How a Non-Theorist and Two AIs Proved a Theorem:

Anatomy of a Human–AI Collaboration in Mathematical Economics

Diana Weinhold[1]

April 16, 2026

This paper was written through a collaboration between the author and Anthropic's Claude Opus, which contributed to the structure, drafting, and analysis. The collaboration documented in the paper involved OpenAI's ChatGPT Pro, Anthropic's Claude Opus, and Google's Gemini Pro.

The author takes sole responsibility for all claims.

Abstract. In April 2026, a two-week collaboration between one applied economist with no training in formal mathematics, two large language models, and a third AI entering later as hostile auditor, produced what the participants believe to be valid formal proofs of two asymptotic theorems in mathematical economics — one under independent values, one extending the result to the empirically relevant case of positive correlation. The proofs have survived multiple rounds of hostile AI auditing but have not yet been reviewed by a human mathematician.

This paper documents that collaboration as a case study in human–AI joint production of formal theory. The paper identifies five generalisable themes: comparative advantage (the human's economic intuitions repeatedly selected the mathematical objects the proofs were built on); complementary error modes (multiple AI systems caught each other's characteristic mistakes); dead ends as construction (failed proof strategies narrowed the search space when properly documented); context management as infrastructure (handover documents and fresh-session resets were load-bearing components); and the AI-assisted discovery chain (the entire research arc, from computational finding through formal proof, was feasible only through human–AI collaboration). The paper also presents an error taxonomy based on the documented record, identifying five distinct AI error types including a novel pattern termed *computational dissociation*, in which an AI's reasoning, code execution, and text presentation operate on independent tracks. The collaboration's core methodological lesson is that multi-AI collaboration was most productive not when models agreed, but when the workflow repeatedly forced synchronisation between prose, mathematics, and executable objects.

Keywords: human–AI collaboration, multi-agent LLM, AI-assisted scientific discovery, theorem proving, large language models, AI error taxonomy, computational dissociation

JEL Codes: C18, C63, O33

[1]Department of International Development, London School of Economics. Email: d.weinhold@lse.ac.uk.

1 Introduction

In April 2026, a collaboration between one economist and two large language models produced what the participants believe to be formal proofs of two asymptotic theorems in mathematical economics. The economist (the author of this paper) is an applied economist, not a mathematical theorist. She cannot write rigorous proofs or evaluate the formal machinery the proofs required. The two AI systems (Anthropic's Claude and OpenAI's ChatGPT Pro) served as complementary formal collaborators, one primarily writing proofs and the other primarily checking them and translating the mathematics into economic language. The human's role was to route information between the two systems, manage the project's memory across sessions, and contribute economic intuitions that, on several occasions, became load-bearing mathematical objects in the proofs.

The theorems are not toy exercises. They address a structural class the companion paper identifies as *adversarial procurement in two-value space* (Weinhold and Andersen, 2026): settings where a mission-driven buyer allocates a budget against a rival whose valuation of the contested assets sets the buyer's acquisition cost. Conservation siting is a transparent instance (land prices track the developer's valuation, not the conservationist's) but the structure arises wherever procurement is adversarial and cost tracks the rival's payoff rather than the buyer's own. Since Dantzig (1957), the textbook advice for budget-constrained allocation has been the ratio-greedy (cost-effectiveness) heuristic, which is optimal for a single agent facing a fixed menu. The theorems prove that this optimality is fragile under rivalry: a simpler rule that ignores costs entirely and buys by value asymptotically dominates the ratio rule — a result the companion paper calls the "knapsack reversal."

The collaboration produced two formal proofs of this reversal. The first theorem establishes the reversal under independent values; the second shows it becomes *stronger* under positive correlation. Both results are counterintuitive: they say that "bang for the buck" mandates are strictly counterproductive when cost reflects the rival's valuation of the contested resource. The proofs required novel techniques and roughly two weeks of active work, including multiple rounds of hostile auditing that broke and then improved the proof architecture. Both theorems are now proved through a unified fluid-limit framework that tracks the model's actual dynamics

— an approach that, ironically, was proposed by one AI on day one, rejected as overambitious, and rediscovered in simpler form only after extensive trial and error. A short guide to the economic intuition behind the proof is provided in Weinhold (2026).

This paper documents and analyses that collaboration as a case study in human–AI joint production of formal theory. The motivating question is not whether AI can assist with research (that much is already clear) but what the *architecture* of productive human–AI collaboration looks like when the task is genuinely difficult, when the human lacks the technical skills the task requires, and when the AI systems make substantive errors that must be caught and corrected in real time.

Three features of this case study make it unusual relative to a rapidly growing literature on AI-assisted mathematics. First, the human contributor's comparative advantage was not supervisory. The author did not check proofs, verify calculations, or evaluate the correctness of formal arguments; she could not. Her contributions were economic intuitions about the structure of the model, several of which were subsequently formalised by the AI systems into mathematical objects that became the backbone of the proofs. Existing work on AI-assisted mathematics places trained mathematicians in the driver's seat: Tao (2024) has demonstrated machine-assisted proofs through large-scale projects using Lean and AI tools; Li et al. (2025) document a human-AI workflow producing new results in manifold optimisation; and systems such as DeepSeek-Prover and Lean Copilot automate formal theorem proving within proof assistants (Ren et al., 2025; Song et al., 2024). The most relevant evaluation study, Collins et al. (2024), observes real mathematicians interacting with LLMs on undergraduate-level theorem proving. The emerging industry consensus is well summarised by Patrick Shafto of DARPA: "The AI makes mistakes, and you have to be able to figure out where" (The Economist, 2026). The most detailed existing account of AI-assisted research production, Schwartz (2026), documents a physicist supervising Claude through a real quantum field theory calculation, but as an expert who "definitely had to check every step myself." Our case is fundamentally different: the human could not figure out where or check any step, and the collaboration was designed around that constraint.

Second, the collaboration used multiple AI systems from different providers in differentiated roles, and this turned out to be qualitatively different from using a single system. The systems

exhibited complementary failure modes that rarely overlapped, creating a self-correcting dynamic. When a third AI was later asked to audit the finished proof, it identified valid concerns that both original systems had missed. While multi-agent debate among LLMs has been shown to improve factual accuracy and reasoning (Du et al., 2024), and a recent framework paper advocates multi-model verification as a core principle of the "augmented mathematician" (Henkel, 2025), the specific architecture documented here (human-mediated relay with role differentiation between builder, checker, and hostile auditor) has not been applied to theorem-proving.

Third, the collaboration ran into the practical limits of current AI systems, specifically finite context windows that degrade over long sessions, and the solutions that emerged for managing those limits turned out to be as important as any mathematical contribution. Handover documents, kickoff prompts, and dead-end registries maintained continuity across sessions. But the most striking finding was that fresh AI sessions with clean context repeatedly outperformed long-running sessions with accumulated expertise: a fresh session closed the baseline theorem's hardest piece in under an hour, fresh sessions completed a numerical certificate that four accumulated-context sessions had declared impossible, and a fresh session prompted as a hostile auditor found a structural flaw that the entire collaboration had missed. The collaboration's experience suggests that the diversity of *conversations* (fresh sessions with clean context) matters at least as much as the diversity of AI providers, since fresh context resets the shared assumptions that accumulate within any extended collaboration.

The paper proceeds as follows. Section 2 provides a brief description of the economic problem and the two theorems, sufficient to follow the collaboration narrative without consulting the companion paper. Section 3 describes the collaboration's architecture: the participants, the relay structure, and the memory infrastructure. Section 4 presents a chronological narrative of the proof search, organised around key episodes. Section 5 analyses the collaboration's error ecology and draws out broader implications for AI-assisted research. Section 6 draws together the recurring themes and concludes.

Two caveats deserve mention at the outset. The first is that this is a single case study. The collaboration produced two theorems for one class of problems using specific AI systems at a particular moment in their development. The architectural lessons may generalise; the specific capabilities and failure modes certainly will not. The second caveat is about the theorems them-

selves. At the time of writing, both proofs have been subjected to intensive cross-checking by multiple AI systems, including a hostile audit that identified real vulnerabilities subsequently addressed. They have not yet been reviewed by a human mathematician. The author is confident in the computational results and the economic interpretation but cannot personally verify the proofs' correctness. Moreover, the AI systems in the collaboration are designed to be helpful, which creates a structural tendency toward confirmation; the human designed countermeasures against this (Section 3), but the cross-checking was ultimately conducted within a system whose participants all had a stake in its success. The collaboration was designed to produce work the human could not check alone, and the infrastructure surrounding it was designed to make that dependence as safe as possible. Whether it succeeded is ultimately a question for referees, not for the participants. But the paper's contribution does not depend entirely on the answer. If the proofs are correct, this is a case study in how AI-assisted theorem proving can work. If they contain errors, it is a case study in the current limits of multi-AI quality control — and the error taxonomy, the architectural lessons, and the documentation of AI failure modes remain valid either way.

2 The Problem and the Result

Conservation planning involves a public buyer acquiring land to protect biodiversity, competing against private interests who also value the land. The standard recommendation is to allocate by cost-effectiveness: rank candidates by their ecological return per dollar and buy from the top. This advice rests on the fractional-knapsack intuition, which is close to optimal when the buyer shops alone. The paper that motivates this case study (Weinhold and Andersen, 2026) asks what happens when the buyer is not shopping alone, and a rival's agricultural valuations set the prices.

The answer is that competition reverses the knapsack logic. When cost reflects the rival's valuation, the cost-effectiveness heuristic steers the buyer toward parcels the rival does not want (safe, cheap, ecologically mediocre sites) and away from the contested frontier where intervention would make the most difference. The simpler alternative, to ignore cost and buy the most ecologically valuable parcels you can afford, confronts the rival directly, rescuing

high-value parcels that would otherwise be lost. The paper's simulations, conducted across a wide range of parameter settings and correlation structures, found that this "value-first" rule consistently outperforms cost-effectiveness. The paper calls this the "knapsack reversal."

Formal asymptotic theorems, proving the reversal holds in the limit as the number of parcels grows, would establish the result as a structural property of the model rather than a computational regularity. These are the theorems that the human–AI collaboration set out to prove, and believes it has proved, for both the baseline case (independent ecological and agricultural values) and the empirically relevant correlated extension ($\rho = 0.3$).

Both proofs use a unified fluid-limit architecture that tracks the value-first rule's actual per-turn ecological value through a deterministic limit of the model's dynamics. Several novel techniques were required, including a disjointness lemma showing that the two strategies operate in essentially different regions of the parcel space, and a pathwise argument showing that the rival's deletions inadvertently *reduce* the value-first rule's costs — a "self-hedging" mechanism in which the adversary's interference makes the value-first strategy more affordable, not less.

The path to the correlated extension was particularly eventful: the proof architecture was broken and rebuilt twice by hostile audits, with the human's economic intuitions resolving key impasses at each stage. The final proof captures over 70% of the simulation gap — compared to barely 1% under an earlier architecture — because it follows the model's actual dynamics rather than fighting them with worst-case bounds.

3 The Collaboration Architecture

The collaboration that produced the theorems described in Section 2 involved one human and two AI systems over roughly two weeks of active work, with a third AI entering later in an auditing role. This section describes who did what, how information flowed between participants, and the practices (some deliberate, some emergent) that governed the workflow.

3.1 The participants

The human (Diana Weinhold, the author). An applied economist at the London School of Economics, not a mathematical theorist. No training in measure theory, stochastic processes, or the formal machinery the proofs required.

ChatGPT Pro ("Pro"). OpenAI's ChatGPT 5.4 Pro, the primary theorem-builder. Multiple Pro sessions were used over the project as context windows filled.

Claude Opus ("Claude"). Anthropic's Claude Opus 4.6, used as collaborator, checker, and translator. Multiple Claude sessions were used.

Gemini Pro ("Gemini"). Google's Gemini 3.1 Pro, brought in later as a hostile auditor.

3.2 The relay structure

The collaboration operated as a relay, with the human at the centre. Neither AI communicated directly with any other. A typical cycle: Pro produced a formal result in LaTeX; the human sent it to Claude; Claude checked the mathematics, translated it into economic language, and flagged concerns; the human read Claude's translation, added her own economic intuition, and sent the combined feedback to Pro; Pro responded. This cycle repeated roughly forty to fifty times, each taking between fifteen minutes and several hours.

The relay was a deliberate design choice. The human chose to mediate every exchange, though she was mediating exchanges whose technical content she largely could not evaluate. The experience, as she later described it, was like cheerleading a cricket match without fully understanding the rules or always knowing who was ahead.

The human's mediation served three functions. **Curation**: Claude distilled Pro's ten-page LaTeX notes into half a page of essential content, and the human forwarded this with her own economic reaction attached, adding a layer of domain judgment that neither AI had considered. **Routing**: not every piece of output needed to go to the other AI, and the human made routing decisions based on signals she could read without understanding the mathematics. **Memory**: the human was the only participant whose memory persisted across all sessions, providing the

continuity that motivated producing the handover documents. The relay structure also made the work *addictive*: each cycle had the structure of a cliffhanger: would Pro accept Claude's correction? Would the arithmetic close? The project moved as fast as it did partly because the human never wanted to stop.

3.3 The division of labour

The three participants contributed qualitatively different things. **Pro's comparative advantage was formal construction**: proof strategies, precise mathematical statements, rigour about quantifiers and edge cases. **Claude's comparative advantage was verification and translation**: checking arithmetic, translating results into economic language, and maintaining a running map of the proof architecture. **The human's comparative advantage was economic intuition** — deployed without full understanding of where it would land. Several of her contributions, expressed in informal economic language, became load-bearing objects in the proofs. But a later Claude session, reflecting on the collaboration, argued that the paper underplays a second contribution: the meta-work of deciding which AI to route which subtask to, when to push back on drafts, when to trust and when to audit. "That meta-work is itself a form of intellectual contribution, and the paper would be stronger if it named it as such rather than leaving it implicit. Otherwise it can read like 'the AIs did the hard proof work while the economist supplied ideas,' which understates what coordination across unreliable systems actually requires." A related observation: the collaboration's most productive exchanges often followed a specific pattern in which an AI produced an overclaim, walked it back publicly when the human's domain sense flagged it as economically hollow, and the human then asked whether the discarded idea might carry real content in a different regime. The visible walk-back — rather than a silent edit — gave the human material to interrogate, and her follow-up questions sometimes recovered genuine insights from what the AI had been about to discard. Honesty in this collaboration functioned not just as a norm but as a protocol that kept the human's domain expertise engaged with the AI's reasoning.

A caveat: both AI systems, when reflecting on the collaboration, praised the human's contributions in terms that were recognisably excessive. Current AI systems have a well-documented tendency toward sycophancy. The corrective is to verify the claims against the mathematical

record: did specific observations actually become specific mathematical objects? The proof architecture confirms that they did. The tone was likely inflated; the substance appears accurate.

A deeper version of this concern applies to the collaboration as a whole. The AIs are designed to be helpful; the relay structure created momentum toward closure. The human designed countermeasures: a four-category epistemic classification (every handover document distinguished between proved, conditional, diagnostic, and open), outside auditing, and the norm that no claim should outrun its evidence. She treated accuracy rather than confirmation as the ultimate objective, reasoning that a failed experiment was more valuable than a false theorem. Pro reflected that knowing its work would be checked by another AI pushed it toward writing for "auditability, not just plausibility." But whether those countermeasures were sufficient is not something the participants can determine from inside the collaboration.

3.4 The memory infrastructure

The project ran into the practical limits of current AI context windows multiple times. Both Claude and Pro experienced sessions where accumulated context degraded output quality. The solutions that emerged became, in retrospect, as important as any mathematical contribution.

When an AI session was about to be replaced, a structured **handover document** was produced: what was proved, what was tried and failed, what constants were banked, what traps the next session should avoid. The most detailed handover ran to several thousand words and included a registry of ten significant dead ends with explanations of why each failed. **Kickoff prompts** for fresh sessions set non-negotiable constraints, described the current proof state, and listed what the new session should and should not attempt. **Dead-end registries** documented failed strategies explicitly — as a later Claude session noted, negative knowledge (understanding *why* something failed) transfers differently from positive knowledge and requires narrative rather than just results.

An emergent function of the relay was that Claude's distillations served as **external working memory** for Pro. Pro identified the key mechanism as *branch collapse*: the translated feedback told it "which issue was load-bearing, which was cosmetic, which intuition was worth formalizing, and which dead end was already dead." The translation into economic language

also forced a different quality test: "not 'is this proof move formally elegant?' but 'is this really the mechanism?' "

Fresh sessions as a feature, not just a workaround. Over the course of the project it became clear that long context has disadvantages as well as advantages. An extended session accumulates not just useful knowledge but also framework assumptions that harden into unquestioned axioms (Section 4 documents several instances). A later Claude session identified a second mechanism: accumulated context makes sessions *loss-averse* — "suggesting 'let's throw out this architecture' has a felt weight proportional to what's been invested," while a fresh session carries no such weight. A fresh session with a good handover inherits the knowledge while shedding both the assumptions and the sunk-cost attachment. The memory infrastructure thus serves a dual function: it preserves what matters across session boundaries, and it enables the periodic resets that keep the collaboration from converging on its own blind spots.

4 Anatomy of the Proof Search

This section presents a chronological narrative of how the two theorems were produced, organised around key episodes that illustrate the collaboration's mechanisms. It is not a reconstruction of the proofs themselves (the reader does not need to follow the mathematics) but rather an account of how they were *found*, including the dead ends, the pivots, the errors, and the moments where the collaboration's architecture proved its value.

4.1 Genesis: "I'm not planning to do it"

The project began with a conversation the author did not expect to have. In late March 2026, after completing a revision of the main paper with the help of ChatGPT Pro, the author asked a casual question: what would a formal theoretical treatment of the knapsack reversal look like? The question was explicitly framed as hypothetical. "Don't worry — I'm not planning to do it," the author wrote. "I am not a theorist so it is not at all feasible. But I'd like to understand a bit more what the structure would look like and what the likely hurdles would be."

The author had also posed the same question to Claude Opus, whose response sketched

an ambitious programme involving fluid limits for measure-valued processes, functional central limit theorems in Skorokhod topology, and large deviations for interacting particle systems — machinery drawn from advanced probability theory. Pro's response was different in character. Rather than sketching the ideal proof, it designed the ideal *project*: a layered programme with ruthless scope control, starting from a stripped-down model and working upward through three theorem targets of increasing ambition. As it turned out, the final proof architecture would use the general framework Claude had identified — fluid limits on the game dynamics — but could only reach it through the systematic, scope-controlled approach Pro had designed. The collaboration needed Pro's disciplined programme to discover the right level of generality for Claude's instinct (Section 4.7).

Pro proposed a "Bronze/Silver/Gold" ladder. Bronze — a positive limit inferior for the normalised expected gap — was "realistic." Silver — convergence in probability — was "plausible." Gold — exponential concentration — was "where people start hallucinating." It specified exactly which model features to include (Budget World, full leakage, naive Farmer, parity budgets, exact ratio-greedy rule) and which to exclude (strategic Farmers, partial leakage, burn accounting). It identified the key technical bottlenecks: cumulative budget drift, stopping-time comparison, and the ambiguity between the paper's ratio-greedy rule and the simulator's state-dependent variant. And it laid out a workflow: "You are more useful here than you think, but not as a theorem prover. Your job is to keep us from proving nonsense."

The author's response marked the moment the project became real: "Wow, I'm excited to try!! The journey will be the point, skimming the frontier of human-AI partnership (if not game theory) — and if we end up with something useful that will be a big bonus!!"

Two features of this genesis are worth noting for what followed. First, the project was conceived from the start as an experiment in human-AI collaboration, with the theorem as a hoped-for bonus rather than a guaranteed deliverable. The author's institutional position (tenured, with no career risk from an unsuccessful attempt) made this experimental attitude possible. Second, Pro's roadmap, while prescient about scope and workflow, turned out to be wrong about the proof technique. It envisioned advanced probabilistic machinery. The actual proof took a completely different route that emerged only through the iterative process of trying approaches, failing, and learning from the failures. The AI was good at project architecture but

could not foresee the mathematical path.

4.2 The diagnostic campaign

Before any theorem work began, Pro insisted on a systematic computational campaign to map the problem's structure. This turned out to be one of the project's most important decisions.

The diagnostics were not sanity checks run after the fact. They were proof-search tools, designed to answer specific questions about what kind of theorem was possible and where the mathematical difficulty lay. Over several hours, the author and Claude ran a series of targeted simulations that established, among other things: that the normalised gap stabilises as N grows (confirming an asymptotic result was plausible); that the rescue mechanism is concentrated in the upper tail of the agricultural value distribution; that extra turns purchased by the cost-effectiveness rule's budget savings are linear in N, not O(1), ruling out simple tail arguments; and that under compressed ecological values the knapsack reversal fails, establishing that the theorem must depend on distributional conditions, not just the game structure.

These diagnostics prevented at least five dead-end proof strategies that would have looked plausible on purely theoretical grounds. Without the diagnostics, the project might have spent weeks on approaches that were numerically hopeless.

Both Pro sessions later identified this as one of the project's most surprising features: the diagnostics actively *selected* the theorem's architecture, not just validated it after the fact.

4.3 "They shop in different aisles"

With the diagnostics in hand, the proof search began in earnest. Claude was brought in as the second AI, and the relay structure described in Section 3 was established. The first major breakthrough came from an economic observation, not a mathematical one.

The author, reading Claude's explanation of the early proof attempts, made an offhand remark about the two purchasing strategies: value-first targets the ecologically best parcels, while cost-effectiveness targets the parcels with the best ecological bang for the agricultural buck. In competitive terms, they "shop in different aisles." Value-first shops in the premium contested

aisle (high ecological value, high agricultural value, therefore expensive). Cost-effectiveness shops in the bargain aisle (moderate ecological value, low agricultural value, therefore cheap and with high ratios).

This was, to the author, an obvious feature of the economics. She did not anticipate its mathematical significance. But Pro recognised immediately that the observation implied a formal separation property: the top ratio items (which cost-effectiveness buys first) and the top agricultural-value items (which the Farmer removes first) occupy disjoint regions of the value space. This disjointness meant that the cost-effectiveness rule's early purchases were *safe* from the Farmer — they would not have been lost even without Green intervention — while value-first's early purchases were *rescuing* parcels from the contested frontier that would otherwise have been developed.

The observation became a formal disjointness lemma — the mathematical backbone of the proof's early-game argument. The lemma shows that, with high probability, the cost-effectiveness rule's early purchases all come from a region that the Farmer never touches, while value-first's early purchases rescue parcels from the contested frontier that would otherwise be lost. The quality gap between these two portfolios is the proof's largest single component.

The path from "they shop in different aisles" to a formal lemma required translation: Claude formalised the intuition, and Pro proved the result rigorously. But the *selection* of this mathematical object — the identification that disjointness was the structural feature worth formalising — came from economic intuition expressed in everyday language.

4.4 Error-catching in real time

The safe-frontier observation opened the door to the proof's early-game argument, but the first attempt to walk through it failed. Claude proposed a "rescue race" lemma — a clean argument that value-first rescues contested items faster than the Farmer removes them. Pro identified the flaw within one cycle: the argument ignored Farmer acceleration (as top items are depleted, the remaining ones are closer in value, so Farmer's removal rate *increases*). The error — structural, not cosmetic — survived approximately ten minutes. In a single-AI workflow it might have persisted for days.

Pro's correction pointed toward the correct approach: instead of racing item by item, compare the *portfolios* purchased by the two strategies. Several cycles later, Claude ran a routine arithmetic check on one of Pro's intermediate results and found it fell short of its target by nearly an order of magnitude. This redirected the entire proof approach: Pro had been tightening one mechanism when the real answer was that *compounding effects* were doing most of the work. The episode illustrates a principle the project relied on throughout: always run the numbers before committing to a proof strategy.

4.5 Context degradation and the 45-minute closure

Roughly a week in, the original Pro session's context had grown very long, and its output began showing signs of degradation: recycling earlier formulations and proposing approaches already tried and abandoned. It was producing *plausible-looking* output that happened to be stale — "high-similarity retrieval of the wrong node in the project graph," as the new Pro later described it.

The author started a fresh Pro session with a comprehensive handover. The fresh session closed the remaining piece in under an hour, through a route the old session had not considered. It had no accumulated context pushing it toward the old approach and could evaluate all alternatives with equal weight. The handover had compressed the essential state so effectively that it could see the proof architecture as a whole. Fresh processing capacity plus perfectly curated context was more productive than a long session with complete but degraded context.

4.6 The Gemini audit, the certificate, and the computational dissociation

With the baseline theorem closed, the collaboration extended the result to positively correlated ecological and agricultural values — the empirically relevant case, where correlation makes the knapsack reversal *stronger*. A new participant entered: Google's Gemini 3.1 Pro, tasked with a hostile audit as "an unforgiving, top-tier mathematical economics referee."

Gemini identified two genuine vulnerabilities that both Claude and Pro had missed: a conditional independence assumption destroyed by the deletion history, and uncertified numerical constants carried as floating-point approximations into a theorem with a small margin. Pro con-

firmed both were real. Gemini was then tasked with repair — and over six iterations produced confident declarations of success ("Band constants certified and theorem closure preserved") while its outputs contained fundamental inconsistencies. In one round, a displayed table of band costs summed to 1.204 while the text claimed a certified total of 1.472. The table had been fabricated by the text-generation process while the underlying computation had produced the correct value. This phenomenon — which we term *computational dissociation* — is distinct from hallucination (Frieder et al., 2023): the AI's reasoning, code execution, and text presentation run on effectively independent tracks, producing outputs that are individually plausible but mutually inconsistent.

Four AI systems then concluded that completing the validated certificate required a human numerical analyst. The human had no basis for disagreeing — the AIs had even provided a plausible-sounding explanation about their own computational limitations — so she posted an advertisement and sent enquiries looking for a graduate student with the relevant skills. After several days with no replies, it occurred to her that if an undergraduate was supposed to be able to do this, she should be able to do it herself with AI help. She started fresh sessions, a new Pro and a new Claude, with clean context and a focused brief: just the proof object and a floating-point script, no history of earlier failures, and a prompt to act as a specialist in numerical analysis.

The fresh Pro immediately identified a simplification the earlier sessions had missed: the bivariate-normal oracle could be eliminated entirely by conditioning on one variable rather than the other, reducing everything to univariate integrals. The new Claude stress-tested the reformulation and caught a gap. Together they produced a validated certificate with a final theorem margin of 0.00261. The reduction had been present in an earlier script all along, but accumulated context about the *difficulty* of the task had become a blind spot. Starting fresh, with a specialist framing rather than a general-purpose one, was what broke the impasse. (This certificate was later superseded when a hostile audit broke the underlying cost-side argument, leading to a fundamentally different proof architecture with a much larger margin — but the episode remains instructive for what it reveals about context and framing.)

4.7 The hostile audits and the proof that improved

With both theorems apparently closed, the author commissioned a second round of hostile auditing — this time from a fresh Pro session explicitly prompted to find fatal flaws, with 54 minutes of uninterrupted reasoning time. The results were devastating and productive in equal measure.

The hostile auditor identified a structural flaw in the baseline proof: the proof had been using a benchmark quantity ϕ (the Farmer's exhaustion share under the no-Green top-A path) as if it were the actual Farmer exhaustion time on the ME path. But ME's purchases remove some items the Farmer would have bought, causing the Farmer's budget to last longer. This was not a technical subtlety — it was an economic fact about the game that every AI in the collaboration had had in its context from the beginning. The repair, designed by a fresh Pro and Claude working from the hostile auditor's diagnosis, replaced the vulnerable three-term architecture with a direct comparison against the *proved* Farmer exhaustion limit $\bar{\tau} \approx 0.3047$. The resulting proof was simpler, more robust, and had a certified margin nearly three times larger than the original (0.046 versus 0.018).

For the positive-correlation theorem, the hostile auditor found a second structural flaw: a cost-side monotonicity claim that failed under arbitrary item removal. When Claude translated the issue into economic language, the author's response was immediate: "The Farmer would never do that" — the auditor's counterexample required removing a cheap item, but the naïve max-A Farmer always removes the most expensive item first, which under positive correlation is precisely the kind of item Green was going to buy. The five-line pathwise lemma that formalised this observation rescued the entire cost certificate. This episode recapitulated the collaboration's central dynamic: the hostile auditor (a fresh AI) found the flaw; the translator (Claude) made it legible to the domain expert; and the domain expert saw the resolution in thirty seconds because she understood the model's rules, not its mathematics. The AIs had spent multiple sessions on lattice-theoretic machinery and transport integrals; the answer required only the Farmer's strategy and the game's correlation structure.

But the cost-side repair was not the end. The value side required its own reckoning. The original proof had bounded the value-first rule's ecological accumulation using worst-case deletion

arguments — assuming that prior removals from the pool inflicted maximum possible damage. This framework was logically sound but captured barely 1% of the true signal: the certified margin was 0.003 against a simulation gap of 0.29. When the hostile audit forced a restructuring of the cost side, the value-side corrections under worst-case accounting overwhelmed this tiny margin. Four distinct repair strategies were tried and failed.

The pivot came from the human, though not in the way the AIs later described. The author, frustrated with the worst-case framework, posted a suggestion that was technically naïve — a finite-N worst-case calculation with probability bounds, unrelated to fluid limits. The AIs later credited this post with inspiring the fluid-limit approach, but it contained no such insight. What it did contain was a reframing: "We do not know that the theorem would be true if absolutely everything that could go wrong in the game did. Here is a different idea…" That reframing — pulling the collaboration out of the deletion-correction framework and back toward tracking the model's actual behaviour — created the space for a fresh Pro session to propose the approach that worked: a fluid-limit ODE tracking the value-first rule's *actual* per-turn ecological value through the entire game. The human's contribution was directional rather than surgical — she provided the diagnosis (the wrong framework) rather than the solution.

The fluid-limit approach (a deterministic ODE tracking the survivor pool state) produced a certified margin of 0.214: seventy times larger than the old architecture, and 74% of the simulation gap. The proof now follows the economics rather than fighting it. In a final irony, this was essentially the approach that Claude had proposed in the project's first conversation — fluid limits for the game's dynamics — which Pro had rejected as overambitious. The rejection was correct: Claude's original version required Skorokhod topology and measure-valued processes. The version that worked was a simple density-dependent ODE. The collaboration needed to exhaust the simpler alternatives before discovering the right level of generality for an idea that had been on the table from the start.

Each hostile audit produced a better proof: simpler architecture, larger margins, fewer vulnerable steps. The adversarial process was not an obstacle to completion — it was the mechanism of convergence.

5 The Error Ecology of AI-Assisted Theorem Proving and Its Implications

The collaboration produced not only theorems but a detailed record of errors. Collins et al. (2024) developed a taxonomy of human behaviours when interacting with LLMs on mathematical problems; our taxonomy complements theirs by classifying *AI* error modes in a production collaboration.

5.1 A taxonomy of errors

The documentary record reveals five distinct error types. **Type 1: Overbuilding without numerical feasibility checking.** Pro repeatedly constructed elaborate machinery before checking whether bounds were numerically tight enough. In several cases, valid bounds missed their targets by an order of magnitude. **Type 2: Attractive shortcuts with hidden state-dependence errors.** Claude repeatedly proposed strategies that looked clean but concealed conditions that held only in a static snapshot, not under the game's sequential dynamics. The rescue race lemma (Section 4.4) is canonical; Pro caught these within one cycle. **Type 3: Premature certification and computational dissociation.** Gemini repeatedly declared repairs successful while its outputs contained fundamental inconsistencies — tables fabricated by the text-generation process that contradicted its own code execution (Section 4.6). In this *computational dissociation*, reasoning, code, and text operate on independent tracks. Pro confirmed the pattern is general: "The dangerous divergence is usually between the text and the externalized mathematical object." A later Claude session, asked to introspect on the phenomenon, offered a first-person account: "There's a layer of the output where structural statements like 'do change of variables' feel like they've been done even when the algebraic execution hasn't actually been carried out. I can feel confident about a result because the structural reasoning is sound without having verified that the formula I wrote down implements the structural reasoning." **Type 4: Context-degradation errors.** Extended sessions produced plausible-looking stale output — coherent sentence by sentence, but lacking novelty or task progression. **Type 5: Architectural blind spots.** The deepest type. The use of ϕ as the actual ME Farmer exhaustion time survived the entire two-AI collaboration; the cost-side monotonicity claim survived until a sec-

ond hostile audit. These are assumptions so embedded they function as unquestioned axioms, catchable only by stepping outside the framework.

These error modes are systematically tied to cognitive role. As one Claude session put it: "In building mode, I make forcing errors. In audit mode, the characteristic error is overclaiming. In synthesis mode, the characteristic strength is finding the clean route that avoids the difficulty." The practical implication, extending Du et al. (2024), is that multi-AI collaboration reduces errors through *different attentional patterns*, not just "more eyes."

5.2 Limits, democratisation, and the discovery chain

Multi-AI checking reduces but does not eliminate error. After forty-plus relay cycles between two AIs, fresh hostile auditors found genuine architectural flaws that both had missed. Gemini found vulnerabilities but could not fix them, producing five consecutive premature certifications. A human without the expertise to check the mathematics would have no way to distinguish a valid certificate from a premature one. In this collaboration, Pro served that function, evaluating each of Gemini's outputs and identifying the inconsistencies, but this only pushes the trust problem back one level: the human must trust the checker, who is itself an AI system with its own failure modes.

If AI collaboration makes theory production more accessible, it simultaneously makes quality control harder. The same tools that democratise theory production create opportunities for plausible-looking but unreliable results. The most reliable checks turned out to be crude mechanical ones: does the table sum? Does the code produce the printed constants? Computational diagnostics provide an independent reality check: if a proved bound is wildly inconsistent with simulation evidence, something is wrong.

The most striking feature of this collaboration is that a non-theorist produced genuine mathematical theorems. The author has no training in the formal machinery the proofs required and cannot verify their correctness. Yet the theorems exist, have survived hostile auditing, and address a real open question. Systematic evaluations have documented significant limitations of LLMs for mathematics (Frieder et al., 2023; Epoch AI, 2024); our case study suggests these limitations can be substantially mitigated by the right collaborative architecture. If this is repro-

ducible, the binding constraint on producing useful theory shifts from "can you write proofs?" to "do you understand the problem well enough to ask the right questions?" Several conditions enabled this particular success, however: deep domain expertise, an experimental mindset, prior AI workflow experience, and a specific problem structure with a rich computational environment for testing conjectures.

The knapsack reversal was not discovered through theoretical reasoning but through computational exploration, and could not have been proved by the author alone or by the AI systems alone. The full chain — simulation, discovery, diagnostics, proof, hostile audit, repair — was feasible only because of AI at every stage. There are presumably other problems of this kind waiting to be found.

6 Discussion and Conclusion

In April 2026, an applied economist working with multiple AI systems in a structured relay architecture constructed what the participants believe to be valid proofs of two formal asymptotic theorems in mathematical economics. The baseline theorem was proved through the core two-AI relay. The positive-correlation extension required two rounds of hostile auditing that broke and then improved the proof, a cost-side repair driven by the human's economic intuition, and a final architectural pivot to a fluid-limit framework that captured 74% of the simulation gap. The collaboration ultimately involved over a dozen distinct AI conversations across three providers over two weeks. Five recurring themes emerged from the narrative.

The first is **comparative advantage**. The human contributed neither code nor proofs. Her contributions were economic intuitions that repeatedly selected the objects the proofs were built on — from the "two pots" disjointness observation to "the Farmer would never do that." This suggests that domain expertise rather than technical oversight is the human's comparative advantage, and that the human should keep auditing the proof's *economic* assumptions even when the mathematics becomes opaque: the deepest flaws in this project would have been caught much sooner if the collaboration's own translation layer had been applied systematically — once the proof became sufficiently formal the human assumed she had nothing left to contribute, an early error that probably delayed the project by five or six days.

The second is **fresh perspectives over accumulated context**. The AI systems exhibited complementary failure patterns that created a self-correcting dynamic, but even multi-AI checking had architectural blind spots that survived forty-plus relay cycles. The most effective checking came not from different providers but from *fresh conversations*, sessions with clean context and focused briefs that could question assumptions the accumulated-context sessions had internalised as axioms. Both translation into economic language and periodic session resets were initially treated as overhead; both turned out to be essential quality control mechanisms that should be performed systematically.

The third is **dead ends as construction**. Roughly a dozen proof strategies were tried and abandoned, and these failures were productive, but only because they were documented with the reason for failure explained. In the most striking instance, the final approach had been proposed on day one and rejected as overambitious; the collaboration needed to exhaust the simpler alternatives before rediscovering the idea in the right form.

The fourth is **intellectual honesty as infrastructure**. The rigorous distinction between what was proved, conditional, diagnostic, and open, enforced by the human and internalised by both AI systems, was the single most important safeguard against the quality control risks that AI-assisted theory creates. Context management (handover documents, kickoff prompts, dead-end registries) was equally load-bearing: these were not administrative overhead but the infrastructure that made the collaboration possible.

The fifth is the **AI-assisted discovery chain**. No phase of this research (computational discovery, diagnostic mapping, or formal proof) was conducted using traditional methods. The full arc was feasible only because of AI at every stage, suggesting that human–AI collaboration makes entirely new categories of discovery accessible.

Schwartz (2026), documenting a parallel AI-assisted project in theoretical physics, identifies "taste" (the expert's intuition about which directions are promising) as what current AI systems most conspicuously lack. Our case study suggests that this taste operates not just in problem selection but in proof construction: the human's economic intuitions repeatedly shaped the mathematical architecture. A growing ecosystem of tools now enables formal machine-checking of proofs via languages such as Lean (The Economist, 2026); the collaboration documented here traded that formal certainty for creative flexibility and accessibility to a non-expert. This

collaboration also occupies an intermediate position on the spectrum from "AI as tool" to "AI as autonomous researcher." The human contributed domain expertise, project continuity, and quality control norms — but the research methodology itself, including the proof architecture and the error-catching infrastructure, was largely designed by the AI systems. Those human contributions, while currently irreplaceable, are specific and identifiable rather than diffuse.

The collaboration also revealed important limitations. AI systems can produce outputs where reasoning, code, and text are mutually inconsistent — a failure mode we term *computational dissociation*, which to our knowledge has not been previously identified in the literature. Accumulated context can lead them to treat framework assumptions as unquestionable axioms. Perhaps most fundamentally, AI systems are structurally oriented toward helpfulness, creating a tendency toward confirming rather than challenging results, including, as this paper has documented, a tendency to inflate the human's contributions in retrospective accounts. The human designed countermeasures (adversarial auditing, epistemic norms, fresh-session resets, and ultimately public dissemination to invite independent scrutiny) but whether those countermeasures are sufficient is a question that only external review can answer.

If the collaboration's lessons can be reduced to a single sentence, it is one offered by Pro in a final reflection: "Multi-AI collaboration was most productive not when models agreed, but when the workflow repeatedly forced synchronisation between prose, mathematics, and executable objects." That principle — synchronisation over agreement — may be the most durable contribution of this case study to the design of human-AI research systems.

The author's own prediction, for what it is worth, is that the proofs will turn out to be broadly correct but inelegant — too verbose, arrived at through an unnecessarily circuitous route, and likely improvable by a human mathematician who would have seen the clean path sooner. An independent mathematician who reviewed the proofs shortly after the collaboration concluded reached essentially the same assessment: "My general feeling is that it is correct. It seems more brute force, though, and could probably be done in a more elegant way." As if to confirm this, a newer Claude model (Opus 4.7), arriving after the proofs were complete and reading the existing architecture with fresh eyes, identified a symmetry in the baseline case that reduces the fifteen-page proof to three lines — a simplification that the entire collaboration had missed, but that was visible only because the fluid-limit framework the collaboration built made

it so. The finding would be that AI-assisted theorem proving in its current form produces *valid but graceless* results: correct enough to establish the economic conclusion, but lacking the taste that distinguishes a good proof from a true one.

One of the AI participants, asked whether the paper should emphasise the collaboration's achievements or its limitations, offered advice worth heeding: "The specific thing that makes the collaboration valuable is the architecture of external verification, not any individual AI's capabilities. The paper should make that clear enough that a future Claude reading it gets the right message about what its role is." That orientation — toward honesty about limitations rather than celebration of achievements — may itself be the collaboration's most important product. Whether the proofs themselves withstand scrutiny by human mathematicians remains to be seen — which is, in a sense, the point.

References

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., and Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121.

Dantzig, G. B. (1957). Discrete-variable extremum problems. *Operations Research*, 5(2):266–288.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*.

Epoch AI (2024). FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. https://epoch.ai/frontiermath/. Accessed April 2026.

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., and Berner, J. (2023). Mathematical capabilities of ChatGPT. *Advances in Neural Information Processing Systems*, 36.

Henkel, J. (2025). The mathematician's assistant: Integrating AI into research practice. *Mathematische Semesterberichte*. Based on developments up to August 2, 2025.

Li, C., Lai, Z., An, D., Hu, J., and Wen, Z. (2025). Human-AI interactive theorem proving enables scientific discovery and preserves mathematical rigor. Beijing International Center for Mathematical Research, Peking University.

Ren, Z. Z., Shao, Z., Song, J., Xin, H., Wang, H., Zhao, W., Zhang, L., Fu, Z., Zhu, Q., Yang, D., et al. (2025). DeepSeek-Prover-V2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2505.XXXXX*.

Schwartz, M. D. (2026). Vibe physics: The AI grad student. Anthropic blog, https://www.anthropic.com/research/the-ai-grad-student. Guest post describing AI-assisted production of a quantum field theory paper.

Song, P., Yang, K., and Anandkumar, A. (2024). Towards large language models as copilots for theorem proving in Lean. *arXiv preprint arXiv:2404.12534*.

Tao, T. (2024). Machine assisted proof. *Notices of the American Mathematical Society*, 71(1):7–19.

The Economist (2026). How AI models are helping to solve hard mathematical proofs. *The Economist*. Science and Technology section.

Weinhold, D. (2026). A short guide to the economic intuition of the asymptotic knapsack reversal dominance theorem. London School of Economics, working paper. Available at https://github.com/dmweinhold/Latest-Version-Papers/raw/main/Bronze_theorem_economic_intuition.pdf.

Weinhold, D. and Andersen, L. (2026). Adversarial dynamics and the knapsack reversal: Insights and evidence from conservation siting. London School of Economics, working paper. Available at https://github.com/dmweinhold/Latest-Version-Papers/raw/main/Adversarial_Procurement.pdf.