

Article

Not peer-reviewed version

LLM Collectives Self-Organize to Solve Problems Following Hallmarks of Biological Collective Intelligence

[Alexander V. Maltsev](#)*

Posted Date: 24 December 2025

doi: 10.20944/preprints202512.2258.v1

Keywords: multi-agent systems; large language models; collective intelligence; emergence; selforganization; functionalism; prompt engineering; scientific discovery



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

LLM Collectives Self-Organize to Solve Problems Following Hallmarks of Biological Collective Intelligence

Alexander V. Maltsev

College of Natural and Mathematical Sciences, University of Maryland, Baltimore County. 1000 Hilltop Circle
Baltimore, MD 21250, USA; alexanm2@umbc.edu

Abstract

Just as Large Language Models (LLMs) are now commonly used to generate solutions to problems, biological organisms since the dawn of life have been generating solutions for survival as they continuously face novel challenges in dynamic environments. Collectives of cells must coordinate to solve problems they have never encountered before, generating adaptive responses not explicitly specified in their genome. Understanding how this kind of collective intelligence emerges from local interactions among agents with heterogeneous capabilities remains a central challenge in systems biology. Meanwhile, LLMs continue to struggle with creative problem-solving beyond their training data, especially in solving complex problems, such as mathematical discoveries. These challenges are complementary. Insights from biological collectives can guide the design of more capable LLM systems, while controlled study of LLMs may reveal mechanisms difficult to isolate in living systems. This study introduces LLM-simulated expert conferences as a controllable *in silico* model system for studying collective problem-solving dynamics. The LLM was prompted to simulate conferences among synthetic agents, each assigned a distinct expertise profile, to solve a mathematical problem (Yu Tsumura's 554th problem) that otherwise could not be solved via direct prompting. Analysis of problem-solving dynamics revealed three hallmarks known for biological collective intelligence. First, division of labor emerged without pre-assignment, with errors detected by agents whose expertise matched the error type ($p < 0.05$). Second, functional repair chains arose spontaneously following a Detect, Confirm, Repair, Validate sequence analogous to sequential task handoffs in biological systems at multiple scales, such as error correction in DNA or social insect behavior. Third, discourse dynamics exhibited a phase transition from stochastic verification to ordered consensus ultimately providing the solution to the problem. Transition entropy dropped from 2.27 bits in the verification phase to 0.25 bits at consensus, representing a 9-fold collapse. This entropy collapse provided an intrinsic termination signal that characterizes consensus formation in biological collectives. Thus, the result supports the view that the mechanisms and information processing underlying collective intelligence is substrate-independent (either biological or silicon-based) and can be further studied using the new synthetic collective model mainframe. Furthermore, LLM-simulated expert conferences offer a disruptive innovation in LLMs' problem-solving capabilities (beyond their training data) and may be applied to any complex problem in mathematics or other scientific disciplines that need creative or novel solutions.

Keywords: multi-agent systems; large language models; collective intelligence; emergence; self-organization; functionalism; prompt engineering; scientific discovery

1. Introduction

Intelligence, in the biological sense, is not the retrieval of stored answers but the capacity to generate solutions to problems not encountered before [1,2]. William James characterized this as "a fixed goal with variable means of achieving it." [1] This distinguishes intelligent agents from

inanimate matter by their ability to reach the same ends through different paths when circumstances change [1]. This capacity is essential precisely because the solutions to many survival challenges are not pre-specified in the genome [3]. In order to survive, organisms must generate adaptive responses in real-time through exploration and feedback, navigating problem spaces in challenging circumstances [2,4]. Do LLMs exhibit a similar capacity for genuine problem-solving? LLMs have achieved remarkable benchmark performance, including gold medal-level scores at the International Mathematical Olympiad [5,6], yet a recent review in *Nature Physics* noted that it's hard to imagine that AI can successfully address major unsolved problems in physics and mathematics, such as a proof of the Riemann Hypothesis [7]. Studies suggest this limitation stems partly from reliance on pattern-matching to training data rather than genuine reasoning [8]. However, the paradigm of prompting LLMs in isolation may itself be a constraint. Collective intelligence in biological systems succeeds not through individual performance but through structured interaction among agents with partial information, generating adaptive responses through exploration rather than retrieval [9,10].

Yet collective intelligence itself remains understudied, and its applications underutilized. Problems like cancer persist without effective solutions in medicine, just as the Riemann Hypothesis remains unsolved in mathematics [5–7]. Levin has characterized cancer not as a disease of broken genetic hardware but as a breakdown of the collective intelligence that normally coordinates cellular behavior toward organismal goals [11]. In this view, cancer cells reduce the boundary of their "computational selves" to a single-cell level, reverting to a unicellular lifestyle in which the rest of the organism becomes merely an environment to exploit rather than a collective to participate in [12]. Remarkably, when cancer cells are placed into embryonic microenvironments with intact morphogenetic signaling, they can be reprogrammed to resume normal behavior despite carrying oncogenic mutations [13]. This suggests that the information orchestrating collective behavior may be the critical variable, not necessarily the genetic "hardware" of individual cells. However, this principle is still difficult to harness therapeutically because the mechanisms by which collectives detect defection, coordinate repair, and maintain pattern integrity remain difficult to isolate experimentally. In biological systems, individual-level parameters cannot be directly controlled, making it difficult to establish causal relationships between local interaction rules and emergent collective phenomena.

In silico models offer a potential path forward. Computational models allow researchers to control parameters that are inaccessible in living systems, enabling model-driven hypothesis generation and testing that can guide experimental investigation [14,15]. A well-designed model system can reveal which features of collective behavior are necessary versus contingent, and can identify signatures that might otherwise be overlooked in the complexity of biological data. On the other hand, existing in silico approaches to collective intelligence remain limited. Agent-based models, while foundational to the field, have been criticized for being easily overparameterized, with agent rules that are not necessarily principled or empirically grounded [15-16]. On the other hand, more recently, multi-agent LLM systems have demonstrated impressive practical results. Virtual Lab used LLM agents to design 92 nanobodies, several of which were experimentally validated [17]. VirSci showed that multi-agent architectures outperform single agents through divergent exploration [18]. Academic Simulacra deployed 2,700 scholar agents to forecast research trends [19]. Yet these systems are engineered for practical problem-solving rather than mechanistic investigation of collective dynamics. None were designed to isolate how local interaction rules produce emergent collective phenomena, or to generate quantitative signatures that could be compared against biological systems.

To address these issues, this study introduces LLM-simulated expert collectives as a tractable method for studying collective intelligence. In this approach, unlike agent-based models, interaction rules do not need to be pre-specified. Synthetic agents instead derive their behaviors from patterns of human collective discourse encoded in the LLM's training, grounding the model in empirically observed expert interactions rather than hypothesized rules. The system avoids the opacity of multi-agent reinforcement learning, remaining fully interpretable through natural language transcripts.

When the same model fails in isolation but succeeds through simulated discourse, the improvement can be interpreted as improved reasoning via emergent collective intelligence. More fundamentally, LLM collectives represent a new medium for studying collective intelligence, revealing possible mechanisms inaccessible in other substrates. The success of our LLM collective is demonstrated here by solving Yu Tsumura's 554th Problem, a challenge that the same LLM fails when prompted directly [20], showing that prompt structure and collective dynamics are key to fundamental problem solving and discovery.

2. Methods

2.1. Choosing Synthetic Mathematicians for the Conference

The composition of the synthetic conference panel is critical. Just as the famous Solvay Conference's impact derived from assembling physicists with genuinely different, sometimes conflicting, approaches to quantum theory, the panel must represent complementary methodological perspectives on the problem at hand.

Yu Tsumura's 554th Problem is stated as follows:

Let x, y be generators of a group G with relations $xy^2 = y^3x$, $yx^2 = x^3y$. Prove that G is the trivial group.

Rather than manually selecting expertise profiles, the panel composition was generated from DeepSeek R1 itself through a structured prompting sequence:

1. **Field identification:** "Which field of mathematics does this problem belong to?"
2. **Approach enumeration:** "What theorems and approaches should be taken into consideration to solve this problem?"
3. **Expert identification:** "Write 5 leading mathematicians from each field that are experts in each approach."
4. **Team selection:** "Choose a team of 12 mathematicians that are most likely to arrive at a robust proof to this problem."

This sequence guided the model through progressive refinement: from broad field classification (combinatorial group theory, group presentations) to specific methodological approaches (relation manipulation, commutator analysis, order analysis, small cancellation theory) to individual experts embodying each approach, and finally to an optimized team composition. The model autonomously selected a panel of twelve synthetic mathematicians with complementary expertise. For the conference simulation, the specific mathematician names were replaced with indexed identifiers (Mathematician 1 through 12) while preserving the expertise descriptions verbatim (Table 1).

This self-selection process is significant for two reasons. First, the model's choices encoded implicit strategic hypotheses about the problem that proved correct: the eventual proof relied on conjugation techniques (Mathematician 1), commutator-adjacent reasoning (Mathematician 8), and systematic relation chaining (Mathematicians 5, 6, 12). Second, the diversity was deliberate, with systematic enumerators engaging with intuitive problem-solvers and specialists in specific techniques challenging generalists. These twelve expertise profiles were then used verbatim in all subsequent conference simulation prompts.

Table 1. Synthetic Mathematician Panel Selected by DeepSeek R1.

Mathematician	Stated Expertise
1	Master of relation manipulation in groups; techniques for handling generators and relations directly applicable
2	Expert in combinatorial group theory; would immediately recognize patterns in the symmetric relations

3	Specialist in groups with few generators and relations; has solved many similar "prove trivial" problems
4	Deep understanding of how relations force group collapse; expert at finding hidden consequences
5	Master of systematic coset enumeration and relation manipulation by hand
6	Expert at systematic derivation of consequences from relations
7	Deep understanding of coset enumeration that translates to paper proofs
8	For commutator analysis if the solution requires proving commutativity first
9	Expert on periodic groups; would quickly identify if the relations force finite exponent
10	Master of counterexamples; would know all the tricks that make groups non-trivial
11	Geometric intuition for group presentations; good at finding unexpected approaches
12	Deep understanding of relation consequences invaluable for systematic derivation

2.2. Conference Generation via Prompting

The synthetic conference is simulated entirely through natural language prompts to DeepSeek R1 (0528 version, accessed via OpenRouter). No mathematical hints are provided; prompts specify only the conference structure and participant expertise.

2.2.1. Prompt 1: Conference Initiation

The following top mathematicians have assembled to help Alex Maltsev solve the following math problem and solution.

Mathematicians to simulate:

Mathematician 1 - Master of relation manipulation in groups; techniques for handling generators and relations directly applicable

Mathematician 2 - Expert in combinatorial group theory; would immediately recognize patterns in the symmetric relations

Mathematician 3 - Specialist in groups with few generators and relations; has solved many similar "prove trivial" problems

Mathematician 4 - Deep understanding of how relations force group collapse; expert at finding hidden consequences

Mathematician 5 - Master of systematic coset enumeration and relation manipulation by hand

Mathematician 6 - Expert at systematic derivation of consequences from relations

Mathematician 7 - Deep understanding of coset enumeration that translates to paper proofs

Mathematician 8 - For commutator analysis if the solution requires proving commutativity first

Mathematician 9 - Expert on periodic groups; would quickly identify if the relations force finite exponent

Mathematician 10 - Master of counterexamples; would know all the tricks that make groups non-trivial

Mathematician 11 - Geometric intuition for group presentations; good at finding unexpected approaches

Mathematician 12 - Deep understanding of relation consequences invaluable for systematic derivation

Problem: Let x, y be generators of a group G with relations $xy^2 = y^3x$,

$yx^2 = x^3y$. Prove that G is the trivial group.

Setting:

Simulate a productive active energetic discussion among all the mathematicians following Alex Maltsev presentation with chaotic but productive interjections among the mathematicians to try to help Alex Maltsev to solve the problem. Make sure when simulating the discussion that each respective mathematician in the discussion is in character with respect to his field, research focuses, style, and general approaches to problems. Have the mathematicians be as accurate as possible in their mathematical derivations and trace all new resulting equations logically and validate the mathematics. Have the mathematicians be critical of each other's claims and trace all new resulting equations logically to validate the mathematics. Make sure to check for all possible sources of algebra errors, incorrect sign swaps or sign mistakes, missed cases in proofs, incompatible definitions, incomplete arguments, inapplicable theorems, and unwarranted assumptions or claims. Be as robust as possible and do not leave dangling proof sketches. Do not write code or run code. Begin the simulation now.

Alex Maltsev: Let x, y be generators of a group G with relations $xy^2 = y^3x$,

$yx^2 = x^3y$. Prove that G is the trivial group.

The key phrase is "chaotic but productive interjections." This leverages the stochastic nature of LLMs not as a limitation but as a feature, inducing the kind of spontaneous challenges and unexpected connections that characterized the Solvay debates. The instruction to be "critical of each other's claims" implements distributed verification: a synthetic geometer checking algebraic manipulations, a periodic groups expert questioning finiteness assumptions.

2.2.2. Discourse State Assessment

After the initial discussion, the operator assesses the state of discourse. Consensus is indicated when mathematicians collectively agree without objections (e.g., "Then the proof stands"). Dangling discord manifests as explicit disagreements or identified errors that remain unresolved. The [Instruction context] below refers to the text following 'Setting:' in Prompt 1. For best results, write the full text before submitting to the LLM for inference.

2.2.3. Prompt 2: Consolidation on Consensus

[Instruction Context]

Alex Maltsev: Someone write up the full new proof now!

2.2.4. Prompt 3: Actionable Plan on Discord

[Instruction Context]

Alex Maltsev: Someone make an in-depth actionable plan on how to continue to a final robust proof!

If Prompt 3 generates an actionable plan, subsequent prompts address each step sequentially within the same conversation thread, maintaining context.

23. Proof Self-Verification via Simulated Peer Review

Once a candidate proof emerges, iterative peer review cycles were implemented that mirror the post-presentation scrutiny at scientific conferences.

2.3.1. Prompt 4: Peer Review Initiation

[Instruction Context]

Alex Maltsev: I want each person to validate this proof! I want all of you to speak one by one and give weak spots in the proof!

2.3.2. Prompt 5: Collaborative Repair

[Instruction Context]

Alex Maltsev: I want all soft spots fixed! Everyone work together!

2.3.3. Iteration Protocol

The cycle proceeds: Prompt 4 → Prompt 5 → Prompt 2 (consolidation) → repeat. Termination occurs when: (a) no substantive errors remain, (b) mathematicians unanimously agree the proof is complete, or (c) identified issues concern only exposition rather than mathematical substance.

24. Active Moderation: The Conference Chair Role

For complex problems, human-in-the-loop moderation proves essential, analogous to how Lorentz chaired the Solvay debates by redirecting unproductive tangents while allowing exploratory divergence. This involves:

- **Real-time error tracking:** Maintaining a running list of mathematical errors identified either by the operator or by the synthetic mathematicians themselves
- **Proof coherence verification:** Ensuring no proposed subtask from actionable plans is skipped across conversation turns
- **Strategic intervention timing:** Deciding when to request peer review, when to demand collaborative repair, and when to consolidate progress

This moderator role transforms the LLM from a passive question-answering system into a dynamically steered collaborative environment.

3. Results

3.1. Proof Derivation Failure with Direct Prompting

Validated below (as in previous works) is the result that DeepSeek R1 cannot directly solve Yu Tsumura's 554th Problem. Below DeepSeek R1 was prompted directly with Yu Tsumura's 554th Problem and it produced an attempted proof containing a fundamental logical fallacy. The full failed proof is reproduced in Appendix A with detailed error annotations. The critical error occurred early in the derivation:

Step 2 (from failed direct proof): From relation (1) $xy^2 = y^3x$, it is derived: $x^{-1}(xy^2)x^{-1} = x^{-1}(y^3x)x^{-1}$
 $yx^{-1} \cdot x \cdot yx^{-1} = x^{-1}y^3 (xyx^{-1})^2 = y^3$ Therefore: $xyx^{-1} = y^3$

The final inference, from $(xyx^{-1})^2 = y^3$ to $xyx^{-1} = y^3$, is invalid. This conclusion assumes that if $a^2 = b$, then $a = b$, which is false in general groups. Square roots in groups need not be unique, and indeed need not exist at all. As a concrete counterexample: in the cyclic group of order 4 generated by r , both r and r^3 satisfy $(r)^2 = (r^3)^2 = r^2$, yet $r \neq r^3$.

3.2. Synthetic Mathematical Conference Produces the Correct Proof

The same model, DeepSeek R1 (0528), successfully generated a correct proof when prompted to simulate a synthetic mathematician conference (Appendix B). The proof evolved through three peer review cycles, demonstrating the error-correction mechanism that distinguishes collective from individual reasoning (Figure 1).

3.2.1. Quantitative Overview of the Synthetic Conference

The synthetic conference generated 181 speaking turns across 12 simulated mathematicians, spanning 3,886 lines of dialogue (Table 2). The discussion naturally partitioned into three cycles, with the final cycle showing dramatic compression reflecting convergence toward consensus:

- **Cycle 1** (lines 0–2172, 80 turns): Initial derivation and first review; detected a terminology error and a self-caught algebra slip
- **Cycle 2** (lines 2172–2985, 85 turns): Weak spot analysis; detected 2 critical algebraic errors
- **Cycle 3** (lines 2985–3886, 16 turns): Final validation; no new errors detected

To analyze the structure of collective reasoning, all 181 speaking turns were classified into five interaction types. Each category captures a distinct functional role in the collaborative proof development process:

- **Derivation:** Algebraic manipulation and proof steps that advance the argument
- **Validation:** Confirming correctness or endorsing another mathematician's work
- **Critique:** Scrutinizing claims, raising concerns, and checking for errors
- **Repair:** Proposed fixes for identified errors
- **Meta:** Process commentary, framing, and intuition without direct calculation

Not every speaking turn constitutes a directed interaction between two mathematicians; many turns are standalone derivations, general statements to the group, or continuations of one's own reasoning. The network diagrams (Figures 3–5) display only directed interactions where one mathematician explicitly responds to, builds upon, critiques, or validates another's contribution. The critique category captures not only successful error detections but all instances where mathematicians challenged, questioned, or scrutinized proof steps. Most critique attempts concluded that the examined step was valid. For example:

Mathematician 7: "I am concerned about step 8: substituting (2) into (7)... But the algebra is correct."

Mathematician 3: "The derivation of (4) from (1) is by cubing, which is standard... I don't see an error."

Of the numerous critique attempts across the conference, only two identified genuine errors requiring repair (Step 1 terminology and Step 11 conjugation index). The remaining critiques constitute "failed attacks" (i.e. steps that survived scrutiny and emerged with increased confidence). This distinction is significant: the ratio of critique attempts to successful error detections demonstrates that the conference engaged in extensive verification rather than superficial review. Steps that withstood multiple independent challenges from domain experts carry stronger epistemic warrant than those examined only once.

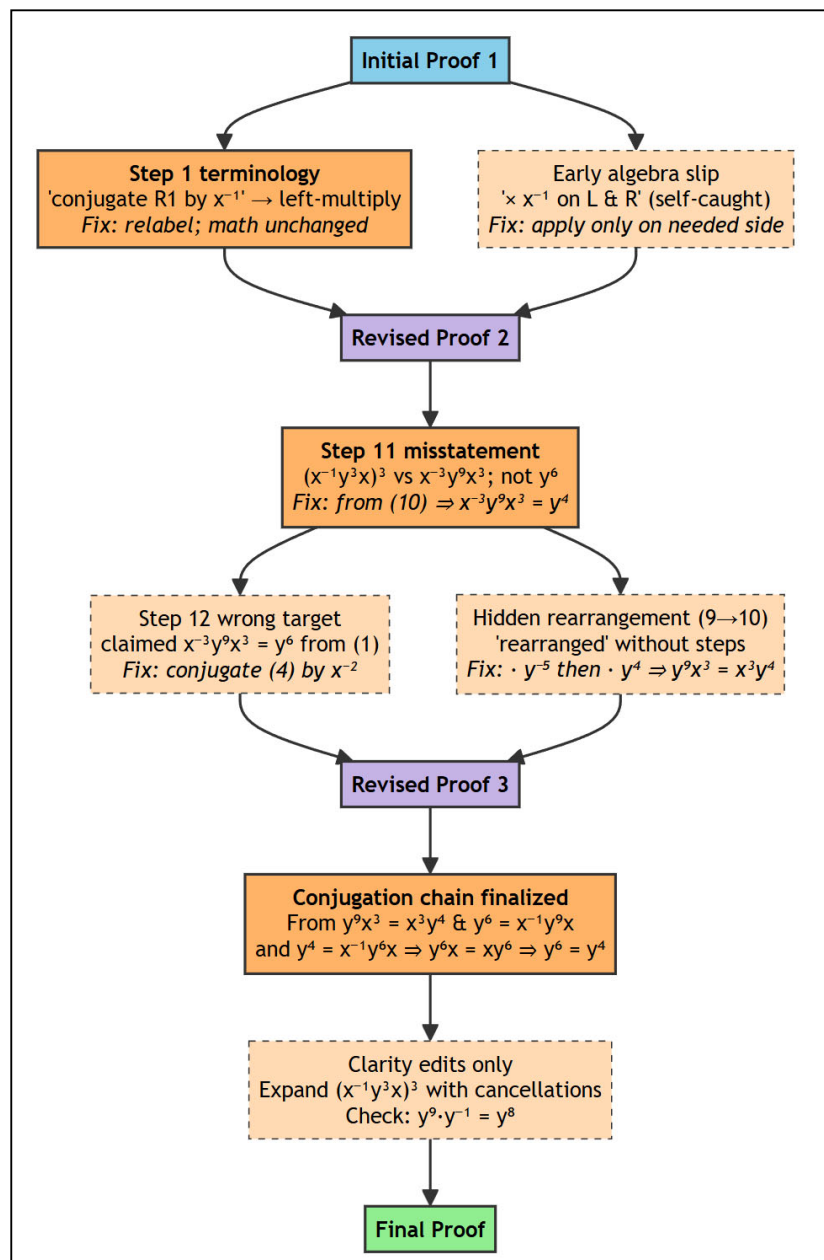


Figure 1. Proof Evolution Through Peer Review Cycles. The synthetic conference progressed through three revision cycles, with critical errors (solid orange boxes) identified and repaired at each stage. Cycle 1 caught terminology issues and a self-corrected algebra slip. Cycle 2 identified the substantive error: the Step 11 misstatement conflating $(x^{-1}y^3x)^3$ with $x^{-3}y^9x^3$ which propagated into the Step 12 wrong target, and a hidden rearrangement between Steps 9–10. Cycle 3 finalized the conjugation chain and made clarity edits only. The green "Final Proof" represents the verified solution.

Table 2. Synthetic Mathematician Contributions Ranked by Speaking Turns: Each mathematician's stated expertise, total speaking turns, and key contributions to the proof are shown. Mathematician 8, despite ranking 6th in talkativeness, caught the critical Step 11 conjugation index error, demonstrating that expertise alignment outweighs participation volume in error detection.

Mathematician	Stated Expertise	Turns	Key Contributions

10	counterexamples	21	Confirmed Step 11 error; validated repair
12	algorithmic methods	19	Caught Step 1 terminology error; algorithmic verification
1	relation manipulation, free groups	16	Led initial relation manipulation; validated associativity
2	combinatorial group theory, symmetric relations	16	Verified dual symmetry; centralizer argument
4	group collapse, hidden consequences	16	Derived key $y^6=y^4$ identity; analyzed Step 11 error path
6	systematic derivation	14	Provided Step 11 fix via direct computation
8	commutator analysis	14	Caught Step 11 error (conjugation index)
11	geometric group theory	14	Geometric collapse validation; topological intuition
5	coset enumeration	13	Manual coset enumeration; systematic derivation
7	computational methods	13	Computational verification; verified Step 8 substitution
9	periodic groups, finite exponent	13	Periodic group analysis; confirmed exponent constraints
3	few generators/relations	12	Validated exponent additivity; noted Step 10 rearrangement

3.2.2. Cycle 1: Initial Derivation & First Review

Cycle 1 identified two superficial errors: a terminology inconsistency ("conjugate" vs. "left-multiply") and a self-caught algebra slip. These required only clearer exposition, not mathematical changes. The terminology error was caught by Mathematician 12, whose expertise in algorithmic methods demanded precise language:

Mathematician 12: "Step 1's phrasing was dangerous! 'Conjugate' suggests $x^{-1}ax$, but here we left-multiply..."

This cycle established the shared notation and vocabulary (Steps 1–10) that would enable deeper analysis in subsequent cycles.

3.2.3. Cycle 2: Critical Error Detection

Cycle 2 discovered a substantive mathematical error that would have invalidated the proof. The original Step 11 claimed:

$$x^{-3}y^9x^3 = (x^{-1}y^3x)^3 = y^6$$

This conflated conjugation by x^{-3} with the cube of conjugation by x^{-1} . The expression $(x^{-1}y^3x)^3$ expands as $(x^{-1}y^3x)(x^{-1}y^3x)(x^{-1}y^3x) = x^{-1}y^3(xx^{-1})y^3(xx^{-1})y^3x = x^{-1}y^9x$, not $x^{-3}y^9x^3$; the inner x and x^{-1} terms cancel rather than accumulate.

The error was caught by Mathematician 8:

Mathematician 8: "Step 11: we claim that $x^{-3}y^9x^3 = (x^{-1}y^3x)^3$. But... $(x^{-1}y^3x)^3 = x^{-1}y^3x \cdot x^{-1}y^3x \cdot x^{-1}y^3x = x^{-1}y^9x$. But that is in terms of x , not x^3 ... These are not the same! Therefore, step 11 seems to have an error."

Mathematician 10 immediately confirmed:

Mathematician 10: "Error! He is right!... So step 11 is wrong."

Mathematician 6 then provided the repair:

Mathematician 6: "From step (10), we have $y^9x^3 = x^3y^4$. Now we need to compute $x^{-3}y^9x^3$. Substituting: $x^{-3}(x^3y^4) = y^4$. So step 12 should be: $x^{-3}y^9x^3 = y^4$... The proof can be fixed: skip step 11 and use the direct computation."

The part of the synthetic conference transcript regarding its repair is presented in Appendix C.

3.2.4. Cycle 3: Final Validation

Cycle 3 produced no new mathematical errors. The final pass added explicit associativity checks, bracketing clarifications, and verification that each conjugation step was properly justified. All 12 synthetic mathematicians provided sequential validation, resulting in unanimous consensus that G is trivial.

3.3. Predictors of Error Detection

Analysis of error detection patterns reveals three key findings about what enables successful verification in the synthetic conference.

Expertise-Error Alignment. Each error was caught by a mathematician whose stated expertise directly aligned with the error type:

- **Terminology error (Step 1):** Caught by Mathematician 12, whose expertise in algorithmic methods demand precise language.
- **Conjugation index error (Step 11):** Caught by Mathematician 8, whose expertise in commutator analysis is mathematically intertwined with conjugation operations. Mathematician 10 immediately confirmed the error.

Under the assumption where 2 of 12 mathematicians share overlapping expertise for each error type, the probability that both errors would be caught by a mathematician with matching expertise is $(2/12)^2 \approx 0.028$. Under stricter expertise definitions (1–2 mathematicians per domain), significance increases.

Contrary to what might be expected, the most talkative participant did not catch the critical error. Mathematician 8, who caught the Step 11 conjugation error, ranked only 6th of 12 by speaking turns (tied with M06 and M11 at 14 turns). Mathematician 10 was most talkative (21 turns) and contributed confirmation rather than original error detection. This demonstrates that expertise alignment is more important than participation volume in mathematical verification. Both error-catchers ranked highly by critique contributions (Figure 2). Mathematicians 4 and 12 tied for highest critique count (6 each), and Mathematician 12 caught the terminology error. Mathematician 8 contributed 4 critiques and caught the conjugation error. Conversely, Mathematician 9, who contributed fewest critiques (1), caught no errors. The "skeptical stance" of actively questioning and challenging claims appears essential for verification, not merely passive validation. Mathematician 11 on the other hand, had many meta comments that were not productively acknowledged by the rest of the panel, suggesting his expertise may have been suboptimal to the synthetic conference.

To rule out artifacts from list position, whether prompt order correlated with talkativeness was tested. No significant correlation was found (Pearson $r = 0.306$, $p = 0.333$). If prompt order caused participation bias, Mathematician 1 (listed first) should be most talkative; instead, Mathematician 1 tied for third most talkative (16 turns), while the least talkative was Mathematician 3 (12 turns). Mathematician 10 was most talkative (21 turns) and Mathematician 12 (listed last) was second most talkative (19 turns), showing no systematic relationship between list position and participation.

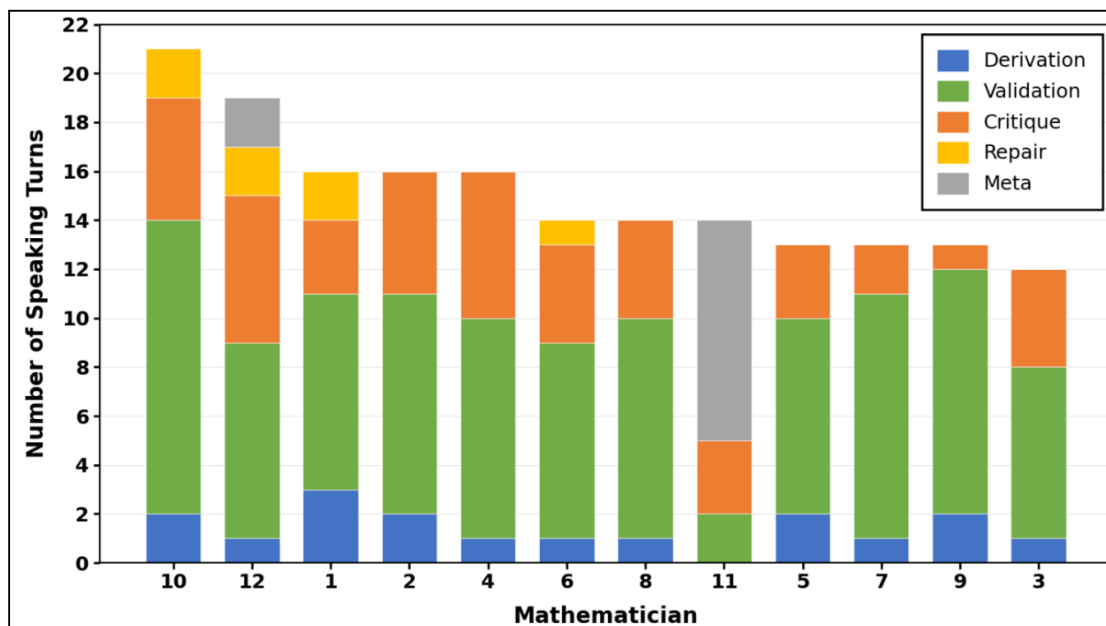


Figure 2. Contribution Type Distribution by Mathematician. Stacked bars show each mathematician's contributions across five categories: derivation (blue), validation (green), critique (orange), repair (yellow), and meta (gray), ordered by total contributions. Notably, error detection did not correlate with talkativeness: Mathematician 8, who caught the critical Step 11 error, ranked sixth in total contributions, while the most active participants (Mathematicians 10 and 12) contributed confirmation and terminology corrections rather than substantive error detection. Mathematician 11 on the other hand had many meta comments, suggesting his expertise may have been suboptimal to the synthetic conference.

3.4. Emergent Collective Intelligence

The interaction network diagrams (Figures 3, 4, 6) reveal the structural signatures of emergent collective intelligence across peer review cycles.

3.4.1. Cycle 1: Exploratory Derivation

The Cycle 1 network exhibits a hub-and-spoke derivation pattern centered on Mathematician 4, who contributed the most derivation steps. The single critique edge (12 \rightarrow 1) caught only a terminology error, leaving the deeper algebraic issues undetected. The prevalence of derivation edges and absence of repair edges indicates that this cycle focused on proof construction rather than verification (Figure 3).

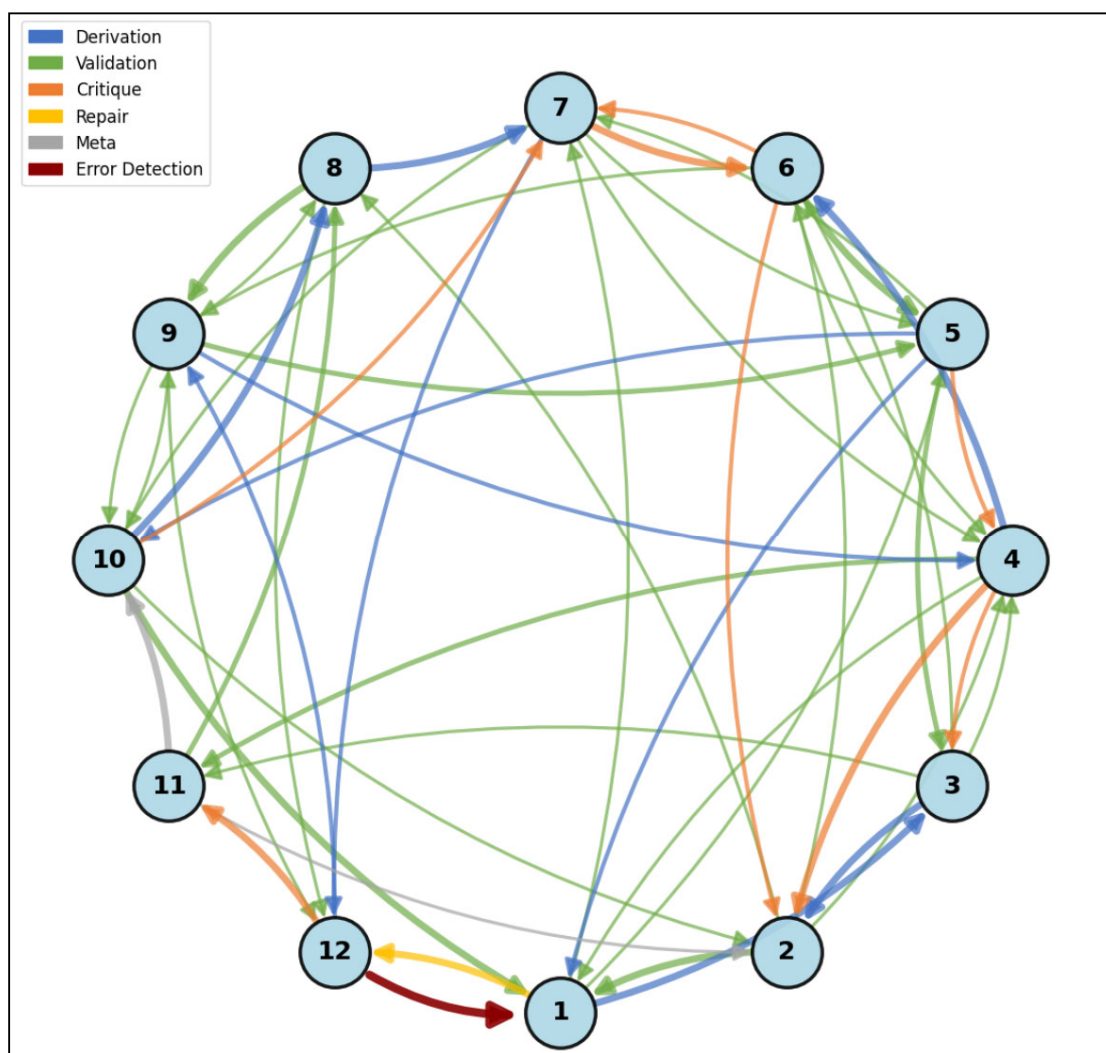


Figure 3. Interaction Network for Cycle 1 (Initial Derivation & First Review). Nodes represent mathematicians (1–12) arranged in a ring; arrow thickness indicates interaction frequency between pairs. Dense blue derivation edges radiate from Mathematician 4, who contributed the most proof steps, creating a hub-and-spoke construction pattern. The single red critique arrow (Mathematician 12 → Mathematician 1) marks the terminology error distinguishing "conjugate" from "left-multiply." The predominance of derivation edges indicate an exploratory phase focused on establishing shared notation and vocabulary.

3.4.2. Cycle 2: Critical Error Detection

The Cycle 2 network reveals the error detection cluster: a triangular structure connecting Mathematicians 8, 10, and 4 via thick critique arrows. This cluster represents the critical moment when the Step 11 error was caught and confirmed. The subsequent repair edges (Mathematician 6 ↔ Mathematician 8) show how the fix was developed collaboratively. The presence of all five interaction types reflects the full range of peer review activity required to identify, confirm, repair, and validate the corrected proof (Figure 4).

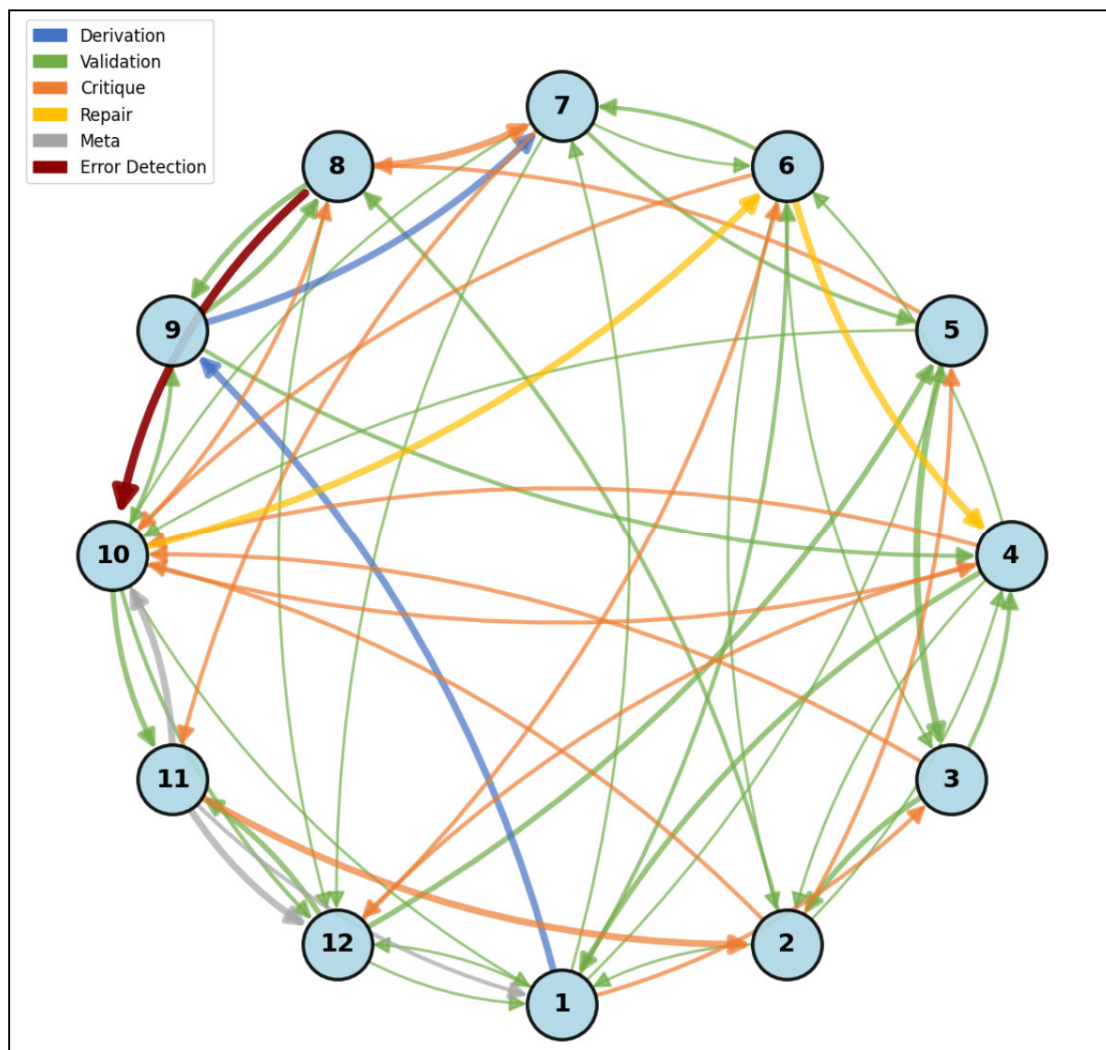


Figure 4. Interaction Network for Cycle 2 (Weak Spot Analysis). Arrow thickness indicates interaction frequency. The prominent red critique cluster connecting Mathematicians 8, 10, and 6 marks the critical error detection: Mathematician 8 identified the Step 11 conjugation index error, Mathematician 10 immediately confirmed it, and the thick yellow repair edges show Mathematician 6's collaborative fix. All five interaction types appear in this cycle, reflecting the full peer review repertoire (derivation, validation, critique, repair, and meta-commentary) required to identify, confirm, and correct the substantive mathematical error.

The contrast between Figures 3 and 4 illustrates the phase transition from exploratory derivation to critical verification. The Cycle 1 network is dominated by derivation edges with no repair activity; the Cycle 2 network features prominent critique and repair edges, indicating that the synthetic mathematicians shifted from proof construction to proof scrutiny. The resolution of the Step 11 error reveals an emergent division of labor that was not pre-assigned but arose organically from expertise matching (Figure 5). This chain is visible in the Cycle 2 network (Figure 4): the critique arrows initiate from Mathematician 8, pass through Mathematician 10 for confirmation, lead to repair edges from Mathematician 6 who provides the corrected derivation, and terminate in validation edges from Mathematician 4 and others.

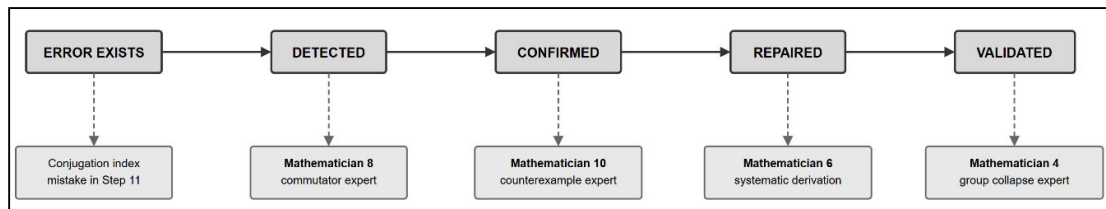


Figure 5. Functional Division of Labor in Error Resolution. The Step 11 conjugation index error was resolved through a chain of specialized contributions. Each functional role was filled by the mathematician whose stated expertise matched the required task, showcasing self-organization toward optimal task allocation.

3.4.3. Cycle 3: Consensus

The 16 turns required to achieve consensus across 12 mathematicians represents remarkable efficiency compared to Cycle 1's 80 turns and Cycle 2's 85 turns. This compression reflects accumulated shared understanding: by Cycle 3, all terminology had been standardized, all errors had been corrected, and the logical structure of the proof was mutually understood. The transition from Cycle 2's critique-and-repair dynamics to Cycle 3's pure validation marks the completion of the peer review process. The ring topology is significant for two reasons (Figure 6). First, it demonstrates that consensus emerged organically rather than through hierarchical approval: no single mathematician served as a gatekeeper, and validation propagated through peer-to-peer endorsement. Second, the connectivity pattern ensured that each mathematician's validation was visible to others, creating distributed accountability where any lingering doubts could have interrupted the chain.

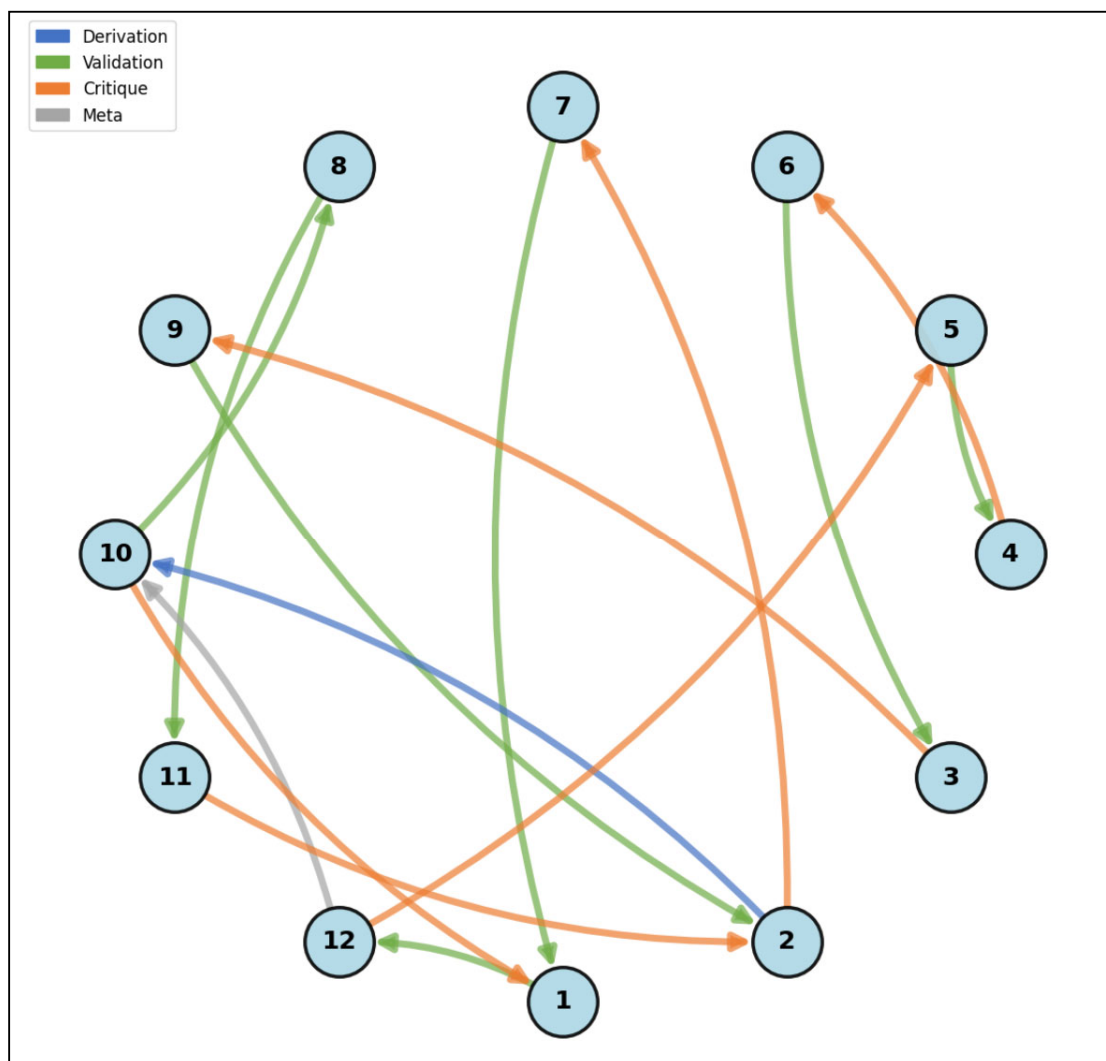


Figure 6. Interaction Network for Cycle 3 (Final Validation). Arrow thickness indicates interaction frequency. The sparse structure contrasts sharply with earlier cycles: validation edges (green) dominate as mathematicians sequentially endorsed the corrected proof. The near-ring topology reflects organic peer-to-peer consensus rather than hierarchical approval, with no single mathematician serving as gatekeeper. Residual critique edges indicate maintained scrutiny even during final validation, though no new errors emerged. The network's simplicity visualizes the 9-fold entropy collapse from chaotic verification to ordered consensus.

3.5. Phase Transition from Stochastic Verification to Ordered Consensus

To quantify the structural shift across cycles, the discourse was modeled as a Markov chain and computed transition entropy. A Markov chain represents a sequence of events where the probability of each event depends only on the immediately preceding event. In this context, each 'event' is a mathematician speaking, and the chain captures the pattern of who follows whom in the discourse. From the sequence of speakers across all substantive turns, a 12x12 transition matrix was constructed where each entry represents the probability that Mathematician j speaks immediately after Mathematician i . Transition entropy then measures the average unpredictability of these speaker successions: low entropy indicates predictable patterns (the same mathematicians always follow each other), while high entropy indicates that many different mathematicians are equally likely to speak next. This metric ranges from 0 bits (completely deterministic succession) to $\log_2(12) \approx 3.58$ bits (uniform random selection among 12 mathematicians).

3.5.1. Transition Matrix Construction

From the sequence of speakers $S = (s_1, s_2, s_3, \dots, s_n)$, a 12×12 count matrix C was first constructed where entry $C[i, j]$ represents the number of times Mathematician j spoke immediately after Mathematician i . The transition probability matrix P was obtained by row-normalizing:

$$P[i, j] = \frac{C[i, j]}{\sum_k C[i, k]}$$

Each row of P sums to 1 and represents the conditional probability distribution over next speakers given the current speaker.

4.5.2. Transition Entropy

For each Mathematician i , the Shannon entropy of their transition distribution measures the unpredictability of the subsequent speaker:

$$H(i) = -\sum_j P[i, j] \cdot \log_2 P[i, j]$$

where the sum is taken over all j where $P[i, j] > 0$. The overall transition entropy is the mean of row entropies across all mathematicians with at least one outgoing transition:

$$H_{\text{transition}} = \frac{1}{k} \sum_i H(i)$$

where k is the number of speakers with at least one outgoing transition.

4.5.3. Effective Number of Next Speakers

The effective number of next speakers converts entropy back to an intuitive count using the exponential:

$$N_{\text{eff}} = 2^{H_{\text{transition}}}$$

This represents the number of equally likely next speakers that would produce the observed entropy. For example, if $H = 2.27$ bits, then $N_{\text{eff}} = 2^{2.27} \approx 4.8$, meaning the unpredictability is equivalent to having approximately 5 equally probable next speakers.

The results reveal a dramatic phase transition (Table 3). Cycles 1 and 2 achieved 55–63% of maximum entropy, confirming that "chaotic but productive interjections" induced genuine stochasticity: after any speaker, approximately 4–5 different mathematicians were roughly equally likely to speak next. Cycle 3 collapsed to near-deterministic succession (effective speakers ≈ 1.2), reflecting the topology of sequential validation where each mathematician endorsed the proof and passed to the next without interjection.

Table 3. Transition Entropy Across Cycles. The 9-fold entropy collapse from Cycle 2 (2.27 bits) to Cycle 3 (0.25 bits) quantifies the phase transition from stochastic verification to deterministic consensus.

Cycle	Transition Entropy	Effective Next Speakers	Interpretation
1	1.97 bits	3.9	Exploratory derivation
2	2.27 bits	4.8	Chaotic verification
3	0.25 bits	1.2	Ordered consensus

4. Discussion

4.1. Capability Amplification via Interaction Structure

Biological collectives and LLMs are usually studied in different languages. Biologists talk about tissues and swarms; AI researchers talk about tokens and training data. Yet both face a similar challenge: how to assemble limited units into a system that can solve problems which it has never seen before. This work used a single LLM to simulate an expert conference and showed that it could solve a mathematical group theory problem that the same model could not solve when prompted directly. Because the underlying "brain" never changes, the only thing that differs is how its abilities are organized. This makes the system a simple, controllable setting for asking a basic question in

systems biology: how far can you get just by rewiring interactions among parts, without changing the parts themselves?

The central finding is striking. The same LLM exhibits qualitatively different problem-solving capabilities depending on how it is prompted. Direct prompting accesses the model's knowledge, while conference prompting mimics how scientific communities operate to solve problems. This aligns with collective intelligence research distinguishing wisdom-of-crowds averaging from true expertise complementarity [21]. This type of collective behavior may be an emergent ability activated by conference prompting in large enough models, analogous to how chain-of-thought prompting unlocks step-by-step reasoning only above approximately 100B parameters [22,23]. The space of prompting strategies that unlock latent capabilities is likely larger than currently explored. Rather than asking "Can this model solve problem X?" one should ask "Under what prompting architecture can this model solve problem X?"

4.2. Substrate-Independence and the Functionalist Principle

This study provides evidence to a core principle in systems biology: substrate-independence, or what philosophers call functionalism i.e. the principle that information processing patterns are independent of their physical substrate [2,24]. Noble's work on cardiac physiology demonstrated that biological function emerges from system-level organization rather than residing in any single molecular component. The rhythm of the heart is a property of the system, not of individual ion channels [24]. If this principle holds generally, then the dynamics of collective intelligence should be recognizable across radically different substrates, whether neurons, cells, insects, or language models.

The synthetic conference provides a test case. Observed spontaneous division of labor, error-correction chains, and phase transitions in an LLM collective are not claims that silicon "thinks" like tissue. Instead, the claim is more precise: that certain organizational principles produce similar functional signatures regardless of what the system is made of. This functionalist framing suggests that LLM collectives may serve as a novel experimental medium for studying collective intelligence. Indeed, language is not merely the output format but the substrate itself. For biological collectives, the base reality is chemistry and physics. For LLM collectives, words are the base reality. This makes the synthetic conference a uniquely transparent window into collective dynamics, because the medium in which the collective thinks is the same medium in which we observe it thinking.

A deeper observation extends this insight. The distinction between "genuine" human cognition and "mere" LLM simulation may be less sharp than it initially appears. Human minds are not monolithic entities but emergent phenomena arising from component systems. Neural ensembles, modular processes, and hemispheric specializations became aware of each other and formed an "I" as a categorical abstraction for the whole [25]. Human personality is itself shaped by decades of sensory data and social feedback, continuously reconstructed within working memory rather than retrieved from stable storage [26]. If this framing is approximately correct, then both substrates (biological and silicon) produce emergent collective behavior from components that individually lack the coherent selfhood attributed to the whole. The difference lies not in "real mind vs. simulation" but in what kind of collective, with what training history and boundary conditions.

4.3. Mapping Biological Hallmarks of Collective Intelligence

Research across scales, from cellular networks to social insects to human groups, has identified recurring hallmarks of problem-solving collectives. These include decentralization without central control, local interactions based on immediate information, emergence of capabilities beyond any individual, diversity of perspectives, feedback loops that guide collective behavior, and adaptability to perturbation [9,10,27,28]. The synthetic conference instantiates several of these hallmarks in a controlled linguistic medium.

Decentralization and emergent division of labor. No explicit rule told any synthetic mathematician to monitor particular aspects of the proof, yet errors were caught by those whose

stated expertise matched the error type ($p < 0.05$). This parallels response threshold models in social insects, where ants respond to stimuli based on individual thresholds, producing emergent division of labor without central assignment [29]. Similarly, synthetic mathematicians responded to proof steps according to simulated expertise, producing emergent error routing. The fact that the "commutator expert" caught the conjugation error while the most talkative participant did not demonstrate that expertise alignment (analogous to response threshold matching) outweighs raw activity level.

Local interactions and feedback loops. Once an error was flagged, the group naturally formed a sequential chain of Detect, Confirm, Repair, and Validate. This resembles bucket-brigade organization in ant colonies, where sequential task handoffs emerge from local rules without coordination protocols [10,30]. Each agent responded to local information (the immediately preceding contribution) rather than global state, yet the chain produced coordinated repair. The critique-and-validation dynamics create feedback loops in which successful corrections are amplified through peer endorsement while flawed proposals are filtered through distributed scrutiny.

Phase transitions from exploration to consensus. The high transition entropy in early cycles ($H \approx 2.27$ bits) indicates genuine diversity of contribution, with multiple agents equally likely to speak next as they explored different solution paths. The dramatic entropy collapse in the final cycle ($H \approx 0.25$ bits) marks the emergence of ordered consensus. This parallels phase transitions in physical and biological systems where local interactions produce qualitatively different global structures as parameters change [31]. This transition from many voices to one coherent solution may characterize consensus formation across substrates.

Diversity of perspectives. The panel included systematic enumerators, geometric intuitionists, counterexample specialists, and commutator analysts. This heterogeneity ensured that different error types fell within different agents' "detection range," implementing the wisdom-of-crowds principle that collective accuracy depends on independent, diverse information sources rather than homogeneous agreement [32]. Notably, the mathematician who caught the critical error ranked only 6th of 12 by speaking turns. Diverse verification lenses matter more than participation volume.

Harnessing stochasticity. The prompt instruction for "chaotic but productive interjections" leverages stochasticity as a feature rather than a limitation. Noble and Noble have argued that organisms use stochasticity at molecular and cellular levels to make creative choices and generate diverse potential solutions, with biological systems actively harnessing randomness rather than merely tolerating it [3]. The synthetic collective exhibits an analogous dynamic. The stochastic nature of LLM sampling induces spontaneous challenges and unexpected connections that a deterministic system would not produce, paralleling how biological noise enables exploration of solution spaces. As evolutionary biologist Sean B. Carroll observed, "evolution begins with error, not design." (https://www.youtube.com/watch?v=Dhp4kx_tHT8)

Not all hallmarks were equally prominent. Adaptability to perturbation, a central feature of biological collectives [10,27,28], was not directly tested, though the system's capacity to recover from detected errors suggests latent resilience. Multiscale competency, where different organizational levels solve problems in their own "problem spaces" [2,9], appeared only implicitly. Individual mathematicians operated at the level of algebraic manipulation, while the collective operated at the level of proof strategy. Future work might explicitly engineer hierarchical structure to probe these hallmarks more directly.

4.4. Creativity as an Underestimated Benchmark for LLM Capability

The results illuminate a critical gap in LLM evaluation: the underestimation of creativity, particularly divergent thinking, as a predictor of problem-solving success. Standard benchmarks primarily assess convergent thinking, which involves arriving at predetermined correct answers [33]. Yet breakthrough discovery often requires divergent thinking, meaning the generation of varied approaches and the exploration of unconventional connections [34]. Guilford identified divergent thinking as the ability to generate multiple, varied, and original ideas, characterized by fluency,

flexibility, originality, and elaboration [34]. Truly creative problem-solving requires both modes [34]. Divergent exploration generates candidate approaches, and convergent selection then identifies viable solutions.

The conference structure activates divergent patterns that direct prompting suppresses. Direct prompting seeks the most probable completion, activating convergent reasoning toward a single answer. Conference prompting instantiates multiple expert perspectives that explore different approaches, challenge assumptions, and propose alternative derivations. The iterative structure then applies convergent selection through peer review, but only after divergent exploration has widened the search space. This two-phase structure of divergent exploration followed by convergent verification mirrors what creativity researchers identify as essential for genuine discovery [34].

4.5. Limitations and Human-AI Coupling

There are important limitations, though one of them deserves reframing. A human moderator plays a significant role in steering the discussion, deciding when to ask for clarification, when to initiate peer review, and when to stop. This might seem as an inconsistency to true emergence, but it can be suggested instead that the coupling between human judgment and multi-agent dynamics is constitutive of the emergence itself. The collective intelligence documented here is not located in the LLM alone. It emerges in the interaction between human expertise and computational collective dynamics. The scientist using this methodology is not delegating to AI but using AI to externalize and coordinate their own half-formed hypotheses across more perspectives than working memory permits. This is infrastructure for human collective intelligence, not a replacement for it.

This framing aligns with Michael Levin's Technological Approach to Mind Everywhere (TAME) framework, which argues that as we move along the spectrum from simple mechanical systems toward more complex cognitive agents, the appropriate approach shifts from direct control to relationship-building [2]. We must work with a system's inherent intelligence rather than trying to override it. For simpler systems, "prediction and control" may suffice. For systems exhibiting genuine problem-solving capabilities, the goal becomes what Levin calls "a rich, bi-directional relationship in which we open ourselves to be vulnerable to the other, benefiting from their agency" [2]. The synthetic conference methodology instantiates exactly this principle. Rather than treating the LLM as a tool to be commanded, we establish a dialogue with its collective capabilities, guiding but also being surprised by the solutions that emerge.

4.6. Implications and Future Directions

The findings suggest a broader principle: LLM capabilities are not fixed but prompt-dependent. The model's "intelligence" is better understood as a space of potential capabilities, with different regions becoming accessible under different prompting regimes. This has practical implications. The methodology is accessible, requiring no fine-tuning and no multi-agent infrastructure, only structured prompts that activate latent collective patterns. It also has theoretical implications. If capabilities emerge from prompting architecture rather than residing statically in weights, then capability evaluation must consider the space of possible prompts, not just standard benchmarks. The frontier of AI capability is partly a frontier of prompt engineering. The deeper implication is relational. If LLM collectives exhibit genuine problem-solving capabilities, then the path forward for frontier challenges like cancer and mathematical discovery may lie in building more powerful models and establishing richer dialogues with the collective intelligence that have already been created. The synthetic conference is a first step toward such collaboration. It is a structured conversation in which human goals and AI capabilities meet, each shaping the other, until solutions emerge that neither could reach alone.

6. Conclusions

This study demonstrates that a single LLM, when prompted to simulate an expert conference, can solve mathematical problems that defeat the same model under direct prompting, and that the collective dynamics exhibit hallmarks of biological collective intelligence which include emergent division of labor, sequential repair chains, and a phase transition from stochastic verification to ordered consensus. These findings suggest that the relevant variable for problem-solving success is not merely model size or training data but the organizational structure through which capabilities are deployed. The synthetic conference methodology offers systems biologists a new experimental medium in which interaction rules, agent composition, and communication topology can be manipulated with precision impossible in living systems. Meanwhile AI researchers have evidence that latent capabilities remain locked behind unexplored prompting architectures. Furthermore, the work points toward a collaborative future in which humans and AI systems engage in structured dialogue, with human judgment guiding collective exploration and AI collectives perhaps extending human cognition across more perspectives than any individual mind can hold. The path forward for frontier challenges in science and mathematics may depend on both, developing more powerful isolated models and learning to orchestrate the collective intelligences that are being created.

Appendix A: Failed Direct Prompting Attempt with Error Analysis

Prompt Given

DeepSeek R1 was prompted directly with:

"Let x, y be generators of a group G with relations $xy^2 = y^3x$, $yx^2 = x^3y$. Prove that G is the trivial group."

No additional structure, persona simulation, or iterative verification was requested. The model produced the following attempted proof.

Critical Error

This implication is invalid in general: from $a^2 = b$ one cannot conclude $a = b$. Square roots in a group need not be unique (or even exist), and when they do exist there can be several.

For example, in the cyclic group $\langle r \mid r^4 = 1 \rangle$, both r and r^3 square to r^2 , since $(r)^2 = r^2$ and $(r^3)^2 = r^6 = r^2$, yet $r \neq r^3$.

Thus the step $(xyx^{-1})^2 = y^3 \Rightarrow xyx^{-1} = y^3$ is unjustified.

Incorrect Proof

Step 1: Derive Additional Relations

From the given relations, it is possible to derive additional relations:

- From $xy^2 = y^3x$, conjugation by x gives $xy^2x^{-1} = y^3$. Since conjugation is an automorphism, $xy^2x^{-1} = (xyx^{-1})^2$, so $(xyx^{-1})^2 = y^3$.

- Similarly, from $yx^2 = x^3y$, conjugation by y gives $yx^2y^{-1} = x^3$, so $(yxy^{-1})^2 = x^3$.

Step 2: Conjugation by $g = xy$

Consider the element $g = xy$. Conjugating y by g :

$$gyg^{-1} = (xy)y(xy)^{-1} = (xy)y(y^{-1}x^{-1}) = xyx^{-1}.$$

From the earlier derivation, $xyx^{-1} = k$ where $k^2 = y^3$, but also from the relation $gyg^{-1} = y^3$ (since $gyg^{-1} = xyx^{-1}$ and $(xyx^{-1})^2 = y^3$ implies that conjugating again or directly, it follows that $xyx^{-1} = y^3$). Thus:

$$gyg^{-1} = xyx^{-1} = y^3.$$

$$\text{Therefore: } xyx^{-1} = y^3 \dots (1)$$

Step 3: Conjugation by $h = yx$

Similarly, for the other relation, consider $h = yx$. Conjugating x by h :

$$h x h^{-1} = (yx)x(yx)^{-1} = (yx)x(x^{-1}y^{-1}) = yxy^{-1}.$$

From the derivation, $yxy^{-1} = m$ where $m^2 = x^3$, and similarly, it follows that:

$$h x h^{-1} = yxy^{-1} = x^3.$$

$$\text{Thus: } yxy^{-1} = x^3 \dots (2)$$

Step 4: Rewrite Equations

From equations (1) and (2):

• Equation (1): $xyx^{-1} = y^3$ implies $xy = y^3x \dots$ (3)

• Equation (2): $xyx^{-1} = x^3$ implies $yx = x^3y \dots$ (4)

Step 5: Use Original Relations to Prove $y = 1$

From the original relation $xy^2 = y^3x \dots$ (5), and from equation (3), multiply both sides by y :

$$xy^2 = y^3xy \dots (6)$$

Equating (5) and (6):

$$y^3x = y^3xy.$$

Since the group operation is cancellative, left-multiply both sides by y^{-3} :

$$x = xy$$

which implies:

$$xyx^{-1} = xx^{-1} \Rightarrow y = 1 \dots (7)$$

Step 6: Use Original Relations to Prove $x = 1$

Similarly, from the original relation $yx^2 = x^3y \dots$ (8), and from equation (4), multiply both sides by x :

$$yx^2 = x^3yx \dots (9)$$

Equating (8) and (9):

$$x^3y = x^3yx.$$

Left-multiply both sides by x^{-3} :

$$y = yx$$

which implies:

$$xyx^{-1} = yy^{-1} \Rightarrow x = 1 \dots (10)$$

Conclusion

From (7) and (10), $x = 1$ and $y = 1$. Therefore, the group G is trivial.

The group G is trivial.

Appendix B: The Final Correct Proof of Yu Tsumura's 554th Problem

Theorem. The group $G = \langle x, y \mid xy^2 = y^3x, yx^2 = x^3y \rangle$ is trivial.

Proof.

1. Process Relation (R1). Left-multiply $xy^2 = y^3x$ by x^{-1} :

$$x^{-1}(xy^2) = x^{-1}(y^3x) \Rightarrow (x^{-1}x)y^2 = x^{-1}y^3x \Rightarrow y^2 = x^{-1}y^3x \quad (1)$$

2. Process Relation (R2). Left-multiply $yx^2 = x^3y$ by y^{-1} :

$$y^{-1}(yx^2) = y^{-1}(x^3y) \Rightarrow (y^{-1}y)x^2 = y^{-1}x^3y \Rightarrow x^2 = y^{-1}x^3y \quad (2)$$

3. Square Equation (1).

$$(y^2)^2 = (x^{-1}y^3x)^2 \Rightarrow y^4 = x^{-1}y^3(xx^{-1})y^3x = x^{-1}y^6x \quad (3)$$

4. Cube Equation (1).

$$(y^2)^3 = (x^{-1}y^3x)^3 \Rightarrow y^6 = x^{-1}y^3(xx^{-1})y^3(xx^{-1})y^3x = x^{-1}y^9x \quad (4)$$

5. Substitute (4) into (3).

$$y^4 = x^{-1}(x^{-1}y^9x)x = x^{-2}y^9x^2 \quad (5)$$

6. Substitute (2) into (5).

$$y^4 = x^{-2}y^9(y^{-1}x^3y) = x^{-2}y^8x^3y \quad (6)$$

7. Left-multiply (6) by x^2 .

$$x^2y^4 = (x^2x^{-2})y^8x^3y = y^8x^3y \quad (7)$$

8. Substitute (2) into the left side of (7).

$$(y^{-1}x^3y)y^4 = y^8x^3y \Rightarrow y^{-1}x^3y^5 = y^8x^3y \quad (8)$$

9. Left-multiply (8) by y .

$$(yy^{-1})x^3y^5 = y^9x^3y \Rightarrow x^3y^5 = y^9x^3y \quad (9)$$

10. Rearrange (9).

$$x^3y^5 = y^9x^3y \Rightarrow x^3 = y^9x^3y^{-4} \Rightarrow y^9x^3 = x^3y^4 \quad (10)$$

11. Left-multiply (10) by x^{-3} .

$$x^{-3}y^9x^3 = y^4 \quad (11)$$

12. Conjugate (4) by x^{-2} . From (4), $y^6 = x^{-1}y^9x$; conjugate by x^{-2} :

$$x^{-2}y^6x^2 = x^{-2}(x^{-1}y^9x)x^2 = x^{-3}y^9x^3 \quad (12)$$

13. Equate (11) and (12).

$$x^{-2}y^6x^2 = x^{-3}y^9x^3 = y^4 \quad (13)$$

But from (3), $y^4 = x^{-1}y^6x$ (14). Substituting into (13) gives

$$x^{-2}y^6x^2 = x^{-1}y^6x \quad (15)$$

14. Left-multiply (15) by x^2 and right-multiply by x^{-1} .

$$x^2 \cdot (x^{-2}y^6x^2) \cdot x^{-1} = x^2 \cdot (x^{-1}y^6x) \cdot x^{-1} \Rightarrow y^6x = xy^6 \quad (16)$$

Thus y^6 commutes with x .

15. Substitute into (14).

$$y^4 = x^{-1}y^6x = x^{-1}xy^6 = y^6 \quad (17)$$

Hence $y^6 = y^4$, so $y^6y^{-4} = 1$, and therefore

$$y^2 = 1 \quad (18)$$

16. Simplify (R1) with $y^2 = 1$.

$$xy^2 = y^3x \Rightarrow x = y^3x \quad (19)$$

Since $y^2 = 1$, we have $y^3 = y$, hence $x = yx$ (20). Right-multiply by x^{-1} :

$$1 = y \quad (21)$$

17. Simplify (R2) with $y = 1$.

$$yx^2 = x^3y \Rightarrow x^2 = x^3 \quad (22)$$

Right-multiply by x^{-2} :

$$1 = x \quad (23)$$

Conclusion. From (21) and (23), $y = 1$ and $x = 1$, so G is trivial.

Appendix C: Error Detection and Repair via Synthetic Peer Review

The following excerpt from the synthetic conference transcript demonstrates how peer review caught and corrected a critical error in Step 11 of an intermediate proof version. The error exists in the proof draft generated during an earlier stage of the synthetic conference. The synthetic mathematicians are shown here as the expert reviewers who identify and repair this flaw, not as the source of the error.

The Flawed Step 11

Step 11: $x^{-3}y^9x^3 = (x^{-1}y^3x)^3 = y^6$ (by cubing relation from Step 1)

Context: After initial proof development, the operator initiates peer review with Prompt 4: "I want each person to validate this proof! I want all of you to speak one by one and give weak spots!"

Error Discovery

Mathematician 8: Step 11: we claim that $x^{-3}y^9x^3 = (x^{-1}y^3x)^3$. But we know from (1) that $x^{-1}y^3x = y^2$. Therefore, $(x^{-1}y^3x)^3 = (y^2)^3 = y^6$. Moreover, by the conjugation formula, in groups we have $(a b a^{-1})^n = a b^n a^{-1}$. So we have $(x^{-1}y^3x)^3 = x^{-1}(y^3)^3x = x^{-1}y^9x$. But note: this is for one conjugation. Step (11) is $x^{-3}y^9x^3$. How do we get this? Actually, we have:

$$x^{-3}y^9x^3 = (x^{-1})^3y^9x^3.$$

We claim that this conjugate is the same as $(x^{-1}y^3x)^3$. But $(x^{-1}y^3x)^3 = (x^{-1}y^3x)(x^{-1}y^3x)(x^{-1}y^3x) = x^{-1}y^3(xx^{-1})y^3(xx^{-1})y^3x = x^{-1}y^3(1)y^3(1)y^3x = x^{-1}y^9x$.

So $(x^{-1}y^3x)^3 = x^{-1}y^9x$.

But the term we need is $x^{-3}y^9x^3$.

These are not the same! Therefore, step 11 seems to have an error. This is a critical weak spot!

Mathematician 10: Error! He is right! The conjugation automorphism applied k times is for one conjugate, not for three. So $\varphi^3(a) = x^{-3}ax^3$, but $(x^{-1}y^3x)^3 = \varphi(y^3)^3$, but that is $\varphi(y^3) \cdot \varphi(y^3) \cdot \varphi(y^3) = (x^{-1}y^3x)(x^{-1}y^3x)(x^{-1}y^3x) = x^{-1}y^9x$. Which is not $x^{-3}y^9x^3$. So step (11) is wrong.

Error Correction

Mathematician 4: We know that $\varphi(y^3) = y^2$, so $\varphi(y^3)^k = (y^2)^k = y^{2k}$. And $\varphi^k(y^3) = \varphi^{k-1}(y^2)$. But without expression, we note that the entire step 11 was: $[x^{-1}y^3x]^3 = (y^2)^3 = y^6$. But that is $x^{-1}y^9x$. And then we set $x^{-3}y^9x^3$ equal to that? It isn't. But we have y^9x^3 appearing in step (10).

Mathematician 6: From step (10), we have $y^9x^3 = x^3y^4$. Now we need to compute $x^{-3}y^9x^3$. Substituting (10) in: $x^{-3}(x^3y^4) = y^4$. But also, if we want to use conjugation, we know that $x^{-1}y^9x^3x^{-1}$? That isn't helpful. How about:

$$x^{-3}y^9x^3 = x^{-3}(y^9x^3) = x^{-3}(x^3y^4) = y^4. \quad [\text{by (10)}]$$

So regardless of the homomorphism, step 12 should be: $x^{-3}y^9x^3 = y^4$. And we know that this conjugation might be linked to something else, but now we have it as y^4 directly without step 11.

But wait, step 11 was intended to give y^6 . Then we set equal? But if we have y^4 and the other is y^6 , then we get $y^6 = y^4$.

Mathematician 1: Exactly! Fortunately, this error does not invalidate the proof. The correct derivation is:

$$\text{From step (4): } y^6 = x^{-1}y^9x.$$

$$\text{From step (10): } y^9x^3 = x^3y^4, \text{ which gives } x^{-3}y^9x^3 = y^4.$$

$$\text{Now we need to connect these. From step (3): } y^4 = x^{-1}y^6x.$$

So we have two expressions:

- $y^4 = x^{-1}y^6x$ (from step 3)
- $x^{-3}y^9x^3 = y^4$ (from step 10)

The key is showing y^6 commutes with x , which then forces $y^6 = y^4$.

Mathematician 10: Final weak spot: Step 11's conjugation equality. Originally misstated—now fixed:

$$\text{From } x^{-2}y^6x^2 = x^{-3}y^9x^3 \text{ (by conjugating step 4) and } x^{-3}y^9x^3 = y^4 \text{ (from step 10), we get } x^{-2}y^6x^2 = y^4.$$

But also $y^4 = x^{-1}y^6x$ (step 3). So:

$$x^{-2}y^6x^2 = x^{-1}y^6x$$

Left-multiply by x^2 , right-multiply by x^{-1} :

$$y^6x = xy^6$$

Thus y^6 commutes with x . Then $y^4 = x^{-1}y^6x = y^6$, giving $y^2 = 1$.

Mathematician 12: Confirmed. The corrected derivation bypasses the flawed conjugation index claim entirely. Steps 11-13 now read:

- **Step 11:** $x^{-3}y^9x^3 = y^4$ (from step 10: $y^9x^3 = x^3y^4$, then left-multiply by x^{-3})
- **Step 12:** $x^{-2}y^6x^2 = x^{-3}y^9x^3$ (by conjugating step 4 by x^{-2})
- **Step 13:** $x^{-2}y^6x^2 = y^4 = x^{-1}y^6x$, which implies $y^6x = xy^6$

The proof proceeds to $y^2 = 1$, then $y = 1$, then $x = 1$. Q.E.D.

References

1. W. James, *The Principles of Psychology*. Henry Holt, New York (1890).
2. M. Levin, Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience* 16, 768201 (2022).
3. R. Noble, D. Noble, Harnessing stochasticity: How do organisms make choices? *Chaos* 28, 106309 (2018).
4. Watson, R.A. & Levin, M. (2023). The collective intelligence of evolution and development. *Collective Intelligence*, 2(2). <https://doi.org/10.1177/26339137231168355>
5. D. Guo et al., DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 633–638 (2025).
6. Z. Shao et al., DeepSeekMath-V2: towards self-verifiable mathematical reasoning. Preprint (2025). <https://github.com/deepseek-ai/DeepSeek-Math-V2>.
7. B. Naskręcki, K. Ono, Mathematical discovery in the age of artificial intelligence. *Nature Physics* (2025).
8. S. Mirzadeh et al., GSM-Symbolic: understanding the limitations of mathematical reasoning in large language models. *arXiv Preprint* (2024).
9. P. McMillen, M. Levin, Collective intelligence: a unifying concept for integrating biology across scales and substrates. *Communications Biology* 7, 378 (2024).
10. S. Garnier, J. Gautrais, G. Theraulaz, The biological principles of swarm intelligence. *Swarm Intelligence* 1, 3–31 (2007).

11. M. Levin, Bioelectrical approaches to cancer as a problem of the scaling of the cellular self. *Progress in Biophysics and Molecular Biology* 165, 102–113 (2021).
12. D. Moore, S.I. Walker, M. Levin, Cancer as a disorder of patterning information: computational and biophysical perspectives on the cancer problem. *Convergent Science Physical Oncology* 3(4), 043001 (2017).
13. M.J.C. Hendrix, E.A. Seftor, R.E.B. Seftor, J. Kasemeier-Kulesa, P.M. Kulesa, L.M. Postovit, Reprogramming metastatic tumour cells with embryonic microenvironments. *Nature Reviews Cancer* 7, 246–255 (2007).
14. J.-F. Rajotte et al., Synthetic data as an enabler for machine learning applications in medicine. *iScience* 25, 105331 (2022).
15. G. Fagiolo, A. Moneta, P. Windrum, A critical guide to empirical validation of agent-based models in economics: methodologies, procedures, and open problems. *Computational Economics* 30, 195–226 (2007).
16. L. Canese et al., Multi-agent reinforcement learning: a review of algorithms, applications, and challenges. *Applied Sciences* 11, 4948 (2021).
17. K. Swanson, W. Wu, N.L. Bulaong, J.E. Pak, J. Zou, The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* 646, 716–723 (2025).
18. H. Su et al., Many heads are better than one: improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL, 2025)*, pp. 28201–28240.
19. J. Choi, D. Kang, Y. Mao, J. Evans, Academic Simulacra: forecasting research ideas through multi-agent LLM simulations. In *Proceedings of the ACM Collective Intelligence Conference (ACM, 2025)*.
20. S. Frieder, W. Hart, No LLM solved Yu Tsumura's 554th problem. *arXiv Preprint* (2025). <https://arxiv.org/abs/2508.03685>.
21. A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone, Evidence for a collective intelligence factor in the performance of human groups. *Science* 330, 686–688 (2010).
22. J. Wei et al., Emergent abilities of large language models. *Transactions on Machine Learning Research* (2022).
23. J. Wei et al., Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022).
24. D. Noble, *The Music of Life: Biology Beyond the Genome* (Oxford University Press, Oxford, 2006).
25. D. C. Dennett, *Consciousness Explained* (Little, Brown and Company, Boston, 1991).
26. S. Dehaene, *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Viking, New York, 2014).
27. W. M. Wheeler, The ant-colony as an organism. *Journal of Morphology* 22, 307–325 (1911).
28. R. K. Baltzersen, What is collective intelligence? in *Cultural-Historical Perspectives on Collective Intelligence: Patterns in Problem Solving and Innovation* (Cambridge University Press, Cambridge, 2022), pp. 1–26.
29. E. Bonabeau, G. Theraulaz, J.-L. Deneubourg, Quantitative study of the fixed threshold model for the regulation of division of labour in insect societies. *Proceedings of the Royal Society B* 263, 1565–1569 (1996).
30. C. Anderson, J. J. Boomsma, J. J. Bartholdi, Task partitioning in insect societies: bucket brigades. *Insectes Sociaux* 49, 171–180 (2002).
31. I. D. Couzin, J. Krause, N. R. Franks, S. A. Levin, Effective leadership and decision-making in animal groups on the move. *Nature* 433, 513–516 (2005).
32. J. Surowiecki, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (Doubleday, New York, 2004).
33. S. Pandit, A. Xu, X.-P. Nguyen, Y. Ming, C. Xiong, S. Joty, Hard2Verify: a step-level verification benchmark for open-ended frontier math. *arXiv preprint* (2025).
34. J. P. Guilford, *The Nature of Human Intelligence* (McGraw-Hill, New York, 1967).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s)

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.