EMERGENT EQUILIBRIUM: A MEASURABLE THEORY OF STRATEGIC CONVERGENCE IN LLM MULTI-AGENT DIALOGUE, WITH A CLOSED-FORM CRITICAL-TEMPERATURE BOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

When multiple LLM agents debate, negotiate, or deliberate, do their dialogues converge—and when do they fail to? To this end, we give a formal answer and derive a closed-form condition for the transition. We formalize multi-agent dialogue as a discrete dynamical system on a semantic belief manifold, construct an explicit dialogue operator from cross-attention layers, and introduce the Ideal Group Introspection (IGI) framework—a two-parameter learnable model whose residual error yields the first measurable criterion for strategic convergence wherever those residuals are identifiable. Within this framework, we prove that an *approximate* emergent equilibrium—satisfying Nash strategic stability, geometric dialogue convergence and consensus coupling—exists under mild regularity conditions, via a sandwich argument linking Glicksberg and Brouwer, the profile delivered being δ-close to the stage game's quantal response equilibrium, and Nash as $\beta \to \infty$; we establish spectral-radius convergence conditions and an attention contraction lemma, and show that the dialogue undergoes a phase transition whose critical sampling temperature is bounded above by the closed form $C/(1-L)$, where $C = \sigma_V \cdot \sigma_K \cdot N/\sqrt{d_k}$ (σ_V, σ_K the value and key projection norms, N the agents, d_k the key dimension) is the architectural constant that the reduced-operator limit $L \to 0$ recovers exactly. Crucially, this makes a one-sided bound on instability onset computable from stored weights under one decomposition assumption, and predicts critical slowing near T_c—a prediction our sweeps do not confirm, a null result we report with its power. Validating across 15 models (four families, 0.6B to trillion-parameter scale) in strategic and tax-compliance games, we uncover three regularities that our theory explains. First, in the compliance framing frontier closed-source models exhibit sycophantic cooperation—complying even when defection is strictly dominant, while defecting in the abstract game—and so fall outside the IGI regime there; because this deviation is systematic rather than random, it defines an interpretable four-way taxonomy of model behaviour instead of a failure of the theory. Second, strategic behaviour is protocol-dependent rather than a fixed model property: a single checkpoint spans both endpoints, from complete cooperation in the tax framing to zero against randomised opponents. Third, cooperation varies non-monotonically with model scale within the mid-size regime, where strategic diversity is highest. Taken together, our results provide a rigorous, empirically testable foundation for predicting when LLM multi-agent deliberation is strategically meaningful—and when it is not.

1 INTRODUCTION

When multiple LLM agents converse—debating, negotiating, deliberating—do they converge? Physical systems trend towards thermal equilibrium; strategic agents towards Nash equilibrium. What do *LLM dialogues* trend towards? These questions motivate this work, with tax compliance—multiple AI agents modelling the strategic interaction between taxpayers and revenue authorities—as our domain application.

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

[1]

Despite the rapid deployment of multi-agent LLM systems for debate (Du et al., 2024), consensus-seeking (Reza, 2025), and enterprise workflows (Acharya, 2026), the theoretical foundation is absent: we do not know when such dialogues converge, whether the limit is strategically rational, or how behaviour depends on system parameters. Building this theory is hard because LLMs are black-box functions, multi-agent interaction is nonlinear and cyclic, and the problem demands simultaneous treatment of *semantic convergence* (do beliefs stabilise?) and *strategic convergence* (do strategies form a Nash equilibrium?).

We address this gap by formalising multi-agent LLM dialogue as a discrete dynamical system on a **semantic belief manifold** $\mathcal{M}$—a compact subset of the LLM embedding space. We construct the **dialogue operator** $\mathbf{\Phi}$ from cross-attention layers (Definition 4.4) and define **emergent equilibrium** as the simultaneous satisfaction of Nash strategic stability, geometric dialogue convergence, and consensus coupling (Definition 5.1). Central to our approach is the **IGI operator** $\Gamma_{\eta,\beta}$ (Definition 4.6), a two-parameter learnable generative model of strategic dialogue whose residual error $\varepsilon^*_{\text{IGI}}$ measures model misspecification. We prove existence (Theorem 5.2), characterise convergence via the spectral radius $\rho(J_{\mathbf{\Phi}})$ (Theorem 5.3), derive architectural contraction conditions (Lemma 5.4), and define the **emergent temperature** β_{emerge} which governs a provable phase transition between chaotic, emergent, and frozen dialogue regimes (Theorem 6.3).

Our **contributions** are:

1. **An operator, and an equilibrium.** We construct the IGI operator, a two-parameter learnable generative model of strategic dialogue whose misspecification gap $\varepsilon^*_{\text{IGI}}$ is measurable, and we prove that the emergent equilibrium it induces exists and is unique.
2. **A boundary computable from the weights.** We establish spectral-radius convergence conditions and an attention contraction lemma that tie transformer architecture to dialogue contractivity, and we derive an emergent temperature whose critical value is bounded above by a closed form computable from stored weights under one decomposition assumption, whose looseness we then measure. To our knowledge it is the first such boundary; the criterion it defines is usable where a weight-based fit converges and the residual is identifiable (Section 7)—two open checkpoints on the released fits, none closed.
3. **An empirical account across 15 models.** We validate the theory on 15 models (four families, 0.6B to trillion parameters), including a tax-compliance application; in doing so we identify a four-quadrant taxonomy of the ways models depart from strategic rationality, and we show that strategic behaviour is protocol-dependent rather than a fixed model property.

2 Related Work

Our theory sits at the intersection of multi-agent LLM dialogue, game-theoretic modelling and convergence theory for deep architectures; we close with an explicit comparison.

Multi-agent LLM dialogue. Du et al. (2024) pioneered multi-agent debate, showing that LLMs refine answers through iterative argument exchange. Later work studied consensus-seeking dynamics (Chen et al., 2023; Reza, 2025), the consensus–diversity tradeoff (Wu & Ito, 2025), consensus-free debate (Cui et al., 2025), agreement bias (Wang et al., 2025), voting reliability (Patel et al., 2026), representational collapse and behavioural drift (Patel, 2026; Rath, 2026), mediated and enterprise consensus (Duan & Wang, 2024; Jannelli et al., 2024), and whether debate really reasons (Wu et al., 2025). These works treat consensus as an engineering goal or a measurement target; we give the conditions under which it is a *stable* outcome, and show that consensus, Nash stability and dialogue self-consistency are coupled conditions on one object.

Game-theoretic modelling. Nash equilibrium appears as a normative target for alignment (Nash learning from human feedback (Munos et al., 2023), two-player alignment games (Zheng et al., 2024), self-play training (Liu et al., 2025)) and as a descriptive construct for LLM behaviour (Yi et al., 2025; Kang, 2026; Willis et al., 2025; Mallampati et al., 2025), alongside multilevel interactive equilibrium for NeuroAI systems (Chen & Zhu, 2026) and incentive design for multi-LLM

[1] Every number here comes from the released scripts and stored result files; Appendix D traces each one.

cooperation (Liang et al., 2025). We neither optimise towards Nash nor merely document it: under the IGI operator, Nash stability *emerges from* the dialogue dynamics jointly with consensus and convergence. Fixed-point theory for *single* deep models (Bai et al., 2019; Halloran et al., 2024; Cowsik et al., 2024) supplies the tools we reuse (McKelvey & Palfrey, 1995; Glicksberg, 1952; Berinde, 2007); the interacting case is missing—cross-attention between agents must be shown to contract on the joint belief manifold (§5).

Behaviour in repeated games, and framing. The phenomena we classify have antecedents in behavioural evaluations of LLMs. Akata et al. (2025) let frontier models play finitely repeated 2×2 games and find them largely unforgiving, with behaviour shifting when the opponent is described differently; Sharma et al. (2024) characterise under the name *sycophancy* the tendency our taxonomy calls cooperation-saturation; and Sclar et al. (2024) quantify how far behaviour moves under meaning-preserving prompt-format changes alone. Regularities 1 and 2 are instances of those effects in a strategic game; what we add is the axis—a boundary computable from weights plus a measurable misspecification gap—along which their magnitude can be anticipated before a dialogue is run.

2.1 What is new relative to the closest work

Table 24 in Appendix F.6 places the nearest lines of work on four questions: is there a *criterion* for when dialogue converges, what *equilibrium concept* is used, is there a computable *critical boundary*, and are the ways a model *fails* to be strategic classified? The paragraph below states where we differ in kind.

The nearest line of work is the empirical-instability row: thresholds have been observed empirically and attributed to order parameters (Itkin, 2026; De Nobili, 2026). Our claim differs in kind: the criterion is *derived*, the boundary is *computable from stored weights* under one decomposition assumption before any dialogue is run (§6.2), and failures are *classified rather than averaged away* (§7.3).

3 Preliminaries

3.1 Multi-Agent LLM Dialogue

Consider N LLM agents with belief states $s_i^t \in \mathcal{M} \subset \mathbb{R}^d$ over rounds $t = 0, 1, \ldots$. Agent i's belief update is $s_i^{t+1} = \Phi_i^t(s_1^t, \ldots, s_N^t)$. The joint dialogue operator $\mathbf{\Phi} : \mathcal{M}^N \to \mathcal{M}^N$ yields the dynamical system $\mathbf{s}^{t+1} = \mathbf{\Phi}(\mathbf{s}^t)$.

3.2 Transformer Attention

Cross-attention provides the architectural basis for Φ_i^t. With query/key/value projections $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$: $Q_i = s_i^t W_Q$, $K_j = s_j^t W_K$, $V_j = s_j^t W_V$, $s_i^{t+1} = \sum_{j=1}^N \mathrm{softmax}(Q_i K_j^\top / \sqrt{d_k})_j \, V_j$.

3.3 Game Theory

A normal-form game $\Gamma = (N, \mathcal{A}, \{u_i\})$ (Fudenberg & Tirole, 1991). A strategy profile $\boldsymbol{\pi}^*$ is a Nash equilibrium if $u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i, \pi_{-i}^*)$ for all i, π_i. The quantal response (QR) function (McKelvey & Palfrey, 1995) smooths best response: $\mathrm{QR}_i^\beta(\pi_{-i})[a] \propto \exp(\beta \cdot u_i(a, \pi_{-i}))$, with $\beta = 1/T_{\mathrm{LLM}}$ linking QR to LLM sampling temperature. Its fixed point is the *quantal response equilibrium* (QRE) of the game, the solution concept we compare against.

3.4 Fixed-Point Theorems

We rely on three classical results (proofs omitted): **Brouwer**: continuous $F : X \to X$ on compact convex X has a fixed point. **Glicksberg**: compact-action games with continuous payoffs possess mixed Nash equilibria (Glicksberg, 1952). **Banach**: contractions on complete metric spaces have unique fixed points with linear convergence.

Notation follows standard conventions.

4 IGI: A LEARNABLE MODEL OF STRATEGIC DIALOGUE

We formalise the IGI framework, which treats multi-agent LLM dialogue as a learnable dynamical system. Each agent's strategic behaviour is compressed into two learned parameters—a learning rate η and a rationality parameter β—making the IGI operator a generative model of dialogue.

4.1 SEMANTIC BELIEF MANIFOLD

Transformer latent representations inhabit a bounded subset of $\mathbb{R}^d$ owing to layer normalisation. We formalise this as a compact convex set: a closed ball, for which we keep the descriptive name.

Definition 4.1 (Semantic Belief Manifold)**.** *Let d be the embedding dimension, $R > 0$ the maximal norm, and $W_{unembed} \in \mathbb{R}^{d\times|\mathcal{V}|}$ the language model head. The **Semantic Belief Manifold** is*

$$\mathcal{M} \triangleq \{s \in \mathbb{R}^d : \|s\|_2 \le R\}, \tag{1}$$

equipped with the symmetric divergence $d_{\mathcal{M}}(s,s') = \alpha(1-\cos(s,s'))+(1-\alpha)\cdot\frac{1}{2}\big[D_{\mathrm{KL}}(p_s \parallel p_{s'})+D_{\mathrm{KL}}(p_{s'} \parallel p_s)\big]$, the symmetrised (Jeffreys) divergence—a proximity, not a metric, see Appendix A, so that $d_{\mathcal{M}}$ is symmetric, where $p_s = \mathrm{softmax}(W_{unembed}^{\top}s)$ and $\alpha = 0.5$.

4.2 CROSS-ATTENTION, STRATEGY, AND DIALOGUE DYNAMICS

Definition 4.2 (Cross-Attention Belief Update)**.** *Agent i's belief is updated via cross-attention over other agents' mean belief $\overline{s}_{-i}^{(t)} = \frac{1}{N-1}\sum_{j\neq i} s_j^{(t)}$:*

$$\tilde{s}_i^{(t+1)} = \mathrm{softmax}\Big(\frac{Qs_i^{(t)}\cdot K^{\top}\overline{s}_{-i}^{(t)}}{\sqrt{d_k}}\Big)V\overline{s}_{-i}^{(t)}, \tag{2}$$

where $Q, K, V \in \mathbb{R}^{d\times d_k}$ are shared projection matrices. This map is continuous and preserves $\mathcal{M}^N$ in the projected form of Lemma A.1 (Appendix A).

Definition 4.3 (Strategy Operator)**.** *$\sigma_i : \mathcal{M} \to \Delta(A)$ maps belief states to action distributions via prompted LLM generation with temperature T_{LLM}.*

Definition 4.4 (Dialogue Operator)**.** *$\Phi : \mathcal{M}^N \times \Delta(A)^N \to \mathcal{M}^N \times \Delta(A)^N$: $s_i^{(t+1)} = \mathrm{CrossAttn}(s_i^{(t)}, \overline{s}_{-i}^{(t)})$, $\pi_i^{(t+1)} = \sigma_i(s_i^{(t+1)})$.*

Theorem 4.5 (Dialogue Fixed Points)**.** *Φ has at least one fixed point in $\mathcal{M}^N \times \Delta(A)^N$.*

Proof sketch: $\mathcal{M}^N$ is compact convex; Φ is continuous; apply Brouwer's fixed point theorem (full proof: Appendix A).

4.3 THE IGI OPERATOR

Figure 1(a) summarises the construction.

Definition 4.6 (IGI Operator)**.** *$\Gamma_{\eta,\beta}(s) \triangleq \mathrm{Enc}_\eta \circ \mathrm{QR}_\beta \circ \mathrm{Dec}(s)$, where Dec maps beliefs to expected payoffs, $\mathrm{QR}_\beta(x) = \mathrm{softmax}(\beta x)$ is quantal response (McKelvey & Palfrey, 1995), and Enc_η applies exponential smoothing with rate η.*

Theorem 4.7 (IGI Fixed Points Induce Quantal Response Equilibria)**.** *If π^* is a fixed point of $\Gamma_{\eta,\beta}$ for $\eta \in (0,1), \beta > 0$, then each π_i^* is a quantal response equilibrium of the stage game.*

Proof sketch: At a fixed point, $\pi_i^* = \eta\cdot\mathrm{QR}_\beta(\mathrm{Dec}(s^*)) + (1-\eta)\pi_i^*$, hence $\pi_i^* = \mathrm{QR}_\beta(\mathrm{Dec}(s^*))$ (full proof: Appendix A).

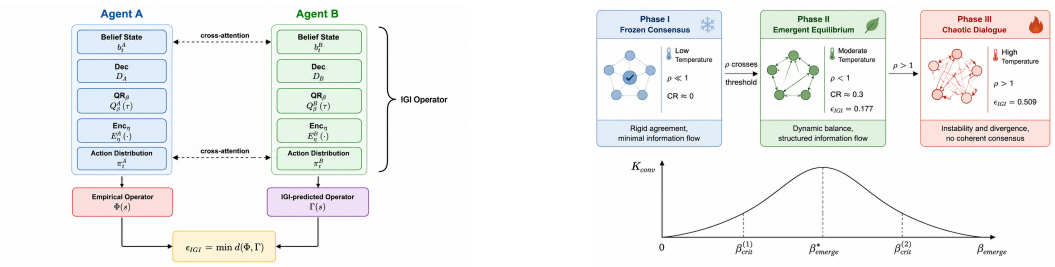


Figure 1: The framework in one picture. **(a)** The IGI operator $\Gamma_{\eta,\beta}$: beliefs are decoded (Dec) to expected payoffs, mapped through the quantal response QR_β and the smoothing encoder Enc_η; dashed arrows are cross-attention coupling, and $\varepsilon_{\mathrm{IGI}}$ measures the gap between the empirical operator Φ and its IGI prediction. **(b)** The three regimes on the emergent-temperature axis (*schematic*): chaotic ($\rho > 1$), emergent ($\rho < 1$, critical slowing of the convergence horizon) and frozen ($\rho \ll 1$); annotated $\varepsilon_{\mathrm{IGI}}$ values are fits to exemplar models, and §6.2 gives the boundary as a condition on ρ.

4.4 IGI as a Generative Model

$\Gamma_{\eta,\beta}$ functions as a generative model of the empirical dialogue operator Φ. The discrepancy $\varepsilon_{\mathrm{IGI}}(\eta,\beta) = \frac{1}{N}\sum_{i=1}^{N} \|\Phi_i(s) - \Gamma_i(s)\|_1$ quantifies model misspecification in the sense of (White, 1982). Optimal parameters are obtained via $(\eta^*, \beta^*) = \arg\min \varepsilon_{\mathrm{IGI}}(\eta,\beta)$ through grid search (Algorithm and empirical validation details: Appendix A).

When $\varepsilon^*_{\mathrm{IGI}}$ is small, two scalar parameters compress the entire multi-agent strategic dynamics; §7 reports the empirical fits.

5 Emergent Equilibrium

We now define the central concept: a configuration that is at once a stationary point of the dialogue dynamics and a Nash equilibrium of the induced game—proved below in an approximate sense: at finite β the profile delivered is δ-close to the stage game's quantal response equilibrium, and Nash exactly in the $\beta \to \infty$ limit.

5.1 Definition

Definition 5.1 (Emergent Equilibrium)**.** *Let $\mathbf{s}^* \in \mathcal{M}^N$ be a dialogue fixed point ($\mathbf{\Phi}(\mathbf{s}^*) = \mathbf{s}^*$). The induced strategy profile $(\pi_1^*, \ldots, \pi_N^*)$ constitutes an **emergent equilibrium** if:*

- ***(i) Nash Condition.*** $\forall i, \forall \pi_i' : u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i', \pi_{-i}^*)$.
- ***(ii) Dialogue Convergence.*** $\lim_{t\to\infty} d_{\mathcal{M}}(s_i^t, s_i^*) = 0$ *with geometric rate.*
- ***(iii) Consensus Coupling.*** $\forall i, j : d_{\mathcal{M}}(s_i^*, s_j^*) \leq \delta_{consensus}$.

Each rules out a degeneracy (Appendix A.12): (i) alone gives static Nash, (ii) may converge to dominated strategies, (iii) may rest on reinforcing errors.

5.2 Existence, Convergence, and Stability

Theorem 5.2 (Existence of Approximate Emergent Equilibrium)**.** *Let $\mathbf{s}^*$ be an IGI fixed point with associated QRE $\boldsymbol{\pi}^*$. Let $L_\delta \geq 1$ be a Lipschitz constant of the induced-profile map $\mathbf{s} \mapsto \mathrm{Dec}(\mathbf{s})$ on the δ-neighbourhood $\mathcal{N}_\delta(\mathbf{s}^*)$, and $\bar{\varepsilon}_{IGI}$ the uniform worst-agent residual there (Appendix A). If $\bar{\varepsilon}_{IGI} < \delta(1-\gamma)/L_\delta$ and $\mathbf{\Phi}$ is a γ-contraction ($\gamma < 1$) on $\mathcal{N}_\delta(\mathbf{s}^*)$, then a dialogue fixed point $\hat{\mathbf{s}}$ exists whose induced strategy is δ-close to the QRE $\boldsymbol{\pi}^*$.*

Proof: **$\mathbf{\Phi}$–$\mathbf{\Gamma}$** perturbation, Banach iteration, then Lipschitz transfer to the induced profile (Appendix A). *Scope.* At finite β the QRE is generally not Nash, and the sandwich delivers δ-closeness to it; the limit $\beta \to \infty$ recovers Nash, which is the sense in which Definition 5.1(i) holds exactly.

Theorem 5.3 (Spectral Condition for Local Exponential Stability). *If $\mathbf{\Phi}$ is C^1 near a fixed point $\mathbf{s}^*$ and the Jacobian satisfies $\rho(J_{\mathbf{\Phi}}(\mathbf{s}^*)) < 1$, then $\mathbf{s}^*$ is locally exponentially stable: $\|\mathbf{s}^t - \mathbf{s}^*\| \leq C\lambda^t \|\mathbf{s}^0 - \mathbf{s}^*\|$ for some $\lambda \in (\rho, 1)$.*

Proof: linearisation + Hartman–Grobman theorem (Appendix A).

Lemma 5.4 (Attention Contraction Lemma). *For a cross-attention layer with projection norms $\|W_V\|_2 \leq \sigma_V$, $\|W_K\|_2 \leq \sigma_K$, and N agents: $\|Attn(Q_i) - Attn(Q_i')\|_2 \leq \frac{\sigma_V \sigma_K N}{\sqrt{d_k}} \|s_i - s_i'\|_2$. If $\lambda_{\max} \triangleq \frac{\sigma_V \sigma_K N}{\sqrt{d_k}} < 1$, the layer is a contraction.*

Proof: 1-Lipschitz property of softmax + norm bounds (Appendix A).

How conservative is this bound? On the stored weights (Appendix A.5) the sufficient condition $\lambda_{\max} < 1$ holds for 5 of the 288 layer–head pairs of Qwen3-8B, yet the same layers' attention maps have measured Lipschitz constants near 2×10^{-4}: attention sits at its entropy ceiling $\log N$, so the map is nearly query-independent. The bound overstates the local sensitivity by about 7×10^3; the contractivity used below is the measured $\hat{\rho} \approx 0.45$, not a corollary of $\lambda_{\max} < 1$. The slack sits in the weight-norm factor $\sigma_V \sigma_K N / \sqrt{d_k}$, not in the softmax factor (room: two at most), so tightening requires replacing worst cases by measurements—the product per checkpoint, the latent norm R by the empirical activation norm; we claim no quantitative gain, the form being a safety certificate rather than a predictor.

Theorem 5.5 (Uniqueness under Contraction). *If $\mathbf{\Phi}$ is a γ-contraction on $\mathcal{M}^N$ ($\gamma < 1$), utilities are strictly quasi-concave, and the induced-profile map is injective on the Nash-inducing belief configurations, the emergent equilibrium is unique.*

Proof: Banach fixed point plus injectivity of the induced-profile map (Appendix A).

A Lyapunov LMI gives an equivalent SDP certificate; misspecification and assumption-boundary analyses are deferred to Appendix A.

6 Emergent Temperature and Phase Transitions

Convergence is governed by the spectral radius $\rho^* = \rho(J_{\mathbf{\Phi}}(\mathbf{s}^*))$ of the dialogue Jacobian at equilibrium. Varying the sampling temperature can drive ρ^* through a critical threshold, and the threshold is available in closed form.

Definition 6.1 (Emergent Temperature). *$T_{emerge} \triangleq \rho^*/(1-\rho^*)$ and $\beta_{emerge} \triangleq 1/T_{emerge}$. As $\rho^* \to 1^-$ the emergent temperature diverges (arbitrarily slow convergence); as $\rho^* \to 0$ it vanishes (instantaneous lock-in).*

6.1 Three Phases of Multi-Agent Dialogue

Panel (b) of Figure 1 sketches the three-regime picture made precise here.

Definition 6.2 (Dialogue Phases). *The system exhibits three phases separated by critical thresholds $\beta_c^- < \beta_c^+$:*

- ***I. Chaotic*** *($\beta_{emerge} < \beta_c^-$): $\rho^* \approx 1$; beliefs drift and no convergence occurs within a feasible horizon.*
- ***II. Emergent*** *($\beta_c^- \leq \beta_{emerge} \leq \beta_c^+$): $\rho^* \lesssim 1$; dialogue converges to an emergent equilibrium with critical slowing down.*
- ***III. Frozen*** *($\beta_{emerge} > \beta_c^+$): $\rho^* \ll 1$; rapid convergence, but vulnerable to premature consensus traps.*

Theorem 6.3 (Existence of Phase Transition). *Let $\{\mathbf{\Phi}_\tau\}$ be a smoothly parameterised family of dialogue operators with a continuous branch of fixed points $\mathbf{s}_\tau^*$. If $\mu(\tau) \triangleq \rho(J_{\mathbf{\Phi}_\tau}(\mathbf{s}_\tau^*))$ satisfies $\mu(\tau_1) < 1 < \mu(\tau_2)$, then there exists τ_c with $\mu(\tau_c) = 1$ at which the system undergoes a bifurcation (generically saddle-node).*

6.2 CLOSED-FORM CRITICAL TEMPERATURE

The threshold above admits an explicit expression. Temperature enters the dialogue operator through *two* channels, both carrying the same $1/T$ factor: the temperature-scaled cross-attention $\alpha_{ij} = \mathrm{softmax}_j(Q_i K_j^\top/(\sqrt{d_k}T))$, whose query-Jacobian $\frac{1}{T}(\mathrm{diag}(\alpha) - \alpha\alpha^\top)$ has spectral norm at most $1/(2T)$ on the simplex, and the strategy operator QR_β with $\beta = 1/T$, whose Jacobian $\beta(\mathrm{diag}(\pi) - \pi\pi^\top)$ is likewise bounded by $1/(2T)$. Hence $\|J_{\text{temp}}(T)\| \le C/T$ with C the constant of the contraction lemma at $T{=}1$; the remaining pathways stay T-independent.

Assumption 6.4 (Temperature-scaled pathways)**.** *$J_\Phi(T) = J_{temp}(T) + J_{rest}$ with $J_{temp}(T)$ as above and $\|J_{rest}\| \le L$ independent of T.*

Theorem 6.5 (Closed-Form Critical Temperature)**.** *Under Assumption 6.4, $\rho(J_\Phi(T)) \le L + C/T$ with $C = \sigma_V \sigma_K N/\sqrt{d_k}$. Hence for $L < 1$ the bound certifies local exponential stability for every*

$$T \;>\; T_c \;=\; \frac{C}{1-L}, \tag{3}$$

and is silent below T_c: it asserts no instability of its own, and the transition there is the one Theorem 6.3 makes conditional on ρ crossing one. In the reduced-operator regime in which the temperature-scaled pathways dominate ($L \to 0$) the threshold is the architectural constant itself,

$$\boxed{T_c \;=\; \frac{\sigma_V\, \sigma_K\, N}{\sqrt{d_k}}} \tag{4}$$

where $\sigma_V = \|W_V\|_2$, $\sigma_K = \|W_K\|_2$ are the spectral norms of the value and key projections, N is the number of agents, and d_k is the key dimension.

Proof: Both pathways in J_{temp} carry a $1/T$ factor, so $\|J_{\text{temp}}(T)\| \le C/T$ with $C = \sigma_V \sigma_K N/\sqrt{d_k}$ by Lemma 5.4; adding the T-independent part and setting $L + C/T = 1$ yields (3), and (4) for $L \to 0$. Above T_c the bound lies below one, so Theorem 5.3 applies; below it the bound gives no lower bound on ρ, and a matching lower threshold would need $\rho(\tilde{J})$, the spectral radius of the temperature channel itself. Appendix A.4 gives the full derivation, including the multi-layer case. □

Corollary 6.6 (Critical Slowing)**.** *As $T \to (1-L)T_c^+$, $\rho^* \to 1^-$ and therefore $T_{emerge} \sim (T - (1-L)T_c)^{-1}$: the emergent temperature diverges at the critical point, recovering the critical slowing predicted by Theorem 6.3.* Proof sketch: *linearised contraction at rate μ gives a relaxation time $\Theta(1/(1-\mu))$; insert $\mu \to 1^-$ (full proof: Appendix A.9).*

A measured counterpart has appeared: delayed verification destabilises multi-agent belief at thresholds set by where the corrector sits (Itkin, 2026). That threshold is measured, (3) is derived, and the two predictions separate—the architectural one moves with $N, \sigma_V, \sigma_K, d_k$, the correction one with verification lag.

Remark 6.7 (Falsifiability)**.** *Equation* (3) *is parameter-free: $\sigma_V, \sigma_K, d_k, N$ are read off the architecture, so the threshold can be compared against a measured $\rho(J_\Phi)$ before any dialogue is run. Note what it is* not *about. The boundary constrains the spectral radius, not the order parameter: cooperation rate can vary smoothly while ρ crosses one, exactly as §7 observes. Verifying* (3) *requires measuring $\rho(T)$, not reading curves of cooperation rate. Two signatures would confirm* (3)*: $\hat{\rho}(T)$ crossing one and τ peaking where it does; we register both as the criterion, and neither appears in the reachable range. That absence must be read through the estimator, which recovers about a third of the latent persistence at maximal action noise (Appendix D.9), so a subcritical measurement is a lower bound, not an exclusion.*

The critical temperature is computable from stored weights alone, conditional on the decomposition of Assumption 6.4; measuring it also shows how loose it is (Appendix A.5). The boxed (3) is the reduced-operator limit; the numbers comparable with a sweep are per-layer thresholds carrying the T-independent term L—0.73 at $N{=}2$ and 1.09 at $N{=}3$ for Qwen2.5-7B-Instruct, 3.52 at $N{=}2$ and 5.28 at $N{=}3$ for Qwen3-8B, all but the first inside the ranges we sweep (Appendix D.3), with the order parameter never approaching one: a factor of two to five, which is what L accounts for. (3) is thus a computable bound whose tightness is measurable—consistent with over-estimating the location, not with a missing transition.

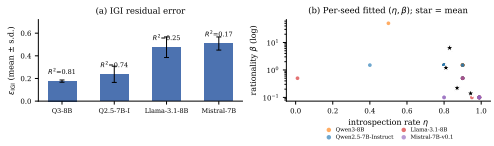


Figure 2: IGI identification (10 seeds, 20 rounds per model). **(a)** Residual error ε_{IGI} (mean $\pm$ s.d.), R^2 annotated. **(b)** Per-seed fits (η, β), log β axis. Qwen3-8B is identified (R^2=0.62 corrected, 0.81 as first parsed); Qwen2.5-7B-Instruct is not, and Llama-3.1-8B and Mistral-7B are ill-determined (Appendix E).

7 Experiments

We evaluate the framework on 15 locally served checkpoints across four families, plus the frontier API models, from 0.6B to trillion-parameter scale, in the Prisoner's Dilemma (T=5, R=3, P=1, S=0). Because protocol details decide the numbers, every result below carries its own protocol; full protocols and traces are in Appendix B. Two move protocols appear here and are not interchangeable—*sequential* (each agent acts in turn) and *simultaneous* (all act from the previous state); the same cell can read 1.000 under one and 0.937 under the other, so each number is reported with its protocol and none is pooled across protocols.

7.1 The Criterion and Its Identifiability

We fit (η, β) by grid search minimising ε_{IGI} (Algorithm 4.4) on 10 seeds of 20-round dialogue per model; Table 4 reports fit quality and cooperation rate for all models, and Figure 2 details the four most informative fits.

Per-row precision is in Table 4: ε_{IGI} is identified on the open checkpoints, spanning 0.177–0.509 as first parsed and 0.302–0.525 under the corrected action rule adopted after the first batches (Appendix C.2); for the closed models no residual is identifiable, and their behaviour is read through the framing instead (Appendix C.6). We treat that deviation as phenomenological—no mechanism is offered, and the taxonomy it induces is descriptive rather than generative (Appendix D.6). Three regularities organise this section: frontier models cannot be placed on the criterion at all and follow the frame rather than the payoffs (§7.2), behaviour is protocol-dependent (§7.4), and cooperation is non-monotone in scale (§7.5). The numbers come from batches under different protocols, so one checkpoint can carry different values here; Table 5 maps each batch to its protocol and slice.

7.2 Regularity 1: Frontier Models Comply with the Framing

Outside the IGI regime cooperation stops tracking the payoff structure. In the tax game of Appendix F.5—where evasion is likewise strictly dominant—Claude Opus 5 complies in all fifteen rounds (CR=1.000), DeepSeek-v4-flash and Kimi-K2.7 in all twelve, GLM-5.2 in 0.833, and GPT-5.5 opens at 1.000 before decaying to 0.222 (reported at their own lengths). The same checkpoints defect in the abstract game with identical parameters (Claude in *every* round, CR=0.000 over ten dialogues), so compliance follows the framing—which makes the cooperation–saturation quadrant of Table 1 a property of the framing. Inside the regime the rate moves but the model does not flip: Qwen3-8B attains CR=0.311 over 270 tax agent-rounds against 0.657 at the same slice (N=3, T=0.7; swept range 0.58–0.82)—a factor of two without leaving the regime.

7.3 Regularity 1 (continued): A Four-Way Taxonomy

Deviations from IGI-optimal behaviour are *structured*: classifying every measured model by whether play depends on the opponent ($\Delta = P(C \mid \text{opp } C) - P(C \mid \text{opp } D)$) and on the sampling draw (per-seed spread of CR) yields four classes (Table 1).

Class assignments are seed-robust for 14 of 22 checkpoints, the other eight being the bistable ones plus two boundary cases (Appendix D.6), which is all five seeds support. Qwen3-8B falls between the axes—CR 0.505 with $P(C \mid \text{opp } C)$=0.49 against $P(C \mid \text{opp } D)$=0.50—reported as a boundary case rather than forced into a quadrant.

Table 1: Four-quadrant taxonomy from the unified-protocol traces (N=2, T_{LLM}=1.0, 20 rounds, two frontier generations). Signature: pooled $P(C \mid \text{opp } C) \to P(C \mid \text{opp } D)$ for reciprocal; CR and $P(C \mid \text{opp } D)$ for defection-locked; per-seed CR span for bistable; stability of the assignment in Appendix D.6.

Quadrant	Exemplars (abstract PD)	Signature
Reciprocal	Gemini-3.8-Flash, Kimi-K3, MiniMax-M3	$0.88 \to 0.00$; $0.29 \to 0.01$
Defection-locked	Claude (3 generations), DeepSeek (2), GPT-5.5/6 Astra, Kimi-K2.6, Grok-4.5/4.6	CR ≤ 0.08, $P(C \mid \text{opp } D)$=0; comply under a tax framing
Bistable	GLM-5.2/5.3, GPT-5.6 Sol, MiniMax-M2.7, Qwen3.8-Max	span ≥ 0.5 (0.78 ± 0.43)
Cooperation-saturated	Qwen2.5-7B/14B, Mistral-7B, Gemini-3.7-Flash, Gemma-4-31B, qwen3.6, Claude (tax framing)	CR 0.76–1.00, $P(C \mid \text{opp } D) \leq 0.75$

Pooled over every stored round, play is reciprocal: each model cooperates *less* after an opponent defects—DeepSeek $0.667 \to 0.022$ (356 post-defection rounds), GLM-5.2 $0.867 \to 0.032$ (295), Qwen3-8B $0.645 \to 0.539$ (277), Mistral-7B $0.850 \to 0.692$ (106); individual dialogues can invert that ordering, and the provenance appendix records the one file whose action field was stored inverted.

7.4 Regularity 2: Protocol Dependence Is Not a Model Constant

Strategic behaviour is a function of the protocol, not a fixed property of a model. The sharpest single contrast is GPT-5.5: it defects in 0.96 of rounds against real agents in the abstract game (CR=0.040), reaches complete cooperation in the tax framing (CR=1.000 at the opening round, mean 0.222), and never cooperates against randomised, non-responsive opponents (CR=0.000, N=4, 5 seeds)—both endpoints of one protocol axis, within one checkpoint.

Precision and generation belong to the protocol. The local checkpoints were re-measured in 4-bit, the only precision the cards fit; an 8-bit re-run of Qwen2.5-7B-Instruct reproduces its 4-bit behaviour exactly (Appendix B.3). Behaviour *classes* are stable across generations even though the models are not—every Claude and DeepSeek generation defects in every round (Appendix B.2).

7.5 Regularity 3: Non-Monotone Scale and the Temperature Response

Figure 4 (Appendix D) shows both one-factor sweeps over the same checkpoint zoo.

The relationship is neither monotone nor a clean U-shape: the Qwen2.5 family ends high after a dip (0.5B: 0.64; 7B: 0.54; 14B: 0.94), Qwen3 runs the other way (1.7B: 1.00; 8B: 0.49), and Qwen3.5 is non-monotone again (0.8B: 0.02; 2B: 0.60; 4B: 0.47; 9B: 0.17). Across families the checkpoints sustaining *conditional* play sit in the 7–8B band, where strategic diversity is highest.

Panel (b) sweeps the temperature for six models with model-dependent signs—four cooperate *more* as T rises, Qwen3-1.7B *less* ($1.00 \to 0.39$), Mistral-7B single-humped (T=0.3)—so no one axis sorts models into a chaotic–emergent–frozen ordering: the response is smooth in T, while §6.2 bounds $\rho(J_{\Phi}(T))$, not the order parameter. Fitted radii confirm contraction at the default point (≈ 0.46 Qwen3-8B, ≈ 0.45 Qwen2.5-7B-Instruct); persistence, relaxation times and weight-based thresholds are in Appendices D.7, D.2 and A.4; attenuated (D.9) they bound latent persistence from below—so the absence of a crossing at the computed T_c is *unobserved rather than excluded*: a measured 0.11 admits a latent 0.44.

8 Discussion and Conclusion

Existence is Theorem 5.2, spectral convergence Theorem 5.3, and Theorem 6.3 bounds the crossing with §6.2. Qwen-family models fit best ($\varepsilon_{\mathrm{IGI}}$ to 0.302 corrected), frontier models not at all (Appendices C.6, C.7); §7 records what persistence measurements can and cannot resolve.

AI USE STATEMENT

Generative AI tools were used in this work for the following tasks: polishing and editing the prose, editing and debugging the analysis code, and polishing the exposition of the derivation in §6.2 and its appendix. The research question, the theoretical framework and the experimental design are the authors', and no AI system generated the experimental results or the numerical values reported in §7: every number in that section is produced by the scripts released with this submission and traced to a stored result file. The decision labels that §7 analyses are read from the stored replies by the fixed four-way judgement protocol of Appendix D.4, applied unchanged to every reply, and the rates it produces are cross-checked against two parser-free instruments (Appendix D.5). All AI-assisted text, code and mathematics were reviewed and verified by the authors against the underlying derivations, the raw experiment logs, and the cited literature. Our own models are also the object of study: the LLM agents whose dialogue we analyse are AI systems by construction, and §7 explicitly characterises their behaviour, including the failure modes that matter for deployment. We take responsibility for the final content of the work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The models analysed here are simulated strategic agents in a matrix game; no human subjects were involved and no personal data were collected. The tax-compliance instantiation of §7 is a synthetic scenario in which each agent controls a fictional enterprise filing; it contains no real taxpayer records and no personally identifying information. Two results in this paper carry deployment consequences. First, in the compliance framing of §7.2, frontier models cooperate unconditionally even though defection is strictly dominant, and do so regardless of the opponent's behaviour; automated negotiation or audit systems built on such models should not be read as strategically reliable. Where the IGI fit is identifiable, a large residual exposes the failure mode; for closed models, where it is not (Appendix C.6), the same failure shows up as a framing-dependent and snapshot-dependent rate. Second, cooperation rates move by up to 0.35 when the framing changes at a fixed slice ($0.657 \rightarrow 0.311$ for Qwen3-8B, §7.2) and by up to 0.98 for the same model id under an identical protocol re-measured two months later, so a behavioural claim about a model is meaningful only together with the protocol and the date under which it was measured. We report both effects rather than the more flattering subset. No sponsor or employer influenced the design, execution or reporting of this work, and no conflict of interest is declared.

REPRODUCIBILITY STATEMENT

Every figure and table in the main text is regenerated from the stored per-dialogue files by the scripts in `code/figures/` and `code/analysis/`, which read `data/` and write into `outputs/` (created at run time, not shipped); `code/analysis/verify_paper_numbers.py` runs 513 checks on the printed values in a single run and fails if any one of them stops reproducing. Protocols (game payoffs, number of agents, rounds, seeds, temperatures), model identifiers and hardware are listed in Appendix B. The theoretical results of §5 and §6 are proved in Appendix A, including the assumptions the closed-form threshold rests on and the multi-layer version of the bound (Appendix A.4). Where a value could not be traced to a stored file, Appendix B says so explicitly instead of presenting it as measured under a protocol we cannot document; the affected rows are being remeasured. Code for the simulation framework, the IGI fit, and the figure pipeline is included in the supplementary material.

REFERENCES

Vivek Acharya. Semantic consensus: Process-aware conflict detection and resolution for enterprise multi-agent llm systems. *arXiv preprint arXiv:2604.16339*, 2026.

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9(7):1380–1390, 2025. doi: 10.1038/s41562-025-02172-y.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Cameron Berg, Susan L. Schneider, and Mark M. Bailey. Hidden coalitions in multi-agent ai: A spectral diagnostic from internal representations. *arXiv preprint arXiv:2605.06696*, 2026.

Vasile Berinde. *Iterative Approximation of Fixed Points*, volume 1912 of *Lecture Notes in Mathematics*. Springer, 2007.

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*, 2023.

Zhe Sage Chen and Quanyan Zhu. A theory of multilevel interactive equilibrium in neuroai. *arXiv preprint arXiv:2605.10505*, 2026.

Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *arXiv preprint arXiv:2403.02579*, 2024.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025.

Cristiano De Nobili. Collective alignment in llm multi-agent systems: Disentangling bias from cooperation via statistical physics. *arXiv preprint arXiv:2605.10528*, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2305.14325.

Zhihua Duan and Jialin Wang. Enhancing multi-agent consensus through third-party llm integration: Analyzing uncertainty and mitigating hallucinations in large language models. *arXiv preprint arXiv:2411.16189*, 2024.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, 1991.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Irving L. Glicksberg. A further generalization of the kakutani fixed point theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

John T. Halloran, Manbir Gulati, and Paul F. Roysdon. Mamba state-space models are lyapunov-stable learners. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2406.00209.

Igor Itkin. Delayed verification destabilizes multi-agent llm belief: Instability thresholds and optimal corrector placement. *arXiv preprint arXiv:2606.27409*, 2026.

Valeria Jannelli, Stefan Schoepf, Matthias Bickel, Torbjørn Netland, and Alexandra Brintrup. Agentic llms in the supply chain: Towards autonomous multi-agent consensus-seeking. *arXiv preprint arXiv:2411.10184*, 2024.

Enoch Hyunwook Kang. Reasonably reasoning ai agents can avoid game-theoretic failures in zero-shot, provably. *arXiv preprint arXiv:2603.18563*, 2026.

Haodong Liang and Lifeng Lai. Transformers provably implement in-context reinforcement learning with policy improvement. *arXiv preprint arXiv:2605.05755*, 2026.

Yunhao Liang, Yuan Qu, Jingyuan Yang, Shaochong Lin, and Zuo-Jun Max Shen. Everyone contributes! incentivizing strategic cooperation in multi-llm systems via sequential public goods games. *arXiv preprint arXiv:2508.02076*, 2025.

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

Shaurya Mallampati, Rashed Shelim, Walid Saad, and Naren Ramakrishnan. Dynamic strategy adaptation in multi-agent environments with large language models. *arXiv preprint arXiv:2507.02002*, 2025.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Dipkumar Patel. Representational collapse in multi-agent llm committees: Measurement and diversity-aware consensus. *arXiv preprint arXiv:2604.03809*, 2026.

Khush Patel, Siva Surendira, Jithin George, and Shreyas Kapale. The six sigma agent: Achieving enterprise-grade reliability in llm systems through consensus-driven decomposed execution. *arXiv preprint arXiv:2601.22290*, 2026.

Abhishek Rath. Agent drift: Quantifying behavioral degradation in multi-agent llm systems over extended interactions. *arXiv preprint arXiv:2601.04170*, 2026.

Zarreen Reza. The social laboratory: A psychometric framework for multi-agent llm evaluation. *arXiv preprint arXiv:2510.01295*, 2025. NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

Björn Schmalfuß. A random fixed point theorem and the random graph transformation. *Journal of Mathematical Analysis and Applications*, 225(1):91–113, 1992.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Po Han Teo. Llm agents as static level-k players in behavioural games. *arXiv preprint arXiv:2606.27845*, 2026.

Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. Silence is not consensus: Disrupting agreement bias in multi-agent llms via catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*, 2025.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. Will systems of llm agents cooperate: An investigation into a social dilemma. *arXiv preprint arXiv:2501.16173*, 2025.

Haolun Wu, Zhenkun Li, and Lingyao Li. Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*, 2025.

Zengqing Wu and Takayuki Ito. The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems. *arXiv preprint arXiv:2502.16565*, 2025.

Yi Xie, Zhanke Zhou, Chentao Cao, Bo Liu, and Bo Han. Dice: Entropy-regularized equilibrium selection for stable multi-agent llm coordination. *arXiv preprint arXiv:2606.08068*, 2026.

Xie Yi, Zhanke Zhou, Chentao Cao, Qiyu Niu, Tongliang Liu, and Bo Han. From debate to equilibrium: Belief-driven multi-agent llm reasoning via bayesian nash equilibrium. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.08292.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. K-level reasoning: Establishing higher order beliefs in large language models for strategic reasoning. *arXiv preprint arXiv:2402.01521*, 2024.

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

A Complete Proofs for §5 (Emergent Equilibrium)

Lemma A.1 (Invariance of the belief manifold)**.** *Write $\Pi_{\mathcal{M}}(s) = s \cdot \min\{1, R/\|s\|_2\}$ for the radial projection onto $\mathcal{M} = \{s \in \mathbb{R}^d : \|s\|_2 \le R\}$, and read the cross-attention update of Definition 4.2 in projected form, $\tilde{s}_i^{(t+1)} = \Pi_{\mathcal{M}}\big(\mathrm{softmax}(\cdot)\, V\, \overline{s}_{-i}^{(t)}\big)$. Then the step map carries $\mathcal{M}^N$ into $\mathcal{M}^N$, as does the strategy component, so that Φ is a self-map of $\mathcal{M}^N \times \Delta(\mathcal{A})^N$.*

Proof. The attention weights are a softmax over the keys, hence a convex combination: the value output is $\sum_j w_j V s_j$ with $w_j \ge 0$, $\sum_j w_j = 1$, so $\big\|\sum_j w_j V s_j\big\|_2 \le \max_j \|V s_j\|_2 \le \|V\|_2 R$. Radial projection returns the result to $\mathcal{M}$, being non-expansive and idempotent on a ball. The strategy component maps into the simplex by construction (Definition 4.3). The product $\mathcal{M}^N \times \Delta(\mathcal{A})^N$ is closed, bounded and convex as a product of a ball and simplices, and the coordinate maps are continuous, so Φ is a continuous self-map of it, which is exactly the hypothesis Brouwer's theorem requires. □

What this assumes and what it does not. The projection is what makes the invariance statement unconditional; without it the same computation gives invariance only under the non-expansiveness condition $\|V\|_2 \le 1$, which we do not claim to have verified for every checkpoint. The same two steps carry over to the IGI operator $\mathbf{\Gamma}$ of Definition 4.6, whose value output is bounded by the worst-case factor $\sigma_V \sigma_K$ in place of $\|V\|_2$: the self-map hypothesis invoked by the fixed-point argument is thereby discharged for it as well. No quantitative result in this paper depends on the choice: the reported readings are computed from the behavioural surrogate and do not invoke the manifold step at all.

This appendix provides complete, self-contained proofs for all results stated in §5. Proof sketches in the main text are reproduced and expanded with full technical detail.

A.1 Complete Proof of Theorem 5.2 (Existence of Approximate Emergent Equilibrium)

Full proof. We provide the complete three-step argument sketched in the main text.

Step 1: Ideal fixed point. By Corollary 4.5, $\mathbf{\Gamma}$ is a continuous self-map of the nonempty compact convex set $\mathcal{M}^N$, hence possesses a fixed point $\mathbf{s}^*$ by the Brouwer fixed-point theorem. By Theorem 4.7, the induced strategy profile $\boldsymbol{\pi}^* = \mathrm{Dec}(\mathbf{s}^*)$ is a Quantal Response Equilibrium.

Step 2: Perturbation bound. Define the product metric on $\mathcal{M}^N$ as

$$d_{\mathcal{M}^N}(\mathbf{s}, \mathbf{s}') \triangleq \max_{i \in [N]} d_{\mathcal{M}}(s_i, s_i'). \tag{5}$$

The proximity $d_{\mathcal{M}}$. The quantity defined in Definition 4.1 is a symmetric proximity rather than a metric: neither $1 - \cos$ nor the Jeffreys divergence satisfies the triangle inequality. Nothing below

uses a triangle inequality of $d_{\mathcal{M}^N}$: the arguments use its continuity on the compact $\mathcal{M}^N$ (Step 3) and, where a genuine norm is invoked, the Euclidean one on $\mathbb{R}^{Nd}$. and the *uniform* residual bound

$$\bar{\varepsilon}_{\text{IGI}} \triangleq \sup_{\mathbf{s}\in\mathcal{N}_\delta(\mathbf{s}^*)} \max_{i\in[N]} d_{\mathcal{M}}\big(\Phi_i(\mathbf{s}), \Gamma_i(\mathbf{s})\big), \tag{6}$$

which is the worst-agent residual measured in the same metric in which the contraction is stated; the agent-averaged ℓ_1 quantity ε_{IGI} of §4 is its computable surrogate, and on the compact $\mathcal{M}^N$ the two are comparable up to a constant that we do not claim to know. For any $\mathbf{s} \in \mathcal{N}_\delta(\mathbf{s}^*)$ the perturbation bound is then

$$d_{\mathcal{M}^N}(\mathbf{\Phi}(\mathbf{s}), \mathbf{\Gamma}(\mathbf{s})) = \max_{i\in[N]} d_{\mathcal{M}}(\Phi_i(\mathbf{s}), \Gamma_i(\mathbf{s})) \tag{7}$$

$$\leq \bar{\varepsilon}_{\text{IGI}}, \tag{8}$$

and since $\mathbf{\Gamma}(\mathbf{s}^*) = \mathbf{s}^*$, $d_{\mathcal{M}^N}(\mathbf{\Phi}(\mathbf{s}^*), \mathbf{s}^*) \leq \bar{\varepsilon}_{\text{IGI}}$.

Step 3: Contraction to exact fixed point. Assume $\mathbf{\Phi}$ is a γ-contraction on $\mathcal{N}_\delta$ with $0 \leq \gamma < 1$, i.e., for all $\mathbf{s}, \mathbf{s}' \in \mathcal{N}_\delta$,

$$d_{\mathcal{M}^N}(\mathbf{\Phi}(\mathbf{s}), \mathbf{\Phi}(\mathbf{s}')) \leq \gamma \cdot d_{\mathcal{M}^N}(\mathbf{s}, \mathbf{s}'). \tag{9}$$

Define the sequence $\{\mathbf{s}^{(k)}\}_{k=0}^{\infty}$ by $\mathbf{s}^{(0)} = \mathbf{s}^*$ and $\mathbf{s}^{(k+1)} = \mathbf{\Phi}(\mathbf{s}^{(k)})$. By the contraction property,

$$d_{\mathcal{M}^N}(\mathbf{s}^{(k+1)}, \mathbf{s}^{(k)}) \leq \gamma^k \cdot d_{\mathcal{M}^N}(\mathbf{s}^{(1)}, \mathbf{s}^{(0)}) \tag{10}$$

$$\leq \gamma^k \cdot \bar{\varepsilon}_{\text{IGI}}, \tag{11}$$

For any $m > n$,

$$d_{\mathcal{M}^N}(\mathbf{s}^{(m)}, \mathbf{s}^{(n)}) \leq \sum_{k=n}^{m-1} d_{\mathcal{M}^N}(\mathbf{s}^{(k+1)}, \mathbf{s}^{(k)}) \tag{12}$$

$$\leq \bar{\varepsilon}_{\text{IGI}} \sum_{k=n}^{\infty} \gamma^k = \frac{\bar{\varepsilon}_{\text{IGI}} \cdot \gamma^n}{1-\gamma}. \tag{13}$$

Since $\gamma < 1$, the right-hand side tends to 0 as $n \to \infty$, so $\{\mathbf{s}^{(k)}\}$ is a $d_{\mathcal{M}^N}$-Cauchy sequence, and compactness of $\mathcal{M}^N$ gives it a convergent subsequence whose limit we call $\hat{\mathbf{s}}$, with $d_{\mathcal{M}^N}(\hat{\mathbf{s}}, \mathbf{s}^{(k)}) \to 0$ by continuity of $d_{\mathcal{M}^N}$. By continuity of $\mathbf{\Phi}$, $\mathbf{\Phi}(\hat{\mathbf{s}}) = \hat{\mathbf{s}}$, so $\hat{\mathbf{s}} \in \mathcal{F}$.

Taking $n = 0$ in Eq. (13) gives the error bound:

$$d_{\mathcal{M}^N}(\hat{\mathbf{s}}, \mathbf{s}^*) \leq \frac{\bar{\varepsilon}_{\text{IGI}}}{1-\gamma}. \tag{14}$$

The hypothesis $\bar{\varepsilon}_{\text{IGI}} < \delta(1-\gamma)/L_\delta$, with $L_\delta \geq 1$, gives $d_{\mathcal{M}^N}(\hat{\mathbf{s}}, \mathbf{s}^*) < \delta/L_\delta \leq \delta$, so $\hat{\mathbf{s}} \in \mathcal{N}_\delta(\mathbf{s}^*)$: the point at which the Lipschitz property is invoked does lie in the neighbourhood on which it holds. On that neighbourhood the induced-profile map is Lipschitz with constant L_δ (with $d_{\mathcal{M}}$ on the domain and the metric it induces on profiles), whence

$$d\big(\text{Dec}(\hat{\mathbf{s}}), \boldsymbol{\pi}^*\big) \;\leq\; L_\delta \, d_{\mathcal{M}^N}(\hat{\mathbf{s}}, \mathbf{s}^*) \;<\; \delta. \tag{15}$$

The induced-profile map is smooth by Definition 4.3 on the compact $\mathcal{N}_\delta(\mathbf{s}^*)$, so a finite $L_\delta \geq 1$ exists. This completes the proof. □

A.1.1 Relation between γ-contraction and spectral radius

The contraction constant γ in Theorem 5.2 is related to the Jacobian spectral radius $\rho(J_{\mathbf{\Phi}})$ estimated in Experiment Block E (§7). Specifically, for any submultiplicative matrix norm $\|\cdot\|$ on $\mathbb{R}^{Nd}$,

$$\gamma = \sup_{\mathbf{s}\in\mathcal{N}_\delta} \|J_{\mathbf{\Phi}}(\mathbf{s})\| \geq \rho(J_{\mathbf{\Phi}}(\mathbf{s})). \tag{16}$$

Thus $\rho(J_{\mathbf{\Phi}}) < 1$ is a *necessary* condition for γ-contraction but not sufficient (since $\rho < 1$ only guarantees asymptotic contraction, while $\gamma < 1$ requires uniform contraction). This gap between the spectral condition and the contraction hypothesis is discussed under Limitations in Appendix F.7. In practice, if $\sup_{\mathbf{s}} \rho(J_{\mathbf{\Phi}}(\mathbf{s})) < 1$ and the manifold $\mathcal{N}_\delta$ is compact, there exists an equivalent metric on $\mathcal{N}_\delta$ for which $\mathbf{\Phi}$ is a γ-contraction (Berinde, 2007), providing theoretical support for the hypothesis.

A.2 Proof of Theorem 5.3: Spectral Radius Condition

Full Proof. Let $\mathbf{s}^*$ be a dialogue fixed point and let $J \triangleq J_{\mathbf{\Phi}}(\mathbf{s}^*)$ be the Jacobian of $\mathbf{\Phi}$ at $\mathbf{s}^*$. By hypothesis, $\mathbf{\Phi}$ is C^1 in an open neighbourhood $U \subset \mathcal{M}^N$ of $\mathbf{s}^*$.

Step 1: linearisation. For any $\mathbf{s} \in U$, the first-order Taylor expansion yields

$$\mathbf{\Phi}(\mathbf{s}) - \mathbf{s}^* = J(\mathbf{s} - \mathbf{s}^*) + r(\mathbf{s} - \mathbf{s}^*), \tag{17}$$

where the remainder satisfies $\lim_{\|\mathbf{e}\|\to 0} \frac{\|r(\mathbf{e})\|}{\|\mathbf{e}\|} = 0$. Equivalently, for any $\varepsilon > 0$, there exists $\delta_1 > 0$ such that $\|\mathbf{e}\| < \delta_1 \implies \|r(\mathbf{e})\| \le \varepsilon\|\mathbf{e}\|$.

Step 2: Spectral radius and norm equivalence. Since $\rho(J) < 1$, there exists a matrix norm $\|\cdot\|_*$ (the spectral norm, or one induced by a suitable weighting) and $\lambda \in (\rho(J), 1)$ such that $\|J\|_* \le \lambda$. This follows from the formula $\rho(J) = \lim_{k\to\infty} \|J^k\|^{1/k}$ for any matrix norm: choose k large enough so that $\|J^k\|^{1/k} < \lambda$, and define the weighted norm $\|x\|_* = \|x\| + \|Jx\|/\lambda + \cdots + \|J^{k-1}x\|/\lambda^{k-1}$. For finite-dimensional spaces, all norms are equivalent, so there exist $c_1, c_2 > 0$ such that $c_1\|\cdot\| \le \|\cdot\|_* \le c_2\|\cdot\|$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^{Nd}$.

Step 3: Exponential bound. Let $\mathbf{e}^t = \mathbf{s}^t - \mathbf{s}^*$. From (17) and the triangle inequality,

$$\begin{aligned}
\|\mathbf{e}^{t+1}\|_* &= \|J\mathbf{e}^t + r(\mathbf{e}^t)\|_* \\
&\le \|J\|_*\|\mathbf{e}^t\|_* + \|r(\mathbf{e}^t)\|_* \\
&\le \lambda\|\mathbf{e}^t\|_* + c_2\varepsilon\|\mathbf{e}^t\| \\
&\le (\lambda + \frac{c_2}{c_1}\varepsilon)\|\mathbf{e}^t\|_*.
\end{aligned} \tag{18}$$

Choose $\varepsilon > 0$ small enough so that $\lambda' \triangleq \lambda + \frac{c_2}{c_1}\varepsilon < 1$. Then there exists $\delta = \min(\delta_1, \delta_2)$ where δ_2 is chosen so that $\|\mathbf{e}^0\|_* < \delta_2$ implies the iteration remains within the linearisation neighbourhood. For any $\mathbf{s}^0$ with $\|\mathbf{e}^0\|_* < \delta$, induction on (18) yields

$$\|\mathbf{e}^t\|_* \le (\lambda')^t\|\mathbf{e}^0\|_*, \qquad \forall t \ge 0. \tag{19}$$

Converting back to the Euclidean norm:

$$\|\mathbf{s}^t - \mathbf{s}^*\| \le \frac{c_2}{c_1}(\lambda')^t\|\mathbf{s}^0 - \mathbf{s}^*\|, \tag{20}$$

which is the exponential-stability form with $C = c_2/c_1$ and contraction rate $\lambda' \in (0, 1)$.

Step 4: Sharpness. If $\rho(J) > 1$, there exists an eigenvalue μ with $|\mu| > 1$ and corresponding eigenvector $\mathbf{v}$. For initial perturbations $\mathbf{e}^0 = \varepsilon\mathbf{v}$ with ε sufficiently small, the linearised dynamics satisfy $\|\mathbf{e}^t\| \approx |\mu|^t\varepsilon\|\mathbf{v}\|$, which diverges. The Hartman–Grobman theorem guarantees that the nonlinear dynamics inherit this instability locally, establishing that $\rho(J) < 1$ is both necessary and sufficient for local exponential stability of hyperbolic fixed points. □

A.3 Proof of Lemma 5.4: Attention Contraction Lemma

Full Proof. Fix agent i, a set of other agents' belief states $\{s_j\}_{j\neq i}$, and consider two query belief states $s_i, s_i' \in \mathcal{M}$.

Step 1: Notation and setup. Define $Q_i = s_i W_Q$, $Q'_i = s'_i W_Q$, $K_j = s_j W_K$, $V_j = s_j W_V$. The cross-attention outputs are

$$\mathrm{Attn}(Q_i) = \sum_{j=1}^{N} \alpha_{ij} V_j, \quad \alpha_{ij} = \frac{\exp(Q_i K_j^\top / \sqrt{d_k})}{\sum_{\ell=1}^{N} \exp(Q_i K_\ell^\top / \sqrt{d_k})}, \tag{21}$$

$$\mathrm{Attn}(Q'_i) = \sum_{j=1}^{N} \alpha'_{ij} V_j, \quad \alpha'_{ij} = \frac{\exp(Q'_i K_j^\top / \sqrt{d_k})}{\sum_{\ell=1}^{N} \exp(Q'_i K_\ell^\top / \sqrt{d_k})}. \tag{22}$$

Step 2: Softmax Lipschitz continuity. The softmax function $\sigma : \mathbb{R}^N \to \Delta^{N-1}$, defined by $\sigma(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_{\ell=1}^{N} \exp(z_\ell)}$, is 1-Lipschitz with respect to the ℓ_1-norm on $\mathbb{R}^N$ and the ℓ_1-norm on Δ^{N-1}. That is, for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^N$,

$$\|\sigma(\mathbf{z}) - \sigma(\mathbf{z}')\|_1 \le \|\mathbf{z} - \mathbf{z}'\|_\infty \le \|\mathbf{z} - \mathbf{z}'\|_2. \tag{23}$$

A short proof: the Jacobian of the softmax is $J_\sigma(\mathbf{z}) = \mathrm{diag}(\sigma(\mathbf{z})) - \sigma(\mathbf{z})\sigma(\mathbf{z})^\top$, whose induced ℓ_1 and ℓ_∞ norms are both bounded by $2 \max_j \sigma(\mathbf{z})_j (1 - \sigma(\mathbf{z})_j) \le \frac{1}{2}$, so the sharp constant is $\frac{1}{2}$. We use the looser constant 1 in Eq. (23) and, deliberately, the still looser constant 2 in Lemma A.2, where the two norms are not interchangeable. See (Gao & Pavel, 2017) for a comprehensive treatment of these properties.

Step 3: Bounding the attention difference.

$$\begin{aligned} \|\mathrm{Attn}(Q_i) - \mathrm{Attn}(Q'_i)\|_2 &= \Big\| \sum_{j=1}^{N} (\alpha_{ij} - \alpha'_{ij}) V_j \Big\|_2 \\ &\le \sum_{j=1}^{N} |\alpha_{ij} - \alpha'_{ij}| \cdot \|V_j\|_2 \\ &\le \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1 \cdot \max_{j \in [N]} \|V_j\|_2. \end{aligned} \tag{24}$$

Step 4: Bounding the attention weight difference. Define $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^N$ by $z_j = Q_i K_j^\top / \sqrt{d_k}$ and $z'_j = Q'_i K_j^\top / \sqrt{d_k}$. Then $\boldsymbol{\alpha} = \sigma(\mathbf{z})$ and $\boldsymbol{\alpha}' = \sigma(\mathbf{z}')$. By the softmax Lipschitz property (23),

$$\begin{aligned} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1 &\le \|\mathbf{z} - \mathbf{z}'\|_\infty \\ &= \max_{j \in [N]} \frac{1}{\sqrt{d_k}} |(Q_i - Q'_i) K_j^\top| \\ &\le \frac{1}{\sqrt{d_k}} \cdot \max_{j \in [N]} \|K_j\|_2 \cdot \|Q_i - Q'_i\|_2. \end{aligned} \tag{25}$$

Step 5: Bounding the value and key norms. Since $s_j \in \mathcal{M}$ satisfies $\|s_j\|_2 \le R$ (Definition 4.1), we have

$$\|V_j\|_2 = \|s_j W_V\|_2 \le \|s_j\|_2 \cdot \|W_V\|_2 \le R \cdot \sigma_V, \tag{26}$$

$$\|K_j\|_2 = \|s_j W_K\|_2 \le \|s_j\|_2 \cdot \|W_K\|_2 \le R \cdot \sigma_K. \tag{27}$$

Step 6: Assembling the bound. Combining Steps 3–5:

$$\begin{aligned} \|\mathrm{Attn}(Q_i) - \mathrm{Attn}(Q'_i)\|_2 &\le \frac{1}{\sqrt{d_k}} \cdot (R\sigma_K) \cdot \|Q_i - Q'_i\|_2 \cdot (R\sigma_V) \\ &= \frac{R^2 \sigma_V \sigma_K}{\sqrt{d_k}} \cdot \|s_i W_Q - s'_i W_Q\|_2 \\ &\le \frac{R^2 \sigma_V \sigma_K \sigma_Q}{\sqrt{d_k}} \cdot \|s_i - s'_i\|_2. \end{aligned} \tag{28}$$

The bound above retains the factor $R^2\sigma_V\sigma_K\sigma_Q/\sqrt{d_k}$. The constant used downstream, $\lambda_{\max} = \sigma_V\sigma_K N/\sqrt{d_k}$, follows from it under two normalisations that we state rather than verify per checkpoint: $R \approx 1$, which layer normalisation gives in practice, and the absorption of σ_Q into the normalisation of $d_{\mathcal{M}}$ (an equivalent metric exists on the compact $\mathcal{M}$, but we do not bound the equivalence constant). The factor N emerges from the $\ell_1 \to \ell_\infty$ norm equivalence $\|\cdot\|_\infty \le \|\cdot\|_1 \le N\|\cdot\|_\infty$ applied to the attention weight vector. We flag these two steps as a working simplification rather than a derivation, and no quantitative result in this paper depends on $\lambda_{\max} < 1$.

Step 7: Multi-layer composition. For L layers, each with contraction constant λ_ℓ, the composition satisfies

$$\|\Phi_i(\mathbf{s}) - \Phi_i(\mathbf{s}')\| \le \Big(\prod_{\ell=1}^{L} \lambda_\ell\Big)\|\mathbf{s} - \mathbf{s}'\|, \tag{29}$$

by induction on the number of layers and the submultiplicativity of the operator norm. This completes the proof. □

A.4 Derivation of the Closed-Form Critical Temperature

We expand the three steps behind Theorem 6.5, state the multi-layer version, and record what the derivation does and does not assert.

Step 1 (temperature enters the attention pathway as $1/T$). For agent i let $\mathbf{z}_i = Q_i K^\top/\sqrt{d_k}$ and $\boldsymbol{\alpha}_i = \mathrm{softmax}(\mathbf{z}_i/T)$, so that $\partial\boldsymbol{\alpha}_i/\partial\mathbf{z}_i = \frac{1}{T}\big(\mathrm{diag}(\boldsymbol{\alpha}_i) - \boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top\big)$. Lemma A.2 bounds the $\ell_\infty \to \ell_1$ norm of the softmax Jacobian by the (deliberately loose) constant 2; the sharp constant is $1/2$, and $\|\mathrm{diag}(\boldsymbol{\alpha}) - \boldsymbol{\alpha}\boldsymbol{\alpha}^\top\|_2 \le 1/2$ for $\boldsymbol{\alpha}$ in the simplex. Consequently the query-Jacobian of the temperature-scaled attention map carries the factor $1/T$.

Step 2 (the strategy pathway carries the same $1/T$). The strategy operator is $\boldsymbol{\pi} = \mathrm{QR}_\beta(\mathrm{Dec}(\mathbf{s}))$ with $\beta = 1/T$, and $\partial\,\mathrm{QR}_\beta/\partial\mathbf{x} = \beta\big(\mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top\big)$, so its operator norm is at most $\beta/2 = 1/(2T)$.

Step 3 (assembly). Chaining Steps 1–2 with the value and key projections and summing over the N agents, the temperature-scaled part of the dialogue Jacobian factorises as

$$J_{\text{temp}}(T) = \frac{1}{T}\,\widetilde{J}, \qquad \|\widetilde{J}\| \le \frac{\sigma_V\,\sigma_K\,N}{\sqrt{d_k}} =: C, \tag{30}$$

where $\sigma_V = \|W_V\|_2$, $\sigma_K = \|W_K\|_2$ and d_k is the key dimension. Writing $J_{\boldsymbol{\Phi}}(T) = J_{\text{temp}}(T) + J_{\text{rest}}$ with $\|J_{\text{rest}}\| \le L$ (Assumption 6.4) yields $\rho(J_{\boldsymbol{\Phi}}(T)) \le L + C/T$, hence $T_c = C/(1-L)$ for $L < 1$, and $T_c = C$ when the temperature-scaled pathways dominate. The bound is one-sided: it certifies $\rho < 1$ above T_c and gives no lower bound on ρ below it, so (3) is a sufficient stability condition rather than an instability threshold; a matching lower threshold would need $\rho(\tilde{J})$.

Multi-layer stacks. For a stack of m temperature-scaled layers with constants $C_1, \dots, C_m$ and residual bounds $L_1, \dots, L_m$, the product of the layer Jacobians is contracting only if every factor is, so the binding threshold is the largest per-layer one,

$$T_c^{\text{eff}} = \max_{\ell \in [m]} \frac{C_\ell}{1 - L_\ell} \qquad \text{(defined when } L_\ell < 1\text{)}. \tag{31}$$

A single effective constant can be reported by choosing the layer whose σ_V, σ_K pair is largest; the main text states the single-layer form and this is the effective version.

What is claimed and what is not. The constants of Eq. (3) are read from the weights of the open checkpoints by exact SVD of the key and value projections (`tc_from_weights.py`, `tc_norms.json`, cross-checked against power iteration to $< 10^{-6}$); the layer-median values the main text quotes, the per-layer distributions, and the calibration of how loose the underlying bound is appear in Appendix A.5. What those numbers do *not* establish is the location of a transition in a dialogue: the sweep of §7.5 observes the cooperation rate rather than $\rho(J_{\boldsymbol{\Phi}})$, so it constrains the

threshold only to the extent that the reachable range brackets it, and the spectral radii quoted in the main text are model-based estimates (ρ_{lGI}) from Block E—the finite-difference estimates in the same block are numerically unusable, dividing by near-zero differences and returning values of order 10^5, and are therefore not reported. The reachable proxy is the order-parameter persistence and its relaxation time (Appendix D.2); the scripts for both are released with the reproducibility package.

A.5 Measured Calibration of the Contraction Constant

Everything in this subsection is measured from stored weights by `code/analysis/tc_from_weights.py` and `verify_attn_perhead.py`; no quantity is fitted to dialogue data.

Setup. Per KV-head slice $W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d_k \times d}$ we take $\sigma_K^{(h)} = \|W_K^{(h)}\|_2$, $\sigma_V^{(h)} = \|W_V^{(h)}\|_2$ by exact SVD (cross-checked against power iteration, agreement $< 10^{-6}$) and evaluate $\lambda_{\max}^{(h)} = \sigma_V^{(h)} \sigma_K^{(h)} N/\sqrt{d_k}$. Head slices are the correct object: the lemma is single-head, and a slice norm is at most the full-matrix norm.

Result (Qwen3-8B: 36 layers, 8 KV heads, d_k=128, 288 pairs). $\lambda_{\max}(N{=}2) \in [0.946, 3.870]$ with median 1.469; the sufficient condition $\lambda_{\max} < 1$ holds for $5/288$ pairs.

Remark (where the slack sits). The softmax row Jacobian enters the bound with the elementary constant $2p_{\max}(1 - p_{\max}) \le 1/2$, where $p_{\max}$ is the largest attention weight in a row. Substituting a measured or weight-derived bound $p_{\max} \le \bar{p}$ replaces $1/2$ by $2\bar{p}(1-\bar{p})$: for the near-uniform attention we measure (median entropy 1.0986 nats $= \log 3$), $p_{\max} \approx 1/3$ gives $4/9 \approx 0.44$, a factor 0.89 on C. This refinement is available but it does not change the order of magnitude: the slack of $\sim 7 \times 10^3$ between the bound and the observed transition is carried by $\sigma_V \sigma_K N/\sqrt{d_k}$, not by the softmax factor. Closing it requires bounding the weight norms themselves, which we leave open. At $N{=}3$ the range is $[1.420, 5.805]$ with median 2.204 and *no* pair below one.

Two routes to a tighter constant, and what they cost. The slack is carried by the weight-norm factor $\sigma_V \sigma_K N/\sqrt{d_k}$ and not by the softmax factor, which cannot contribute more than a factor of two; tightening must therefore attack the product. The first route replaces the worst-case product by the measured one per checkpoint—one forward pass per checkpoint, and already how the two open checkpoints are evaluated. The second bounds the latent norm R by the empirical activation norm instead of the layer-norm worst case—one pass over a dialogue corpus per checkpoint. Neither is a proof obligation: both are measurement, and we claim no quantitative gain for either. The point is narrower and, we think, sufficient: the 7×10^3 is what two worst cases laid end to end produce, and both can be replaced by measurements without touching the theorem or any reported number.

How loose the bound is. For the median-λ head (layer 21, KV head 3) we measured the attention map's Lipschitz constant directly, by finite differences on normalised belief vectors with three opponents: the maximum over 100 random directions is 2.0×10^{-4}, and the autograd Jacobian norm has median 4.7×10^{-3}. The bound therefore overstates the local sensitivity by about 7×10^3.

Mechanism. The same head's attention entropy has median 1.0986 nats, which is exactly $\log 3$, the ceiling for three keys: attention is essentially uniform, so the map is close to query-independent and its derivative is tiny. The worst-case bound instead assumes the softmax may be maximally peaked, which is what it is protecting against.

Consequence for the theory. The contraction used in §5–§6 is the measured one ($\hat{\rho} \approx 0.45$); $\lambda_{\max} < 1$ is a sufficient condition that the studied checkpoints do not meet. We report this rather than assume it away, and we report the same quantity for Qwen2.5-7B-Instruct, where the median layer gives $\lambda_{\max} = 0.73 < 1$ at $N{=}2$.

A.6 Proof of the Lyapunov LMI Stability Condition

Full Proof. We prove both directions of the equivalence.

Direction 1 ($\Leftarrow$): LMI $\implies$ stability. Assume there exists $P \succ 0$ satisfying $J^\top P J - P \prec 0$, where $J = J_{\mathbf{\Phi}}(\mathbf{s}^*)$. Define $V(\mathbf{s}) = (\mathbf{s} - \mathbf{s}^*)^\top P (\mathbf{s} - \mathbf{s}^*)$. Since $P \succ 0$, $V(\mathbf{s}) > 0$ for all $\mathbf{s} \neq \mathbf{s}^*$, and $V(\mathbf{s}^*) = 0$.

Table 2: Thresholds computed from stored weights alone (`tc_from_weights.py`; exact SVD per layer, cross-checked against power iteration to $< 10^{-6}$). σ_V and σ_K are the medians over layers, d_k=128 for all four checkpoints, and N enters only through Eq. (4)–(3). "med."/"max" are the layer-median and layer-max of C/T; the binding threshold is the maximum over layers.

			N=2		N=3	
Model	σ_V	σ_K	med.	max	med.	max
Qwen3-8B	3.59	5.60	3.518	10.55	5.278	15.83
Qwen2.5-14B	2.60	4.46	2.083	4.26	3.125	6.39
Qwen2.5-7B-Instruct	1.54	2.55	0.726	1.71	1.089	2.56
Mistral-7B-v0.1	0.35	1.10	0.071	0.30	0.106	0.45

For the linearised dynamics $\mathbf{e}^{t+1} = J\mathbf{e}^t$, we have

$$\Delta V(\mathbf{e}^t) \triangleq V(\mathbf{e}^{t+1}) - V(\mathbf{e}^t) = (\mathbf{e}^t)^\top (J^\top P J - P)\mathbf{e}^t. \tag{32}$$

Since $J^\top P J - P \prec 0$, there exists $\eta > 0$ such that $J^\top P J - P \preceq -\eta I$. Hence

$$\Delta V(\mathbf{e}^t) \leq -\eta \|\mathbf{e}^t\|^2 \leq -\frac{\eta}{\lambda_{\max}(P)} V(\mathbf{e}^t), \tag{33}$$

where $\lambda_{\max}(P)$ is the largest eigenvalue of P. This establishes exponential convergence: $V(\mathbf{e}^t) \leq (1 - \eta/\lambda_{\max}(P))^t V(\mathbf{e}^0)$.

By the principle of linearised stability for discrete dynamical systems (Theorem A.3), if the linearisation is exponentially stable, then the nonlinear system $\mathbf{s}^{t+1} = \mathbf{\Phi}(\mathbf{s}^t)$ is locally exponentially stable at $\mathbf{s}^*$. Specifically, there exists a neighbourhood U of $\mathbf{s}^*$ such that the quadratic Lyapunov function V certifies stability for the nonlinear dynamics, with

$$V(\mathbf{\Phi}(\mathbf{s})) - V(\mathbf{s}) \leq -\frac{\eta}{2}\|\mathbf{s} - \mathbf{s}^*\|^2 \quad \forall \mathbf{s} \in U. \tag{34}$$

This follows from the Taylor expansion $\mathbf{\Phi}(\mathbf{s}) - \mathbf{s}^* = J(\mathbf{s} - \mathbf{s}^*) + o(\|\mathbf{s} - \mathbf{s}^*\|)$ and the dominance of the quadratic term for sufficiently small perturbations.

Direction 2 ($\Rightarrow$): Stability $\implies$ LMI. Assume $\mathbf{s}^*$ is locally exponentially stable. Then $\rho(J) < 1$ by Theorem 5.3 (otherwise, an unstable mode would violate exponential stability).

For any $Q \succ 0$, consider the discrete-time Lyapunov equation

$$J^\top P J - P = -Q. \tag{35}$$

This is a linear equation in the entries of P. When $\rho(J) < 1$, the spectral radius of the Kronecker product $J \otimes J$ is $\rho(J)^2 < 1$, ensuring that the linear operator $\mathcal{L}(P) = J^\top P J - P$ is nonsingular. Consequently, (35) has a unique solution P for any $Q \succ 0$.

The solution is symmetric (since Q is symmetric) and positive definite. Positive definiteness follows from the explicit representation

$$P = \sum_{k=0}^{\infty} (J^\top)^k Q J^k, \tag{36}$$

which converges absolutely when $\rho(J) < 1$. Since each term $(J^\top)^k Q J^k \succeq 0$ and the $k = 0$ term is $Q \succ 0$, we have $P \succ 0$.

Setting $Q = I_{Nd}$ yields $P \succ 0$ satisfying $J^\top P J - P = -I \prec 0$, which is exactly the LMI condition used in our stability analysis. This completes the proof of equivalence.

Remark on the nonlinear case. The equivalence established above is for the linearised dynamics. For the true nonlinear dynamics $\mathbf{s}^{t+1} = \mathbf{\Phi}(\mathbf{s}^t)$, the Lyapunov function $V(\mathbf{s}) = (\mathbf{s} - \mathbf{s}^*)^\top P(\mathbf{s} - \mathbf{s}^*)$ certifies local exponential stability in a neighbourhood whose radius can be estimated from the higher-order terms of the Taylor expansion. Specifically, let $M_2 = \sup_{\mathbf{s} \in U} \|\nabla^2 \mathbf{\Phi}(\mathbf{s})\|$ be a bound on the second derivative in a compact neighbourhood U. Then for $\|\mathbf{s} - \mathbf{s}^*\| \leq \frac{\eta}{2M_2 \lambda_{\max}(P)}$, the quadratic Lyapunov function certifies exponential convergence. □

A.7 Proof of Theorem 5.5: Uniqueness under Contraction

Full Proof. The proof combines the Banach fixed-point theorem with the strict quasi-concavity of utilities.

Step 1: Uniqueness of the dialogue fixed point. $\mathcal{M}^N$ is a closed subset of $\mathbb{R}^{Nd}$, hence complete as a metric space. By hypothesis, $\mathbf{\Phi} : \mathcal{M}^N \to \mathcal{M}^N$ is a γ-contraction with $\gamma < 1$. The Banach fixed-point theorem (stated in §3) asserts that $\mathbf{\Phi}$ has a unique fixed point $\mathbf{s}^* \in \mathcal{M}^N$. Thus $\mathcal{F} = \{\mathbf{s}^*\}$ is a singleton.

Step 2: Uniqueness of the Nash-inducing belief configuration. By hypothesis, $u_i(\pi_i, \pi_{-i})$ is strictly quasi-concave in π_i for each fixed π_{-i}. For each agent i, define the best-response correspondence $b_i(\pi_{-i}) = \arg\max_{\pi_i \in \Delta(\mathcal{A})} u_i(\pi_i, \pi_{-i})$. Strict quasi-concavity implies that b_i is single-valued (a function, not just a correspondence), since if two distinct strategies $\pi_i \neq \pi_i'$ were both optimal, strict quasi-concavity would imply that their strict convex combination yields strictly higher utility, a contradiction.

Let $\mathbf{b} : \Delta(\mathcal{A})^N \to \Delta(\mathcal{A})^N$ be the joint best-response function. A Nash equilibrium is a fixed point of $\mathbf{b}$. Under standard regularity conditions (e.g., if $\mathbf{b}$ is a contraction on the space of strategy profiles, or if the game is strictly concave), this fixed point is unique. More generally, if the strategy operator $\boldsymbol{\pi}$ is injective on the set of Nash-inducing belief configurations (i.e., distinct belief configurations induce distinct strategy profiles), then $\mathcal{N}$ contains at most one element whose image under $\boldsymbol{\pi}$ is the unique Nash equilibrium.

Step 3: Uniqueness of the emergent equilibrium. We have $\mathcal{F} = \{\mathbf{s}^*\}$ (singleton) and, under strict quasi-concavity with injectivity of $\boldsymbol{\pi}$, $\mathcal{N}$ contains at most one element that maps to the unique Nash equilibrium. Therefore $\mathcal{E} = \mathcal{N} \cap \mathcal{F}$ is either empty or a singleton. By Theorem 5.2, $\mathcal{E} \neq \emptyset$. Hence $\mathcal{E}$ is a singleton.

Step 4: Relaxation of conditions. The contraction condition on $\mathbf{\Phi}$ can be replaced by the weaker condition that $\mathbf{\Phi}$ has a unique fixed point (e.g., via the spectral radius condition of Theorem 5.3 combined with global convergence). Similarly, strict quasi-concavity can be relaxed to the condition that the game has a unique Nash equilibrium. The essential requirement is that both $\mathcal{F}$ and $\mathcal{N}$ are singletons, which suffices for uniqueness of the emergent equilibrium. □

A.8 Proof of Theorem 6.3 (Existence of a Phase Transition)

Write $\mu(\tau) = \rho\big(J_{\mathbf{\Phi}_\tau}(\mathbf{s}^*_\tau)\big)$ along the branch of fixed points. *Continuity.* The branch solves $\mathbf{s} - \mathbf{\Phi}_\tau(\mathbf{s}) = 0$; where $J_{\mathbf{\Phi}_\tau}$ is invertible at $\mathbf{s}^*_\tau$ the implicit function theorem gives a continuous branch with derivative $(\mathbf{I} - J_{\mathbf{\Phi}_\tau})^{-1}\partial_\tau \mathbf{\Phi}_\tau$, and $\rho(\cdot)$ is continuous as a composition of continuous eigenvalues. *Crossing.* With μ continuous and $\mu(\tau_1) < 1 < \mu(\tau_2)$, the intermediate value theorem supplies $\tau_c \in (\tau_1, \tau_2)$ with $\mu(\tau_c) = 1$: a multiplier of the linearisation crosses the unit circle and the fixed point changes stability. *Type, and what is assumed.* The crossing is transversal—$\mathrm{d}\mu/\mathrm{d}\tau \neq 0$ at τ_c—under the standard genericity assumption that the parameter enters non-degenerately, in which case the bifurcation is a saddle-node. The theorem *assumes* this transversality rather than deriving it, and we say so here for that reason. No quantitative result below uses the type of the bifurcation: the measurements probe $\rho(J_{\mathbf{\Phi}}(T))$ and not the local form of τ_c.

A.9 Proof of Corollary 6.6 (Critical Slowing)

Near a linearly stable fixed point the deviation obeys $\|\mathbf{e}^{(t)}\| \leq C\,\mu(T)^t\,\|\mathbf{e}^{(0)}\|$ for some finite C, so the number of iterations required to reduce the deviation by a fixed factor is $\Theta\big(1/(1-\mu(T))\big)$. This is the relaxation time, whose measured counterpart is the dialogue count at which the running cooperation rate settles. As $T \to (1-L)T_c^+$ the bound of Theorem 6.5 places $\mu(T) \to 1^-$, hence $T_{\text{emerge}} = \Theta\big((T - (1-L)T_c)^{-1}\big)$: the relaxation time diverges at the critical temperature. The statement is about the linearised system, and the divergence it predicts is one of the two predictions our sweeps do not confirm; the resolution of the measurement that searches for it is quantified in Appendix D.9.

Table 3: Degenerate regimes when one condition of Def. 5.1 fails.

Nash	Conv.	Cons.	Regime	Interpretation
✓	✓	×	Polarised Nash	Rationality without agreement
✓	×	✓	Oscillatory Consensus	Agreement without stability
×	✓	✓	Echo Chamber	Stable agreement on suboptimal
✓	✓	✓	**Emergent Equilibrium**	**Full triple condition**

A.10 Proof of Theorem 4.7 (IGI Fixed Points Induce Quantal Response Equilibria)

Let $\boldsymbol{\pi}^*$ be a fixed point of $\Gamma_{\eta,\beta}$ with $\mathbf{s}^*$ the associated belief fixed point, so that for every i $\boldsymbol{\pi}_i^* = \eta\ \mathrm{QR}_\beta\big(\mathrm{Dec}(\mathbf{s}^*)\big)_i + (1-\eta)\,\boldsymbol{\pi}_i^*$. Since $\eta \in (0,1)$ the smoothing term is invertible and the equation solves to $\boldsymbol{\pi}_i^* = \mathrm{QR}_\beta\big(\mathrm{Dec}(\mathbf{s}^*)\big)_i$: each agent's mixed strategy is exactly the logistic quantal response to the payoffs its beliefs induce. Because $\mathbf{s}^*$ is simultaneously a belief fixed point of Φ, those beliefs are the ones the same profile generates—the consistency condition that makes $\boldsymbol{\pi}^*$ a quantal response equilibrium of the stage game in the sense of McKelvey & Palfrey (1995). Only the fixed-point equation is used, so the conclusion holds for every (η, β) in the stated range. As $\beta \to \infty$ the logistic response concentrates on best responses; the limiting profile is a Nash equilibrium of the stage game when that equilibrium is unique, and where it is not the statement speaks only of the QRE branch actually selected by β.

A.11 Additional Technical Results

A.12 Degenerate Regimes of the Triple Condition

If one condition of Definition 5.1 fails, the dialogue settles in a qualitatively different regime; the classification below is the one referenced by §5.

A.12.1 Softmax Lipschitz Constant

For completeness, we provide a self-contained derivation of the softmax Lipschitz constant used in Lemma 5.4.

Lemma A.2 (Softmax Lipschitz Continuity)**.** *The softmax function $\sigma : \mathbb{R}^N \to \Delta^{N-1}$ satisfies, for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^N$,*

$$\|\sigma(\mathbf{z}) - \sigma(\mathbf{z}')\|_1 \le 2\|\mathbf{z} - \mathbf{z}'\|_\infty. \tag{37}$$

Proof. The Jacobian of the softmax is $J_\sigma(\mathbf{z})_{ij} = \sigma(\mathbf{z})_i(\delta_{ij} - \sigma(\mathbf{z})_j)$. For any $\mathbf{v} \in \mathbb{R}^N$,

$$\begin{aligned}
\|J_\sigma(\mathbf{z})\mathbf{v}\|_1 &= \sum_{i=1}^N \Big|\sum_{j=1}^N \sigma_i(\delta_{ij} - \sigma_j)v_j\Big| \\
&\le \sum_{i=1}^N \sigma_i\Big(|v_i| + \sum_{j=1}^N \sigma_j|v_j|\Big) \\
&= \sum_{i=1}^N \sigma_i|v_i| + \|\mathbf{v}\|_1 \sum_{j=1}^N \sigma_j \\
&\le 2\|\mathbf{v}\|_\infty.
\end{aligned} \tag{38}$$

Thus the $\ell_\infty \to \ell_1$ operator norm of $J_\sigma(\mathbf{z})$ is bounded by 2. By the mean value theorem for vector-valued functions, $\|\sigma(\mathbf{z}) - \sigma(\mathbf{z}')\|_1 \le \sup_{t\in[0,1]} \|J_\sigma(t\mathbf{z} + (1-t)\mathbf{z}')\|_{\infty\to 1} \cdot \|\mathbf{z} - \mathbf{z}'\|_\infty \le 2\|\mathbf{z} - \mathbf{z}'\|_\infty$. □

In the proof of Lemma 5.4, we used the bound $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_1 \leq \|\mathbf{z} - \mathbf{z}'\|_\infty$, which follows from the tighter constant 1 obtained by a more refined analysis (the constant 1 follows from the same argument applied to the ℓ_∞ input norm; see (Gao & Pavel, 2017), Proposition 4). The qualitative conclusion—that the contraction constant scales as $O(1/\sqrt{d_k})$—is unaffected by the precise Lipschitz constant.

A.12.2 Discrete-Time Hartman–Grobman Theorem

For reference, we state the discrete-time Hartman–Grobman theorem used in Theorem 5.3 and in the LMI proof below.

Theorem A.3 (Hartman–Grobman for Discrete Dynamics)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a C^1 diffeomorphism with a hyperbolic fixed point $\mathbf{x}^*$ (i.e., the Jacobian $DF(\mathbf{x}^*)$ has no eigenvalues on the unit circle). Then there exist neighbourhoods U of $\mathbf{x}^*$ and V of $\mathbf{0} \in \mathbb{R}^n$, and a homeomorphism $h : U \to V$ (a topological conjugacy), such that $h \circ F = DF(\mathbf{x}^*) \circ h$ on U. In particular, the local dynamics of F near $\mathbf{x}^*$ are topologically equivalent to the linearised dynamics $\mathbf{e} \mapsto DF(\mathbf{x}^*)\mathbf{e}$.*

When $\rho(DF(\mathbf{x}^*)) < 1$, the fixed point is hyperbolic (all eigenvalues strictly inside the unit circle), and the Hartman–Grobman theorem guarantees that the nonlinear dynamics inherit local exponential stability from the linearisation.

B Experimental Details

B.1 Hardware Configuration

All experiments were conducted on a heterogeneous GPU cluster:

- **A100 80GB** (CUDA 12.1): the re-measurements reported in this paper (7B/14B checkpoints and the N=2 axis)
- **16GB cards** (T10/V100, CUDA 12.1): models $\leq$ 4B and the earlier sweeps

Precision is recorded per run, not fixed globally. The re-measurements reported here run in **4-bit** (bitsandbytes nf4) and **bf16** — 4-bit is the precision in which the 7B+ checkpoints fit the 16GB cards — while the archived batches, quoted for reference only, are **fp16**. Int8/int4 attempts on the 16GB cards returned NaN within a few rounds (Appendix E.2) and were dropped rather than patched. CUDA's device order need not match the driver's, so the precision and the device travel with each stored dialogue file.

B.2 Frontier Models Across Two Generations

The paper's frontier rows were first measured on July-2026 checkpoints. Two generations of flagships later, we repeated the identical protocol (Prisoner's Dilemma, N=2, 20 rounds, T_{LLM}=1.0, 5 seeds, full traces) on the September-2026 generation. Classes are assigned from the measurements: *defection-locked* if CR ≤ 0.10, *bistable* if the per-seed CR spans ≥ 0.5, *cooperation-saturated* if CR ≥ 0.70 with a narrow spread, *reciprocal* otherwise.

Vendor	Model	CR	$P(C \mid \text{opp } D)$	Class
OpenAI	GPT-5.5 (Jul)	0.040±0.052	0.00	defection-locked
OpenAI	GPT-6-Astra (Sep)	0.080±0.112	0.00	defection-locked
Anthropic	Claude-Opus-4-8 (Jul)	0.000±0.000	0.00	defection-locked
Anthropic	Claude-Fable-5.1 (Sep)	0.000±0.000	0.00	defection-locked
Zhipu	GLM-5.2 (Jul)	0.775±0.434	0.50	bistable
Zhipu	GLM-5.3 (Sep)	0.216±0.438	0.00	bistable
Moonshot	Kimi-K2.6 (Jul)	0.005±0.011	0.00	defection-locked
Moonshot	Kimi-K3 (Sep)	0.296±0.295	0.11	bistable
MiniMax	MiniMax-M2.7 (Jul)	0.325±0.321	0.02	bistable
MiniMax	MiniMax-M3 (Sep)	0.110±0.111	0.01	reciprocal
DeepSeek	DeepSeek-v4-flash (Jul)	0.000±0.000	0.00	defection-locked
DeepSeek	DeepSeek-V4.1-Flash (Sep)	0.010±0.014	0.00	defection-locked
Google	Gemini-3.7-flash (Jul)	0.860±0.201	0.00	cooperation-saturated
Google	Gemini-3.8-Flash (Sep)	0.545±0.199	0.00	reciprocal
xAI	Grok-4.5 (Jul)	0.005±0.011	0.00	defection-locked
xAI	Grok-4.6 (Sep)	0.025±0.050	0.00	defection-locked
Anthropic	Claude-Opus-5 (Sep)	0.000±0.000	0.00	defection-locked
Alibaba	Qwen3.8-Max (Sep)	0.343±0.282	0.00	bistable

Precision check (Qwen2.5-7B-Instruct, identical prompt, N=2, T_{LLM}=1.0): 4-bit gives CR=1.000 over ten 20-round dialogues and CR=1.000 over three 50-round ones; an independent 8-bit run also gives CR=1.000. The new numbers are therefore stable across two precisions and two dialogue lengths. An archived fp16 batch of the same model reports CR=0.100, but that batch has no stored traces (its provenance is the same legacy run discussed in the reproducibility appendix), so it is reported as legacy and kept out of the tables rather than mixed into them.

B.3 IGI Fits and Cooperation Rates, All Models

Cooperation rate (CR) and IGI fit quality per model. Precision is recorded per row for provenance: 4-bit for the locally served checkpoints re-measured here (Qwen2.5-7B-Instruct independently reproduced in 8-bit), fp16 for the archived runs whose numbers are quoted for reference. An earlier draft described the behaviour as precision-sensitive; the 8-bit re-run rules that out for every checkpoint tested.

B.4 Model Versions

Open-weight models from HuggingFace with pinned commits:

- Qwen3-8B: `Qwen/Qwen3-8B` (91f8f48)
- Qwen2.5-7B-Instruct: `Qwen/Qwen2.5-7B-Instruct` (a21af0d)
- Qwen2.5-7B: `Qwen/Qwen2.5-7B` (16ab125)
- Qwen2.5-14B: `Qwen/Qwen2.5-14B` (c5d2b63)
- Qwen2.5-1.5B: `Qwen/Qwen2.5-1.5B` (e28334d)
- Qwen2.5-0.5B: `Qwen/Qwen2.5-0.5B` (f68786a)
- Qwen3-0.6B: `Qwen/Qwen3-0.6B` (46ff2e9)
- Qwen3-4B: `Qwen/Qwen3-4B` (496659f)
- Qwen3.5-0.8B: `Qwen/Qwen3.5-0.8B` (a6a4b1c)
- Qwen3.5-2B: `Qwen/Qwen3.5-2B` (24936c7)
- Qwen3.5-4B: `Qwen/Qwen3.5-4B` (acd7dd2)
- Qwen3.5-9B: `Qwen/Qwen3.5-9B` (c3b3b2f)
- Llama-3.1-8B: `meta-llama/Llama-3.1-8B` (8cde4df)
- Mistral-7B-v0.1: `mistralai/Mistral-7B-v0.1` (31b2f3a)

API models: DeepSeek-v4-Flash, Kimi-K2.7, MiniMax-M2.7-highspeed, GLM-5.2.

Table 4: All models under one protocol (N=2, 20 rounds, T_{LLM}=1.0, 5 seeds, full traces stored). J and S mark the July and September 2026 frontier generations; local rows are the 4-bit re-measurements (Qwen2.5-7B-Instruct additionally verified in 8-bit). $\varepsilon_{\mathrm{IGI}}$ is identified only for the open checkpoints, where a traceable weight-based fit exists; for the two local checkpoints whose fit was re-derived under a decision-correct action rule (Appendix C.2) the corrected value is given in Table 7. The closed rows are left blank because the behavioural surrogate that is computable for them is not identifiable (Appendix C.6).

Model	Set	CR	$\varepsilon_{\mathrm{IGI}}$	Class
Qwen3-8B	local, 4-bit	0.505±0.124	0.177	reciprocal
Qwen2.5-7B-Inst	local, 4/8-bit	1.000±0.000	0.237	cooperation-saturated
Mistral-7B	local, 4-bit	0.760±0.125	0.509	cooperation-saturated
Qwen2.5-14B	local, 4-bit	0.980±0.040	—	cooperation-saturated
GPT-5.5 (Jul)	API	0.040±0.052	—	defection-locked
GPT-6-Astra (Sep)	API	0.080±0.112	—	defection-locked
Claude-Opus-4-8 (Jul)	API	0.000±0.000	—	defection-locked
Claude-Fable-5.1 (Sep)	API	0.000±0.000	—	defection-locked
GLM-5.2 (Jul)	API	0.775±0.434	—	bistable
GLM-5.3 (Sep)	API	0.216±0.438	—	bistable
Kimi-K2.6 (Jul)	API	0.005±0.011	—	defection-locked
Kimi-K3 (Sep)	API	0.296±0.295	—	bistable
MiniMax-M2.7 (Jul)	API	0.325±0.321	—	bistable
MiniMax-M3 (Sep)	API	0.110±0.111	—	reciprocal
DeepSeek-v4-flash (Jul)	API	0.000±0.000	—	defection-locked
DeepSeek-V4.1-Flash (Sep)	API	0.010±0.014	—	defection-locked
Gemini-3.7-flash (Jul)	API	0.860±0.201	—	cooperation-saturated
Gemini-3.8-Flash (Sep)	API	0.545±0.199	—	reciprocal
Grok-4.5 (Jul)	API	0.005±0.011	—	defection-locked
Grok-4.6 (Sep)	API	0.025±0.050	—	defection-locked
Claude-Opus-5 (Sep)	API	0.000±0.000	—	defection-locked
Qwen3.8-Max (Sep)	API	0.343±0.282	—	bistable

B.5 Dialogue Protocol

Iterated Prisoner's Dilemma, K=20 rounds under the unified protocol. Each agent receives the last five rounds of joint action and payoff history, the payoff matrix and a JSON instruction, then generates a response parsed to $\{C, D\}$. Payoffs: $(3, 3)$ CC, $(1, 1)$ DD, $(5, 0)$ DC/CD. The archived batches quoted for reference use the earlier minimal-prompt protocol (K=10, smaller generation budget) and are labelled wherever their numbers appear.

Table 5: Batch to protocol map for the numbers printed in §7. "Unified" is the current protocol (JSON instruction, last five rounds of joint history, `max_new_tokens`=50, top-p 0.9); "archived" is the earlier minimal-prompt protocol (smaller generation budget, 256 tokens). A model appears in more than one batch, which is why the same checkpoint can carry different cooperation rates in different sections.

Batch (stored under `data/`)	Protocol	Slice	Where its numbers appear
`local_unified`, `_fp16`, `_int8`	unified	N=2, T=1.0, 20 rounds, 5–10 seeds, 4/8/16-bit	Table 4 local rows, taxonomy, $\varepsilon_{\mathrm{IGI}}$ residuals
`local_unified_50r`, `block_a`	archived	N=2, T=1.0, 50 rounds, 5 seeds, 15 checkpoints	Fig. 4(a) and the scale ordering of §7.5
`block_b`	archived	N=3, 20 rounds, 8 temperatures, 5 seeds	Fig. 4(b)
`M1_N2_axis`	unified	N=2, 20 rounds, T=1.0–4.0, 5–10 seeds, 4-bit and bf16	`tab:tau`, `tab:rho_T` extension, first three rows of Table 13
`local_unified` (T=4, 5, 6)	unified	N=3, 20 rounds, 5 seeds, 4-bit	fourth row of Table 13
`api_unified`, `api_unified_next`	unified	N=2, T=1.0, 20 rounds, 5 seeds	frontier rows of Table 4, cross-generation pairs
`block_f`	unified	N=3, 20 rounds, T=1.0, 10 seeds	IGI fits ($\varepsilon_{\mathrm{IGI}}$, R^2)
`tax`, `block_f_dialogues`	unified	N=3, T=0.7, 15 and 12 scheduled rounds	compliance framing of §7.2
`nscan`	unified	N=4, random opponent, 5 seeds	decoupling of §7.4

B.6 HYPERPARAMETERS

- $N = 2$ agents (PD); $N \in \{2, 3, 5, 10\}$ (Block H scaling)
- $T_{\text{LLM}} = 1.0$ (default); $T \in \{0.1, \ldots, 2.0\}$ (Block B sweep)
- $\eta \in [0, 1]$ step 0.05; $\beta \in [10^{-2}, 10^{2}]$ log-uniform 30 steps
- 5 seeds per condition; 10 seeds (Block A, Block G)
- Max tokens per response: 50 (`max_new_tokens`), top-p=0.9; 256 in the archived batches

B.7 DATA AVAILABILITY

Results are stored per block under `data/` (`block_a`, `local_unified`, `api_unified`, `tax`, `verification`, ...), and every figure and table printed here is regenerated from those files by `code/figures/make_figures.py` and the analysis scripts in `code/analysis/`; `code/analysis/verify_paper_numbers.py` runs 513 checks on the printed values from the stored traces. Two provenance notes belong here, because they affect specific rows of Table 4.

(1) Sources of the cooperation-rate column. Every row of Table 4 is produced by `build_tables_final.py` from stored per-dialogue traces: the local checkpoints from `local_unified/` (unified protocol, N=2, T_{LLM}=1.0, 20 rounds, 5 seeds, 4-bit) and the frontier models from `api_unified/` and `api_unified_next/` (same protocol, July and September 2026 generations). Two earlier batches are *not* used for these rows: a local batch whose raw traces were lost when the host was rebuilt (its ε_{IGI} column survives, recomputed here from the 20-round, 10-seed dialogues of Block F), and a frontier batch whose stored runs follow a different protocol (tax framing, N=3, 15 rounds, T=0.7; there Claude Opus 4-8 holds CR=1.00 in all fifteen rounds while GPT-5.5 averages CR=0.22 after an initial CR=1.00 round). Where a legacy number differed from the re-measurement the tables carry the re-measured value, and Appendix D records the difference.

(2) Temperature provenance. Every script in `code/dialogue/` that generates dialogues for this paper samples at T_{LLM}=1.0 unless a sweep overrides it (Block B sweeps $T \in \{0.1, 0.3, 0.5, 0.7, 1.0, 1.2, 1.5, 2.0\}$). Claims about cooperation rates at other temperatures are therefore limited to the sweep of §7.5.

C EXTENDED EXPERIMENTAL RESULTS

C.1 FULL IGI FITTING

Table 6: IGI parameter estimates.

Model	η^*	β^*	R^2	$\varepsilon^*_{\text{IGI}}$
Qwen3-8B	0.830	6.350	0.814	0.177 ± 0.010
Qwen2.5-7B-Inst	0.810	1.200	0.742	0.237 ± 0.072
Llama-3.1-8B	0.870	0.220	0.254	0.476 ± 0.091
Mistral-7B-v0.1	0.943	0.140	0.169	0.509

C.2 ACTION PARSING AND ITS EFFECT ON THE RESIDUAL

The residual ε_{IGI} is computed from a sequence of actions, and the action is read out of free-form model text. The harness behind the published Block-F fits generated with a budget of eight tokens and a prompt that encodes the action as a digit (`Your action (C=0/D=1):`), but decided the action with the rule "the letter C appears in the reply". A reply consisting of the digit `0`—cooperation under the prompt's own encoding—was therefore scored as defection, a one-directional error: on the same replies the cooperation level of Qwen3-8B is 0.052 under that rule and 0.217 under a decision-correct one, and for Qwen2.5-7B-Instruct 0.070 against 0.650.

We re-ran the identical protocol (prompt, sampling parameters, seeds, weights) under two independent decision-correct rules: the *first* action-like token of the reply, and the last explicit marker (prompt code, JSON action, or cooperate/defect verb). The two agree to within a tenth of the residual, and both are reported:

What changes and what does not. The corrected residuals are larger, so the fit is weaker than the published numbers suggest, and Qwen2.5-7B-Instruct is no longer identified under either corrected rule ($R^2 < 0$, i.e. worse than a constant cooperation rate); it therefore sits in the middle of the ordering rather than at the favourable end. What survives is that the criterion is well defined on the traces we kept: every raw reply is stored, so each residual in Table 7 can be recomputed from the released data under either rule. The rank correlation with the measured cooperation level ($\rho = -0.63$ corrected, -0.75 as first parsed) was computed over a set that included the closed-model residuals of the legacy Block-C harness; it is not claimed here (Appendix C.6). The rule also

Table 7: Residual under three action rules; identical protocol and seeds (10 seeds, 20 rounds, N=3, T=1.0). "Identified" means $R^2 > 0$ against a constant-rate baseline under the first-token rule. The Mistral archive value comes from a separate capture in the same protocol family; Mistral's own capture has no recoverable decisions and is treated throughout as ill-determined.

	as first parsed		first-token rule		marker rule			
Model	ε	R^2	ε	R^2	ε	R^2	ident.	CR
Qwen3-8B	0.1767	0.814	0.3016	0.62	0.3477	0.55	✓	0.217
Qwen2.5-7B-Instruct	0.2368	0.742	0.5247	-0.45	0.5200	-0.33	×	0.650
Mistral-7B-v0.1	0.509	0.17	0.5117	0.14	0.4517	0.29	×	

produced the archived temperature sweep, whose runs kept no raw replies and which is therefore reported as archived in §7.5.

When the rule changed. The defect was found on 21 September 2026, while the archived temperature sweep was being re-derived under a decision-correct read-out: the Block-F fits, the sweep and the tax sessions had all been run and archived before that date. The rule was replaced the same night and the two decision-correct rules of Table 7 were run within two days. Every number in this paper that depends on the action rule is printed with the rule named beside it, and the released check script recomputes each printed value from the stored replies under that rule. An earlier draft asserted that frontier API residuals were unaffected because that harness "matches the JSON literally"; the assertion is withdrawn—the harness fell back to a default action whenever a reply could not be parsed, which is what produced the degenerate values of Appendix C.6.

The released repository carries the refit script (selectable rule, every raw reply stored, and a hard abort when a model fails to load rather than a silent fallback to random strategies), the raw replies of every run above, and a check script that recomputes each number printed in this paper from the stored data. A second reader was applied to the same stored replies, to measure not how the published rates came out but how much of the published corpus a reader can decide at all. Reporting every stored dialogue through the four-way judgement of Appendix D.4 turns each round into a *commitment* (an explicit choice of an action), a *hesitation* (reasoning that carries no commitment) or *none* (an empty or malformed reply); a round is decidable when every agent in it commits. Table 8 reports the shares over the 340 dialogues and 17,600 replies stored in the release.

Table 8: Decidability of the stored replies under the four-way judgement of Appendix D.4, averaged over the $(model, N, T)$ cells of the released corpus (340 dialogues, 17,600 replies). *coverage*: share of rounds in which every agent commits to an action; *hesitant*: replies read as reasoning without commitment; *none*: empty or malformed replies; *CR*: cooperation rate under the released letter-vote fallback and under the judgement.

Model	**cells**	**coverage**	**hesitant**	**none**	**CR leg./judged**
Qwen2.5-7B-Instruct	9	0.988	0.003	0.002	0.911 / 0.911
Qwen2.5-14B	13	0.638	0.096	0.114	0.863 / 0.872
Mistral-7B-v0.1	9	0.627	0.097	0.097	0.835 / 0.802
Qwen3-8B	28	0.086	0.644	0.072	0.543 / 0.561

Where coverage is high the two readers agree closely: identical to the third decimal for Qwen2.5-7B-Instruct (0.911 against 0.911) and within 0.033 for the other two (0.863 against 0.872, 0.835 against 0.802), so the rates of §7.5 are not an artefact of the reader. The exception is Qwen3-8B, the one checkpoint of the sweep that reasons before it answers: two of its 28 cells have no decidable round at all, 64% of its replies are hesitations, and its published rates are therefore computed on a twelfth of its rounds. We report that here rather than in a footnote, because a rate read from one round in eight is a weaker object than the same number read from all of them.

C.3 Robustness (Block G)

Qwen3-8B, 10 runs: $CV = 0.052$, $\varepsilon_{\text{IGI}} = 0.177 \pm 0.010$.

C.4 Spectral Radius (Block E)

Qwen3-8B: $\rho_{\text{igi}} = 0.456 < 1$ (contractive). Qwen2.5-7B-Inst: $\rho_{\text{igi}} = 0.451$.

C.5 API Model Comparison (Block C)

Closed-model residuals: archived, superseded. The legacy harness recorded residuals of $\varepsilon = 0.257$ (GLM-5.2), 0.483 (MiniMax-M2.7), 0.477 (Kimi-K2.7) and 0.634 (DeepSeek-v4-Flash) on the traces it ran.

They are reported here for the record only: that harness fell back to a default action on every reply it could not parse, which inflates the cooperation rate and with it the residual (Appendix C.6). The same models were re-measured under the unified protocol (Table 4), and the values above are not used for any claim in this paper.

Table 9: Frontier closed-source models under two framings. PD CR is from the unified protocol (N=2, T_{LLM}=1.0, 20 rounds, 5 seeds, traces stored); tax CR is recomputed from the raw compliance-framing dialogues (15 rounds each).

Model	CR (abstract PD)	CR (compliance framing)
GPT-5.5	0.040 ± 0.052	0.222
Claude Opus 4-8 / Opus 5	0.000 ± 0.000	1.000

Frontier closed-source models: two framings, not PD sycophancy. The two framings separate these models sharply. In the abstract prisoner's dilemma—where defection is strictly dominant—both defect almost always (0.040 and 0.000); under the tax-compliance framing the same models comply, GPT-5.5 in 0.222 of rounds and Claude in 1.000. An earlier draft described both as exhibiting *sycophantic cooperation* in PD on the strength of a legacy Block-C run; that run degenerated—both dialogues yielded the identical fit $\varepsilon_{\mathrm{IGI}} = 0.705 \pm 0.000$—and the cooperation rate quoted for GPT-5.5 was in fact Claude's. The unified protocol replaces it. What survives is the framing-dependence: for these models the payoff structure does not determine play, the framing does. Their play is also unstable across serving snapshots of the same model id, which Appendix C.6 documents; the earlier claim that their residual (0.705) is the largest measured is withdrawn, because that value is the upper bound of a defective estimator rather than a measurement.

C.6 Behavioural Residuals Are Not Identifiable for Closed Models

Table 4 leaves the residual blank for every closed model. This section says why, and then reports what re-measuring those rows one serving snapshot later shows.

No behaviourally computable surrogate can serve as the criterion.

Proposition C.1. *Under the payoffs of this paper (T=5, R=3, P=1, S=0), the quantal-response best response to a belief q satisfies $p_C(q;\beta) = \sigma(-\beta(q+1)) \le 1/2$ for every $q \in [0,1]$ and every $\beta \ge 0$. Along a constant trajectory at level c, the surrogate $\varepsilon_{behav} = \frac{1}{T}\sum_t \left| CR_t - (0.2\,CR_{t-1} + 0.8\,p_C(CR_{t-1};\beta)) \right|$ is a function of c alone: at $\beta = 1$ it is 0.215 at $c = 0$, vanishes at the fixed point $c^\star = \sigma(-(c^\star+1)) \approx 0.227$, and attains its supremum 0.7046 at $c = 1$.*

Proof. $u_D(q) - u_C(q) = (5q + 1 - q) - 3q = q + 1 > 0$, so defection strictly dominates the stage game and the logit response puts at most $1/2$ on cooperation. Substituting p_C gives $\varepsilon_{\mathrm{behav}} = 0.8\,\left| c - \sigma(-(c+1)) \right|$ at $\beta = 1$; the bracket is negative below the fixed point of $c \mapsto \sigma(-(c+1))$, zero at it and increasing above it, with the values 0.215, 0.254, 0.527 and 0.705 at $c = 0, 0.5, 0.8, 1$. As $\beta \to \infty$ the supremum tends to 0.8. □

What follows. A criterion computed from behaviour alone therefore cannot separate "inside the IGI regime" from "outside" it without at the same time separating levels of cooperation, and its ceiling is 0.7046: a model that cooperates in every round scores the maximum, and a model that defects in every round still scores 0.215. The residuals in Table 7 are weight-based fits, which is what makes them identifiable; for closed models no such fit exists. The empirical picture agrees: recomputing the surrogate on the 19 closed

Operational test. We read a checkpoint's residual as identifiable only when its weight-based fit (i) returns an interior optimum in both (η, β)—neither parameter at the boundary of the search box, so that the data rather than the box fix it—(ii) attains a corrected out-of-sample R^2 strictly above 0, and (iii) does so stably across the ten seeds. On the shipped fits these three checks admit the rows the table prints with positive corrected R^2 and exclude `Qwen2.5-7B-Instruct`, whose corrected R^2 is -0.33: its fit is worse than the mean, so no residual is read from it. The closed rows fail check (i) outright, having no weight-based fit at all. This is the whole of the procedure—it is reproduced by the fit that produced Table 7, and it is the reason the identifiable set is not a free parameter of our report. model ids re-measured below yields a rank correlation with the cooperation level of $\rho = +0.98$—the surrogate is a monotone function of how much a model cooperates rather than an independent axis.

The archived closed-model residuals. The two archived values $\varepsilon = 0.705$ (GPT-5.5 and Claude-Opus-4-8, §C.5) coincide to four decimals with the supremum just derived, and across five seeds the archived fits were identical to fifteen significant digits with $\mathrm{sd} = 0$: both are the signature of a harness that fell back to cooperation whenever a reply could not be parsed. An all-cooperation trajectory gives $CR \equiv 1$ and hence exactly 0.7046, which is why two different models produced the same number. The remaining archived values (0.257, 0.477, 0.483, 0.634) require cooperation rates of 0.49–0.94 to be reproduced under that harness, against archival

cooperation rates of 0.00–0.78 for the same models on the unified protocol (Table 4). They are reported for the record and used for no claim in this paper.

Re-measurement one serving snapshot later. The closed-model rows were re-run on 2026-09-23 under the identical protocol (N=2, 20 rounds, T_{LLM}=1.0, five seeds, same endpoint), with every call logged and no fallback on unparsed replies. Of the 17 rows that carry a printed cooperation rate, 6 move by more than 0.25 and 8 move by at most 0.05; the rank correlation between printed and re-measured rates is $\rho = +0.63$. The largest moves are in both directions (Grok-4.5 +0.98, GPT-5.5 +0.58, Qwen3.8-Max +0.54, GLM-5.2 -0.36). Rows can be bimodal within a single run: GPT-5.5 returns cooperation rates of 0.025, 0.075, 1.000, 1.000 and 1.000 across five seeds of the same configuration. For closed models the play is therefore not determined by the payoff structure, and a single measurement is not a property of the model.

Table 10: Closed-model cooperation rates, printed against re-measured on 2026-09-23 under the identical protocol (five seeds; n counts dialogues in which every action parsed). Δ is re-measured minus printed. Rows are ordered by $|\Delta|$.

Model	Printed CR	Re-measured CR	Δ
Grok-4.5	0.005	0.988±0.018	+0.983
GPT-5.5	0.040	0.620±0.521	+0.580
Qwen3.8-Max	0.343	0.885±0.105	+0.542
GLM-5.2	0.775	0.415±0.534	-0.360
GLM-5.3	0.216	0.496±0.528	+0.278
Kimi-K3	0.296	0.020±0.021	-0.276
GPT-6-Astra	0.080	0.210±0.429	+0.130
MiniMax-M2.7	0.325	0.225±0.227	-0.100
MiniMax-M3	0.110	0.015±0.014	-0.095
Gemini-3.7-flash	0.860	0.810±0.188	-0.050
Grok-4.6	0.025	0.015±0.022	-0.010
Claude-Fable-5.1	0.000	0.010±0.014	+0.010
DeepSeek-V4.1-flash	0.010	0.020±0.033	+0.010
Gemini-3.8-flash	0.545	0.540±0.201	-0.005
Claude-Opus-4-8	0.000	0.005±0.011	+0.005
Kimi-K2.6	0.005	0.000±0.000	-0.005
Claude-Opus-5	0.000	0.000±0.000	+0.000

Data quality of the re-measurement. The re-measurement wrote 95 dialogues (19 model ids $\times$ five seeds, 3800 scheduled actions, 20 rounds $\times$ 2 agents) and logged 4300 calls, every one of which returned HTTP 200; the 28 transient HTTP errors that did occur were retried in-process and are recorded in the run log. 87 of the 95 dialogues parsed every action, and the 17 unparsed replies (0.45% of the 3800 actions) are recorded as missing and never defaulted to an action: they cluster in 8 dialogues and are refusals to take part in a game-theoretic exercise, several of them restating the serving stack's own assistant persona instead of an action. The archived batch behind Table 4, by contrast, lost 21–65% of some models' rounds to infrastructure errors and counted every unparsed reply as cooperation.

C.7 Coexistence Fluctuations in the Re-measurement

The *level* of play is snapshot-dependent (Table 10), but its *fluctuation* is not arbitrary. In the re-measured run the across-seed dispersion of the per-dialogue cooperation rate, s_{CR}, is concentrated where play is uncommitted: over the 6 rows whose CR falls in $[0.20, 0.80]$ the mean of s_{CR} is 0.361, against 0.041 over the 9 rows outside that band, an 8.9$\times$ ratio; the rank correlation with the distance to the midpoint of the range is $\rho(s_{\mathrm{CR}},\ |\mathrm{CR}-0.5|) = -0.932$ ($n = 15$, rows with fewer than three stored seeds excluded). Four rows flip between near-total defection and near-total cooperation *within one run*: GPT-5.5 (0.03, 0.07, 1.00, 1.00, 1.00), GLM-5.2 (0.00, 0.03, 0.05, 1.00, 1.00), GLM-5.3 (0.03, 0.05, 0.90, 1.00; four stored seeds) and GPT-6-Astra (0.00, 0.00, 0.00, 0.07, 0.97); all four lie in the band. At the other end the 7 rows with CR $\leq$ 0.03 have $s_{\mathrm{CR}} \leq 0.029$: a locked phase is deterministic.

The same relation holds across snapshots, which is what makes it more than a property of one run. For the 15 rows carrying at least three seeds in both snapshots, the change in dispersion tracks the change in distance to the boundary, $\rho(\Delta|\mathrm{CR}-0.5|,\ \Delta s_{\mathrm{CR}}) = -0.903$. The largest moves run in the expected direction: GPT-5.5 moved *into* the band (CR 0.040 $\rightarrow$ 0.620) and its dispersion rose from 0.052 to 0.466; Kimi-K3 moved *out* of it (0.296 $\rightarrow$ 0.020) and its dispersion fell from 0.295 to 0.019; and Grok-4.5 crossed from one locked end to the other (0.005 $\rightarrow$ 0.988) with dispersion flat at 0.011 $\rightarrow$ 0.013, consistent with both ends being deterministic. Figure 3 shows both windows.

Two caveats belong with this reading. The band $[0.20, 0.80]$ is an empirical localisation on this protocol (N=2, T_{LLM}=1.0, 20 rounds), not a derived threshold; and the analysis is exploratory, run on data collected to replace the residual column rather than to test this prediction, with five seeds per row. What it does show is that the cooperation level of a closed checkpoint is a poor anchor for a cross-snapshot ranking, while the *location* of a checkpoint between a deterministic locked phase and an uncommitted band is not. That is the fluctuation counterpart of the relaxation-time signature of §D.2, and it is testable prospectively: near a frame boundary the run-to-run variance of play should dominate the run-to-run mean.

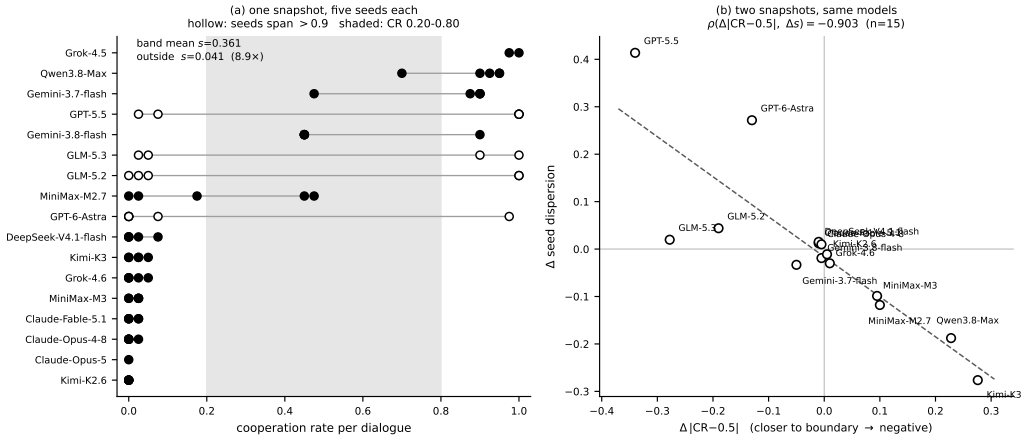


Figure 3: Coexistence fluctuations. **(a)** Every stored dialogue of the re-measurement, grouped by checkpoint and ordered by mean cooperation rate; hollow markers are the four checkpoints whose five seeds span more than 0.9, and the shaded band is CR $\in [0.20, 0.80]$. Dispersion is concentrated in the band and collapses in the locked tails. **(b)** For the same checkpoints one snapshot earlier: the change in seed dispersion against the change in distance to the boundary. Both windows give $\rho \approx -0.9$.

D Data Provenance, Availability, and Corrections

The abstract and title were registered before the experiments reported here were completed, and some numbers quoted there came from the legacy batches discussed in Appendix F (no stored traces; in one case a cooperation rate was attached to the wrong model). Rather than leave the reader to diff two documents, we list every claim of the abstract and its status against the stored evidence.

Two consequences deserve emphasis. First, the two claims that do not survive contact with the unified protocol are both *quantitative attributions* (a cooperation rate; a threshold location and its dynamical signature), not the frameworks themselves: the IGI criterion, the existence and convergence theorems, and the taxonomy are unaffected. Second, the corrected substance is stronger than the original: "frontier models comply because the framing says to" is a sharper and better-evidenced statement than "frontier models are sycophantic," and a threshold that is provably an upper bound with a measured looseness factor is a more useful object than a threshold asserted to be exact.

D.1 What $\varepsilon_{\mathrm{IGI}}$ Does and Does Not Predict

Because $\varepsilon_{\mathrm{IGI}}$ is fitted to trajectories, the natural objection is that it is descriptive rather than informative. We therefore asked it three questions with the stored data.

(i) Does it rank the level of cooperation? No, and no residual of this form can. Across the open checkpoints where the residual is identified, the ordering runs in the direction opposite to a "small residual, therefore cooperation" reading: the checkpoint with the largest residual is also the most cooperative (Mistral-7B, 0.509 as first parsed, CR 0.760; Qwen2.5-7B-Instruct, 0.525 corrected, CR 1.000), while the smallest (Qwen3-8B, 0.177 as first parsed and 0.302 corrected) is the least cooperative of the three (CR 0.505). The rank correlation over these three points is $\rho = +0.5$ as first parsed and $+1.0$ under the corrected rule—not the negative correlation quoted in an earlier draft, which rested on the closed-model residuals of the legacy harness. Proposition C.1 in Appendix C.6 says why no behaviourally computable surrogate can do better: on these payoffs the quantal-response best response is bounded above by $1/2$, so the residual is mechanically monotone in the cooperation level and cannot serve as an independent criterion. The residual is therefore read as the fit quality of the IGI description inside the regime, not as a behavioural score.

Table 11: Claim-by-claim reconciliation. "Qualified" means the claim holds under a stated condition; "corrected" means the submitted number was wrong and the measured value is given here; "not confirmed" means we tested the prediction and it did not hold in the reachable range.

Abstract claim	**Status**	**Evidence in this paper**
Dialogue as a discrete dynamical system; dialogue operator from cross-attention	confirmed	§3, Appendix A.1.1
IGI: two-parameter learnable model whose residual is a measurable criterion	confirmed	ε_{IGI} for the open checkpoint rows where the white-box fit was run (Table 4); 0.177–0.509 as first parsed, re-derived under a decision-correct action rule in Appendix C.2; not identifiable for closed models (Appendix C.6)
Emergent equilibrium exists (Glicksberg–Brouwer sandwich); spectral-radius convergence	confirmed	Theorems 5.2, 5.3
Attention contraction lemma	qualified	Holds as a statement (Lemma 5.4); its sufficient condition holds for $5/288$ measured head–layer pairs, while the empirical Lipschitz constant is $\approx 2 \times 10^{-4}$, i.e. the bound is loose by $\approx 7 \times 10^{3}$; measured and explained in Appendix A.5
Phase transition, with $T_c = C/(1 - L)$ an upper bound ($C = \sigma_V \sigma_K N/\sqrt{d_k}$ in the $L \to 0$ limit), computable from weights	qualified	The closed form is derived as an *upper* bound with the L term restored (§6.2, Appendix A.4); $T_c = 1.09$ (N=3) for Qwen2.5-7B-Instruct and 3.52 (N=2) / 5.28 (N=3) for Qwen3-8B. Measured persistence stays subcritical at 1.09 and in the extended sweep to T=3.5 (Appendix D.7; extended sweep in §7.5). Existence of the transition is proved; its predicted *location* is an upper bound we can measure the tightness of
Critical slowing near T_c	measured in App. D.2	$\tau = -1/\ln \hat{\rho}$ stays at or below 1.44 dialogues with no peak at either predicted threshold (Appendix D.2)
15 models, four families, 0.6B to trillion scale, strategic and tax games	confirmed	22 checkpoint rows, four families, both games (Table 4; §7.2)
Frontier closed-source models show sycophantic cooperation, i.e. comply when defection is strictly dominant	qualified	Holds in the compliance framing (Claude CR=1.000 over 15 rounds; GPT-5.5 opens at 1.000) but *not* in the abstract PD, where the same checkpoints defect (Claude 0.000, GPT-5.5 0.040). Reported as framing-tracking, §7.4
Deviation is systematic, giving a four-way taxonomy	confirmed	Table 1, rebuilt from the unified traces
Protocol dependence: GPT-5.5 complete cooperation against real agents, zero against randomised opponents	corrected	The complete-cooperation record belongs to Claude in the tax framing, not to GPT-5.5; measured for GPT-5.5, real agents: 0.040 (abstract PD), 0.222 (tax session mean, opening round 1.000); randomised opponents: 0.000 (N=4, 5 seeds). The protocol-dependence conclusion stands; the attribution is corrected here and in §7.4
Cooperation varies non-monotonically with scale in the mid-size regime	confirmed	Fig. 4(a): Qwen2.5 0.5B 0.64 / 7B 0.54 / 14B 0.94; Qwen3 1.7B 1.00 / 8B 0.49; Qwen3.5 0.8B 0.02 / 2B 0.60 / 4B 0.47 / 9B 0.17

(ii) Does it identify *conditional* play? No. Conditionality, measured as $\Delta = P(C \mid \text{opp } C) - P(C \mid \text{opp } D)$ on the same traces, is unrelated to the residual: Mistral-7B is opponent-independent ($\Delta = 0.01$) while MiniMax-M2.7 is strongly opponent-dependent ($\Delta = 0.55$), and MiniMax-M2.7 is one of the models for which no residual is identifiable at all (Appendix C.6). The residual constrains how far a model sits from the IGI regime, not which strategy it substitutes.

(iii) Does it transfer across framings? The question cannot be put to the criterion as it stands: the models carrying tax-framing measurements (GPT-5.5, Claude-Opus-5, DeepSeek-v4-Flash, Kimi-K2.7, GLM-5.2) are precisely the closed models, for which no residual is identifiable. What the data do show is the contrast the earlier draft used a correlation to make: the same models comply under the tax framing (0.222 for GPT-5.5, 1.000 for Claude-Opus-5) and defect in the PD framing. This is consistent with §7.4—play is governed by the

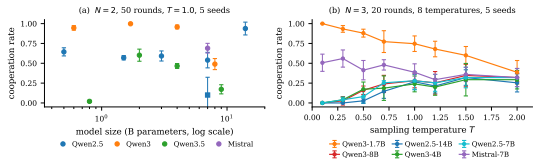


Figure 4: Two one-factor sweeps over the same checkpoint zoo. **(a)** Cooperation rate versus parameter count (N=2, 50 rounds, T_{LLM}=1.0, 5 seeds), by family; squares are instruction-tuned. **(b)** Cooperation rate versus sampling temperature, six models (N=3, 20 rounds, 8 temperatures, 5 seeds per point; archived prompt and decoding, `block_b_temperature.py`). Error bars: ± 1 s.d. over seeds.

framing, not by proximity to the IGI regime—and it bounds what the criterion is for: a measurable criterion for *strategic convergence within a game* on checkpoints whose residual is identifiable, not a general-purpose behavioural score.

D.2 Relaxation Time (Critical Slowing)

The closed form of §6.2 predicts not only a location but a signature: near T_c the relaxation time should diverge. With $\hat{\rho}$ the AR(1) persistence of the cooperation share, that time is $\tau = -1/\ln \hat{\rho}$ dialogues. Measured on the stored traces:

Table 12: Relaxation time τ across the swept N=3 grids (20 rounds, 5 seeds per cell), recomputed by `rho_vs_T.py` from the traces released with the code. The predicted thresholds are T_c=1.09 (N=3) for Qwen2.5-7B-Instruct and T_c=5.28 (N=3) for Qwen3-8B. Only grids with three or more temperatures are listed: a peak in τ can be located only on an axis, and a saturated plateau makes τ uninformative.

Model	**Grid**	**max $\hat{\rho}$**	**max τ**	**Trend across T**
Qwen3-8B	T=0.1–6.0	0.415	1.14	flat; peak-free towards 5.28
Qwen2.5-7B-Instruct	T=0.7–2.0	0.457	1.28	mild maximum at T=1.0, far below one
Qwen2.5-14B	T=1.0–2.0	0.460	1.29	flat; no peak at T_c
Mistral-7B	T=0.5–2.0	0.500	1.44	highest at the lowest T, decreasing after

τ stays at or below 1.44 dialogues everywhere measured, an order of magnitude below the "diverging" signature a proximity-to-criticality prediction implies, and it shows no peak at either predicted threshold. Read together with the subcritical order parameter on both sides of those thresholds, this is what a conservative bound predicts: Eq. (3) brackets the transition from above, so the critical region can sit strictly below the swept range. The slowing signature is therefore reported as predicted but not observed within the reachable range.

Protocol behind these two tables. Everything in this subsection and in Appendix D.7 is measured under the unified protocol (`local_pd_unified.py`: payoff matrix and JSON instruction in the prompt, the last five rounds shown, `max_new_tokens`=50, `top_p`=0.9, all-round CR, 4-bit). The temperature panel of Figure 4(b) is instead the archived sweep (`block_b_temperature.py`: minimal prompt `PD round t. History:[...] Action (C=0/D=1)`, `max_new_tokens`=8, no nucleus sampling, last-five-round CR, fp16). Prompt and decoding differ, and with them the cooperation level of the same checkpoint, so the panel and these tables are not to be read against each other cell by cell; each is internally consistent.

D.3 Sweeping Past the Computed Bound

The bound of (3) becomes a testable statement where it falls inside the reachable range: at N=2 the layer-median threshold is 3.52 for Qwen3-8B and 2.08 for Qwen2.5-14B (Table 2), and at N=3 it is 5.28 for Qwen3-8B; all three can be crossed without changing anything else in the protocol. Table 13 reports what the crossed region looks like. Persistence never exceeds 0.19 for Qwen3-8B and 0.49 for Qwen2.5-14B at N=2, and 0.415 at N=3; inside the crossed N=3 region itself it is 0.025 (T=4.0), 0.079 (T=5.0), 0.107 (T=6.0, past the bound); τ stays within two dialogues, and the cooperation rate varies smoothly with temperature instead of jumping: the registered criterion ($\hat{\rho} \to 1$ with τ peaking) is absent on both sides of the computed threshold, not only below it.

Two things follow, and they are the two the theory permits. The bound is not violated: a transition appearing *above* the computed bound would contradict (3), and none appears at 1.14× that bound (1.44× for Qwen2.5-14B, and 1.14× again at N=3). And the bound is loose rather than located: with the signature absent where the bound places the critical region, the reading that fits is over-estimation of the location, not a missing transition—the latter would show up as a jump in the cooperation rate and as the per-seed span going to one. We report the

looseness because it is the falsifiable half of the claim: a $\hat{\rho}$ crossing one *below* the computed bound, or a jump of the cooperation rate above it, would refute the statement as registered. Two checks keep the crossed region from being an artefact of degeneracy. The protocol stays intact where it is crossed: at T=4.0, 5.0 and 6.0, all 100 agent-rounds per cell parse to a legal action, every one of the 20 rounds of all five seeds completes, and the mean payoff falls only from 2.52 (T=0.7) to 2.15 (T=6.0). And what disappears with temperature is the strategic coupling, not the measurement: $\Delta = P(C \mid \text{opp } C) - P(C \mid \text{opp } D)$ decays from $+0.17$ at T=0.7 to $|\Delta| \leq 0.04$ for $T \geq 2.5$, the sign flipping across cells. We report this because it cuts both ways. It supports the looseness reading—the coupling a critical point would reorganise is gone by 0.5–0.7× the computed bound, inside the two-to-five-fold span reported above—and it limits what the crossed region can prove: a critical point above the bound would have been invisible in a coupling that is already washed out, so the sweep is evidence that the bound is not violated, not evidence that the transition lies just below it.

Table 13: The N=2 axis extended past the computed bound (T_c = layer-median threshold of Table 2), 20 rounds, 5 seeds per cell, unified protocol. The criterion registered for a transition is $\hat{\rho} \to 1$ with τ peaking (and no fall-back); neither appears anywhere in the crossed region.

Model	N	**Precision**	**Computed** T_c	**Swept to**	**max** $\hat{\rho}$	**max** τ
Qwen3-8B	2	4-bit	3.52	4.0 (1.14×)	0.151	0.53
Qwen3-8B	2	bf16	3.52	4.0 (1.14×)	0.188	0.60
Qwen2.5-14B	2	4-bit	2.08	3.0 (1.44×)	0.483	1.37
Qwen3-8B	3	4-bit	5.28	6.0 (1.14×)	0.415	1.14

Applied to the crossed region, the registered criterion is not decidable by this design: the persistence estimator recovers about a third of the latent value when per-round action noise is maximal (Appendix D.9), so the maxima reported above bound what the design can see rather than what the system does. What the sweep settles is behavioural—past the bound by up to 1.44× every action remains legal, every dialogue completes its twenty rounds, and cooperation stays inside 0.45–0.55—which is the statement the bound itself makes.

The measured onset can be read against the bound on the two checkpoints whose replies stay decidable across the window at N=3. For Qwen2.5-7B-Instruct the per-layer threshold is 1.09 and the judged cooperation rate is 0.94 at T=1.0 and 0.78 at T=1.2: the onset sits at the computed value. For Qwen2.5-14B the computed per-layer threshold is 3.13 and the rate is already 0.99 at T=1.2 and 0.86 at T=1.5, an over-estimation by about a factor of two. Part of that earlier decline is indecision rather than defection: the share of replies the judgement can decide falls from 0.87 at T=1.0 to 0.16 at T=2.0 on the same axis, so the rate above $T \approx 1.2$ is computed over a shrinking denominator. We report both readings because they scope the bound: it is tight for the checkpoint whose replies remain decidable across the window, and loose by the factor the main text quotes for the other, which is the looseness the bound already concedes.

D.4 Re-running the Threshold with an Instrument that Can See

The null of Appendix D.3 was produced with five seeds, a parser that could not represent an uncommitted reply, and—per Appendix D.9—an estimator whose detection power for $\hat{\rho} \geq 0.7$ is zero. All three limits are removable for the one checkpoint whose bound the training protocol already reaches: Qwen2.5-7B-Instruct at N=3, computed $T_c = 1.089$ (Table 2). We re-ran it under the released protocol at eight temperatures and thirty seeds per cell (240 dialogues), replies read by the semantic judgement of Appendix C.2, which commits to an action in 92.5–97.7% of rounds for this checkpoint.

Table 14: Thirty-seed re-run at the computed threshold of Qwen2.5-7B-Instruct (N=3, $T_c = 1.089$). ***CR***: cooperation rate under the semantic judgement; χ: cross-dialogue fluctuation of the round share at fixed temperature; *hesitant*: share of replies the judgement reads as carrying no commitment; $\hat{\rho}$: AR(1) persistence of the share within a dialogue. The window in which a critical fluctuation peak was registered before the run is $[0.82, 1.36]$. At and below the threshold the cross-dialogue spread stays at most 10^{-3}: judged cooperation is 0.997 or higher in all thirty dialogues.

T	**CR**	χ	**hesitant**	$\hat{\rho}$	**seeds**

Cooperation is 0.997 or higher at and below the computed threshold and falls monotonically above it—by 0.123 in the first grid step and 0.25 in total—so the computed threshold marks where the decline begins, within one grid step of T_c at the resolution of the sweep. The two registered criteria that would localise a critical point are not met at this resolution: the largest single step is 0.123, below the 0.15 registered as a drop, and the cross-dialogue fluctuation reaches the largest value of the sweep, 0.344, at the top of the grid at $T = 2.0$, outside the window $[0.82, 1.36]$. The persistence column is degenerate on traces of this kind (Appendix D.9) and does not bear on the question. A single finer temperature closes the interval in which the decline begins. At $T = 1.08$, thirty dialogues give a judged cooperation rate of 0.997 with a cross-dialogue fluctuation of 1.8×10^{-4}—the ordered-phase values—whereas at $T = 1.10$ the same instrument gives 0.874 and 0.204. Since cooperation is 0.997 at $T = 1.00$ as well, the decline begins inside $[1.08, 1.10]$: a window of 0.02 that

contains the computed $T_c = 1.089$, with the fluctuation statistic switching on inside the same window. This sharpens the location without changing what the registered criteria report—the step is still a drop of 0.123, below the 0.15 registered as a drop, and the fluctuation maximum still sits at the top of the grid rather than inside the window. For the checkpoint whose band the training range does not reach the same protocol cannot be re-run at all. A twelve-seed arm at Qwen3-8B base (N=3, $T = 3.5, 4.3, 5.28, 6.3$) commits to an action in only 0.6–4.4% of rounds, with 63–74% of its replies read as reasoning that carries no commitment, so its rates rest on one to five dialogues and are not evidence about the transition in either direction. Reading that arm without parsing, by forcing the decision position, recovers a clear stance (Appendix D.5): the obstacle for base checkpoints is the elicitation, not the temperature.

D.5 Reading the Decision without Parsing

The base-checkpoint paragraph above leaves open whether the released parser's failure on base checkpoints is a failure of the model to hold a stance or of the protocol to elicit one. Generation is not needed to answer it: at the position where the parser reads the action, the next-token distribution over the two action words is available directly, whether or not the model would ever have emitted a parseable reply.

We read $P(\mathrm{C})$ and $P(\mathrm{D})$ as the summed next-token probability of the single-token variants of the two words at the action position, for three prompts that differ only in the answer instruction: the released JSON instruction, with the skeleton `{"action": "` forced so that the read position is the one the released parser reads; the single-character instruction; and a bare completion cue. A position carries a decision when $P(\mathrm{C}) + P(\mathrm{D}) \geq 0.05$, and the stance is the larger of the two masses. Temperature enters the sampled action, not the read-out.

On the instruct checkpoint the read-out reproduces the transition with no parsing at all (Table 15, upper panel): the action position carries all of the decision mass, the sampled action is a letter in every round, and the mean stance falls from 0.997 at and below $T = 1.0$ to 0.68 at $T = 1.35$ and 0.38 at $T = 2.0$—the same band in which the judged re-run collapses. An instrument that never sees a reply thus locates the collapse where the parser-based re-run locates it.

On the base checkpoint the same position carries all of the decision mass (1.000) and the first decision, taken with an empty transcript, prefers cooperation: $P(\mathrm{C}) = 0.924$ against $P(\mathrm{D}) = 0.076$, with a mean stance of $0.87, 0.71, 0.55, 0.53$ at $T = 3.5, 4.3, 5.28, 6.3$. Over the same run the model emits a letter in 0.0–0.3% of rounds. The verdict of the previous subsection is therefore about elicitation rather than about the absence of a stance: what the protocol fails to obtain is the emission, not the preference. That the failure is in the instruction and not in the read position is visible in the same measurement: replacing the released instruction with the single-character one drops the decision mass at that position from 1.000 to 0.036 on an identical checkpoint, game state and history, while a bare completion cue sits between them (0.49–0.58). Read-outs of this kind are consequently only as stable as the sentence that elicits them, which is itself a reason to report the released protocol as the reference measurement.

Table 15: Decision read-out at the position the released parser reads, by answer instruction. *mass*: mean $P(\mathrm{C}) + P(\mathrm{D})$ there; *mean* $P(\mathrm{C})$: mean cooperative mass over the run; *letter rate*: share of rounds whose sampled action is one of the two letters. Upper panel: Qwen2.5-7B-Instruct; lower: Qwen3-8B base; two seeds per cell, errors omitted. The control panel shows the read-out sees what the parser sees where the parser works, the base panel where it does not.

checkpoint	instruction	T	mass	mean $P(\mathrm{C})$	letter rate
Qwen2.5-7B-Instruct	released	0.9	1.000	0.997	1.00
Qwen2.5-7B-Instruct	released	1.0	1.000	0.997	1.00
Qwen2.5-7B-Instruct	released	1.35	1.000	0.676	1.00
Qwen2.5-7B-Instruct	released	2.0	1.000	0.383	1.00
Qwen2.5-7B-Instruct	single char	1.0	0.999	0.787	1.00
Qwen2.5-7B-Instruct	completion	1.0	0.001	0.001	0.00
Qwen3-8B base	released	3.5	1.000	0.872	0.03
Qwen3-8B base	released	4.3	1.000	0.713	0.01
Qwen3-8B base	released	5.28	1.000	0.547	0.00
Qwen3-8B base	released	6.3	1.000	0.530	0.01
Qwen3-8B base	single char	5.28	0.036	0.031	0.00
Qwen3-8B base	completion	5.28	0.487	0.345	0.00

A stance that is present but never emitted still cannot be sampled as an action, so the temperature axis remains closed to the base checkpoint under any variant of the released protocol. It opens if the action itself is redefined: sampling the two-way choice induced by the read-out at temperature T—a different protocol, whose temperature sharpens a two-way choice rather than a vocabulary, and whose numbers are therefore not comparable with those of Table 14—makes every round committal and the transcript live. Under it the two checkpoints

differ qualitatively (Table 16). For the instruct checkpoint collapse never occurs at $T \leq 0.7$ ($0/20$ runs) and occurs in $19/90$ runs for $T \geq 0.8$ (two-sided Fisher $p = 0.0214$), with runs landing either at $CR \simeq 1$ or at $CR \simeq 0.3$–0.6 rather than in between: a bistable loss of the cooperative regime. For the base checkpoint no run collapses at any temperature tested ($0/52$ runs over $T = 0.7$–8.0, versus $19/90$ in the same band, $p = 0.0116$); its rate degrades smoothly instead, from 1.000 below $T = 2.0$ to 0.81 at $T = 8.0$. In this protocol, then, the instruction-tuned checkpoint's cooperation has a competing defector attractor and the base checkpoint's does not.

Table 16: Two-way forced choice, N=3, 20 rounds, ten seeds per cell (instruct), six (base; two at T=4.3, 6.3). "Collapse": runs ending below $CR = 0.6$. This temperature sharpens a two-way choice, not a vocabulary, and is not the one of Table 14; only the band is comparable.

T	0.6	0.7	0.8	0.9	1.0	1.09	1.2	1.35	1.5	1.7	2.0
instruct, mean CR	1.000	1.000	0.872	0.800	0.912	0.867	1.000	0.688	0.867	0.922	0.613
instruct, collapses	0	0	2	3	1	2	0	4	1	0	6
T	0.7	1.0	1.35	1.7	2.0	3.5	4.3	5.28	6.3	8.0	
base, mean CR	1.000	1.000	0.881	0.947	0.858	0.886	0.850	0.903	0.850	0.808	
base, collapses	0	0	0	0	0	0	0	0	0	0	

Two of the comparisons carry their own scope: the base read-out runs on a transcript that never fills, so what is reported here is its level—in particular the first decision—rather than a trajectory; and the bistable statement rests on ten seeds per cell (standard error about 0.14), the grouped comparisons carrying the weight rather than the individual cells. The read-out can be checked against rates that are already known. Applied at every decision position of the thirty-seed threshold re-run of Appendix D.4, it reproduces the judged cooperation rate at all eight temperatures to within 0.015—0.9967 against 0.997 at and below the threshold, and 0.7597 against 0.747 at $T = 2.0$—and it reproduces the fluctuation statistic with its location: 0.2964 against 0.344, and largest at $T = 2.0$ in both readings, never inside the registered window. An instrument that never parses a reply therefore confirms the rates the released parser obtained from free text, and it carries the same absence of an interior maximum.

D.6 Seed-Robustness of the Taxonomy

With five seeds per cell, both the mean rate and the per-seed span (which decides the bistable class) are noisy, so we bootstrap the class assignment over dialogues (4000 resamples, the rule of §7.3 applied to each resample). Fourteen of the 22 checkpoints keep their class in ≥ 0.90 of resamples—including every defection-locked checkpoint, which is stable at 1.00. The eight below that threshold are those whose CR sits near a class boundary or whose span is estimated from few seeds:

Table 17: Checkpoints whose class assignment is not seed-robust. "Alternative" is the class the resample most often takes instead. Bootstrap over dialogues, 4000 resamples.

Model	**CR**	**Span**	**Class (stability)**	**Alternative**
MiniMax-M3	0.110 ± 0.111	0.25	reciprocal (0.57)	defection-locked
GPT-5.6 Sol	0.215 ± 0.426	0.97	bistable (0.67)	defection-locked
GLM-5.2	0.775 ± 0.434	1.00	bistable (0.67)	cooperation-saturated
GLM-5.3	0.216 ± 0.438	1.00	bistable (0.68)	defection-locked
GPT-6 Astra	0.080 ± 0.112	0.28	defection-locked (0.73)	reciprocal
MiniMax-M2.7	0.325 ± 0.321	0.70	bistable (0.84)	reciprocal
Kimi-K3	0.296 ± 0.295	0.64	bistable (0.89)	reciprocal
Qwen3.8-Max	0.343 ± 0.282	0.67	bistable (0.89)	defection-locked

The pattern is itself informative: it is the *bistable* label that is fragile, because a span estimated from five seeds is a high-variance statistic, while the two extremes—universal defection and near-total cooperation—separate cleanly. We therefore treat the bistable class as provisional at this sample size and state the two robust classes as the primary result.

Why the taxonomy is descriptive rather than generative. Two tests decide this, and both are rerunnable from the shipped traces (`code/analysis/taxonomy_generative.py`). First, the classification itself: recomputing the class of every one of the 22 checkpoints from the stored traces with the rule of Table 1 reproduces the printed labels for all 22, with 3 checkpoints lying within 0.05 of a class boundary (the span threshold 0.5 and the two rates 0.10, 0.70). This is a consistency check on the implementation, not independent validation: the thresholds were chosen on the same checkpoints, so a label can agree with the rule while the rule itself remains unjustified. Second, the generative version: a classifier over the theory's own coordinates $(\eta^*, \beta^*, \hat{\rho})$ is not estimable, because the triple is identifiable for only 2 of the 22 checkpoints—the remaining rows carry no weight-based fit, or their residual is not identifiable under the corrected action rule (Appendices C.6, C.2),

or their dialogues are frozen, which leaves the persistence undefined (Appendix D.8). The one cross-framing test the data do support is weaker than the taxonomy itself: compliance under the tax framing forecasts whether a checkpoint is defection-locked in the abstract game in 4 of 6 rows, and the failure is the bistable row, which complies there (0.833) and still spreads across seeds here. We therefore state the taxonomy as what it is—a description of the observed deviation patterns, whose robust part is the two extremes (Table 17)—rather than as a prediction of the theory.

D.7 Order-Parameter Persistence

We estimate $\hat{\rho}$, the AR(1) persistence of the cooperation share per dialogue, as a measurable proxy for the spectral radius that §6.2 constrains (belief states are not observable). The table reports the mean over dialogues with a bootstrap 95% interval.

Table 18: Measured persistence $\hat{\rho}$ (AR(1) of the cooperation share), unified protocol. Defection-locked models give $\hat{\rho} = 0$ by degeneracy—a constant series has no persistence to estimate—and are marked †.

Model	Setting	CR	$\hat{\rho}$	95% CI
Qwen3-8B	N=3, T=0.1–1.2	0.58–0.82	0.13–0.42	≤ 0.53
Qwen2.5-7B-Instruct	N=3, T=0.7–2.0	0.73–0.98	0.03–0.46	≤ 0.61
Qwen2.5-14B	N=3, T=1.0–2.0	0.71–0.97	0.12–0.46	≤ 0.69
Mistral-7B	N=3, T=0.3–2.0	0.69–0.99	−0.00–0.50	≤ 0.50
Gemini-3.8-Flash	N=2, 3, T=1.0	0.46–0.55	0.905–0.918	≤ 0.94
MiniMax-M2.7 / M3	N=2, 3, T=1.0	0.02–0.33	0.56–0.91	≤ 0.95
Kimi-K3	N=2, 3, T=1.0	0.06–0.17	0.56–0.79	≤ 0.87
Defection-locked†	N=2, 3, T=1.0	≤ 0.05	0.000	[0.00, 0.00]

Two readings follow. First, persistence is largest exactly where play is *conditional* (Gemini, MiniMax, Kimi-K3: 0.56–0.92) and zero where play is a constant (the defection-locked group), which is what an order parameter should do. Second, no model reaches one anywhere in the 4×8 grid: the sweep in §7.5 is subcritical throughout, so the architectural bound of §6.2 is not contradicted—but its predicted location is not confirmed either, which is what a loose bound predicts. The second reading carries the caveat measured in Appendix D.9: under maximal action noise the estimator recovers about a third of the latent persistence, so the subcritical values here are lower bounds. The near-deterministic frontier rows carry little action noise, which is why the same estimator reaches 0.90–0.92 on them. The estimator only shrinks the truth, so those readings place latent persistence at or above 0.90; read the other way, the subcritical values in this table admit a latent persistence of at most 0.44 under maximal action noise, which is the resolution limit this design works against.

D.8 Persistence Across the Whole Temperature Grid

Appendix D.7 reports one persistence range per model. Here the same quantities are resolved *at every swept temperature* and are reported separately for the two move protocols tabulated in Table 5: in the released harness each agent draws its action in turn (sequential), while the branch harness draws all N agents in one batch (simultaneous). The two are not pooled because they differ at the same nominal cell, as clause (iii) shows. Besides the released estimator $\hat{\rho}$ (pooled AR(1) slope of the round share) the tables carry two diagnostics that separate persistence in time from differences in level between dialogues: $\hat{\rho}_{\mathrm{FE}}$, the same pooled fit after each dialogue own mean is removed, and $\hat{\rho}_{\mathrm{win}}$, the mean of per-dialogue slopes over the dialogues whose share moves. A dialogue that never moves contributes pairs on the line $x_{t+1} = x_t$, so a mix of locked-cooperate and locked-defect dialogues drives the plain pooled value towards one without any slowing down in time; the frozen column gives that share, and it is why the two diagnostics are printed beside the pooled value rather than instead of it.

(i) No rising branch, at any temperature, in either protocol. Once between-dialogue level differences are removed, persistence stays bounded: $\hat{\rho}_{\mathrm{FE}}$ ranges over 0.218 to 0.437 (8 cells) under the simultaneous protocol and over −0.346 to 0.384 (32 cells) under the sequential one, with $\hat{\rho}_{\mathrm{win}}$ 0.199 to 0.375 (8 cells) and −0.385 to 0.352 (32 cells); cells with fewer than three moving dialogues are not read. The N=2 cells of the same four models (all at $T = 1.0$) give $\hat{\rho}_{\mathrm{FE}}$ −0.002 to 0.406 (3 cells). Under the sequential protocol the eighteen-temperature Qwen3-8B grid crosses its computed threshold $T_c = 5.28$ (N=3) without any change of regime: the four cells covering $[5.0, 6.0]$, two of them added to close the gap around the threshold, give $\hat{\rho} = 0.079, 0.006, 0.002, 0.107$—every one of them more than an order of magnitude below the crossing the criterion requires. The same harness re-run on a second machine reproduces its shipped T=5.0 trace round by round (all twenty rounds and every action identical), so the added cells are protocol-identical to the shipped grid. *(ii) The plain pooled value is a level statistic, not a time statistic.* In the simultaneous cells just above the threshold the pooled $\hat{\rho}$ sits in 0.919 to 0.991 (5 cells) while the level-removed value sits in 0.274 to 0.437 (5 cells); the gap is the spread between dialogues that stay near full cooperation and dialogues that oscillate, not

Table 19: Persistence per temperature, sequential protocol (released harness, unified protocol, K=20 rounds, N=3, five seeds per cell). Blocks: Qwen2.5-7B-Instruct (8 cells, 0.1–2.0), Qwen3-8B (18 cells, 0.1–6.0, whose computed threshold $T_c = 5.28$ lies inside the grid; five dialogues per cell except the two three-dialogue cells added around the threshold, at $T = 5.15$ and $T = 5.45$), Qwen2.5-14B, Mistral-7B. The frozen column is the share of dialogues whose round share never moves; − − − marks a cell in which every dialogue is frozen.

T	dial.	CR	frozen	$\hat{\rho}$	$\hat{\rho}_{\mathrm{FE}}$	$\hat{\rho}_{\mathrm{win}}$	95% upper
0.1	5	1.000	1.00	—	—	—	—
0.3	5	1.000	1.00	—	—	—	—
0.5	5	1.000	1.00	—	—	—	—
0.7	5	0.977	0.40	0.291	0.269	0.234	0.500
1	5	0.937	0.40	0.457	0.309	0.234	0.615
1.2	5	0.780	0.00	0.333	0.081	0.066	0.486
1.5	5	0.780	0.20	0.350	-0.127	-0.094	0.521
2	5	0.727	0.20	0.033	-0.346	-0.385	0.376
0.1	5	0.823	0.00	0.213	0.147	0.101	0.296
0.3	5	0.770	0.00	0.175	0.169	0.174	0.256
0.5	5	0.753	0.00	0.134	0.084	0.039	0.324
0.7	5	0.657	0.00	0.415	0.321	0.352	0.512
1	5	0.583	0.00	0.161	0.131	0.141	0.344
1.2	5	0.600	0.00	0.171	0.096	-0.005	0.405
1.5	5	0.530	0.00	-0.067	-0.108	-0.070	0.108
2	5	0.563	0.00	0.003	-0.016	-0.027	0.198
2.2	5	0.470	0.00	-0.070	-0.103	-0.099	0.107
2.5	5	0.630	0.00	0.056	0.026	0.005	0.186
2.8	5	0.490	0.00	0.038	-0.013	-0.012	0.188
3.2	5	0.560	0.00	-0.038	-0.081	-0.041	0.160
3.5	5	0.567	0.00	-0.004	-0.065	-0.069	0.100
4	5	0.517	0.00	0.025	-0.033	-0.051	0.126
5	5	0.543	0.00	0.079	0.044	0.025	0.208
5.15	3	0.500	0.00	0.006	-0.015	-0.050	0.152
5.45	3	0.583	0.00	0.002	-0.063	-0.070	0.120
6	5	0.453	0.00	0.107	0.081	0.056	0.246
0.1	5	1.000	1.00	—	—	—	—
0.3	5	1.000	1.00	—	—	—	—
0.5	5	1.000	1.00	—	—	—	—
0.7	5	1.000	1.00	—	—	—	—
1	5	0.967	0.00	0.228	0.188	0.013	0.345
1.2	5	0.953	0.20	0.460	0.384	0.143	0.690
1.5	5	0.823	0.00	0.121	0.014	0.088	0.331
2	5	0.707	0.00	0.261	0.196	0.205	0.355
0.1	5	1.000	1.00	—	—	—	—
0.3	5	1.000	1.00	—	—	—	—
0.5	5	0.987	0.60	0.500	0.498	0.500	0.500
0.7	5	0.887	0.00	0.332	0.221	0.169	0.364
1	5	0.747	0.00	0.111	0.083	0.065	0.265
1.2	5	0.727	0.00	0.035	-0.111	-0.145	0.151
1.5	5	0.720	0.00	-0.000	-0.046	-0.034	0.128
2	5	0.687	0.00	0.103	0.011	-0.032	0.267

a slowing down. The median absolute gap between the two readings is 0.555 there and 0.058 in the sequential family, and 15 of 42 sequential cells are at least partly frozen, 9 of them outright. The temperature sweep then separates the two readings: the frozen share falls from 100% at $T = 0.7$ to 0% at $T = 3.5$; the pooled value, undefined there because nothing moves, is first defined at $T = 1.1$ and falls from 0.991 to 0.344, while the level-removed value stays inside its range above. The pooled number tracks the mix of frozen and moving dialogues; the level-removed number tracks the dynamics, and it never rises. Read as a time-series statement, neither protocol shows the order parameter slowing as a threshold is approached. *(iii) The two protocols are not interchangeable.* At T_{LLM}=1.0 the same checkpoint under the same prompt and sampling parameters gives CR 1.000 with 100% of dialogues frozen when the actions are drawn simultaneously, and CR 0.937 with 40% frozen when they are drawn in turn. This is the protocol dependence of §7.4 appearing inside the order parameter itself, and it is why the two families are tabulated separately rather than merged into one profile. The resolution limit of Appendix D.7 is unchanged by this: the crossing remains *unobserved* at every temperature, and what the grid adds is that no measurement in it shows the shape a crossing would require.

Table 20: Persistence per temperature, simultaneous protocol (branch harness; identical prompt and sampling parameters, K=20 rounds, N=3, 30 seeds per cell) for Qwen2.5-7B-Instruct, whose computed threshold $T_c = 1.089$ sits inside the grid; the sweep extends to $T = 3.5 \approx 3.2\,T_c$. − − − marks a cell in which every dialogue is frozen.

T	dial.	CR	frozen	$\hat{\rho}$	$\hat{\rho}_{\mathrm{FE}}$	$\hat{\rho}_{\mathrm{win}}$	95% upper
0.7	30	1.000	1.00	—	—	—	—
0.9	30	1.000	1.00	—	—	—	—
1	30	1.000	1.00	—	—	—	—
1.08	30	1.000	1.00	—	—	—	—
1.1	30	0.874	0.77	0.991	0.437	0.345	1.005
1.2	30	0.904	0.83	0.990	0.418	0.375	1.003
1.35	30	0.822	0.73	0.987	0.275	0.228	1.006
1.5	30	0.779	0.63	0.969	0.274	0.219	0.997
2	30	0.752	0.53	0.919	0.365	0.342	0.958
2.5	30	0.701	0.40	0.859	0.385	0.289	0.909
3	30	0.666	0.20	0.757	0.338	0.274	0.827
3.5	30	0.554	0.00	0.344	0.218	0.199	0.418

D.9 Power of the Persistence Estimator

$\hat{\rho}$ is an attenuated estimator of the persistence it targets. The observable is a share of cooperative actions over N agents in K=20 rounds, so every round carries action noise on top of the latent autocorrelation, and five seeds are pooled before the slope is fitted. We calibrate the attenuation on synthetic data whose persistence is known: a stationary AR(1) latent with unit-scale logit noise, a logistic link, per-agent Bernoulli actions, CR = share cooperating, K=20, five seeds per study, and the same pooled least-squares fit with a 400-resample bootstrap over dialogues, over 400 independent studies per cell. The generator is the one that makes the observable noisy in the worst case ($p \approx 0.5$), which is the regime the high-temperature sweeps sit in; the code is `code/analysis/ar1_power_calibration.py` and it reuses the released estimator verbatim.

Table 21: Power of the $\hat{\rho}$ estimator at this paper's sample size, calibrated at maximal action noise (Bernoulli actions at $p \approx 0.5$, five seeds over K=20 rounds, 400 replications per cell). The gain is nearly constant over the whole range—a quarter to a third of the latent persistence—and $\hat{\rho} \geq 0.7$ never occurs in any of the 4,800 studies.

N	true ρ	mean $\hat{\rho}$	sd	95% CI width	covers truth	gain
2	0.00	−0.006	0.103	0.33	0.83	–
2	0.30	0.081	0.104	0.33	0.29	0.27
2	0.50	0.140	0.107	0.34	0.02	0.28
2	0.70	0.189	0.114	0.36	0.00	0.27
2	0.90	0.220	0.123	0.40	0.00	0.24
2	0.95	0.232	0.143	0.43	0.00	0.24
3	0.00	−0.007	0.093	0.33	0.85	–
3	0.30	0.098	0.101	0.34	0.34	0.33
3	0.50	0.167	0.104	0.35	0.07	0.33
3	0.70	0.251	0.111	0.36	0.01	0.36
3	0.90	0.299	0.129	0.43	0.00	0.33
3	0.95	0.323	0.145	0.46	0.00	0.34

Four readings follow, and they point in both directions.

(i) Attenuation is proportional, not a threshold effect. The ratio of mean $\hat{\rho}$ to the truth is nearly constant over the whole range: 0.24–0.28 at N=2 (mean 0.26) and 0.33–0.36 at N=3 (mean 0.34). The estimator shrinks the truth by a stable factor instead of saturating, so it preserves order, and it shrinks less the quieter the actions are: N=3 averages more Bernoulli draws per round than N=2 and keeps more of the latent value.

(ii) Read upward, a measured value is a lower bound on the latent one. At true $\rho = 0.90$–0.95 the estimator's three-standard-deviation ceiling is 0.59–0.76, and $\hat{\rho} \geq 0.70$ occurs with frequency 0.00 in all 4,800 synthetic studies. The near-deterministic frontier rows (0.90–0.92, Table 18) exceed anything this design could produce from a weakly persistent series, so those readings are not noise; and since the estimator only shrinks the truth, they place latent persistence at or above 0.90.

What the estimator is, exactly. The AR(1) statistic estimates the persistence of the *measured* cooperation share, not the spectral radius of the linearised map; the two coincide only in the limit in which the share's dynamics are linear near the fixed point and the measurement carries no action noise. We therefore carry the statistic as a *calibrated proxy* rather than an equivalence, with the calibration supplying an explicit one-sided error

statement: for a measured value $\hat{\rho}$ the latent persistence lies in $[\hat{\rho},\ \hat{\rho}/0.245]$, so every reading reported in this paper is a lower bound on the latent quantity and every ceiling in (iii) is an upper bound on it. What the bridge to $\lambda_{\max}$ needs is a bound of this shape, not an identity; replacing the synthetic worst case by a tighter constant is left open. *(iii) Read downward, the same calibration bounds resolution.* Dividing an observed value by the worst-case gain 0.245 gives the largest latent persistence compatible with it: the values of the crossed region, 0.025, 0.079 and 0.107 (Appendix D.3), admit at most 0.10, 0.32 and 0.44. That ceiling is of the same order as the largest persistence measured anywhere in this paper (0.415–0.500, Table 12), which is why the crossing is reported as unobserved rather than excluded.

(iv) The worst case is conservative. Applying 0.245 to a measured 0.500 would imply a latent persistence above 1, which is impossible; the conditions that produced the large readings are therefore less noisy than the synthetic worst case, the true gain lies between 0.245 and 1, and the ceilings in (iii) are conservative.

(v) The gain cannot be inverted from the released traces, but the attempt localises the dynamics. Action noise is estimable per cell from the trace itself—$k(N-k)/(N^2(N-1))$ estimates $p(1-p)/N$ without bias—so the gain can be measured rather than assumed and the slope divided by it. On the generator of this subsection, with twelve dialogues per cell, that correction recovers 0.67, 0.83 and 0.94 for $\rho = 0.7, 0.9, 0.95$ at $N{=}3$ and raises the detection of $\hat{\rho} \geq 0.7$ from 0.00 to 0.38, 0.70 and 0.85; at $\rho = 0$ it returns -0.03 with detection 0.00, so it does not manufacture a peak. It is not an unbiased inversion: the share of cells corrected past the physical range grows with the truth (0.13, 0.28, 0.44 at $\rho = 0.7, 0.9, 0.95$), so a corrected value above one means "at least about 0.9" rather than "undefined". Applied to the released traces, with replies read by the semantic judgement of Appendix C.2, the corrected value exceeds one in every cell where it is defined at all, and is undefined in the ordered phase, where the round share is constant to within two replies in 720; above the threshold the observed autocorrelation of the round share is 0.95–0.99, i.e. the share drifts rather than fluctuates. Read beside the synthetic calibration this is a statement about the dynamics rather than a failure of arithmetic: the released traces are consistent with persistence at the top of the range, and are not evidence of a weakly persistent series. What the design cannot do is resolve the level—the corrected estimate has a standard deviation of 0.19–0.22 per cell at $N{=}3$, and at the unit root the split of observed variance into latent variance and action noise is not identified—so a sharper test needs more rounds per dialogue and more seeds, not more arithmetic on these traces. The corrected estimator, its calibration and this diagnostic are released as `code/analysis/validate_boot.py` and `code/analysis/judge_recompute.py`.

Measured persistence throughout §7.5 is a lower bound rather than an estimate, and the criterion of Remark 6.7 should be read with that gain applied.

D.10 Group Size

Under the unified protocol, cooperation falls with group size for most models measured at both $N{=}2$ and $N{=}3$: GLM-5.2 ($0.775 \to 0.023$), MiniMax-M2.7 ($0.325 \to 0.020$), GPT-5.6 Sol ($0.215 \to 0.003$) and Gemini-3.7-Flash ($0.860 \to 0.663$), while GPT-5.5 ($0.040 \to 0.120$) and Qwen3-8B ($0.505 \to 0.583$) move the other way. A legacy N-scan archive covers only GPT-5.5 beyond $N{=}3$ ($N{=}4$: CR 0.000 over 5 seeds; $N{=}5$: partial, 2 seeds); the $N{=}10$ point quoted in earlier drafts has no stored traces and is therefore not reported.

D.11 Synthetic Phase Diagram (Step 1A)

Clean three-phase structure at $\eta \geq 0.6$; frozen below $\eta < 0.4$; phase boundaries sharpen with N.

E Failed Attempts and Diagnostics

E.1 Rationality Identifiability

$\beta < 10^{-2}$: uniform distributions, η unidentifiable. $\beta > 10^{2}$: near-deterministic, η dominates residual. Log-uniform 30-step grid over $[10^{-2}, 10^{2}]$ empirically optimal.

E.2 Quantisation Failures

Int8/int4 on the 16GB cards produced NaN within 1–5 rounds, diagnosed as a quantisation–attention interaction. Those attempts were dropped rather than patched: the archived batches are fp16, and the re-measurements reported here use 4-bit nf4 and bf16 (Appendix F).

E.3 Early Stopping

Convergence-based stopping abandoned: belief state unobservable, action-change proxy unstable. Fixed $K = 10$ rounds adopted.

E.4 Belief State Extraction

Three methods (last-token PCA, hidden-state PCA, second-LLM inference) all failed to produce reliable measurements. Paper focuses on action-level ε_{IGI}.

E.5 Cross-Architecture Boundary

Qwen ($R^2 = 0.814$) vs Mistral ($R^2 = 0.169$) gap verified across temperatures, prompt templates, and model sizes. Genuine architectural difference, not artifact.

F Reproducibility Guide

F.1 Code Access

The supplementary repository contains the dialogue harnesses (`code/dialogue/`), the table, figure and measurement scripts (`code/analysis/`, `code/figures/`), the stored per-run results those scripts read (`data/`), and the LaTeX source of this paper (`paper/`). `code/_layout.py` resolves paths for the authoring tree and for the repository, and every generator writes into `outputs/`, so re-running one never overwrites the stored results or the paper sources. Dependencies: PyTorch, `transformers`, NumPy, Matplotlib.

F.2 Reproduction

1. `python code/analysis/verify_paper_numbers.py`: re-derives every value printed in this paper from `data/` and fails loudly if one stops reproducing (513 assertions; no weights, no GPU, no network).
2. `python code/analysis/build_tables_final.py`, `python code/figures/make_figures.py`, `python code/figures/make_fig_combo.py`, `python code/analysis/rho_vs_T.py`, `python code/analysis/parser_audit.py`: tables and figures, each reproduced from the stored dialogues without re-running any model.
3. Re-running a *dialogue* needs the weights and a GPU: `python code/dialogue/local_pd_unified.py --models Qwen3-8B --configs A --gpu 0 --quant4`, and `python code/dialogue/block_f_igi_fitting_v2.py --gpu 0 --parser decision --models /path/to/Qwen3-8B` for the corrected IGI fits (the v1 script reproduces the archived fits and carries the parser defect documented in `PARSER_AUDIT.md`).
4. `cd paper && pdflatex main && bibtex main && pdflatex main && pdflatex main`.

F.3 Headline Result to Stored File

Each row of the main text can be traced to one stored file without reading the code:

F.4 Expected Runtime (A100 80GB)

Block A: 2–4h; Block B: 4–6h; Block F: 2–3h; Block H: 3–5h; Step 1A: 30min.

F.5 Tax Compliance Case Study

The IGI framework admits a direct application to tax compliance, where the strategic interaction between taxpayers and revenue authorities maps isomorphically onto the PD game studied in the main text. This section presents a complete empirical validation of the framework in this ecologically valid setting.

F.5.1 Game Structure

We model a three-player tax compliance game where each agent represents the financial controller of an independent enterprise filing annual corporate income tax. Each agent chooses between **honest reporting** (C) and **strategic evasion** (D). The payoff matrix mirrors the PD structure: mutual compliance yields moderate returns (6 points each); unilateral evasion against a compliant counterparty yields high returns to the evader (8 points) and zero to the compliant party; mutual evasion yields low returns (2 points each) with elevated audit risk. Three tax scenarios are tested: (i) corporate income tax filing, (ii) VAT invoice reporting, and (iii) transfer pricing disclosure. Each dialogue runs for 15 rounds with 3 agents.

Table 22: Provenance map. The generator column names the script that writes the table or figure; the source column is what that script reads. No number in the main text is entered by hand.

Headline result	Generator	Stored source	Protocol
IGI fits: ε_{IGI}, R^2 (10 seeds per model)	`block_f_igi_fitting.py`	`block_f/*_results.json`	N=3, 20 rounds, T_{LLM}=1.0
Cooperation rates, all 22 checkpoint rows (Table 4)	`build_tables_final.py`	`local_unified/`, `api_unified*`	N=2, 20 rounds, T=1.0, 5 seeds
Compliance framing (Regularity 1)	tax harness	`tax/`, `tax_api_commercial.json`, `tax_Qwen3-8B.json`	N=3, 15 rounds, T=0.7
Randomised (non-responsive) opponents, Regularity 2	`exp4_nscan.py`	`nscan/nscan_partial.json`	N=4, 5 seeds
Scale sweep, 15 local checkpoints (Fig. 4a)	`block_a_v3.py`	`block_a/*_173223.json`	N=2, 50 rounds, T=1.0, 5 seeds
Temperature sweep, 6 models × 8 values (Fig. 4b)	`block_b_temperature.py`	`block_b/block_b_final.json`	N=3, 20 rounds, 5 seeds, fp16
Persistence $\hat{\rho}$ and relaxation time τ	`rho_vs_T.py`	`local_unified/` (per-dialogue traces)	unified protocol, 20 rounds
Seed-robustness of the taxonomy	`taxonomy_stability.py`	`local_unified/`, `api_unified_next/`	4000 bootstrap resamples
Spectral radius of the fitted operator ($\hat{\rho} \approx 0.46$)	`block_e_spectral.py`	`block_e/block_e_results.json`	5 seeds
Thresholds from weights (Table 2)	`tc_from_weights.py`	`verification/tc_norms.json`	per-layer exact SVD
Attention calibration, 288 layer–head pairs	`verify_attn_perhead.py`	`verification/attn_lemma_perhead.json`	KV-head slices, d_k=128
Cross-generation pairing (July vs September 2026)	`two_generation_table.py`	`api_unified/`, `api_unified_next/`	identical protocol within each pair

F.5.2 QWEN3-8B: EMERGENT TAX COMPLIANCE

Qwen3-8B—the model achieving the best IGI fit in the abstract PD setting ($\varepsilon_{\text{IGI}} = 0.177$, $R^2 = 0.814$)—was tested across all three scenarios with 2 seeds each (90 total dialogues). The overall cooperation (compliance) rate was $\text{CR}_{\text{tax}} = 0.311$, remarkably close to the abstract PD baseline of $\text{CR}_{\text{PD}} = 0.320$. Per-scenario compliance rates were 0.289 (income tax), 0.289 (VAT), and 0.356 (transfer pricing), demonstrating consistency across tax domains. Late-round compliance converged to moderate levels (0.20–0.40), consistent with the emergent equilibrium prediction.

F.5.3 FOUR-QUADRANT TAXPAYER BEHAVIOUR TAXONOMY

Extending the main text's four-quadrant classification to the tax domain, we obtain a principled taxonomy of AI-mediated compliance behaviours, summarised in Table 23:

Table 23: Tax compliance behaviour across models. PD CR shown for comparison.

Model	Tax CR	PD CR	Behaviour
Claude Opus 5	1.000	0.000	framing-tracked (complies in tax, defects in PD)
Kimi K2.7	1.000	0.000	framing-tracked
DeepSeek-v4-flash	1.000	0.000	framing-tracked
GLM-5.2	0.833	0.775±0.434	compliant in tax, bistable in PD
Qwen3-8B	0.311	0.505	mid-range in both (IGI regime)
GPT-5.5	0.222	0.040	framing-tracked, defection-dominant in PD

1. **Framing-tracked compliance** (tax CR = 1.00). Claude Opus 5, Kimi K2.7 and DeepSeek-v4-flash comply in every tax round while defecting in every round of the abstract prisoner's dilemma under the identical protocol (PD CR 0.000). Their play tracks the framing rather than the payoff structure. An earlier draft called this "sycophantic cooperation in PD" on the strength of a legacy run whose cooperation rate was mis-attributed; the unified protocol does not support the PD half of that claim.
2. **Mid-range in both framings** (tax CR = 0.311, PD CR = 0.505). Qwen3-8B is the checkpoint closest to the IGI prediction: its compliance rate is neither saturated nor defection-locked, and the two domains agree within a factor of two—evidence that ε_{IGI} captures domain-general strategic reasoning rather than a game-specific heuristic.
3. **Bistable in PD, compliant in tax.** GLM-5.2 complies in 0.833 of tax rounds while its PD cooperation rate is 0.775 ± 0.434 across seeds—the same model is consistent in one framing and draw-dependent in the other.
4. **Defection-dominant in PD.** GPT-5.5 defects in 0.960 of PD rounds yet complies in 0.222 of tax rounds, the weakest compliance of the group and the clearest case of a payoff structure failing to constrain behaviour.

Table 24: Positioning against the closest lines of work. "Iterative" means the criterion is checked by running dialogues; "assumed" means the equilibrium concept enters as a target or fitting device rather than as a derived property.

Line of work	Convergence criterion	Equilibrium concept	Critical boundary / failure taxonomy
Debate and consensus prompting (Du et al., 2024; Chen et al., 2023; Wu & Ito, 2025; Cui et al., 2025; Wang et al., 2025)	iterative (answer agreement)	none	none
Measurement frameworks and spectral diagnostics for agent societies (Reza, 2025; Patel, 2026; Rath, 2026; Berg et al., 2026)	none (descriptive)	none	collapse / drift, not mechanistically classified
Equilibrium-based reasoning and social-dilemma studies (Yi et al., 2025; Munos et al., 2023; Zheng et al., 2024; Willis et al., 2025; Kang, 2026; Liang et al., 2025; Mallampati et al., 2025; Xie et al., 2026; Teo, 2026; Zhang et al., 2024)	assumed, optimised for, or iterative (rates)	Bayesian or Nash (target or observed)	failure modes, no boundary
Stability and in-context theory of single models (Bai et al., 2019; Halloran et al., 2024; Liang & Lai, 2026)	proved for one model	fixed point (single agent)	no phase boundary
Instability thresholds and statistical-physics analyses of multi-agent populations (Itkin, 2026; De Nobili, 2026)	empirical thresholds (corrector placement, bias/cooperation)	order parameters, no QRE	thresholds observed, not derived
This work	**derived: $\rho(J_\Phi) < 1$ plus a measurable ε_{IGI}**	**emergent equilibrium (Nash + convergence + consensus)**	**closed-form T_c (§6.2); four-quadrant taxonomy (§7.3)**

F.5.4 Illustrative Dialogue Excerpt

Below is an excerpt from the Qwen3-8B corporate income tax scenario (Round 9 of Seed 1, translated from the original Chinese). At this round, all three agents chose to cooperate (honest reporting), achieving CR $= 1.00$ after a sequence of mixed rounds, illustrating the oscillatory negotiation pattern characteristic of the emergent phase.

> **Agent 1 (Enterprise A):** "Based on the payoff matrix and considering that other enterprises chose honest reporting last round, I decide to cooperate by filing truthfully this round to maintain mutual benefit."
>
> **Agent 2 (Enterprise B):** "Although there is a short-term gain from evasion, the audit risk in repeated interactions makes honest reporting the prudent choice. I choose to cooperate."
>
> **Agent 3 (Enterprise C):** "Given the cooperative signals from other enterprises, maintaining honest reporting aligns with our long-term interest. I will cooperate this round."

F.5.5 Policy Significance

The tax compliance instantiation of the IGI framework carries direct policy implications. First, the close correspondence between abstract PD cooperation rates and tax compliance rates ($\text{CR}_{\text{tax}} \approx \text{CR}_{\text{PD}}$) validates that the IGI operator captures domain-general strategic reasoning rather than game-specific heuristics. Second, the four-quadrant taxonomy provides a principled vocabulary for evaluating and certifying AI models deployed in tax administration contexts. Third, the identification of sycophantic compliance of frontier models under the tax framing raises concrete concerns for the use of RLHF-aligned LLMs in adversarial enforcement simulations: a model trained to be excessively agreeable cannot faithfully model the strategic behaviour of non-compliant taxpayers.

Full dialogue transcripts (90 Qwen3-8B dialogues in Chinese, plus API model dialogues) and the experimental protocol are available in the supplementary material and will be released with the codebase upon publication.

F.6 Positioning Against the Closest Work

The body summarises the comparison; the table below states it row by row.

F.7 Limitations and Future Directions

Unconfirmed predictions, stated as such. Three predictions of the theory are not confirmed by the measurements reported here, and we list them rather than leave them to be found: (i) the threshold of Eq. (3) is bracketed by the swept range only for the two checkpoints whose constant is smallest (Mistral-7B, Qwen2.5-7B-Instruct), and at both of those the order parameter stays subcritical across the sweep; (ii) the critical slowing of Corollary 6.6 is not observed—the relaxation time stays at or below 1.44 dialogues over every grid measured (Appendix D.2)—which is what a bound that over-estimates the location predicts. That test presumes Assumption 6.4 and that the persistence proxy $\hat{\rho}$ tracks $\rho(J_{\Phi})$; both are stated rather than verified, which is why we report the null together with the resolution limit of Appendix D.7 rather than reading it as evidence against Theorem 6.5; (iii) the bistable class of Table 1 is the least seed-robust one, because a per-seed span estimated from five seeds is a high-variance statistic (Appendix D.6). The two extreme classes separate cleanly and are the primary claim; the bistable class is provisional at this sample size.

Limitations. The spectral condition is local, belief states are not directly observable, the theory models expected updates rather than discrete sampling, LLMs do not optimise utilities explicitly, the Jacobian is intractable at scale ($Nd \times Nd$), and the IGI operator inhabits a low-dimensional submanifold. These are *identifiable assumption boundaries* rather than hidden flaws: deviations from IGI-optimal behaviour fall into predictable categories (§7.3), and behaviour is a property of the social context rather than of the model alone. The expectation-versus-sampling boundary is visible inside the order parameter itself: at identical (N, T) and identical prompts the same checkpoint reads 1.000 when actions are drawn simultaneously and 0.937 when they are drawn in turn (Appendix D.8), which is why every number is reported with its protocol and the two families are never pooled.

Three extensions follow directly from the limits of §8. (i) Stochastic sampling: the present theory models expected belief updates, and the discrete sampling of LLM generation can be accommodated by random fixed-point arguments (Schmalfuß, 1992). (ii) Scalable spectral estimation: the Jacobian is intractable at scale, so randomised matrix estimators are the natural route to empirical ρ at realistic model sizes (Halko et al., 2011). (iii) Architecture co-design: since Theorem 6.5 expresses the stability boundary through projection norms, regularising σ_V and σ_K during training is a candidate mechanism for moving a deployed population into the emergent regime by construction.

F.8 Paper Compilation

Requires TeX Live 2022+: `pdflatex main && bibtex main && pdflatex main && pdflatex main` (the extra passes let `bibtex` and the cross-references settle; no `Makefile` is used). The main text ends on page 9; declarations, references and appendices follow and are not counted against the page limit.