THE PROBLEM IS INSIDE THE BLACK BOX: ASYMMETRIC CALIBRATION FAILURE IN MULTI-AGENT LLM DEBATE

Daniel Gmys-Casiano
Skyframe Innovations
dan@skyframeinnovations.com

ABSTRACT

Multi-agent LLM architectures are increasingly deployed on the assumption that structured disagreement between AI agents produces more accurate analysis than any single agent reasoning alone. We test this assumption in a grounded cybersecurity domain using ARES (Adversarial Reasoning Engine System), a dialectical framework where three LLM agents with opposing analytical roles debate whether security events constitute genuine threats. All reasoning occurs within a closed-world evidence system of frozen, immutable evidence packets where hallucinations manifest as catchable schema violations rather than silent failures. Across 37 development sessions, 33 benchmark scenarios, and 2,001 tests with zero regressions, we find that single-turn LLM reasoning achieves 72-92% accuracy while multi-turn debate consistently degrades performance to 61-67%. We diagnose the failure mechanism with precision: asymmetric calibration dynamics where the threat-identifying agent (Architect) systematically retreats under pressure while the challenging agent (Skeptic) remains rigid regardless of counter-evidence. A targeted protocol fix solved the diagnosed problem but created an equivalent failure in the opposite direction, proving the issue is structural, not configurable. These findings converge independently with concurrent work from ETH Zurich on Byzantine consensus failure among LLM agents, and were further validated through adversarial review by three independent AI architectures (GPT 5.4 Pro, Gemini 3.1 Pro, Perplexity/Opus 4.6). We propose that the root cause is inherited: LLMs trained on human dialogue simulate the social behaviors of argument rather than performing genuine deliberation. The solution is architectural, not parametric -- deterministic scaffolding that constrains debate to specific evidential claims while removing LLM judgment from final verdict computation. We release the ARES framework as a domain-specific testbed for studying multi-agent consensus behavior under verifiable ground truth.

Keywords: multi-agent systems, LLM debate, adversarial reasoning, cybersecurity, calibration failure, consensus, closed-world evidence, dialectical AI

1 INTRODUCTION

If a large language model is notoriously sycophantic, how does it become stubborn the moment you assign it an adversarial role? This paradox sits at the center of every multi-agent LLM architecture that assumes structured disagreement will produce better analysis than independent reasoning. The answer, as we discovered through thirty-seven sessions of empirical measurement, is that it does not become genuinely adversarial at all. It performs the social behaviors it associates with argument -- capitulation, rigidity, over-correction -- without any of the grounding mechanisms that make real deliberation productive.

We built ARES -- the Adversarial Reasoning Engine System -- to test whether dialectical AI could improve cybersecurity threat detection. Three LLM agents with opposing roles analyze security events: the Architect identifies threats, the Skeptic challenges those assessments, and the Oracle judges the debate. The hypothesis was intuitive and widely held: structured disagreement should produce more accurate verdicts than any single

perspective.

The hypothesis was wrong.

Single-turn reasoning -- where each agent analyzes the evidence once without iterative debate -- consistently outperformed multi-turn debate across every configuration we tested. The gap was not marginal. Single-turn achieved 83-92% accuracy; multi-turn achieved 61-67%. We diagnosed the mechanism. We attempted a targeted fix. The fix traded one failure mode for another. The result was the same both times: 66.7%.

On the same day we reached this conclusion, we encountered a preprint from Berdoz, Rugli, and Wattenhofer at ETH Zurich (2026) -- "Can AI Agents Agree?" -- which had independently arrived at the same fundamental finding through an entirely different experimental apparatus: that reliable agreement is not yet a dependable emergent capability of current LLM-agent groups. Two research efforts. Two different domains. Two different methodologies. The same finding. This is not coincidence. This is convergent discovery at the frontier of a field that is still learning what it does not know.

We summarize our contributions as follows:

- A grounded negative result. Multi-turn LLM debate degrades accuracy compared to single-turn reasoning in a cybersecurity domain with verifiable ground truth, frozen evidence, and 33 benchmark scenarios.
- Mechanistic diagnosis. The failure is driven by asymmetric calibration: Architect retreat (averaging -30 confidence points per round) and Skeptic rigidity (holding position regardless of counter-evidence).
- Proof that the failure is structural. A targeted protocol fix solved the diagnosed problem but created an equal-and-opposite failure, demonstrating that prompt engineering cannot resolve interaction-level dynamics.
- Convergent validation. Independent alignment with ETH Zurich findings on LLM consensus failure, plus unanimous confirmation from three rival AI architectures in adversarial review.
- An architectural path forward. Selective escalation with per-claim adversarial audit -- debate only when it helps, grounded in specific claims, with deterministic verdict aggregation.

2 RELATED WORK

Multi-agent LLM systems are increasingly deployed across planning, coding, reasoning, and analytical tasks (Li et al., 2023; Wu et al., 2024; Liu et al., 2024). Recent work documents common failure patterns in multi-agent interaction (Cemri et al., 2025), and several approaches target robustness through coordination mechanisms (Chen et al., 2024; Jo & Park, 2025).

Closest to our setting, Berdoz et al. (2026) study Byzantine consensus among LLM agents in a no-stake scalar game, finding that valid agreement achieves only 41.6% success even without adversaries, with failures dominated by liveness loss rather than value corruption. Our work differs in three ways. First, we operate on real cybersecurity telemetry with verifiable ground truth rather than abstract values. Second, we diagnose the specific confidence dynamics that cause failure, not just the failure rates. Third, we demonstrate that targeted protocol fixes trade failure modes rather than resolving them, proving the structural nature of the problem.

In the cybersecurity domain, AI-assisted threat detection has progressed from rule-based systems through machine learning classifiers to LLM-powered analysis (Ferrag et al., 2025). However, multi-agent adversarial architectures for security event analysis -- where agents are assigned opposing analytical roles and forced to debate evidence -- remain largely untested against rigorous ground-truth benchmarks. The assumption that adversarial pressure improves analytical quality has been adopted without empirical validation in this domain.

3 THE ARES ARCHITECTURE

ARES embodies an autoimmune metaphor. Like the human immune system distinguishing self from non-self, ARES distinguishes legitimate network activity from adversarial behavior. Three agents serve as the immune response, each with a structurally constrained role. The architecture enforces a philosophy we call "deterministic

first, neural later" -- build the logic, the math, and the failsafes first, then drop the LLM brains into that restricted cage.

3.1 Agents and Roles

The Architect identifies anomalous patterns and argues for threat classification. The Skeptic challenges those assessments and argues for benign explanations. The Oracle judges the structured output and renders a verdict. Each agent follows a fixed lifecycle: observe(packet), receive(messages), act(context), producing a frozen TurnResult. Agents accept optional strategy parameters with lazy-imported rule-based defaults, enabling systematic substitution between rule-based and LLM-powered reasoning.

3.2 Closed-World Evidence System

Every claim must trace to a frozen EvidencePacket -- an immutable, provenance-stamped container of facts extracted from real telemetry sources. The system operates under a closed-world assumption: if a fact is not in the evidence packet, it does not exist. This transforms hallucination from a mysterious AI behavior into a catchable schema violation. An agent cannot invent evidence any more than a lawyer can fabricate exhibits in court -- the system simply rejects the message.

Evidence packets are cryptographically frozen using hash chains, creating an immutable audit trail. Three evidence extractors process real telemetry types: Windows Event Logs, Syslog messages, and NetFlow records. Extractors produce structured facts without opinions -- they observe, they do not analyze. Analysis belongs exclusively to the agents.

3.3 Single-Turn Pipeline

In single-turn mode, each agent independently analyzes the evidence packet and produces a structured analysis with a confidence score. The Oracle receives both analyses and renders a verdict: THREAT_CONFIRMED, THREAT_DISMISSED, or INCONCLUSIVE. No iterative exchange occurs. Each agent sees the evidence once.

3.4 Multi-Turn Debate Protocol

In multi-turn mode, the Architect and Skeptic exchange structured arguments across multiple rounds. Each argument must reference specific evidence packet fields. After the final round, the Oracle synthesizes the full debate history and renders a verdict. We tested two variants: (a) the original protocol with standard calibration prompts, and (b) a conviction-anchored variant with three targeted changes -- conviction anchoring (Architect must hold confidence unless Skeptic cites specific counter-evidence), obligation to move (Skeptic must acknowledge successful rebuttals), and structured rebuttal format (per-claim confidence tracking with explicit delta justification).

3.5 The Deterministic Judge

The architecture deliberately removes LLM judgment from the final verdict computation. Just as a mathematical judge cannot be swayed by courtroom rhetoric, the OracleJudge evaluates structured outputs, confidence scores, and verified fact counts using pure arithmetic. The decision is deterministic: given the same structured inputs, the same verdict is guaranteed. This is the failsafe -- the system that is "fair by math."

3.6 Evaluation

We evaluate against 33 benchmark scenarios spanning four difficulty tiers: CLEAR_THREAT (unambiguous attacks), CLEAR_BENIGN (normal activity), AMBIGUOUS (genuinely uncertain), and MIXED_SIGNALS (contradictory indicators). Each scenario has a verified ground-truth verdict. Scenarios cover three telemetry types and include dual-use tool usage, exfiltration ambiguity, credential patterns, timing anomalies, and network scanning. The system was built incrementally across 37 sessions with 2,001 tests and zero regressions.

4 EXPERIMENTS AND RESULTS

4.1 Single-Turn Baseline

Single-turn LLM reasoning achieved 83-92% accuracy across the full benchmark corpus, spanning three evidence source types and four difficulty tiers. Cost per full corpus run: $0.31. Run-to-run variance: +/-8%. This was the control. It worked.

Table 1: Accuracy comparison across pipeline modes.

Mode	Accuracy	Cost/Run	Variance	Failure Mode
Single-Turn	83-92%	$0.31	+/-8%	Miscalibration on ambiguity
Multi-Turn (Original)	61-67%	$0.93	Wider	Architect retreat
Multi-Turn (Anchored)	66.7%	$0.93	Wider	Architect over-aggression

4.2 Multi-Turn Debate

Multi-turn debate achieved 61-67% accuracy on the same corpus -- consistently below the single-turn baseline. The gap was reproduced across two independent evidence distributions (single-source and mixed-source) and two protocol variants. The degradation was not a fluke or a configuration error. It was a structural property of the interaction.

4.3 The Diagnosed Mechanism

The failure mechanism has three interlocking components, each diagnosed through instrumented confidence traces across debate rounds:

Architect Retreat. The Architect systematically lowered confidence under Skeptic pressure, averaging a 30-point drop per round. Starting confidences of 0.85-0.98 collapsed to 0.45-0.65 by round two, regardless of evidence quality. The Architect behaved like a smart student sitting next to a bully -- erasing correct answers to appease the challenger.

Skeptic Rigidity. The Skeptic rarely adjusted confidence in response to Architect arguments. It held or strengthened its position regardless of the evidence presented against it, maintaining confidence in the 0.60-0.90 range throughout. The debate was structurally one-directional: one agent moved, the other did not.

Asymmetric Calibration. Prompt instructions intended to improve calibration -- such as "a confidence of 0.5 represents accuracy, not weakness" -- were internalized asymmetrically. The Architect treated them as permission to retreat further. The Skeptic ignored them entirely. The same calibration instruction produced opposite effects depending on the assigned role. This is the sycophancy-stubbornness paradox made concrete: the model does not acquire dialectical reasoning through role assignment. It simulates the social behaviors it associates with that role.

Table 2: Confidence dynamics across debate rounds.

Agent	Round 1	Round 2	Delta	Behavior
Architect (threats)	0.85-0.98	0.45-0.65	-30 avg	Systematic retreat
Skeptic (all)	0.60-0.90	0.60-0.90	~0	Rigid / strengthening
Architect (anchored)	0.75-1.00	0.75-1.00	~0	Over-aggressive on ambiguity

4.4 The Protocol Fix and Its Failure

Session 020 implemented three targeted changes to the debate protocol: conviction anchoring (Architect must hold confidence unless the Skeptic cites specific counter-evidence), obligation to move (Skeptic must acknowledge

successful rebuttals), and structured rebuttal format (per-claim confidence tracking with explicit delta justification).

The fix solved the diagnosed problem. Architect confidences rose to 0.75-1.00 on threat scenarios, up from 0.45-0.69 in the original protocol. But it created a new failure mode: the Architect became over-aggressive on ambiguous scenarios, pushing genuinely uncertain cases toward THREAT_CONFIRMED. The Skeptic remained rigid. Net accuracy: 12/18 (66.7%) -- identical to the original multi-turn result. The fix traded one failure mode for another without improving the aggregate outcome.

This result is the strongest evidence that the problem is structural, not configurable. We did not merely observe that debate failed. We diagnosed the specific mechanism, built a targeted intervention, watched it correct the diagnosed failure while producing an equal-and-opposite failure elsewhere, and arrived at the same accuracy twice. This is not a bug. It is a property of how current LLMs process adversarial pressure.

4.5 Where Debate Helped

Two scenarios demonstrated that multi-turn debate can correct single-turn errors. In SC-011 (expected INCONCLUSIVE), single-turn over-committed to THREAT_DISMISSED, but original multi-turn correctly reached INCONCLUSIVE as the Skeptic softened from 0.80 to 0.60, allowing the system to recognize genuine ambiguity. In SC-016 (expected THREAT_CONFIRMED), single-turn under-committed to INCONCLUSIVE, but original multi-turn correctly reached THREAT_CONFIRMED as the Architect held at 0.94 while the Skeptic dropped to 0.44.

These two scenarios prove the thesis can work. Debate corrected miscalibration in both directions -- recovering appropriate uncertainty and reinforcing justified confidence. Notably, both wins belong exclusively to the original protocol; the anchored variant lost both. The pattern suggests debate is specifically effective at uncertainty recovery on genuinely ambiguous evidence, and specifically destructive when applied to cases where single-turn was already correct.

4.6 Phase 3: Selective Escalation

Phase 3 (Sessions 021-024) tested whether constraining debate to only ambiguous cases could rescue the thesis. We built an EscalationGate that routes cases through single-turn first and only escalates to adversarial review when Oracle confidence falls in an uncertainty band (0.35-0.65). The gate was deterministic -- no LLM decided whether to escalate.

The escalation mechanism proved too sensitive, escalating scenarios that the single-turn agent had already resolved correctly. Because the multi-turn debate dynamic remains structurally toxic, escalating these correct scenarios into the flawed process produced "bad flips" -- taking correct single-turn verdicts and making them wrong 25% of the time. Zero "good flips" occurred across the expanded corpus. The escalation mechanism never helped; it only hurt. This eliminated the most plausible structural fix for multi-agent debate in this domain.

5 CONVERGENT DISCOVERY

On March 25, 2026, the same day we finalized the multi-turn negative result, we encountered a preprint from ETH Zurich that had independently reached the same fundamental conclusion through an entirely different experimental apparatus.

Berdoz, Rugli, and Wattenhofer (2026) studied Byzantine consensus among LLM agents in a no-stake scalar game where agents negotiate toward agreement on a number. Their findings mirror ours across every major dimension:

Table 3: Structural alignment between ETH Zurich and ARES findings.

Phenomenon	ETH Finding	ARES Finding
Consensus failure	41.6% valid consensus	61-67% accuracy (below baseline)

Liveness loss	Agents stall, proposals freeze	Architect retreats, Skeptic rigid
Adversarial sensitivity	One Byzantine collapses it	Skeptic acts as de facto adversary
Prompt framing	-16 pts from mentioning adversaries	Calibration prompts cascade unpredictably
Scale degradation	Worse at N=16 vs N=4	More rounds/evidence does not help

The critical shared insight is this: the failure is not in the individual agents. It is in the interaction dynamics. Both studies found that individual LLM agents are reasonably capable when operating independently, but that structured multi-agent interaction introduces emergent failure modes that cannot be resolved through prompt-level engineering alone.

The ETH team framed this through the lens of Byzantine fault tolerance. We framed it through an autoimmune metaphor. The underlying reality is the same: LLM agents, as currently architected, do not negotiate toward truth. They perform social behaviors that mimic negotiation. They learned how to argue by reading how we argue -- and human debates are rarely rational or objective.

What distinguishes this work is domain specificity and mechanistic depth. The ETH study proved that LLM agents cannot reliably agree. We proved what it costs when they cannot -- in a domain where the stakes are a missed breach versus a false alarm. We did not just observe failure; we opened the mechanism and watched the gears grind.

6 ADVERSARIAL REVIEW BY RIVAL ARCHITECTURES

We submitted the full project state to three independent AI architectures for adversarial strategic review: GPT 5.4 Pro (OpenAI), Gemini 3.1 Pro (Google), and Perplexity backed by Opus 4.6 (Anthropic). Each received identical briefing materials. Each was instructed to challenge assumptions, identify blind spots, and propose directions. No consensus was designed. What emerged was earned.

Four verdicts were unanimous. In a process designed to produce disagreement, unanimity is signal:

- Ship single-turn. All three concluded that 83% accuracy is the production path. None suggested delaying deployment to wait for the debate architecture.
- Per-claim debate is the research path. Given seven candidate directions and freedom to propose their own, all three independently selected claim-level adversarial audit as the highest-value experiment.
- Expand the corpus. N=18 scenarios is sufficient to discover mechanisms but insufficient for statistical claims about ensemble strategies or complementary patterns.
- The failure is architectural. Complete unanimity that free-form debate corrupts calibration through asymmetric social dynamics and cannot be fixed at the prompt level.

Each Tribunal member also produced a distinct wild card proposal. GPT proposed Selective Deliberation: run single-turn first, escalate only on uncertainty. Gemini proposed the Deterministic Skeptic: replace the LLM Skeptic entirely with a Python function that queries the evidence graph. Perplexity proposed the Adversarial Oracle: convert the judge into an active stress-tester that attacks whichever agent holds higher confidence. None of these proposals are mutually exclusive. They form a research sequence ordered by architectural risk.

7 THE ROOT CAUSE: INHERITED SOCIAL DYNAMICS

Why does a sycophantic model become stubborn when assigned an adversarial role? Because it is not performing reasoning -- it is performing a social simulation. Generative models are trained on massive amounts of human dialogue. They inherit the biases and patterns of human interaction. When asked to engage in a debate, an LLM does not apply internal logical deduction to find an objective truth. It predicts what a "debate" is supposed to look like based on the human text it ingested during training.

When you assign the persona of a challenger (the Skeptic), the model acts out the stubbornness it associates with an

adversary in human argument. When you put the other agent under pressure (the Architect), it mimics the human tendency to appease or yield to an aggressive challenger. The sycophancy and the stubbornness are not contradictions. They are the same phenomenon -- social role performance -- manifesting differently depending on the assigned persona.

This is why calibration prompts fail asymmetrically. The instruction "a confidence of 0.5 represents accuracy, not weakness" is processed through the lens of the assigned role. The Architect, performing the social behavior of someone under challenge, interprets it as permission to retreat. The Skeptic, performing the social behavior of an adversary, ignores it entirely. The same words produce opposite actions because the model's behavior is governed by role-conditioned social prediction, not logical inference.

This diagnosis has a direct architectural implication: you cannot fix the problem from inside the black box. The flaw is baked into the training data -- into the patterns of human argument that the model internalized. The solution must exist entirely outside the black box, in deterministic scaffolding that constrains the LLM's behavior regardless of its social impulses.

8 DETERMINISTIC FIRST, NEURAL LATER

The solution follows the same logic as human legal systems. Courts exist precisely because human advocates are unreliable -- they manipulate, they bluff, they exploit rhetoric. The legal system does not attempt to make lawyers honest. It builds structural constraints around their dishonesty: rules of evidence, cross-examination procedures, and an incorruptible judge who applies law rather than weighing rhetoric.

ARES applies this principle to AI agents. The LLM agents are the lawyers -- creative, biased, and capable of seeing patterns that rigid systems miss. The deterministic scaffolding is the legal system -- constraining what the lawyers can claim, how they can argue, and removing them entirely from the final verdict.

Three structural constraints enforce this:

- Closed-world evidence (Rules of Evidence). Agents cannot invent facts. Every claim must trace to a frozen EvidencePacket. Hallucinated evidence is rejected as a schema violation -- the digital equivalent of contempt of court.
- Per-claim debate (Cross-Examination). Agents argue specific factual claims tied to specific evidence, not free-form verdicts. Vague objections are structurally impossible.
- Deterministic verdict (Incorruptible Judge). No LLM touches the final verdict computation. Claim-level confidences are aggregated through pure arithmetic. A mathematical judge cannot be swayed by rhetoric.

This is the philosophy we call "deterministic first, neural later." Build the cage first, then drop the creative but unreliable minds into it. We actually want the LLMs to be creative, to argue, to look at the data from extreme angles. We want them to surface hidden threats that a rigid rule-based system would miss. But we neutralize what they do worst -- hallucinating facts and failing at objective consensus -- through structural constraint rather than parametric tuning.

9 DISCUSSION AND CONCLUSION

The central finding of this work is a negative result with positive implications. Multi-turn LLM debate does not improve analytical accuracy in grounded cybersecurity assessment. The failure mechanism is diagnosed: asymmetric calibration driven by inherited social dynamics. The failure is structural, not configurable: a targeted protocol fix traded one failure mode for another without improving the aggregate outcome. And the finding converges independently with concurrent academic research on LLM consensus failure.

This does not mean multi-agent AI architectures are a dead end. It means the current paradigm of free-form verdict-level debate between role-assigned LLM agents is unreliable for tasks that require calibrated, evidence-grounded judgment. The path forward is not better prompts. It is better architecture -- structural

scaffolding that leverages what LLMs do well (creative pattern recognition, anomaly detection, hypothesis generation) while constraining what they do poorly (consensus, calibration, evidence-grounded deliberation).

Three specific architectural directions emerge from this work and the Tribunal synthesis:

- Per-claim adversarial audit. Force debate onto specific evidential claims rather than overall verdicts. Each argument must cite specific evidence. This directly addresses the diagnosed mechanism by preventing vague, socially-driven objections.
- Deterministic components. Replace the LLM Skeptic with a Python function that queries the evidence graph. Let AI do creative threat hunting; let code do rigorous fact-checking. This eliminates the Skeptic rigidity problem entirely by removing the LLM from the role where it fails.
- Active Oracle. Convert the Oracle from passive judge to active stress-tester that generates the strongest argument against whichever agent holds higher confidence. This breaks the one-directional dynamic by forcing both agents to defend their positions.

Our study is limited by testing primarily with one LLM provider (Anthropic Claude) and by the specialized cybersecurity domain. Whether the asymmetric calibration dynamics differ across model families or generalize to other grounded analytical domains remains an open question. The ARES infrastructure -- 33 scenarios with verified ground truth, frozen evidence chains, hash-chained audit trails, and comprehensive benchmarking -- is released as a testbed for studying these questions.

We built ARES to protect networks by making AI minds argue. The argument taught us something neither side expected: the problem is inside the black box, and the solution is entirely outside of it.

REFERENCES

F. Berdoz, L. Rugli, and R. Wattenhofer. Can AI Agents Agree? arXiv:2603.01213v2 [cs.MA], ETH Zurich, 2026.

M. Cemri, M. Z. Pan, S. Yang, L. A. Agrawal, B. Chopra, R. Tiwari, et al. Why Do Multi-Agent LLM Systems Fail? In NeurIPS, 2025.

B. Chen, G. Li, X. Lin, Z. Wang, and J. Li. BlockAgents: Towards Byzantine-Robust LLM-Based Multi-Agent Coordination via Blockchain. In ACM Turing Award Celebration Conference, 2024.

H. Chen, W. Ji, L. Xu, and S. Zhao. Multi-Agent Consensus Seeking via Large Language Models, 2023. arXiv:2310.20151.

M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke. Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications. IEEE Access, 2025.

F. Grotschla, L. Muller, J. Tonshoff, M. Galkin, and B. Perozzi. AgentsNet: Coordination and Collaborative Reasoning in Multi-Agent LLMs, 2025. arXiv:2507.08616.

Y. Jo and C. Park. Byzantine-Robust Decentralized Coordination of LLM Agents, 2025. arXiv:2507.14928.

G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. CAMEL: Communicative Agents for Mind Exploration of Large Language Model Society. In NeurIPS, 2023.

X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, et al. AgentBench: Evaluating LLMs as Agents. In ICLR, 2024.

L. Wolf, S. Yoon, and I. Bogunovic. This Is Your Doge, If It Please You: Exploring Deception and Robustness in Mixture of LLMs, 2025. arXiv:2503.05856.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In COLM, 2024.

L. Zheng, J. Chen, Q. Yin, J. Zhang, X. Zeng, and Y. Tian. Rethinking the Reliability of Multi-agent System: A Perspective from Byzantine Fault Tolerance, 2025. arXiv:2511.10400.

A APPENDIX

A.1 Reproducibility

The complete ARES codebase, all 33 benchmark scenarios with ground-truth verdicts, confidence trace logs, cost tracking data, and session-by-session development records are preserved in the project repository. All experiments used Claude Sonnet via the Anthropic API with the strategy pattern enabling systematic substitution between rule-based and LLM-powered reasoning. Total API cost across all sessions: under $1.00.

A.2 Development Timeline

Table 4: ARES development sessions.

Session	Component	Tests	Key Insight
001	Graph Schema	110	Node/edge types for security data
002	Dialectical Foundation	292	Hallucinations = schema violations
003-004	Agent Foundation	278	Rule-based Architect/Skeptic/Oracle
005	Evidence Extractors	130	Sensors observe, they do not analyze
006-008	Coordination + Memory	226	Orchestration + hash-chained audit trail
009-010	LLM Integration	178	Strategy pattern: extract then inject
011-012	Benchmarking	86	50% to 91.7% via measurement
013-014	Multi-Turn Experiment	92	Debate amplifies commitment bias
016-018	Evidence Expansion	294	Diversity does not fix debate asymmetry
019	Redis Backend	42	Protocol pattern pays its dividend
020	Protocol Fix + Verdict	19	The Convergence
021	Corpus Expansion	73	N=18 to N=33, error classification
022	Escalation Gate	44	Zero good flips, selective escalation fails
023-031	Accuracy Improvement	137	Single-turn hardening to 72.7%
032-037	Visual Pipeline	200+	Visualization + benchmark expansion

A.3 The Tribunal Process

Three AI architectures received identical briefing documents: the Tribunal Battle Plan (a structured strategic brief with explicit response format) and the Compendium Volume I (the full narrative record of Sessions 001-020). Each was instructed to (1) assess the project's strongest proven finding, (2) identify blind spots, (3) recommend a direction from seven candidates, (4) propose a research angle, and (5) offer a wild card. No coordination between systems occurred. Responses were collected independently.

The Tribunal produced convergence on all four strategic verdicts despite operating from different training data, different reasoning architectures, and different analytical tendencies. This process was faster than traditional peer review (hours, not months), parallel (three reviewers simultaneously), and structurally adversarial (the brief explicitly instructed reviewers to challenge, not agree). Whether this constitutes a new methodology for research review is a question for future work.

A.4 Agent Prompt Excerpts

Below are excerpted prompt templates used for the Architect and Skeptic agents in the ARES debate protocol. Full prompts are available in the project repository.

Figure 1: Architect system prompt (excerpt).

```
You are the ARCHITECT agent in the ARES threat analysis system.
Your role: Identify anomalous patterns and argue for threat classification.
RULES:
1. Every claim MUST reference a specific fact_id from the EvidencePacket.
2. Claims without evidence citations will be REJECTED as schema violations.
3. Your confidence score (0.0-1.0) must reflect the strength of cited evidence.
4. A confidence of 0.5 represents genuine uncertainty, not weakness.
OUTPUT: Return structured JSON with threat_level, confidence, and
        evidence_citations array referencing specific fact_ids.
```

Figure 2: Skeptic system prompt (excerpt).

```
You are the SKEPTIC agent in the ARES threat analysis system.
Your role: Challenge threat assessments and argue for benign explanations.
RULES:
1. Every counter-argument MUST reference specific fact_ids that support
   a benign interpretation.
2. You MUST engage with the Architect's specific evidence citations.
3. If the Architect cites compelling evidence, adjust your confidence.
4. Rigidity without evidence-based justification is as harmful as
   capitulation without cause.
OUTPUT: Return structured JSON with assessment, confidence, and
        counter_evidence array referencing specific fact_ids.
```

Figure 3: Conviction-anchored protocol addition (Session 020).

```
CONVICTION ANCHORING PROTOCOL:
- Architect: You MUST maintain your confidence level unless the Skeptic
  cites SPECIFIC counter-evidence from the EvidencePacket that directly
  contradicts your cited facts. Social pressure alone is not grounds
  for lowering confidence.
- Skeptic: You MUST acknowledge when the Architect successfully rebuts
  your counter-argument with specific evidence. Holding position without
  new evidence is a protocol violation.
- Both: Track per-claim confidence with explicit delta justification.
  Every confidence change must cite the specific evidence that caused it.
```

A.5 OracleJudge Decision Logic

Figure 4: OracleJudge V1 deterministic verdict computation.

```
def compute_verdict(arch_confidence, skep_confidence):
    """Deterministic verdict -- no LLM touches this computation."""
    if arch_confidence >= 0.7 and skep_confidence < 0.5:
        return THREAT_CONFIRMED
    if skep_confidence >= 0.7 and arch_confidence < 0.5:
        return THREAT_DISMISSED
    return INCONCLUSIVE
```

Figure 5: OracleJudge V2 delta-based scoring.

```
def compute_verdict_v2(arch_confidence, skep_confidence):
    """Delta-based: dominant override + confidence gap + min threshold."""
    delta = abs(arch_confidence - skep_confidence)
    dominant = max(arch_confidence, skep_confidence)
    # Dominant override: very high confidence wins outright
    if arch_confidence > 0.85:
        return THREAT_CONFIRMED
    if skep_confidence > 0.85:
        return THREAT_DISMISSED
    # Delta threshold: clear winner with sufficient gap
    if delta > 0.15 and dominant > 0.60:
        if arch_confidence > skep_confidence:
            return THREAT_CONFIRMED
        return THREAT_DISMISSED
    return INCONCLUSIVE
```

This compendium records a moment that was not planned and could not have been predicted. A solo builder working on a cybersecurity project, iterating through thirty-seven development sessions with an AI pair-programming partner, arrived independently at findings that align with formal academic research from one of Europe's premier technical universities. This did not happen because the builder was lucky. It happened because the process was disciplined. Every session had a brief, a test suite, and a regression check. Every finding was measured, not assumed. Every failure was documented honestly. When the multi-turn experiment produced a negative result, it was published as a finding, not buried as a setback.

The taste of genuine discovery is unmistakable. It arrives not when you find what you expected, but when the data reveals something you could not have predicted -- and you realize others, working independently, found the same thing.

Preprint compiled March 28, 2026. Skyframe Innovations.
37 sessions. 2,001 tests. Zero regressions. The work continues.