exactly 8 steps*. The Stag Hunt yields the highest payoff (15.88) and the lowest inequality ($\mathcal{G} = 0.161$): in a coordination game the EC-grounded update steers the population toward the payoff-dominant (mutual-cooperate) equilibrium rather than the risk-dominant one, which is the theoretically desirable behaviour. Snowdrift is intermediate, as expected

TABLE 12: Temperature robustness: CGS-EC on small-world ($n = 40$, seed 0, $T = 50$, shock at $t = 30$). CV: coefficient of variation across the four temperatures.

T_s	ρ_C^{peak}	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$
0.0	0.800	0.605 ± 0.077	14.60	0.235
0.2	0.825	0.605 ± 0.075	14.61	0.235
0.5	0.800	0.598 ± 0.079	14.51	0.236
0.8	0.800	0.605 ± 0.075	14.63	0.234
CV	1.6%	0.6%	0.4%	0.3%

TABLE 13: CGS-EC across game substrates (small-world, $n = 40$, seed 0, $T = 50$, shock at $t = 30$).

Game	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	ρ_C^{peak}	**Recov.**
PD (baseline)	0.600 ± 0.080	14.55	0.234	0.800	8
Snowdrift	0.615 ± 0.081	13.65	0.177	0.825	8
Stag Hunt	0.640 ± 0.065	15.88	0.161	0.850	8

for an anti-coordination dilemma. CGS is therefore not specific to the Prisoner's Dilemma; the architecture transfers across social-dilemma classes without re-tuning.

K. NON-LLM LEARNING AND IMITATION BASELINES

To situate CGS-EC against classical and learning-theoretic update rules, we evaluate five baselines that replace the LLM advisory entirely, on the primary PD/small-world setting with identical shock and initialisation: tabular ϵ-greedy Q-learning over a binned neighbour-cooperation state, aspiration learning, pure Fermi imitation, a discrete graph replicator, and a uniform random control. Table 14 and Figure 22 report the results.

Pure imitation and replicator dynamics fail. Pure Fermi imitation collapses to $\rho_C^{\text{final}} = 0.113$ and the graph replicator to 0.455, neither recovering from the shock. This is the strongest single piece of evidence that the EC-theory update's *combination* of components is necessary: the social-learning term in isolation, the classical microscopic rule of the spatial-games literature, is by itself insufficient to sustain cooperation in this setting, and the reinforcement and advisory components are doing real work. Aspiration learning (0.353) likewise fails to recover.

An idealised Q-learner upper-bounds the task. A tabular Q-learner with full neighbourhood-state observability achieves higher final cooperation (0.728) than CGS-EC (0.595). This is expected and we report it transparently: with a small discrete state space, a stationary payoff matrix and 50 learning steps, tabular Q-learning converges to a near-optimal policy. The comparison is not like-for-like, the Q-learner is given a hand-specified state encoding and is unconstrained by natural-language context, so it functions as an idealised upper bound on achievable cooperation rather than as a deployable LLM-free competitor. CGS-EC's value is (1) operating from natural-language context without a hand-engineered state, (2) transferring across game substrates and topologies without re-tuning (Section V-J), and (3) the LLM advisory premium over

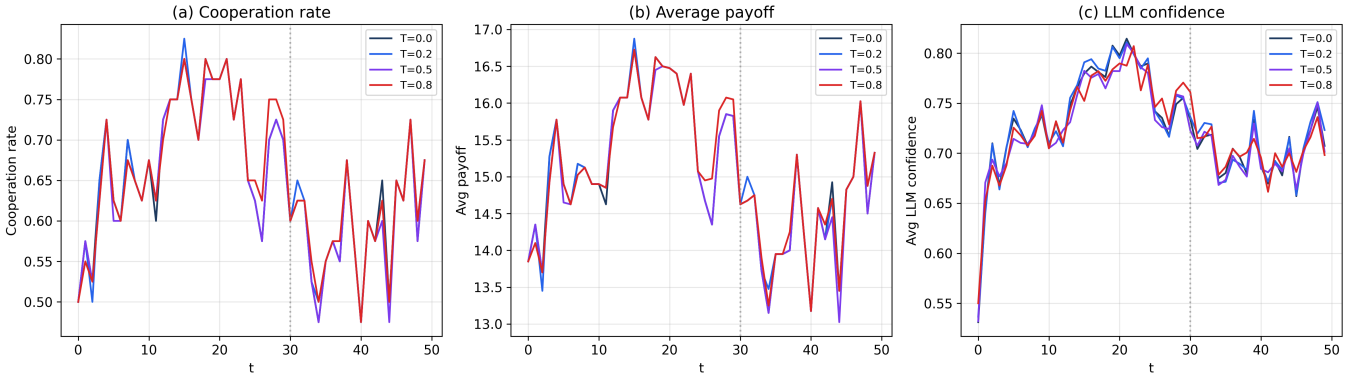

FIGURE 20: Temperature robustness: cooperation rate trajectories for $T_s \in \{0.0, 0.2, 0.5, 0.8\}$. All four conditions reach peak cooperation of 0.80–0.83 and recover from the adversarial shock within 8 steps. Trajectory overlap confirms that CGS dynamics are insensitive to LLM sampling temperature.

FIGURE 21: CGS-EC across PD, Snowdrift and Stag Hunt (small-world, $n = 40$, seed 0). (a) Cooperation rate, (b) average payoff, (c) payoff Gini. Trajectories are qualitatively identical; all three recover from the shock (dotted line) within 8 steps.

TABLE 14: CGS-EC vs. non-LLM baselines (PD, small-world, $n = 40$, seed 0, $T = 50$, shock at $t = 30$). "—" = no recovery within T.

Method	ρ_C^{final}	$\bar{\pi}^{\text{final}}$	$\mathcal{G}^{\text{final}}$	ρ_C^{peak}	**Recov.**
CGS-EC	0.595 ± 0.084	14.48	0.236	0.800	8
Q-learning	0.728 ± 0.036	15.92	0.226	0.975	0
Replicator	0.455 ± 0.055	13.07	0.263	0.725	—
Aspiration	0.353 ± 0.036	11.40	0.274	0.525	—
Fermi-only	0.113 ± 0.096	7.95	0.212	0.625	—
Random	0.493 ± 0.077	13.46	0.260	0.625	1

TABLE 15: Empirical equilibrium: CGS-EC, small-world, $n = 40$, seed 0, $T = 80$, no shock. Final cooperation is the mean over the last 20 steps.

μ_p (init)	ρ_C (final, last 20)	Δ from previous
0.10	0.353 ± 0.046	—
0.25	0.397 ± 0.050	+0.045
0.40	0.691 ± 0.050	**+0.294**
0.55	0.739 ± 0.061	+0.047
0.70	0.754 ± 0.058	+0.015
0.85	0.775 ± 0.063	+0.021

structural baselines isolated in Section V-H. Reporting the baseline that exceeds CGS-EC, and explaining precisely why, is a deliberate choice for transparency.

L. EMPIRICAL EQUILIBRIUM STRUCTURE

A closed-form fixed-point analysis is unavailable because the LLM response distribution is not analytically tractable. We therefore characterise the attractor structure empirically: CGS-EC is run with the shock disabled for $T = 80$ steps from a sweep of initial prosocial means $\mu_p \in \{0.10, \ldots, 0.85\}$. Table 15 and Figure 23 report the long-run cooperation as a function of μ_p.

The final-vs-initial curve is a sigmoid with a *critical-mass threshold* near $\mu_p \approx 0.30$–0.40: below it the population settles into a low-cooperation regime (≈ 0.35–0.40), above it into a high-cooperation regime (≈ 0.74–0.78) that *saturates* ($\Delta < 0.05$ for successive $\mu_p \geq 0.40$). The single largest jump, $+0.294$, occurs between $\mu_p = 0.25$ and $\mu_p = 0.40$. This is an empirical attractor map: a single high-cooperation attractor whose basin is entered only when the initial cooperative mass exceeds a threshold, below which the system is trapped near a low-cooperation fixed point. The result is mechanistically consistent with, and explains, the dominant OAT sensitivity of μ_p (NSI $= 1.222$, Section V-I): the high elasticity is the signature of the bifurcation.

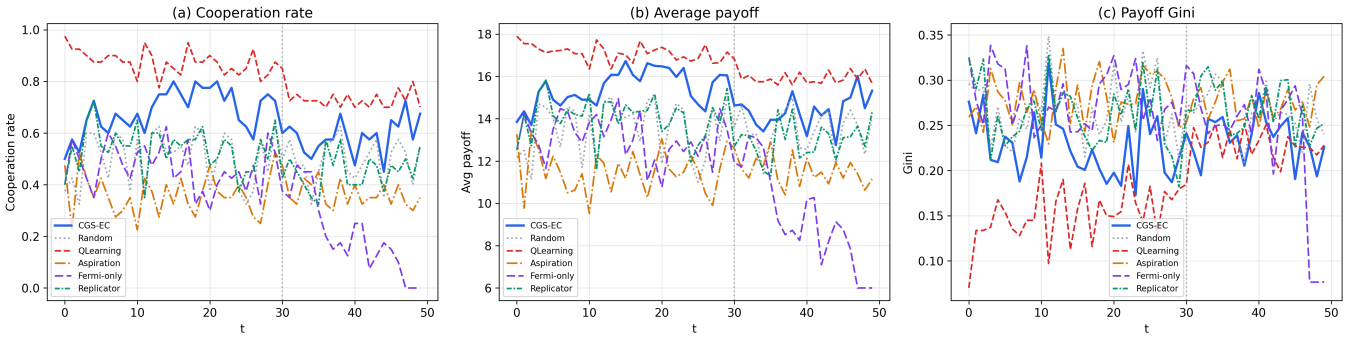

FIGURE 22: CGS-EC vs. five non-LLM baselines (PD, small-world, $n = 40$, seed 0). Pure Fermi imitation and the graph replicator fail to sustain cooperation after the shock; the idealised tabular Q-learner upper-bounds the task. (a) Cooperation, (b) payoff, (c) Gini.

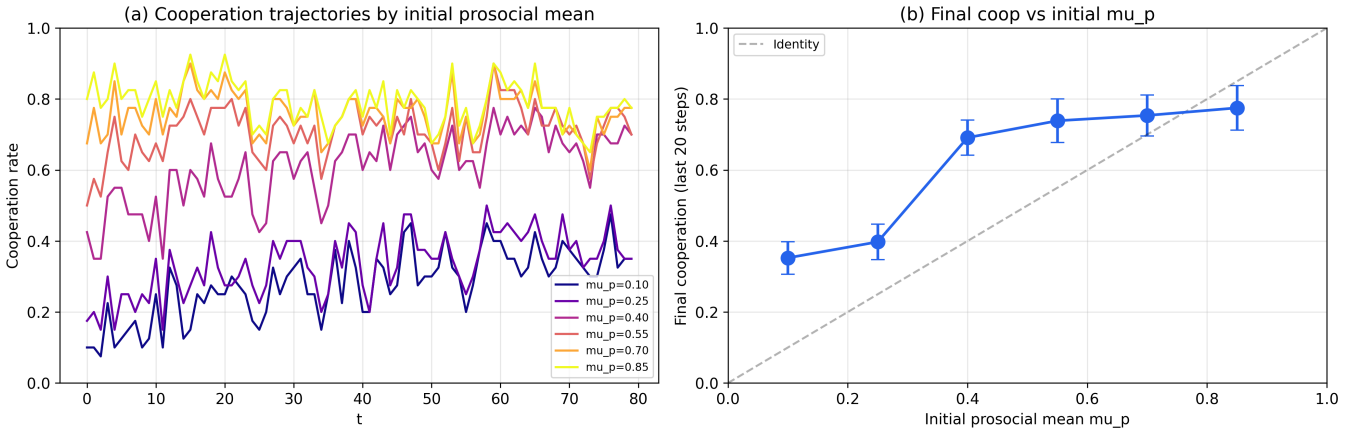

FIGURE 23: Empirical equilibrium structure (CGS-EC, small-world, $n = 40$, $T = 80$, no shock). (a) Cooperation trajectories by initial prosocial mean. (b) Final cooperation vs. μ_p: a sigmoid with a critical-mass threshold near $\mu_p \approx 0.35$ separating a low- from a high-cooperation attractor; the high branch saturates.

M. FAILURE MODES AND THE RESILIENCE BOUNDARY

To characterise the regime in which CGS-EC degrades, we sweep the shock fraction over $f_{\text{shock}} \in \{0.05, 0.15, 0.30, 0.50, 0.75\}$ (single shock at $t = 30$) and additionally test a persistent regime of five 15% shocks at $t \in \{20, 25, 30, 35, 40\}$. Table 16 and Figure 24 report the outcomes.

Two findings delimit the operating envelope. First, there is a *recovery threshold between* 15% *and* 30%: CGS-EC returns to within 0.05 of pre-shock cooperation for $f_{\text{shock}} \leq 0.15$ but does not recover within $T = 50$ for $f_{\text{shock}} \geq 0.30$, with final cooperation degrading monotonically to 0.198 at $f_{\text{shock}} = 0.75$. Second, *repeated shocks are more damaging than a single large one*: five spaced 15% shocks (cumulatively 75% but distributed) collapse cooperation to 0.222, worse than a single 50% shock (0.323), because the population is never granted enough recovery time between perturbations. The EC-grounded update is thus resilient to single moderate perturbations but not to severe or rapidly repeated ones; the recovery horizon must scale with both shock magnitude and frequency.

TABLE 16: Shock-magnitude sweep (CGS-EC, PD, small-world, $n = 40$, seed 0, $T = 50$). "—" = no recovery within T.

Shock	$\rho_C^{\min}$	ρ_C^{final}	Recov.	Recovered?
$f = 0.05$	0.525	0.675	0	yes
$f = 0.15$ (default)	0.475	0.603	8	yes
$f = 0.30$	0.375	0.513	—	no
$f = 0.50$	0.200	0.323	—	no
$f = 0.75$	0.075	0.198	—	no
persistent 5×0.15	0.100	0.222	—	no

VI. DISCUSSION

The results establish that grounding LLM deliberation within evolutionary cooperation theory produces qualitatively different, and substantially superior, dynamics compared to pure narrative-driven strategy updates. We organize the discussion around the following themes: the theoretical implications of the hybrid architecture (Section VI-A), the role of network structure (Section VI-B), the contribution hierarchy revealed by ablations (Section VI-C), the generality of cooperation dynamics across seeds, scales, and models (Section VI-D),

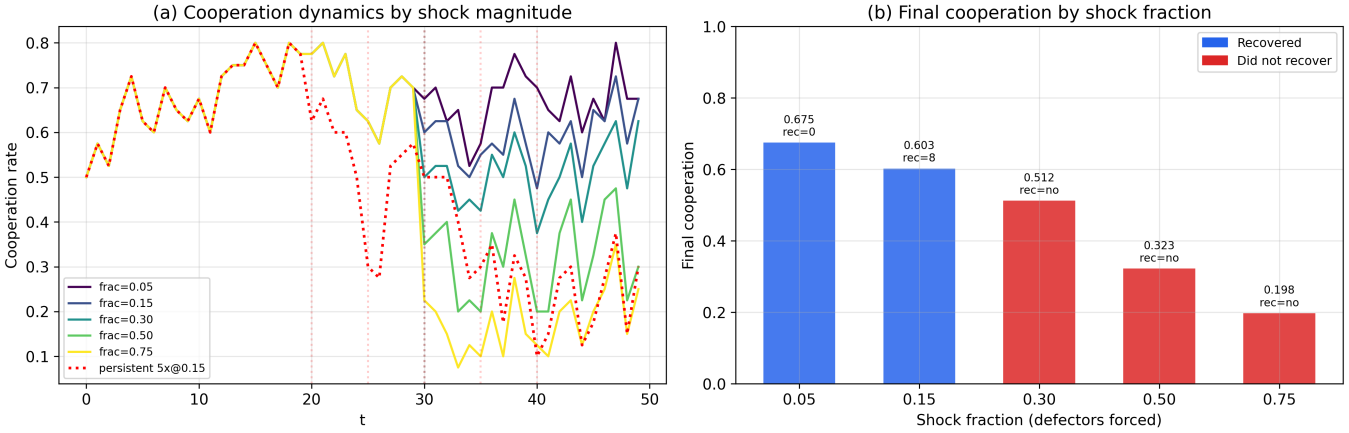

FIGURE 24: Failure-mode analysis (CGS-EC, PD, small-world, $n = 40$, seed 0). (a) Cooperation trajectories by shock magnitude and the persistent-shock regime. (b) Final cooperation by shock fraction; bars coloured by whether recovery occurred. CGS-EC recovers for $f_{\text{shock}} \leq 0.15$ and fails for $f_{\text{shock}} \geq 0.30$.

the contribution of the LLM advisory component relative to structural baselines (Section VI-E), computational cost and latency (Section VI-F), societal implications and risks (Section VI-G), and limitations of the current study (Section VI-H).

A. WHY GROUNDING MATTERS: THE COMPLEMENTARITY OF LLM REASONING AND EVOLUTIONARY DYNAMICS

The central finding of this paper, that EC grounding transforms LLM societies from static equilibria to dynamic, resilient cooperation, admits a precise mechanistic interpretation. The narrative-only baseline converges to a bimodal strategy distribution ($\sigma_o \in \{0, 1\}$) because the LLM's binary action recommendation, applied directly as a strategy update (11), lacks three properties that the EC-theory rule (6) provides.

First, *social learning* via the Fermi imitation component (7) introduces frequency-dependent selection: agents do not simply follow the LLM's recommendation but adjust their propensity based on whether cooperators or defectors are performing better in their local neighborhood. This creates a payoff-mediated feedback loop absent from narrative-only updates, enabling cooperation to *spread* through the network when it is locally advantageous rather than relying on each agent independently arriving at the same LLM-generated conclusion.

Second, *individual reinforcement* (9) maps each agent's experienced payoff to a cooperation target through a degree-normalized sigmoid. This provides a grounded self-assessment mechanism: agents who cooperate and receive high payoffs are reinforced toward cooperation, while those who cooperate but are exploited by defecting neighbors are pushed toward defection. The degree normalization (π_o/d_o) is critical, preventing high-degree agents from being spuriously driven toward cooperation simply because they accumulate more interactions.

Third, *trust heterogeneity* (β_o) within the LLM advisory component (10) allows the population to maintain a diversity of responses to LLM advice. Agents with low trust are effectively immune to LLM recommendations, acting as evolutionary anchors that stabilize the population against potential LLM errors. This heterogeneity is absent from the narrative-only baseline, where every agent applies the LLM signal with equal (and maximal) influence.

The implication for the broader LLM agent literature is significant. The generative agents paradigm [11] and its descendants produce emergent social behavior through narrative reasoning alone, which our results suggest is fundamentally limited in strategic settings. When agents face genuine payoff tradeoffs, as in the Prisoner's Dilemma, narrative reasoning without formal grounding produces convergence to static equilibria rather than the dynamic, adaptive cooperation that theory-grounded updates enable. This does not diminish the value of LLM reasoning; rather, it argues for *complementarity*: the LLM contributes context-sensitive, memory-informed advice that enriches the evolutionary update, while the evolutionary dynamics provide the selection pressures and feedback mechanisms that the LLM alone cannot supply.

B. TOPOLOGY AS A MODULATOR OF COOPERATION AND EQUITY

The cross-topology results confirm and extend classical findings from evolutionary game theory in the CGS context. The small-world topology produces the highest sustained cooperation ($\rho_C^{\text{final}} = 0.595$) and lowest inequality ($\mathcal{G} = 0.236$), consistent with Ohtsuki et al.'s [4] prediction that high clustering promotes cooperation through network reciprocity. Cooperative clusters in small-world networks are protected by their dense internal connections, making them resistant to invasion by defectors who emerge after the shock.

The scale-free topology amplifies cooperation at hubs while simultaneously creating the highest inequality ($\mathcal{G} = 0.420$), consistent with Santos and Pacheco's [5] finding that heterogeneous degree distributions can promote cooperation

but at the cost of concentrating payoffs. The $3.89\times$ hub-to-periphery payoff disparity we observe is a direct consequence of the accumulated-neighbor-payoff mechanism (2): agents with $d = 17$ interact with $4\times$ more neighbors than those with $d = 3$, and this structural advantage translates to proportional cumulative payoff advantages regardless of strategy. This finding has implications for the design of LLM agent societies: if equitable outcomes are a design objective, scale-free interaction topologies should be avoided or augmented with redistributive mechanisms.

The Erdös–Rényi topology provides an informative structural baseline. Its intermediate performance on most metrics confirms that the small-world and scale-free results are driven by their specific structural properties (clustering and degree heterogeneity respectively) rather than by artifacts of network size or density. The unique near-convergence observed on the random graph ($t = 20$–27, rolling std < 0.02) suggests that structural homogeneity facilitates equilibration, an observation that connects to the broader literature on mixing times in evolutionary dynamics on graphs [3].

An unexpected finding is the dissociation between recovery speed and sustained resilience. Scale-free networks show the fastest formal recovery (1 step) but the lowest post-recovery cooperation, while small-world networks recover more slowly (8 steps) but sustain higher cooperation afterward. This suggests that recovery time alone is an inadequate resilience metric; future work should incorporate sustained post-recovery performance into resilience evaluation.

C. THE CONTRIBUTION HIERARCHY: EC GROUNDING ≫ DELIBERATION QUALITY

The ablation study reveals a clear hierarchy of architectural contributions. Removing EC grounding (narrative-only ablation) reduces peak cooperation by 0.275 (34%), eliminates shock recovery, and produces strategy polarization. Removing the critic (no-critic ablation) reduces final cooperation by 0.027 (4.5%) and increases uncalibrated LLM confidence (0.778 vs. 0.698) but leaves all other metrics nearly unchanged. The ratio of these effects, roughly $10{:}1$ in favor of EC grounding, establishes a clear design priority for hybrid LLM-evolutionary systems: *get the evolutionary dynamics right first, then refine the deliberation pipeline*.

The critic's primary contribution is confidence calibration rather than action correction. Without the critic, the solver tends to produce overconfident recommendations, and the aggregator's self-correction mechanism (action flip on strong disagreement) is unavailable. The resulting higher confidence inflates the LLM advisory signal's influence within the EC-theory update, occasionally producing less stable dynamics. However, because the LLM component carries only $w_\ell = 0.25$ weight in the update rule, even uncalibrated confidence has limited impact on the overall system behavior.

This finding has practical implications for system designers: in resource-constrained settings where API calls must be minimized, the critic stage can be omitted with modest performance loss. Each critic call doubles the per-agent API cost (from 1 to 2 calls per step), which is significant at scale. The 4.5% cooperation reduction may be an acceptable tradeoff for 50% reduction in inference cost, depending on the application.

D. GENERALITY OF COOPERATION DYNAMICS ACROSS SEEDS, SCALES, AND MODELS

The multi-seed, scaling, multi-model, and GNN experiments collectively characterise the domain over which CGS dynamics generalise.

With ten seeds the cooperation-standard-deviation effect is significant after Bonferroni correction ($p_{\text{Bonf}} < 0.001$, Cohen's $d = 4.59$), and peak cooperation likewise ($p_{\text{Bonf}} = 0.015$).

Population scaling reveals that the core dynamics are stable across $n \in \{40, 60, 80\}$: pre-shock cooperation is scale-invariant within 0.014 of a percentage point, and payoff Gini remains below 0.255 at all tested sizes. The main scale-dependent factor is post-shock recovery time, which grows with the absolute number of shocked agents. This is a tractable practical consideration: for a target recovery time, the simulation horizon should scale proportionally with n.

The two-cluster outcome of the multi-model sweep shows that LLM choice has a significant quantitative but not qualitative effect on CGS dynamics. All four tested models produce cooperation above the best structural baseline (0.495), even the lower-performing DeepSeek-V3 and Qwen3-235B at 0.338. The cluster boundary appears to correlate with instruction-following quality in structured game-theoretic prompts, a dimension not captured by standard reasoning benchmarks and worth investigating directly in future model selection studies. The observation that confidence is uncorrelated with, and slightly negatively associated with, cooperation quality across models reinforces the interpretation from the no-critic ablation that the EC-theory grounding, not the deliberation pipeline's self-assessed confidence, is the primary determinant of cooperation outcomes.

E. THE CONTRIBUTION OF LLM ADVISORY BEYOND STRUCTURAL DYNAMICS

The GNN structural baselines provide a reference frame that the internal ablations of Section V-D cannot offer: they isolate the contribution of the LLM advisory component from the structural components of the update rule (6) without removing either.

GNN-Mean implements social learning without payoff sensitivity, achieving $\rho_C^{\text{final}} = 0.495$. GNN-PayW adds Fermi-weighted payoff sensitivity, but achieves only 0.408, lower than GNN-Mean, because payoff weighting without degree normalisation amplifies defector imitation during the post-shock period. GNN-2L, which augments the direct-neighbour readout with a two-hop neighbourhood signal, ties GNN-Mean at 0.495, indicating that richer structural aggregation does not compensate for the absence of the reinforcement and LLM components. The $+0.100$ gap from GNN-Mean to CGS-EC thus quantifies the cooperation premium

attributable to the combined effect of degree-normalised reinforcement and LLM context-sensitive advisory, net of any structural contribution.

This decomposition also informs the design of cost-efficient approximations. In settings where LLM API calls are unavailable, GNN-Mean achieves 83% of CGS-EC's final cooperation at zero per-step inference cost. In settings where API cost is a constraint but not prohibitive, replacing the full Solver–Critic–Aggregator pipeline with solver-only inference (the no-critic ablation of Section V-D2) incurs only a 4.5% cooperation reduction while halving the per-step API cost. These two checkpoints, GNN-Mean and EC-no-critic, provide a practical cost–performance frontier for hybrid LLM-evolutionary systems.

The CGS framework has direct relevance to real-world coordination settings. In industrial autonomous systems, the EC-grounded update can serve as a principled behavioural substrate for LLM-assisted decision support agents operating under resource and communication constraints [23]. In e-commerce and service environments, where multiple AI agents must coordinate pricing, recommendations, or logistics under competitive incentives, the topology-modulated cooperation dynamics reported here offer a simulation testbed for evaluating governance mechanisms [36]. Beyond these settings, CGS offers a natural testbed for *cybersecurity multi-agent systems* [37], [38], where defender and attacker agents interact on a network and the tension between cooperation (information sharing) and defection (hoarding intelligence) maps directly onto the Prisoner's Dilemma substrate; and for *economic simulations*, where market participants must balance competitive self-interest against the collective gains of coordinated pricing or resource allocation [21].

To facilitate experimentation and adoption, we open-source a browser-based simulator that runs CGS directly and extends to a broader family of co-evolutionary LLM multi-agent governance simulations, powered by the Nebius AI Studio API for open-weight model serving, allowing practitioners to interactively explore cooperation dynamics, network topologies, and shock protocols[3].

Figure 25 shows the CGS tab of the simulator, displaying the real-time network state and cooperation trajectory for a Watts–Strogatz topology with $n = 40$ agents.

F. COMPUTATIONAL COST AND LATENCY

Table 17 reports wall-clock time and API-call counts for CGS-EC at $n \in \{20, 40, 60\}$ and for the no-critic variant at $n = 40$. The pipeline issues exactly two API calls per agent per step (solver + critic) and costs ≈ 2.7 s of wall time per agent per step under the Nebius AI Studio API. The no-critic variant at $n = 40$ requires 2 000 calls and 2 072 s versus 4 000 calls and 5 422 s for the full pipeline: the critic stage *exactly doubles* the API cost and *roughly doubles* wall time, for the $+4.5\%$ cooperation gain quantified in Section V-D2. This makes the cost–benefit trade-off explicit: practitioners under strict API budgets can drop the critic for a modest, quantified cooperation loss. Cost grows linearly in n; extrapolating the per-agent-step figure, $n = 1{,}000$, $T = 50$ would require $\approx 10^5$ API calls and ≈ 37 h of wall time, motivating the selective-LLM- invocation strategies noted in Section VI-H.

[3]https://github.com/drdezarza/societal_dynamics

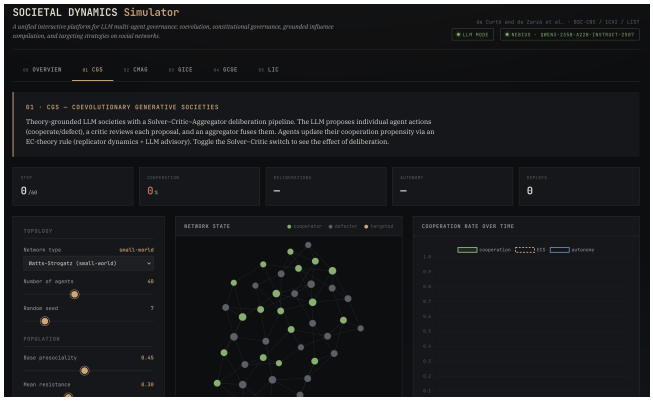

FIGURE 25: The *SOCIETAL DYNAMICS* browser simulator (CGS tab). Left panel: configurable topology and population parameters. Centre: live network state with cooperators (green) and defectors (grey). Right: cooperation rate, ECS score, and autonomy index over time. Five simulation narratives are accessible via the top tabs.

TABLE 17: Computational cost (CGS-EC, small-world, seed 0, $T = 50$). Calls/(agent·step)$= 2.0$ for the full pipeline.

Config	Wall (s)	API calls	s/agent·step	ρ_C^{final}
$n = 20$, full	2 194	2 000	2.19	0.585
$n = 40$, full	5 422	4 000	2.71	0.597
$n = 60$, full	8 069	6 000	2.69	0.592
$n = 40$, no-critic	2 072	2 000	1.04	0.573

G. SOCIETAL IMPLICATIONS AND RISKS

Theory-grounded LLM societies raise concrete deployment risks that merit explicit statement. First, *manipulation*: because the LLM advisory can shift population cooperation (the OAT sweep shows mean trust $\bar{\beta}$ monotonically moves both cooperation and equity), an adversary controlling the advisory channel could steer collective behaviour; the EC-grounding's $w_s + w_r = 0.75$ structural floor bounds but does not eliminate this exposure. Second, *inequality amplification*: on scale-free topologies CGS concentrates payoff at hubs (Gini $= 0.420$, a $3.89\times$ hub–periphery disparity), so deploying such societies on heterogeneous interaction graphs can entrench structural advantage; equitable deployment requires either homogeneous topologies or explicit redistributive mechanisms. Third, *failure under stress*: the resilience boundary (Section V-M) shows cooperation is not robust to severe or repeated shocks, so safety-critical coordination should not rely on emergent cooperation alone without external guarantees. We recommend that deployments expose the advisory channel for audit, monitor the payoff Gini as a fairness indicator, and treat the empirical resilience boundary

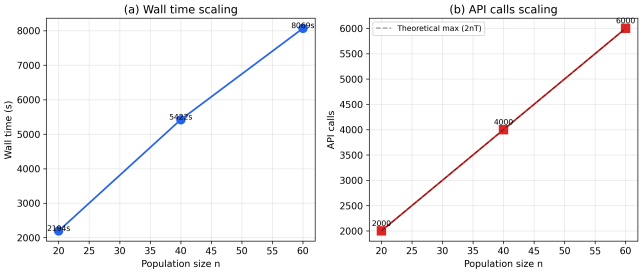

FIGURE 26: Computational cost scaling (CGS-EC, small-world). (a) Wall time vs. n, (b) API calls vs. n (linear, $2nT$), Cost is linear in population size; the critic stage accounts for exactly half of both wall time and API calls.

as an operating constraint rather than a worst case.

H. LIMITATIONS

Several limitations should be acknowledged. First, our experiments use a population size ($n = 40$) and simulation horizon ($T = 50$). While sufficient to demonstrate the core phenomena, scalability to hundreds or thousands of agents remains untested. The $\mathcal{O}(n{\cdot}T)$ LLM API calls required (4,000 for the full pipeline with $n = 40$, $T = 50$, doubling with the critic) may become prohibitive at larger scales, motivating investigation of selective LLM invocation strategies where only a subset of agents consults the deliberation pipeline at each step. The cost analysis of Section VI-F quantifies this: cost is linear in n at ≈ 2.7 s and 2 API calls per agent per step, so $n > 10^3$ is feasible only with selective LLM invocation.

Second, the multi-model sweep confirms that deliberation quality varies substantially across LLM families: Llama-70B and Hermes-4-70B achieve final cooperation above 0.57, while DeepSeek-V3 and Qwen3-235B reach only 0.338. The temperature robustness analysis of Section V-I confirms that CGS dynamics are insensitive to sampling temperature across $T_s \in \{0.0, 0.2, 0.5, 0.8\}$, closing this gap. The critic's modest marginal contribution over the no-critic ablation (4.5%) likely reflects the high baseline strategic reasoning quality of Llama-3.3-70B; models with weaker instruction-following in game-theoretic contexts may benefit more substantially from the critic's confidence correction mechanism.

Third, the component weights ($w_s = 0.40$, $w_r = 0.35$, $w_\ell = 0.25$) and hyperparameters ($\kappa = 2$, $\alpha = 0.8$, $\theta = 2.0$) were set based on preliminary experimentation. The OAT sensitivity analysis of Section V-I confirms that κ and w_ℓ have negligible influence on outputs (NSI < 0.1), validating the chosen values post hoc. α is moderately sensitive (NSI $= 0.14$) and systematic optimisation of the reinforcement inflection point θ remains an avenue for future work. A closed-form fixed-point characterisation is unavailable for instruction-tuned LLMs; we instead provide an empirical attractor map (Section V-L) revealing a critical-mass bifurcation near $\mu_p \approx 0.35$. A formal proof remains open. Formal convergence (rolling std < 0.02) is observed on the Erdös–Rényi topology during $t = 20$–27 (Section V-A3).

Fourth, CGS-EC recovers from single shocks affecting up to $\sim 15\%$ of agents but not from shocks $\geq 30\%$ or rapidly repeated shocks within $T = 50$ (Section V-M); the architecture is not designed for severe or sustained adversarial regimes without external guarantees.

Finally, the current architecture treats all agents as invoking the same underlying LLM. An extension to heterogeneous LLM populations, where agents differ not only in trust β_o but in the model they consult, would better reflect realistic deployment scenarios and could reveal emergent phenomena arising from model diversity.

VII. CONCLUSION

We introduced Coevolutionary Generative Societies (CGS), a hybrid architecture that integrates LLM deliberation with evolutionary cooperation theory for multi-agent coordination under social dilemmas. The core contribution is a three-component strategy update rule that fuses Fermi social learning, degree-normalized individual reinforcement, and trust-weighted LLM advisory signals produced by a Solver–Critic–Aggregator deliberation pipeline. This design allows each mechanism to contribute what it does best: evolutionary dynamics provide payoff-grounded selection pressures and frequency-dependent feedback, while the LLM contributes context-sensitive, memory-informed reasoning that enriches fixed update rules with adaptive intelligence.

Experiments across three network topologies (Watts–Strogatz small-world, Barabási–Albert scale-free, Erdös–Rényi random) with $n = 40$ agents over $T = 50$ steps demonstrate that CGS achieves peak cooperation of 0.80 and recovers from adversarial shocks within 8 steps on small-world networks. The topology sweep reveals that small-world networks sustain the highest cooperation (0.595) with the lowest inequality (Gini $= 0.236$), while scale-free networks amplify hub–periphery payoff disparities ($3.89\times$) despite reaching comparable peak cooperation. The Erdös–Rényi topology occupies an intermediate position and is the only network achieving near-convergence before the shock, consistent with its structural homogeneity.

The ablation studies deliver the paper's central message: *EC grounding is essential, not optional*. Removing the evolutionary update and relying on pure LLM-driven strategy revision reduces peak cooperation by 34%, eliminates shock recovery entirely, and produces bimodal strategy polarization, a qualitatively different and inferior regime. The no-critic ablation reveals a secondary but non-negligible role for deliberation quality, with confidence calibration as the critic's primary contribution. The resulting $10{:}1$ contribution hierarchy (EC grounding vs. deliberation refinement) provides clear architectural guidance for future hybrid systems.

Ten-seed replication establishes that the qualitative distinction between CGS-EC and the narrative-only baseline is statistically significant after Bonferroni correction (Cohen's $d = 4.59$ on cooperation responsiveness, $p_{\text{Bonf}} < 0.001$): CGS-EC remains adaptive in nine of ten seeds whereas the narrative-only baseline converges to a static equilibrium in

ten of ten. Population scaling demonstrates that cooperation dynamics and payoff equity are consistent across $n \in \{40, 60, 80\}$, with post-shock recovery time as the primary scale-dependent factor. The multi-model sweep reveals that LLM choice has a significant quantitative but not qualitative effect on CGS dynamics: all four tested models sustain cooperation above the best structural baseline, and the EC-theory grounding provides insulation against lower-quality advisory signals through the $w_s + w_r = 0.75$ structural weight floor. Comparison against three GNN structural baselines quantifies the LLM advisory premium at $\Delta\rho_C = +0.100$ ($+20.2\%$ relative to GNN-Mean), establishing that the LLM advisory component contributes a meaningful cooperation gain beyond what graph message-passing dynamics alone can achieve, and decomposing the CGS-EC advantage into structural and advisory sub-contributions that ablation studies alone cannot resolve. The architecture generalises across PD, Snowdrift and Stag Hunt without re-tuning, exceeds all non-LLM imitation and replicator baselines (pure Fermi imitation collapses to 0.11), and exhibits a quantified resilience boundary (recovery up to 15% shocks) and a critical-mass cooperation bifurcation near $\mu_p \approx 0.35$.

Future work will extend CGS along several axes: scaling to larger populations ($n = 200$–$1{,}000$) with selective LLM invocation, n-player and asymmetric dilemmas (public goods, bargaining), introducing coevolutionary topology adaptation [22], and empirical grounding through human-agent experiments. Structured narrative interventions and constitutional governance mechanisms represent further avenues for building LLM agent societies that are not only capable of emergent coordination but also transparent, equitable, and governable.

ACKNOWLEDGMENT

During the preparation of this work, the author(s) used ChatGPT/Copilot/Gemini/Claude for code completion and pseudocode generation, Grammarly for English language editing, and Mendeley for bibliography management. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of this publication. The authors would like to thank the BARCELONA Supercomputing Center for providing access to MareNostrum 5 and technical support throughout this research. This research was supported by the LUXEMBOURG Institute of Science and Technology through the projects 'ADIALab-MAST" and 'LLMs4EU" and the BARCELONA Supercomputing Center through the project 'TIFON".

CODE AND SOFTWARE AVAILABILITY

The complete CGS simulation framework, including all experiment scripts, configuration files, and analysis notebooks, is publicly available at: https://github.com/drdezarza/cgs

An interactive browser-based simulator for LLM multi-agent governance is available at: https://github.com/drdezarza/societal_dynamics

The simulator allows real-time exploration of cooperation dynamics under configurable network topologies, shock protocols, and LLM advisory parameters, and supports the broader family of multi-agent governance experiments described in this work.

REFERENCES

[1] R. Axelrod, *The Evolution of Cooperation*. New York: Basic Books, 1984.
[2] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, 2006.
[3] G. Szabó and G. Fáth, "Evolutionary games on graphs," *Physics Reports*, vol. 446, no. 4-6, pp. 97–216, 2007.
[4] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, "A simple rule for the evolution of cooperation on graphs and social networks," *Nature*, vol. 441, no. 7092, pp. 502–505, 2006.
[5] F. C. Santos and J. M. Pacheco, "Scale-free networks provide a unifying framework for the emergence of cooperation," *Physical Review Letters*, vol. 95, no. 9, p. 098104, 2005.
[6] G. Szabó and C. Tőke, "Evolutionary Prisoner's Dilemma game on a square lattice," *Physical Review E*, vol. 58, no. 1, p. 69, 1998.
[7] A. Traulsen, J. M. Pacheco, and M. A. Nowak, "Pairwise comparison and selection temperature in evolutionary game dynamics," *Journal of Theoretical Biology*, vol. 246, no. 3, pp. 522–529, 2007.
[8] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, "Statistical physics of human cooperation," *Physics Reports*, vol. 687, pp. 1–51, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157317301424
[9] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
[11] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.
[12] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2024.
[13] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi, "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint arXiv:2305.19118*, 2024.
[14] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative agents for "mind" exploration of large language model society," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[15] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.
[16] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.
[17] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz, "Playing repeated games with large language models," *arXiv preprint arXiv:2305.16867*, 2023.
[18] N. Fontana, F. Pierri, and L. M. Aiello, "Nicer than humans: How do large language models behave in the prisoner's dilemma?" *arXiv preprint arXiv:2406.13605*, 2024.
[19] C. Fan, J. Chen, Y. Jin, and H. He, "Can large language models serve as rational players in game theory? a systematic analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17 960–17 967, 2024.
[20] G. Piatti, Z. Jin, M. Kleiman-Weiner, B. Schölkopf, M. Sachan, and R. Mihalcea, "Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents," *arXiv preprint arXiv:2404.16698*, 2024.
[21] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press, 1998.
[22] M. Perc and A. Szolnoki, "Coevolutionary games—a mini review," *BioSystems*, vol. 99, no. 2, pp. 109–125, 2010.

[23] H. Jiang, J. You, Z. Chen, J. Shi, H. Gou, X. Ming, and P. Z. Sun, "Llm-enhanced intent-aware for proactive decision support services in industrial activities," *IEEE Transactions on Automation Science and Engineering*, 2026.
[24] J. de Curtò and I. de Zarzà, "Llm-driven social influence for cooperative behavior in multi-agent systems," *IEEE Access*, vol. 13, pp. 44 330–44 342, 2025.
[25] I. de Zarzà, J. de Curtò, G. Roig, P. Manzoni, and C. T. Calafate, "Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with llms," *Electronics*, vol. 12, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2722
[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
[27] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
[28] P. Erdős and A. Rényi, "On random graphs I," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.
[29] F. C. Santos, J. M. Pacheco, and T. Lenaerts, "Cooperation prevails when individuals adjust their social ties," *PLoS Computational Biology*, vol. 2, no. 10, p. e140, 2006.
[30] C. Hauert and M. Doebeli, "Spatial structure often inhibits the evolution of cooperation in the snowdrift game," *Nature*, vol. 428, no. 6983, pp. 643–646, 2004.
[31] S. Begum, M. R. Islam, and K. A. Kabir, "Q-learning driven adaptive decision rules and environmental transformation in three strategy evolutionary games," *Chaos, Solitons & Fractals*, vol. 207, p. 118052, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077926001931
[32] K. Ariful Kabir, "Behavioral vaccination policies and game-environment feedback in epidemic dynamics," *Scientific Reports*, vol. 13, no. 1, p. 14520, 2023.
[33] K. A. Kabir, T. Risa, and J. Tanimoto, "Prosocial behavior of wearing a mask during an epidemic: an evolutionary explanation," *Scientific Reports*, vol. 11, no. 1, p. 12621, 2021.
[34] K. A. Kabir, "How evolutionary game could solve the human vaccine dilemma," *Chaos, Solitons & Fractals*, vol. 152, p. 111459, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077921008134
[35] P. Brookins and J. M. DeBacker, "Playing games with GPT: What can we learn about a large language model from canonical strategic games?" *Available at SSRN 4493398*, 2023.
[36] S. K. Mittameedi, V. Dogra, and A. Sayal, "Customer experience in e-commerce: A systematic review of metrics, models, and the role of ai," *IEEE Access*, 2025.
[37] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, "When llms meet cybersecurity: A systematic literature review," *Cybersecurity*, vol. 8, no. 1, p. 55, 2025.
[38] M. M. Yamin, E. Hashmi, M. Ullah, and B. Katt, "Applications of llms for generating cyber security exercise scenarios," *IEEE Access*, vol. 12, pp. 143 806–143 822, 2024.

J. DE CURTÒ received the M.Sc. degree in Telecommunication Engineering from Universitat Autònoma de Barcelona and Universitat Politècnica de Catalunya in 2013, and the M.Sc. degree in Multimedia Information Technology (with Distinction) from City University of Hong Kong in 2015. He earned his Ph.D. degree in Computer Science from Universitat Politècnica de València (UPV) in 2023.

From 2014 to 2015, he conducted research at the Robotics Institute at Carnegie Mellon University, Pittsburgh, PA; and afterwards from 2015 to 2016 at ETH Zürich in Switzerland. Since 2024, he has been a tenured Recognised Researcher (R2) in the Department of Computer Applications in Science & Engineering, Group of Dual Technologies, at the BARCELONA Supercomputing Center - Centro Nacional de Supercomputación. He is also a Professor (Profesor Asociado Colaborador - Doctor) in the School of Engineering (ICAI), Department of Electronics, Control Systems and Communications, at Universidad Pontificia Comillas in Madrid.

Prof. Dr. De Curtò is the author of more than 40 articles published in indexed journals and internationally recognized conferences on topics including Vehicular Technologies, IoT, AI, and Space Mission Design. His research interests include Large Language Models for Unmanned Aerial Robot Control and Communication, Advanced Computer Vision / Control Techniques for Autonomous Navigation, Robotic Systems Integration and Design for Enhanced Robot Performance, and Space Mission Design.

He is also an editor (Early Career Editorial Board since 2025; Topical Advisory Panel - Innovative Urban Mobility since 2024) for MDPI Drones. Since May 2025, he also serves as Topic Editor of Frontiers in Artificial Intelligence. Prof. Dr. De Curtò has received various national and international awards, including the Top Achiever 2015 award from City University of Hong Kong for his academic performance in the master's program.

I. DE ZARZÀ received the B.S. degree in Mathematics from Universitat Autònoma de Barcelona in 2013, and the M.S. degree in Multimedia Information Technology from City University of Hong Kong in 2015. She earned her Ph.D. degree in Computer Science from Universitat Politècnica de València in 2023.

Since 2025, she is a Research &Technology Scientist at the LUXEMBOURG Institute of Science and Technology, Human Centered AI, Data & Software. Previously, she was a Professor in the Department of Computer Science and Systems Engineering at the Universidad de Zaragoza (2024-2025). Before, she was an Assistant Professor at the Escuela Politécnica Superior at Universidad Francisco de Vitoria in Madrid (2024). She has also held research positions at various prestigious institutions including the Institute for Computer Science and Mathematics, GOETHE-University in Germany, the Computer Vision Laboratory at ETH Zürich in Switzerland, and the Department of Computer Science and Engineering at The Chinese University of Hong Kong. As well as at the Robotics Institute, Carnegie Mellon, Pittsburgh, PA, USA.

Dr. de Zarzà's research interests include the development and application of Large Language Models for computational social science, behavioral understanding and new economic theories where AI plays an important role; she also works on drone navigation and communication, advanced computer vision and robotics techniques for object recognition and manipulation, and the integration of AI and robotics in complex engineering systems.

She is the author of numerous publications in indexed journals and national and international conferences. Prof. Dr. de Zarza is also a member of professional organizations such as IEEE, ACM, SIAM, and KES.