The Fibonacci Jury Protocol

A Multi-Agent Blind Verification Architecture for High-Confidence Proof Validation

Matthew J. Goss, Jr.
Quantiterate Independent Research
June 2026

Abstract

We present the Fibonacci Jury Protocol, a staged multi-agent verification architecture that achieves approximately 19–20× confidence amplification over single-agent evaluation. The protocol sequences four independent verification rounds with panel sizes following the Fibonacci sequence—2, 3, 5, and 8 agents—under strict information isolation between tiers. No round has access to any prior round's deliberations, conclusions, or even the knowledge that prior rounds exist. The protocol is triggered only after a primary proof-hardening engine achieves ≥95% confidence, ensuring that the verification jury never evaluates incomplete work. We analyze the confidence multiplier arising from the interaction of intra-round debate and inter-round blindness, identify the information-theoretic basis for stopping at Fibonacci index 6 (panel size 8), and propose the architecture as a general-purpose verification framework applicable to any domain requiring high-confidence validation of complex claims.

1. Introduction

Single-agent evaluation of complex proofs and claims suffers from well-documented failure modes: confirmation bias in reasoning chains, hallucinated logical steps that appear internally consistent, and the fundamental inability of a system to reliably detect its own errors. These failures are not resolved by scale alone. A larger model with more parameters inherits the same structural blind spots that arise from operating within a single inference context.

The natural response is to deploy multiple agents. Existing multi-agent approaches—ensemble voting, chain-of-thought verification, parallel task decomposition—improve on single-agent performance but fail to address the core vulnerability: when agents share context, training data, or evaluation history, their errors correlate. Five agents that all saw the same flawed reasoning chain are only marginally more reliable than one. The appearance of independent confirmation masks a single point of failure.

This paper introduces the Fibonacci Jury Protocol, which solves the correlation problem through *structured independence*. The protocol enforces strict information isolation between verification rounds, creating genuinely independent confirmation events whose combined confidence compounds multiplicatively rather than additively. The Fibonacci sequence governs panel sizing at each round, and we identify the information-theoretic basis for stopping at eight agents per panel—the point beyond which additional evaluators contribute noise rather than signal.

2. The Pre-Verification Gate

The Fibonacci Jury is never invoked on raw or exploratory work. It sits behind a mandatory confidence gate: a primary proof-hardening engine must first iterate, attack, stress-test, and advance the claim until its internal confidence metric crosses 95%. Below this threshold, the jury never sees the material. The hardening engine continues working.

This gate is not merely procedural. It is architecturally essential. The jury's power derives from evaluating near-complete work where the remaining uncertainty is whether the proof is correct, not whether it is finished. Presenting incomplete proofs to the jury would consume its verification power on problems that are better solved by continued development, not independent review. The 95% gate ensures the jury's independence is spent on the right question: is this correct?

In our implementation, the hardening engine is the War Room—a multi-model proof pipeline that tracks confidence through claim verification rates, proof percentage, and

dead-end documentation. The protocol is agnostic to the specific engine used; any system that produces a calibrated confidence metric can serve as the gate.

3. Protocol Architecture

Once the confidence gate is satisfied, the proof artifact—in its entirety, with no annotations, commentary, or evaluation history—enters four sequential verification rounds. Each round receives an identical, byte-for-byte copy of the original artifact accompanied only by a neutral evaluation prompt. The panel sizes follow the Fibonacci sequence: 2, 3, 5, 8.

3.1 Round One: The Seed Pair

Two agents evaluate the proof independently, then engage in structured debate. Each presents its evaluation, challenges the other's reasoning, and identifies weaknesses. They iterate until reaching a settled position—consensus, qualified consensus, or documented disagreement. The output is a single verdict document with a confidence score and a list of identified issues. This document is sealed and cannot be accessed by any subsequent round.

3.2 Round Two: The Blind Triad

Three completely fresh agents receive the original proof artifact. They have zero exposure to Round One's deliberations, verdict, or even the fact that Round One occurred. The same structured evaluation process unfolds among three agents rather than two. Three-way debate introduces dynamics absent from pairwise exchange: coalition formation, minority dissent preservation, and the possibility of two-against-one pressure that forces deeper justification. The output is again a sealed verdict document.

3.3 Round Three: The Quintet

Five fresh agents, blind to all prior rounds, evaluate the original artifact. At five agents, the probability that all evaluators share the same blind spot drops significantly. The larger panel also increases the surface area for catching subtle errors—an undefined

object, a silent scope shift, an assumption that went unexamined in smaller panels. The sealed verdict document captures the quintet's independent conclusion.

3.4 Round Four: The Octave

Eight fresh agents, blind to all prior rounds, perform the final evaluation. This is the terminal round. The octave represents the highest Fibonacci tier at which additional agents produce meaningful signal. Beyond eight, marginal verification value drops below the noise floor—the ninth through thirteenth agents are statistically unlikely to catch errors that eight independent evaluators missed. The signal-to-noise ratio degrades rather than improves. We analyze this stopping condition in Section 5.

4. Information Isolation Requirements

The confidence multiplier of the Fibonacci Jury depends entirely on the independence of its rounds. If any round's output leaks to a subsequent round, the multiplicative compounding collapses to additive improvement. The following isolation rules are therefore inviolable—breaking any one of them compromises the protocol's guarantees.

First, no forward leakage: agents in Round N never see the output of Round $N–1$. Second, no backward leakage: agents in Round N never learn that prior rounds exist. Third, no shared context: each round's agents receive only the original proof artifact, never any derivative analysis, summary, or annotation produced by any other round. Fourth, no shared model state: if using the same model family, each agent must be a fresh instance with no conversation history and no cached embeddings from related evaluations. Fifth, no meta-information: agents are not told they participate in a multi-round protocol, what round number they occupy, or how many rounds exist.

These requirements must be enforced programmatically. Human discipline alone is insufficient for maintaining isolation at this level of strictness. Round outputs must be stored in sealed containers that are computationally inaccessible to subsequent rounds until all four rounds complete and the aggregation phase begins.

5. Confidence Multiplier Analysis

Consider a single AI agent evaluating a complex proof with an error rate E — the probability of missing a critical flaw. For frontier models on mathematical reasoning, E is estimated at 0.15–0.30 depending on proof complexity.

Within a round, K agents debating and cross-checking reduce the effective error rate. For a pair (K=2), structured debate drives the effective error toward E^2. For a triad (K=3), toward E^3. The quintet and octave drive their respective rounds toward E^5 and E^8. These are conservative upper bounds—structured debate catches correlated errors that pure statistical independence would miss, so actual rates are lower.

Because rounds are fully blind to each other, their error rates compound multiplicatively across rounds. The combined probability that all four rounds miss the same flaw is approximately the product of the four individual round error rates. With intra-round debate reducing each round's effective error significantly, the combined probability of a missed flaw becomes vanishingly small.

The claimed 19–20× confidence amplification represents the effective verification power defined as the ratio of combined detection probability to single-agent detection probability. The exact multiplier depends on the base error rate and the degree to which intra-round debate catches correlated failures, but for typical frontier model performance on mathematical proofs, the protocol consistently achieves this range.

5.1 The Stopping Condition

The Fibonacci sequence continues beyond 8: 13, 21, 34. Why does the protocol stop at the sixth Fibonacci index? The answer is informational. Each additional agent in a panel contributes a diminishing increment of independent perspective. At small panel sizes, each new agent is likely to bring a genuinely novel evaluation angle. By eight agents, the space of plausible independent evaluation strategies for a given proof is substantially covered. The thirteenth agent is unlikely to find an error that eight agents debating among themselves did not. The marginal information gain falls below the noise introduced by coordinating a larger panel—logistical complexity, debate diffusion, and

the tendency of large groups to converge on socially comfortable rather than analytically rigorous conclusions. Eight is the empirically motivated sweet spot where verification power peaks before noise dominates.

6. Verdict Aggregation

After all four rounds complete, a non-participating aggregator—human or designated system—opens the four sealed verdict documents simultaneously. The aggregator never participates in evaluation; its role is purely procedural.

Unanimous confirmation across all four rounds constitutes acceptance at maximum confidence. If three of four rounds confirm while one raises concerns, the proof returns to the hardening engine with the dissenting round's specific objections for targeted strengthening; the proof re-enters the jury after revisions. If two or fewer rounds confirm, the proof is rejected and the hardening engine receives all objections for fundamental rework. If any round identifies a fatal flaw—a logical contradiction, an undefined object, or circular reasoning—the proof is immediately rejected regardless of the other rounds' verdicts.

7. Generality of the Protocol

Although developed for mathematical proof verification, the Fibonacci Jury is domain-agnostic. Any domain requiring high-confidence verification of complex claims can adopt this architecture: legal contract review, clinical diagnostic confirmation, security code audits, financial model validation, and scientific peer review. The 95% confidence gate and the specific Fibonacci panel sizes may require domain-specific calibration, but the core architecture—staged blind rounds with Fibonacci scaling—transfers directly. The protocol's value proposition is strongest in domains where the cost of a missed error is catastrophic and the cost of verification is trivial by comparison.

8. Relationship to Existing Work

Ensemble methods in machine learning deploy multiple models but typically share training data and architecture, creating correlated error patterns. The Fibonacci Jury enforces structural independence through information isolation, not merely model diversity. Constitutional AI uses self-critique within a single model's context window; the Jury extends critique to inter-model blindness across multiple independent sessions. The debate frameworks proposed by Irving et al. (2018) use adversarial debate for AI alignment; the Jury extends this to multi-tier blind debate with Fibonacci-scaled panels. Dynamic workflows, recently introduced by Anthropic (2026), parallelize subtasks across subagents for efficiency; the Jury is orthogonal—it verifies rather than decomposes, and requires sequential blindness rather than parallel execution. Formal verification systems such as Lean 4, Coq, and Isabelle provide mathematical certainty for fully formalized proofs; the Jury operates upstream, evaluating whether an informal or semi-formal proof merits the substantial effort of full formalization.

9. Open Questions

Several questions remain for empirical investigation. Is there a formal information-theoretic proof that the thirteenth Fibonacci panel adds noise rather than signal, or is the stopping condition at eight purely empirical? The current specification prescribes "structured debate" within rounds but does not mandate a specific debate format; whether adversarial, collaborative, or hybrid debate produces the highest-fidelity verdicts may vary by domain. The 95% confidence gate assumes a well-calibrated confidence metric in the hardening engine; overconfident systems would trigger the jury prematurely, wasting its verification power on undercooked proofs. Finally, the verdict aggregation rules described in Section 6 are deliberately coarse; a more nuanced scoring system may be warranted for complex proofs containing multiple independent claims at different confidence levels.

10. Conclusion

The Fibonacci Jury Protocol provides a principled architecture for high-confidence verification of complex claims. Its power derives not from the number of agents

deployed but from the structure of their independence. By enforcing complete information isolation between Fibonacci-scaled verification rounds, the protocol transforms what would otherwise be redundant evaluation into genuinely independent confirmation, achieving multiplicative rather than additive confidence gains. The architecture is simple, the isolation rules are enforceable, and the confidence analysis is grounded in straightforward probabilistic reasoning. We propose it as a general-purpose verification layer for any domain where the cost of being wrong substantially exceeds the cost of being thorough.

Matthew J. Goss, Jr.
Quantiterate — research.quantiterate.com
The Signal Carries Everything