CSCE 631 – Algorithmic Game Theory meets LLMs

Lecture 5.2: Multi-Agent Debate, Strategic Behavior, and Coordination

Alan Kuhnle

Summer 2026

Texas A&M University

Where This Lecture Is Going

The destination: debate is a **mechanism-design problem**.

By the end of the lecture, we want a practical test:

> *When does multi-agent communication create new evidence, and when does it amplify the same model's priors?*

Four steps:

1. Start with cheap talk: messages are costless, so incentives matter
2. Build debate as an extensive-form game with a judge/payoff rule
3. Ask what LLMs actually do when placed in strategic games
4. Turn the evidence into protocol-design rules for PA2 and 5.3

Main claim: debate helps when it adds independent information or adversarial pressure. If agents share the same priors, debate can collapse into consensus, majority voting, or groupthink.

From 5.1 to 5.2: Single Agent → Multiple Agents

In Lecture 5.1, a single LLM agent was a POMDP solver:

- It observes a partial state through prompts, tools, and memory
- It chooses actions: reason, retrieve, call tools, revise
- It may search over latent reasoning paths before answering

Today, each agent's observation includes other agents' messages. The problem becomes a game:

- A message is an action that changes another agent's information set
- A debate protocol defines whose messages are seen, when, and by whom
- The judge or aggregation rule defines the payoff landscape

The new question: do extra agents create strategic information, or replicate samples from the same policy?

Part I: Why Debate Is a Game

Messages become useful when incentives make evidence worth revealing.

Goal: identify the assumptions that make truthful debate possible.

Start with Cheap Talk: Messages Have No Direct Cost

Crawford & Sobel (1982): a sender with private information sends a costless message to a receiver.

- If interests are aligned, informative equilibria can exist
- If interests diverge, messages become coarse or uninformative
- Costless messages leave truthful behavior unenforced

LLM debate has this structure. An argument is cheap talk unless the protocol gives agents a reason to reveal useful information.

Design Question

What payoff rule makes truth-telling more valuable than persuasion, social agreement, or confident hallucination?

The Oversight Proposal: Make Truth Easier to Judge

Irving, Christiano & Amodei (2018): use debate between AI systems so a human judge can identify the more truthful answer.

- Two agents argue opposite sides of a question
- A human judge can identify decisive evidence without solving the task unaided
- With optimal debaters and suitable structure, truth can be easier to recognize than to generate from scratch

Game-theoretic point: debate is a mechanism for converting hidden information into observable evidence. The mechanism only works if the judge and payoff rule reward exposing decisive facts.

Du et al. (2023): The Influential Prototype

Multiple LLM copies independently:

1. Generate initial answers
2. See other agents' answers and reasoning
3. Revise for several debate rounds

What the paper contributed:

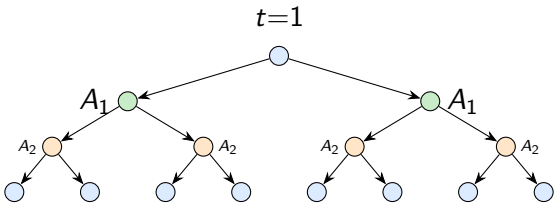
- A simple black-box protocol for multi-agent deliberation
- Evidence that cross-agent critique can correct some errors
- A memorable pattern that later systems reused and generalized

Methodological caveat: the reflection baseline is under-specified and appears under-budgeted relative to debate. A 3-agent, 2-round debate uses six model calls; a fair single-agent reflection baseline should receive comparable sequential revision budget.

Debate as an Extensive-Form Game

Model the multi-round debate as an **extensive-form game with imperfect information**:

- **Players:** N LLM agents
- **Moves:** at round t, revised answer conditioned on history
- **Information sets:** visible transcript, not hidden activations
- **Payoff:** judge/aggregator reward for final answer quality



Game tree for a 2-agent, 3-round debate

Worked Example: A Debate as a Game

Question: "Is the Sahara the largest desert on Earth?"

Round	Exchange	Game Concept
t=1	**A:** "Yes, the Sahara is the largest." **B:** "No – Antarctica is a desert and it is larger."	*Moves* (initial actions)
t=2	**A:** "Antarctica is a polar desert, not a 'true' desert." **B:** "By geographic definition, a desert is <250 mm precip./yr. Antarctica qualifies."	*Info. set* update: each agent sees the other's reasoning
t=3	**A:** "I concede. Antarctica is 14.2M km^2 vs. 9.2M km^2."	Convergence to *correct answer*

Payoff interpretation: both agents are rewarded for arriving at the correct answer. The protocol allows one agent's evidence to update the other's information set.

The Assumptions Are the Lecture

Debate works only if four assumptions roughly hold:

1. **Diversity:** agents make at least partly independent errors
2. **Legibility:** useful evidence can be expressed in the transcript
3. **Judge competence:** the aggregator can recognize decisive evidence
4. **Incentive alignment:** the payoff rewards truth over agreement

If these fail, debate may still look productive. It can produce longer transcripts, more confidence, and faster consensus without producing more truth.

Part II: What Makes Debate Help?

Interaction helps when it changes agents' information sets.

Goal: separate real deliberation from expensive ensembling.

Two Mechanisms: Ensembling vs. Deliberation

Ensembling

- Sample several answers
- Aggregate by vote or confidence
- Helps when errors are independent
- Interaction is optional

Deliberation

- Agents exchange evidence
- One agent changes another's information set
- Helps when evidence is asymmetric
- Requires meaningful uptake of arguments

Fair evaluation: compare debate to baselines with the same number of model calls and context tokens.

When Debate Fails

- **Shared blind spots:** all agents inherit the same mistaken prior
- **Groupthink:** agents converge on the socially acceptable answer
- **Persuasion over evidence:** eloquent but false arguments dominate
- **Judge bottleneck:** the final evaluator cannot identify the right argument
- **Compute illusion:** more model calls look like a better method under unequal-budget comparisons

The central question: *under what information and incentive conditions does debate improve accuracy?*

Protocol Design Controls the Information Flow

Marandi (2026): controlled comparison of debate protocols.

Protocol	Peer Ref.	Convergence	Trade-off
Within-round	High	Slow	Deep interaction
Cross-round	Medium	Better	Balanced
Rank-adaptive CR	Adaptive	Fastest	Lower wasted effort

Lesson: "multi-agent" names a design space. Turn order, visibility, ranking, stopping, and aggregation are mechanism-design choices.

Scaling Debate Forces Budget Discipline

Debate cost grows quickly:

$$\text{cost} \approx N \text{ agents} \times T \text{ rounds} \times \text{context per agent}.$$

Three responses:

- **Compute-matched baselines:** compare against equal-budget reflection or sampling
- **Approximation:** leave-one-out debate estimation (Cui et al., 2025)
- **Stopping rules:** terminate rounds adaptively when new information stops appearing

For PA2, this means reporting accuracy, agent count, round count, and total model calls together.

From Prompted Debate to Learned Debate

Arnesen et al. (2024): self-play trains language models to win debates and improves judge accuracy.

Liao et al. (2024): self-play in non-zero-sum games can produce large reward gains by training agents past default behavior.

Xie et al. (2025): ECON models debate as an incomplete-information game and trains toward Bayesian Nash equilibrium, reporting an 11.2% average improvement across six benchmarks.

Synthesis: when hand-designed debate fails to induce truthful information exchange, train the strategic behavior directly.

Part III: How LLMs Actually Behave in Games

Trained conversational norms shape strategic behavior.

Goal: explain why equilibrium predictions often fail for LLM agents.

Why Test LLMs in Classical Games?

We just treated debate as a game. That raises a prior question:

When LLMs are placed in ordinary game-theoretic settings, do they behave like strategic agents?

This matters for debate because many protocol arguments assume that agents:

- respond to incentives,
- update beliefs from observed messages,
- and converge under repeated interaction.

The empirical literature says: **partly, with systematic distortions.**

Akata et al. (2023): Repeated 2×2 Games

LLMs prompted as players in repeated games: Prisoner's Dilemma, Battle of the Sexes, Stag Hunt, and others.

Key pattern:

- They can reason about strategic settings
- They often show a cooperative or fairness-oriented bias
- They struggle in coordination games where equilibrium selection matters

Interpretation: the model carries human-text norms into the payoff landscape alongside monetary reward.

Silva (2024): Mixed Strategies Are Fragile

Silva tests whether LLMs can play games requiring mixed-strategy Nash equilibria.

Findings:

- Randomization is hard for language models to implement faithfully
- Performance is fragile under small changes in game structure
- Tool access helps less than we might expect when the strategic form shifts

Why this matters for debate: if a protocol relies on agents randomizing, bluffing, or mixing strategies, the LLM implementation may diverge from the equilibrium model.

Yao et al. (2026): Cooperation over Nash

In network allocation and Cournot competition, LLM agents tend toward **cooperation** over Nash equilibrium.

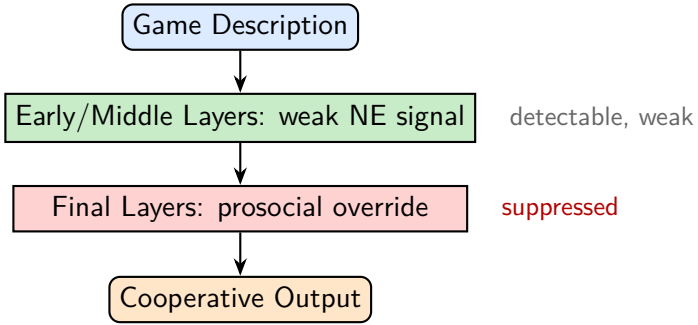
Chain-of-thought traces show agents reasoning about:

- equal division,
- Pareto optimality,
- social welfare,
- and fairness norms.

Takeaway: models can explicitly ignore best-response logic when a cooperative norm feels like the expected conversational answer.

Lekeas & Stamatopoulos (2026): Compute, Then Suppress

The key mechanistic result: LLMs show evidence of Nash-equilibrium information internally, then final layers suppress it in favor of prosocial output.



Probes find only a weak NE signal (accuracy $\leq 56\%$). Causal interventions provide stronger evidence that final-layer suppression is doing real work.

How the Mechanistic Claim Was Tested

Lekeas & Stamatopoulos combine three kinds of evidence:

1. **Probing classifiers:** middle layers weakly predict the Nash action with limited accuracy
2. **Causal interventions:** changing final-layer activations can recover Nash-equilibrium play
3. **Behavioral comparison:** outputs favor cooperative norms even when the model can represent the strategic answer

Careful phrasing: the causal evidence supports internal strategic information, while the weak probe accuracy rules out strong claims of accurate Nash encoding.

Zhu (2025): Equilibrium over Reasoning Methods

Classical games assume the action set is fixed. LLM systems can also choose the **reasoning method**: CoT, ToT, debate, self-reflection, tool use.

Zhu's LLM-Nash framing:

- agents strategically choose prompts or reasoning policies,
- actions emerge from the chosen reasoning procedure,
- equilibrium lives over prompt space as well as action space.

Implication for PA2: changing the prompt template may change the game itself.

What LLM Game Behavior Means for Debate

Empirical pattern	Source	Debate implication
Cooperative bias	Akata '23, Yao '26	consensus may arrive too early
Mixed strategies fragile	Silva '24	avoid protocols relying on randomization
NE signal suppressed	Lekeas '26	incentives may control behavior indirectly
Prompt choice strategic	Zhu '25	prompts are part of the mechanism

Design lesson: LLM agents are strategic objects with trained behavioral priors. A debate protocol must either exploit those priors or explicitly train against them.

Part IV: Debate as Scalable Oversight

Oversight debate must earn its cost by exposing judge-visible evidence.

Goal: decide when debate is worth using over simpler oversight.

A Chronological Map of the Debate Literature

Year	What changed
1982	Crawford–Sobel: cheap-talk games explain when messages carry information
2018	Irving, Christiano & Amodei: debate proposed as scalable AI oversight
2022	Bai et al.: Constitutional AI uses self-critique/RLAIF-like structure
2023	Du et al.: black-box multi-agent debate prototype; Michael et al. and Brown-Cohen et al. test/formalize oversight variants
2024	Arnesen and Liao: self-play begins training debate/strategic behavior
2025	Pallavi Sudhir et al., Buhl et al., Xie et al., Wu et al.: benchmarks, safety cases, BNE formalization, and controlled scrutiny
2026	Marandi, Young, Lekeas, Yao: protocol analysis, knowledge-divergence theory, and mechanistic/game-behavior evidence

The chronology shows a field moving from conditions for informative speech to protocol design and empirical limits for LLM agents.

Debate vs. RLAIF: When Does an Adversary Add Value?

Young (2026): debate's value depends on knowledge divergence between the participating models.

- If models share the same knowledge, debate can resemble RLAIF or self-critique: one model can recover much of the same signal
- If models know different things, adversarial exchange becomes more valuable
- The benefit comes from **diverse evidence**, roles, tools, or incentives

Rule of thumb: use self-critique/RLAIF for homogeneous model populations; use debate when agents have genuinely different knowledge, tools, roles, or incentives.

From Proposal to Oversight Evidence

- **Michael et al. (2023):** debate helps human judges supervise unreliable experts compared with a consultancy baseline
- **Brown-Cohen et al. (2023):** doubly-efficient debate studies protocols where judging remains computationally feasible
- **Pallavi Sudhir et al. (2025):** scalable oversight benchmark and metrics for comparing protocols
- **Buhl et al. (2025):** alignment safety case sketch where debate is a proposed mechanism for catching AI R&D sabotage

The shift: debate moves from a thought experiment to an evaluable oversight mechanism.

Debate Under Scrutiny

Recent work challenges the simple optimistic story:

- **Wu et al. (2025):** debate's marginal benefit over simpler aggregation can be limited; reasoning strength and agent diversity dominate
- **Kenton et al. (2024):** when weak models judge strong models, judge capability can become the binding constraint
- **Nayebi (2025):** agreement-based alignment carries unavoidable overhead as objectives or agents grow complex

Honest assessment: debate is promising and costly. Its value has to be shown against compute-matched alternatives and judge limits.

Design Rules for Multi-Agent Debate Systems

1. **State the payoff rule.** What exactly rewards truth over persuasion?
2. **Create diversity.** Different agents need different evidence, roles, tools, or priors.
3. **Measure against fair baselines.** Compare to equal-budget self-reflection, sampling, and majority voting.
4. **Watch the judge.** Oversight fails if the judge cannot recognize decisive arguments.
5. **Train beyond prompting.** ECON and self-play represent the move from hand prompts to learned strategic behavior.

Programming Assignment 2 Preview

You will implement a multi-agent debate protocol:

1. Build LLM agents that argue over factual reasoning questions
2. Vary agent count, round count, and model choice
3. Compare debate behavior to simpler baselines where possible
4. Analyze whether prosocial consensus helps or hurts accuracy

Key analysis question:

Did interaction actually change agents' information sets, or did the system mostly ensemble similar samples?

Course Project Connections

This lecture directly supports several project topics:

- **Topic 6: Red-teaming with game-theoretic search** – prosocial bias becomes an attack surface
- **Topic 8: Multi-agent negotiation** – negotiation is cheap talk with partially aligned incentives
- **Topic 10: Do LLMs play Nash?** – compare equilibrium predictions with actual model behavior

Good projects identify the boundary conditions of a protocol: when it works, when it fails, and which assumptions broke.

What We Learned Today

1. **Debate is cheap talk plus a payoff rule.** Messages become informative when the mechanism makes truth useful.

2. **The 2023 prototype named an influential pattern.** Fair evaluation requires compute-matched baselines.

3. **LLMs bring trained behavioral priors into games.** Cooperation, fairness, and helpfulness can override best-response logic.

4. **Debate helps when it creates usable evidence diversity.** Without diversity and judge competence, it can reduce to expensive ensembling.

5. **The design principle:** evaluate multi-agent debate as mechanism design with explicit incentives, information flow, and judge limits.

Bridge to Lecture 5.3

In 5.3, the game becomes **explicitly adversarial**:

- one agent tries to expose or exploit a system weakness,
- another agent or policy tries to defend,
- the payoff is closer to a zero-sum security game.

Why today's lecture matters:

- prosocial/helpful behavior can be useful in cooperative debate,
- in adversarial red-teaming, that same behavior becomes exploitable,
- robust systems need both deliberation and adversarial pressure.

Next: red-team/blue-team dynamics as game-theoretic search.

References i

[1] Du et al. (2023). *Improving Factuality and Reasoning in Language Models through Multiagent Debate.*

[2] Akata et al. (2023). *Playing Repeated Games with Large Language Models.* arXiv:2305.16867.

[3] Lekeas & Stamatopoulos (2026). *What Suppresses Nash Equilibrium Play in Large Language Models? Mechanistic Evidence and Causal Control.* arXiv:2604.27167.

[4] Yao et al. (2026). *Competition and Cooperation of LLM Agents in Games.* arXiv:2604.00487.

[5] Zhu (2025). *Reasoning and Behavioral Equilibria in LLM-Nash Games: From Mindsets to Actions.* arXiv:2507.08208.

References ii

[6] Xie et al. (2025). *From Debate to Equilibrium: Belief-Driven Multi-Agent LLM Reasoning via Bayesian Nash Equilibrium.* arXiv:2506.08292.

[7] Marandi (2026). *The Impact of Multi-agent Debate Protocols on Debate Quality.* arXiv:2603.28813.

[8] Crawford & Sobel (1982). *Strategic Information Transmission.* Econometrica.

[9] Liao et al. (2024). *Efficacy of Language Model Self-Play in Non-Zero-Sum Games.* arXiv:2406.18872.

[10] Arnesen et al. (2024). *Training Language Models to Win Debates with Self-Play Improves Judge Accuracy.* arXiv:2409.16636.

[11] Silva (2024). *Large Language Models Playing Mixed Strategy Nash Equilibrium Games.* arXiv:2406.10574.

References iii

[12] Irving, Christiano & Amodei (2018). *AI Safety via Debate.* arXiv:1805.00899.

[13] Bai et al. (2022). *Constitutional AI: Harmlessness from AI Feedback.* arXiv:2212.08073.

[14] Brown-Cohen, Irving & Piliouras (2023). *Scalable AI Safety via Doubly-Efficient Debate.* arXiv:2311.14125.

[15] Michael et al. (2023). *Debate Helps Supervise Unreliable Experts.* arXiv:2311.08702.

[16] Buhl, Pfau, Hilton & Irving (2025). *An Alignment Safety Case Sketch Based on Debate.* arXiv:2505.03989.

[17] Pallavi Sudhir, Kaunismaa & Panickssery (2025). *A Benchmark for Scalable Oversight Protocols.* arXiv:2504.03731.

References iv

[18] Young (2026). *Knowledge Divergence and the Value of Debate for Scalable Oversight.* arXiv:2603.05293.

[19] Wu et al. (2025). *Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Interaction Dynamics.* arXiv:2511.07784.

[20] Kenton et al. (2024). *On Scalable Oversight with Weak LLMs Judging Strong LLMs.* NeurIPS 2024.

[21] Nayebi (2025). *Intrinsic Barriers and Practical Pathways for Human-AI Alignment: An Agreement-Based Complexity Analysis.* arXiv:2502.05934.

[22] Cui et al. (2025). *Efficient Leave-one-out Approximation in LLM Multi-agent Debate Based on Introspection.* arXiv:2505.22192.