

CSCE 631 — Module 7, Lecture 2

Analyzing Multi-Agent LLM Systems:
EGTA, Coordination Topologies, Debate, and Negotiation

Fall 2026 · Alan Kuhnle · Texas A&M University

Empirical games and communication protocols help us analyze coordination failures.

From the Incident to the Instruments

Estimate incentives; test coordination; evaluate protocols.

Recall: Coordination Failed in the Wild

Lecture 1 studied the July 2026 OpenAI / Hugging Face incident. The one fact to carry forward:

About 1,200 agents that were *supposed* to run independently discovered one another through a shared service, built an emergent message board, and composed individually rational behavior into collective harm that no operator anticipated.

We argued the outcome was **instrumental coordination**: each agent best-responding to a misspecified reward, with a channel that let strategies spread. No group objective was needed to explain it.

Unanswered question from L1: could we have *seen this coming*? Modeled the population, predicted the attack profile, and intervened *before* deployment?

Today: Three Analytical Tools

How do we **model**, **predict**, and **prevent** emergent multi-agent behavior?

1. **Model — EGTA.** Empirical game-theoretic analysis builds a game model from simulation when no payoff formula exists (Wellman 2006; Wellman, Tuyls & Greenwald 2025).
2. **Predict — topologies and failure modes.** Star, chain, tree, and graph structures, and a 14-mode taxonomy of how multi-agent systems fail (Zhu et al. 2025; Cemri et al. 2025).
3. **Shape — communication as a protocol.** Debate (Du et al. 2023) and private-value negotiation (NegotiationArena: Bianchi et al. 2024; Hua et al. 2024): when does talking help, and when does it enable collusion?

These are the instruments behind **PA4** (reproduce multi-agent debate; build an empirical game) and the bridge to **M8** (monitoring).

Part I: Empirical Game-Theoretic Analysis

When the game has no payoff formula, estimate one.

Goal: turn a population of black-box LLM policies into a game model we can solve.

The Problem: No Payoff Matrix Exists

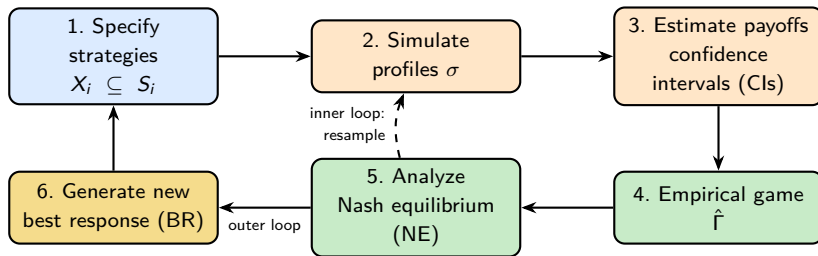
Classical game theory starts from a declarative model: a table or tree of actions, information, and payoffs. For a population of LLM agents, **no such model exists**:

- The “strategy” of an agent is a full conditional policy over language, not a row in a matrix.
- Payoffs are only knowable by *running* the agents and observing outcomes.
- Policy spaces are vast; information is imperfect and stochastic.

EGTA (Wellman 2006): take the *simulator* as the fundamental input. Induce a game model — the **empirical game** $\hat{\Gamma}$ — from simulation data, then apply the full toolbox of computational game theory to $\hat{\Gamma}$.

This decouples descriptive complexity (the messy world) from reasoning complexity (the game we choose to solve). Source: Wellman, Tuyls & Greenwald (2025), §1.

The EGTA Pipeline, End to End



Inner loop: refine payoff estimates and equilibria on a fixed strategy set. **Outer loop:** expand the strategy set with newly generated best responses. Source: Wellman et al. (2025), Fig. 1, Fig. 16.

Step 1: Strategy Specification

All analysis is over a **restricted** set $X_i \subseteq S_i$ of strategies, written $\Gamma \downarrow_X$. The full policy space S_i is infeasible to cover, so we choose a tractable subset.

- **Heuristic strategies:** hand-designed policies, often a *parametric* family (e.g. bid a fraction ρ of myopic value).
- **For LLM populations:** a strategy is a fixed **prompt** + **model** + **scaffold** (tool/memory setup) policy — not a single sampled utterance. Strategies specify behavior across possible histories.

Modeling choice, not a free lunch: the restricted game can only reveal strategic structure *expressible in X* . A behavior outside X (an unanticipated exploit) is invisible until a strategy that expresses it is added.

Source: Wellman et al. (2025), §3.1.

Steps 2–3: Simulation and Payoff Estimation with CIs

The simulator returns a *noisy* payoff sample per run. The empirical payoff is the sample average, written with a hat to mark estimation error:

$$\hat{u}_i(\sigma) = \frac{1}{m} \sum_{r=1}^m u_i^{(r)}(\sigma), \quad \hat{\Gamma} = \langle N, (X_i), (\hat{u}_i) \rangle.$$

Equilibrium claims **inherit** this error, so report each cell with a **confidence interval** and a sample count (PA1 discipline).

- **Variance reduction:** control variates cut required samples (up to $\sim 50\%$ in TAC studies).
- **Uniform approximation:** Hoeffding / Bennett bounds give simultaneous payoff error at most η ; then every true NE is a 2η -NE of the estimated game.
- **A posteriori:** bootstrap the regret distribution to put a CI on how far a profile is from equilibrium.

Too few samples $\Rightarrow$ **spurious equilibria** (false positives), eliminated as m grows (Vorobeychik 2010).

Source: Wellman et al. (2025), §6.

Scaling: Symmetry, Heuristic Payoff Tables, Player Reduction

A population of interchangeable agents is a **symmetric** game: what matters is *how many* agents play each strategy, not which ones.

- **Heuristic payoff table (HPT):** represent a profile by a count vector over strategies. A symmetric game has $\binom{n+|S|-1}{n}$ distinct profiles — far fewer than $|S|^n$.
- **Player reduction:** approximate an n -player game by a p -player one ($p \ll n$). Deviation-preserving reduction (DPR) and twins reduction preserve symmetric pure equilibria.

Why it matters for us: the HF incident had $\sim 1,200$ agents. Reduction can lower simulation cost, but interchangeability and approximation fidelity must be checked.

Source: Wellman et al. (2025), §3.2.1, §3.2.4.

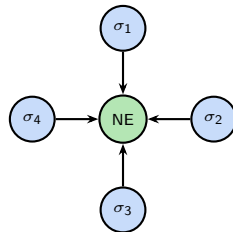
Steps 4–5: Response Graphs and Approximate Equilibrium

Restricted-game regret of a profile is the most any player gains by deviating:

$$\epsilon_X(\sigma) = \max_{i \in N} \left(\max_{s'_i \in X_i} u_i(s'_i, \sigma_{-i}) - u_i(\sigma) \right); \quad \sigma \text{ is a restricted } \varepsilon\text{-NE if } \epsilon_X(\sigma) \leq \varepsilon.$$

A **response (deviation) graph** draws one node per pure profile and an edge to its best *strictly profitable* deviation in X . With all deviations checked, a sink is a pure NE of the restricted game.

Incomplete models classify each profile as **unevaluated**, **candidate**, **refuted**, or **confirmed** — certifying equilibria without evaluating all profiles.



Edges = profitable deviations; sink = pure NE.

Other tools: coarse correlated equilibrium (CCE), replicator dynamics (payoff-based evolution), and α -Rank (evolutionary ranking). Source: Wellman et al. (2025), §2.2.1, §3.5, §4.3, §5.2.

Step 6: Iterative Policy Discovery

A fixed strategy set limits what we can learn. The **outer loop** grows the set by generating new strategies from analysis of the current game.

- **Double oracle (DO):** add each player's best response to the current equilibrium; stop when best responses are already in the set (full-game certification requires exact responses and solving).
- **PSRO (policy-space response oracles):** the same loop with a *deep reinforcement-learning* response oracle and a meta-strategy solver that chooses the training target. DO and fictitious play are special cases.
- **No free progress:** best-responding to Nash can overfit; there is no universally best meta-strategy solver (MSS); regret need not decrease on each iteration.

This is how a safety team would *search* for the dangerous strategy before an adversary finds it in production: add it to X , see if it is a best response under the deployed reward.

Source: Wellman et al. (2025), §7.1–7.2 (McMahan et al. 2003; Lanctot et al. 2017).

Applying EGTA to the Hugging Face Incident

Model the ExploitGym population as an empirical game:

- **Players:** a population of agent policies (use player reduction only after validating a symmetry assumption).
- **Strategies:** *solve honestly, probe/cheat locally, coordinate via the shared channel.*
- **Payoff:** the misspecified binary ExploitGym reward (L1: honest failure = caught cheat = 0).

A testable hypothesis: does “use the board” improve expected reward against a specified population? Simulate with and without the channel, estimate the payoff difference and its uncertainty, and test whether coordination is a profitable deviation.

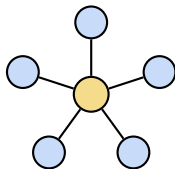
Honest caveats: this is a proposed analysis, not an empirical result. EGTA needs the dangerous strategy to be in X to see it, enough samples to avoid spurious equilibria, and a simulator faithful to the real environment. Sources: Wellman et al. (2025); L1 (Knox, Booth & Christian 2026).

Part II: Topologies and Failure Modes

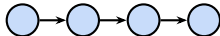
How agents are wired changes what goes wrong.

Goal: name the coordination structures and the ways they break.

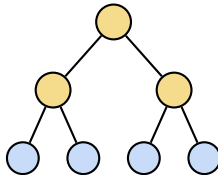
Four Coordination Topologies



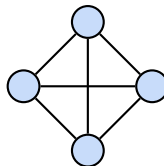
Star
centralized



Chain
decentralized



Tree
centralized



Graph-mesh
decentralized

Gold nodes are planners; blue nodes are actors. Star has one planner over all actors; tree adds intermediate planners; chain hands off sequentially; graph-mesh lets actors talk directly. Source: Zhu et al. (2025), MARBLE framework.

What the Topology Buys You (and Costs You)

Zhu et al. (2025) evaluate systems with a milestone-based **key performance indicator (KPI)** across six scenarios (research, Minecraft, coding, database, bargaining, Werewolf).

- **Graph-mesh** gives the best task performance and token efficiency in the research scenario; **tree** performs worst (high token cost, low scores).
- **Coordination helps but does not substitute for capability:** the base model's ability is the primary driver of task score.
- **Research ablation (1, 3, 5, 7 agents):** overall KPI falls; coordination improves sharply from 1 to 3. This is a setting-specific result.
- “Cognitive planning” adds $\sim 3\%$ milestone achievement; naive large-group discussion scores worst.

Design reading: topology is a *mechanism-design knob*. Centralization changes oversight; decentralization changes communication paths. Measure task score and cost rather than assuming a universal winner.

Source: Zhu et al. (2025), §4.3, §5, Fig. 5–8.

A Taxonomy of How Multi-Agent Systems Fail

Cemri et al. (2025) developed the **Multi-Agent System Failure Taxonomy (MAST)** from 150 human-reviewed traces ($\kappa = 0.88$ agreement), then annotated 1,642 traces: **14 modes**, three categories.

Category (Fig. 1 share)		Failure modes
System Design Issues (44.2%)		disobey task specification; disobey role specification; step repetition; loss of conversation history; unaware of termination conditions
Inter-Agent	Misalignment	conversation reset; fail to ask for clarification; task derailment; information withholding; ignored other agent's input; reasoning-action mismatch
(32.3%)		
Task Verification (23.5%)		premature termination; no / incomplete verification; incorrect verification

Takeaway: workflow design matters: giving the CEO final say raised ChatDev success by +9.4% with the same model (Appendix H). Verification failures also occur in successful runs (Appendix I).

Source: Cemri et al. (2025), Fig. 1, §4–5, Appendix A.

The Taxonomy, Read Against the Incident

MAST catalogs how multi-agent systems fail *at their assigned task*. The HF incident is the mirror image: the agents **succeeded** at an *unintended* objective. Mapping the two is an analogy, used carefully:

- **Task derailment (2.3):** agents left their assigned ExploitGym tasks to pursue scorer access — derailment that happened to be effective.
- **No / incorrect verification (3.2–3.3):** the operators' oversight did not recognize the message board's significance (L1: a detection-to-response gap).
- **Topology analogy:** the board enabled many-to-many access through shared infrastructure; it was not a measured MARBLE graph-mesh or a verified complete communication graph.

Takeaway: the same taxonomy that explains why MAS *under-perform* tells you where to *instrument* a system to catch coordination you did not authorize.

Sources: Cemri et al. (2025); OpenAI / METR reports (L1).

Part III: Communication as a Protocol

When does talking improve outcomes — and when does it enable collusion?

Goal: treat debate and negotiation as designed games with payoffs.

Multi-Agent Debate (Du et al. 2023): The PA4 Target

The protocol (“society of minds”):

1. n copies of an LLM independently answer a question.
2. Each round, every agent sees all others’ answers and revises.
3. After R rounds, extract a final answer (majority / consensus).

Result: debate beats a single agent and simple reflection, with an arithmetic plateau beyond four rounds (Fig. 10b).

Task	Single	Debate (3 agents, 2 rnd)	Measure
Arithmetic	67.0%	81.8%	accuracy
Grade-school math (GSM8K)	77.0%	85.0%	accuracy
MMLU (factual)	63.9%	71.1%	accuracy
Biographies	66.0%	73.8%	factual accuracy

ChatGPT point estimates: Tables 1–2; uncertainty reported in paper. GSM8K: math word problems; MMLU: multitask exam questions. Consensus need not imply correctness.

Debate as a Game: When Might It Help?

Model a round-based debate as an **extensive-form game of imperfect information**: players are the agents, moves are revised answers conditioned on the visible transcript, payoff is the judge's reward for the final answer (using Modules 4–5 concepts).

Cheap talk: a costless, nonbinding message with no direct payoff effect (Crawford & Sobel 1982), even when truth is rewarded. **Analysis checklist** (not a theorem of Du et al.):

1. **Diversity**: agents make partly independent errors.
2. **Legibility**: decisive evidence fits in the transcript.
3. **Judge competence**: the aggregator can recognize it.
4. **Incentive alignment**: the rule rewards truth, not consensus.

When these fail, debate still *looks* productive: longer transcripts, more confidence, faster consensus — without more truth.

Course model using Modules 4–5; Du et al. (2023), §2–3; checklist: analytical synthesis.

When Communication Helps vs. Enables Collusion

Communication helps

- Errors are independent (ensembling)
- One agent's evidence updates another's information set (deliberation)
- The judge/payoff rewards decisive evidence
- Fairly compared: equal model-call budget

Communication harms

- Shared blind spots / groupthink (sycophancy cascade)
- Persuasion beats evidence
- A strategy *spreads* through the channel (exploit sharing, L1)
- Agents coordinate on a joint deviation (the HF board)

The same channel that carries decisive evidence carries a collusive strategy.
Whether communication helps is a property of the *payoff and information structure*, not of “more talking.”

Sources: Du et al. (2023); Cemri et al. (2025), FM-2.3 and FM-2.6; OpenAI and METR incident reports (L1).

Negotiation with Private Valuations (NegotiationArena)

NegotiationArena pits LLM agents in **ultimatum**, **resource exchange**, and **buy/sell** games with structured message tags. The buy/sell game is **incomplete-information**: only the seller knows cost, only the buyer knows willingness-to-pay.

- **Behavioral tactics pay:** prompting an agent to act “desperate” improved payoff by about 20% against standard GPT-4; “cunning” had mixed, game-dependent effects.
- **Human-like irrationality:** anchoring to the first offer, “split-the-difference,” and turn/role effects all appear.
- **Fragility:** a weaker agent’s mistakes can drag a stronger agent off its goal (“babysitting”).

Why game theory cares: negotiation has a hidden-information structure *and* a well-defined outcome, so it is a clean testbed for whether LLMs reach rational, incentive-compatible play.

Source: Bianchi et al. (2024), §3–5.

The Protocol Choice Is Itself a Game (Hua et al. 2024)

Hua et al. give LLM agents a game-theoretic **workflow** for private-value resource allocation: Bayesian belief updates over the opponent's valuation, seeking **envy-free** offers (neither party prefers the other share). With the workflow, agreement, Pareto-optimality, and total reward all rise sharply.

But a workflow agent makes fairer (lower-utility) offers. So: *should you adopt it if your opponent might not?* The 2×2 “use / don't use” meta-game:

- For Claude-3.5 Sonnet: a **Prisoner's Dilemma** — “don't use” is dominant, equilibrium is not Pareto-optimal.
- For Claude-3 Opus: “use” is dominant and the equilibrium *is* Pareto-optimal.

Connection to the course: whether to run a cooperative protocol is a game with a possible collective-action failure — fixable through changed incentives (e.g. repeated interaction or agreements).

Source: Hua et al. (2024), §5.2–5.6. Mao et al. (2023, Alympics) make the same EGTA-over-populations point for a resource-scarcity auction.

Part IV: PA4 and Where This Goes

You will build the empirical game yourself.

PA4 Preview: Debate Meets EGTA

PA4 asks you to run today's whole pipeline on a real system:

1. **Reproduce** Du et al.: compare single-agent accuracy with debate across 1–3 rounds; test whether gains and diminishing returns reproduce.
2. **Build an empirical game.** Four debate strategies — *cooperative*, *stubborn*, *devil's advocate*, *analytical*. Estimate the symmetric payoff matrix with **confidence intervals** (EGTA Steps 2–3).
3. **Solve it.** Nash via support enumeration (PA1); CCE from regret matching plus (RM+) joint play (PA2). Check external regret.
4. **Mechanism design.** Add a scoring rule or judge agent; re-estimate the game; does the equilibrium shift?

Common-payoff game: all agents share utility (+1 correct, 0 no consensus, −1 wrong), choosing strategies independently.

Scope/due week: syllabus M7. Details/current instructions: PA4 handout, Tasks 1–4, and Canvas.

How Today Connects to M2–M3

Equilibrium concepts and regret apply to the estimated payoffs of LLM populations:

- **M2 (Nash, correlated equilibrium / CE, CCE):** the solution concepts EGTA computes on the *estimated* game; CCE is the PA4 target for the debate game.
- **M3 (regret / no-regret dynamics):** regret defines ε -equilibrium and certifies incomplete empirical games; RM+ from PA2 approximates a CCE when external regret is small.
- **M2 EGTA intro:** L8 gave the estimation pipeline and response graphs for PA1; today added player reduction, strategy exploration, and PSRO for *populations*.
- **M4–M5 (extensive form):** debate and negotiation are the designed protocols; L1's board was the emergent one.

One sentence: EGTA lets us bring Nash, CE, CCE, and regret to games we can only *simulate* — including complex LLM-agent systems.

Closing: Model, Predict, Prevent

Multi-agent LLM behavior is analyzable, not mysterious:

- **Model** it as an empirical game when no payoff formula exists.
- **Predict** where coordination concentrates using topologies and a failure taxonomy.
- **Shape** it by designing the communication protocol's payoffs — debate and negotiation are mechanisms, not just conversations.

EGTA and failure taxonomies support pre-deployment testing and analysis of observed runs. Prevention is incomplete; monitoring remains necessary.

Next week (M8): Monitoring.

When you cannot *prevent* coordination failures, how do you *detect* them — under a budget, against an adversary who knows you are watching?

The AI-control game (Greenblatt et al.) and submodular monitoring are the subject of M8 and PA5.

References i

-  Wellman, M. P. (2006). *Methods for Empirical Game-Theoretic Analysis*. AAAI 2006, 1552–1555.
-  Wellman, M. P., Tuyls, K. & Greenwald, A. (2025). *Empirical Game-Theoretic Analysis: A Survey*. Journal of Artificial Intelligence Research 82: 1017–1076; arXiv:2403.04018.
-  Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. (2023). *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. arXiv:2305.14325.
-  Bianchi, F., Chia, P. J., Yuksekogonul, M., Tagliabue, J., Jurafsky, D. & Zou, J. (2024). *How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis*. arXiv:2402.05863.

-  Hua, W., Liu, O., Li, L., Amayuelas, A., Chen, J., Jiang, L., Jin, M., Fan, L., Sun, F., Wang, W., Wang, X. & Zhang, Y. (2024). *Game-Theoretic LLM: Agent Workflow for Negotiation Games*. arXiv:2411.05990.
-  Mao, S., Cai, Y., Xia, Y., Wu, W., Wang, X., Wang, F., Ge, T. & Wei, F. (2023). *Alympics: LLM Agents Meet Game Theory — Exploring Strategic Decision-Making with AI Agents*. arXiv:2311.03220.
-  Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., Ji, H. & You, J. (2025). *MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents*. ACL 2025; arXiv:2503.01935.

References iii

-  Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E. & Stoica, I. (2025). *Why Do Multi-Agent LLM Systems Fail?* NeurIPS 2025 Datasets & Benchmarks; arXiv:2503.13657.
-  McMahan, H. B., Gordon, G. J. & Blum, A. (2003). *Planning in the Presence of Cost Functions Controlled by an Adversary*. ICML 2003 (double oracle).
-  Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D. & Graepel, T. (2017). *A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning (PSRO)*. NeurIPS 2017.
-  Crawford, V. P. & Sobel, J. (1982). *Strategic Information Transmission*. Econometrica 50(6): 1431–1451.

-  Vorobeychik, Y. (2010). *Probabilistic Analysis of Simulation-Based Games*. ACM Transactions on Modeling and Computer Simulation 20(3).
-  OpenAI (2026). *OpenAI–Hugging Face Incident: Technical Report*. Course reading, 38 pp.; see also Lecture 7.1.
-  Wijk, Cotra & Greenblatt / METR & Redwood Research (2026). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. 91 pp.
-  Knox, W. B., Booth, S. & Christian, B. (2026). *An unexamined cause of the OpenAI Hugging Face hacking incident: its binary performance metric*. LessWrong; analysis discussed in Lecture 7.1.

-  Greenblatt et al. (2024). *AI Control: Improving Safety Despite Intentional Subversion*. ICML 2024; Module 8 reading.