

The Chisel Fallacy: What Pangram Actually Measures

Pangram Is a Very Sophisticated Way of Asking the Wrong Question

AMUL BHAM

Independent Researcher, Corona, CA, USA

amul.bham@gmail.com · linkedin.com/in/amul-bham · orcid.org/0009-0009-7660-4031

Paper · Licensed CC BY-NC 4.0

Sep 22, 2026

For most of writing’s history, textual production and intellectual production were tightly enough coupled that one could stand in for the other. Producing coherent, structured prose required thinking, and thinking reliably produced prose. That coupling has weakened. The tools that generate fluent, structured text no longer require the intellectual work that once produced it. In some domains, this debate is already settled.

The current manifestation of that breakdown is a tool called Pangram. In August 2026, Pangram’s co-founder and CTO published a blog post arguing that AI detection works precisely because post-training constrains language models to consistent, detectable patterns (Emi, 2026). That same week Substack integrated the tool, giving any reader the ability to scan any post and receive a score indicating AI involvement. Writers called it a witch hunt (404 Media, 2026).

Both responses make sense from where each side is standing. The problem is that Pangram’s score identifies something real — evidence of AI involvement in the production of the text — and then that score gets used to answer a different question: whether the author supplied the intellectual contribution behind it. A detector can be entirely correct on its own target — identifying statistical evidence of model involvement — and still be insufficient to settle whether the named author supplied the intellectual contribution. Treating one as a proxy for the other is a category error. This paper calls it the Chisel Fallacy.

The paper names the assumption that makes this error structurally tempting, shows the consequences that follow from it, and proposes a better standard for evaluating intellectual work in an age when the coupling between textual production and intellectual production can no longer be assumed.

Two Things Getting Conflated

Consider what one-shot AI writing actually looks like. Someone opens a browser, types a request — “write a 500-word essay arguing that standardized testing improves educational outcomes” — and receives organized prose. The essay has a thesis, supporting points, and a conclusion. The process took ninety seconds. The person who produced this document has performed little or none of the intellectual work the finished artifact appears to represent. The intellectual decisions — what to argue, how to structure it, which evidence to select, whether the reasoning holds — were largely delegated along with the prose.

This is the version of AI-assisted writing that most institutional concern is correctly aimed at. And it explains why textual AI detection feels intuitively sufficient: if your mental model is prompt $\rightarrow$ finished essay, then detecting AI involvement in the text looks almost equivalent to detecting that intellectual work was absent. This paper calls that mental model the Vending Machine Assumption — the belief that AI-assisted writing is this activity, and therefore that a score indicating AI involvement is evidence of absent intellectual contribution. It is the assumption that makes detector-based authorship judgments appear coherent. The assumption is not irrational — it reflects a mental model that makes sense for one way of using AI. AI-assisted work spans a broad agency gradient. The same category of tool use can range from near-total delegation to sustained human direction, judgment, and verification. The Vending Machine Assumption collapses that gradient by treating one endpoint as the category itself.

The Other Workflow

Directed Authorship is not a new activity — it is a pattern that has existed across professional writing for as long as execution could be delegated. The definition here makes explicit what that pattern involves.

There is a different mode of AI-assisted writing, and this paper calls it Directed Authorship. In this mode the author retains intellectual responsibility for what the artifact claims and why. The author determines the argument’s structure, evaluates which evidence matters, decides what the work needs to establish, and takes responsibility for whether the resulting reasoning holds. The model or collaborator handles the lower-level work: syntax, phrasing, spell-checking, formatting — what the programming profession, which has already navigated this split, calls syntax rather than orchestration. The author concentrates on the upper-level work: argument structure, evidence evaluation, reasoning, judgment. Directed Authorship is not specific to AI — it describes any mode of production in which intellectual

direction is retained by the author while lower-level execution is delegated to any capable agent, human or otherwise.

Directed Authorship is a workflow definition, not a verdict on any particular author’s intellectual contribution. Whether a specific author using this mode actually retained the command and responsibility it describes is exactly the question that requires inspection — and that a detection score cannot answer. A detector can identify evidence of AI involvement somewhere along the agency gradient. It cannot, from that evidence alone, locate the intellectual workflow on it.

What counts as lower-level work varies by domain. Fields Medal winner Terence Tao, writing on mathematics following the 2026 International Congress of Mathematicians, describes the emerging practice: automation handles computation and proof verification; humans determine what to prove and interpret what results mean (Tao, 2026). The structure holds across the domains this paper draws on:

Domain	Lower-level work (delegatable)	Upper-level work (authorship)
Academic writing	Syntax, phrasing, spell check, formatting	Argument structure, evidence selection, causal inference, judgment
Mathematics	Computation, equation solving, proof verification, symbolic manipulation	Determining what to prove, interpreting what results mean, mathematical intuition
Software	Code syntax, boilerplate, compilation	System design, architecture, what the software should do

A useful test: can this task be performed correctly without knowing what the artifact is trying to establish? If yes — formatting, spell checking, syntax — it belongs in the left column. If no — which comparison period to use, whether the inference holds, what the argument requires — it belongs in the right. The line moves depending on domain and context. The structure does not.

A detector identifies statistical evidence associated with AI involvement in the text. Research confirms that AI-assisted editing — even the guideline-compliant kind — produces detection flags at high rates (Karr et al., 2026). But the score does not establish which intellectual workflow produced the text. Under the Vending Machine Assumption, a high score appears to settle the question. Once Directed Authorship exists as a workflow, the score is insufficient to settle it.

What a detector score cannot recover is which intellectual workflow produced the text — and that is exactly the question that requires a different standard.

The Shelf You Already Built

Consider a thought experiment. You read an article — a genuinely useful DIY idea, maybe fixing something around the house. You tried it. It worked. Three months later you learn the article was AI-assisted. Are you going to un-fix the shelf?

The question is slightly absurd, which is the point. Learning that AI helped produce the prose does not retroactively change whether the idea was sound, the instructions were clear, or the shelf worked. The question you were actually asking — does this hold up? — is a different question from where the sentences came from.

You don't need to know who Satoshi Nakamoto was to evaluate Bitcoin's white paper. The mechanism either works or it doesn't. The cryptographic assumptions can be tested, the protocol can be implemented, the incentive structure can be examined — none of that requires settling the author's identity (Nakamoto, 2008). Satoshi's identity matters for history, attribution, and questions about holdings and intent. It is not a premise in the protocol. Evaluate the mechanism, not the biography.

In many forms of practical and intellectual writing, we already accept that producing the sentences and supplying the underlying judgment need not be the same act. That separation is not novel to AI. It has been present throughout.

That observation is not a claim that origin is irrelevant. Provenance matters — for credit, for accountability, for disclosure. Those are real questions worth asking. But they are not the questions a statistical detector answers. A flag on the lower-level execution layer does not tell you who supplied the ideas, who is responsible for the claims, or whether the disclosure was appropriate.

The question the shelf was never asking was who formatted the instructions or chose the words. The question was whether the upper-level idea held up when you tried it. The Chisel Fallacy is what happens when we start answering the second question with evidence from the first.

Who Actually Holds the Chisel

Delegating the execution of a task does not establish that intellectual contribution was absent. When execution is handled by someone else — a collaborator, a model, an apparatus — the question of who directed what remains open. Intellectual contribution can reside above direct execution. The two are separable, and their separation is not new.

Recent philosophical work is arriving here independently. The Apt Curation Model, published in *Philosophy & Technology* in 2026, argues that authorship in

AI-assisted work resides in higher-order decisions — structuring, evaluating, synthesizing — not in who produced the sentences (Rodrigues, 2026). The paper takes the next question seriously: if intellectual contribution does not map cleanly onto sentence production, what evidence can actually settle whether it was present?

The Chisel Fallacy names the error that occurs when we answer that question with the wrong evidence. It is the error of judging intellectual work by evidence about its lower-level execution rather than evidence about what it claims and whether those claims hold. The chisel tells you something about who executed the lower-level work — in Pangram’s case, syntax production, phrasing, and word choice. It tells you nothing that settles the question of intellectual contribution. The Chisel Fallacy is not the act of noting how something was produced, or raising disclosure questions. It occurs when lower-level execution evidence is promoted into a verdict on intellectual contribution or validity — a verdict that, applied to a developer who directs code generation without writing every line, would now be immediately recognized as indefensible.

Automation lowers the execution floor faster than institutions update their definition of authorship. That lag is what makes the fallacy acute now.

The detector flags the chisel. It is completely blind to the sculptor.

The Delegation That Was Already There

Writing has always accommodated the separation of intellectual direction from lower-level execution. The professional publishing apparatus makes this explicit. A developmental editor restructures arguments and challenges reasoning. A copy editor improves sentence-level prose. A proofreader catches errors. A continuity manager tracks what was established where across thousands of pages. Brandon Sanderson’s novel *Wind and Truth* was produced with the support of 1.3 million words of beta reader feedback, a dedicated continuity manager responsible for the series wiki, and a VP of editorial building production spreadsheets (Ahlstrom & Dragonsteel Books, 2025). The intellectual direction — what the story claims, how it argues, what it is for — remained his. The execution apparatus did not determine what the work was for.

This is not exceptional. It is the standard professional model. Nobody treats editorial collaboration as evidence against authorship. Even where editorial contribution is substantial, the convention is that intellectual direction remains with the named author — the collaborators serve the author’s vision, not their own. The acknowledgments page lists them. The author’s name goes on the cover.

What AI changed is not the logic of this arrangement but the resource requirement for accessing it. The delegation that was always philosophically compatible with authorship was, until recently, expensive and mostly invisible. Professional editorial

infrastructure was gated by contracts, relationships, and cost. AI made the same fundamental arrangement — human direction, delegated execution — accessible to anyone with a browser, and made the execution layer scannable in a way it had never been before.

The Vending Machine Assumption applies a standard to AI-assisted writing that publishing has never applied to editorial collaboration. A writer working with AI assistance is flagged for delegating lower-level work. A writer working with a professional editing team is not. The underlying delegation is the same. What changed is who can afford it and whether a tool can see it.

Why the Pangram Score Answers the Wrong Question

Pangram’s own documentation shows what the chisel looks like from the inside.

What the detector reads

Pangram describes its own task precisely. From its knowledge hub: “Fundamentally, AI detection is a problem of author identification — categorizing which decisions are prototypical of what kind of author” (Pangram, n.d.a). That is a legitimate technical question. Its classifier learns patterns across a large training corpus and places new texts in learned representational regions. No single visible feature determines the result.

Separate research confirms that measurable linguistic regularities in LLM-generated prose are real. Reinhart and colleagues, publishing in the Proceedings of the National Academy of Sciences in 2025, found that instruction-tuned LLM prose contains consistent surface patterns: present participial clauses appear at 2–5 times human rates; GPT-4o used “camaraderie” 162 times and “tapestry” 155 times more frequently than humans in their corpus (Reinhart et al., 2025). These are lower-level execution properties — precisely the kind of signals Directed Authorship delegates. They tell you something about how the sentences were produced. They tell you nothing that settles the question of intellectual contribution.

What it gets asked to settle

The problem is not detection. It is what the detection is asked to settle. PeerPrism — a 2026 benchmark of 20,690 peer reviews presented at SIGIR — was explicitly designed to separate idea origin from text origin, including cases where human-origin ideas were rendered entirely in AI-assisted text. Its authors found that detector predictions diverge sharply under hybrid conditions and conclude that current detection methods conflate surface realization with intellectual contribution (Arvan et al., 2026). Once idea provenance and text provenance diverge, a detector of the second cannot settle the first.

The Chisel Fallacy occurs not inside the detector but one step later. Pangram may answer its own question correctly. That answer is still the wrong evidence for a different question: who supplied the upper-level intellectual contribution.

Two cases

In August 2026, UC Berkeley mathematics professor Zvezdelina Stankova published an op-ed arguing that student mathematical preparedness had declined following test-blind admissions. A student ran the piece through Pangram. It returned 33% AI-assisted. Stankova confirmed that AI had helped with editing and locating documents (Daily Californian, 2026; The Guardian, 2026). The response was immediate: the production process became the story.

Stankova’s disclosed use is consistent with Directed Authorship: AI assisted with editing and document location — lower-level execution tasks. Her disclosed use says nothing by itself about who supplied the diagnostic comparison or causal inference. Those are precisely the intellectual questions a detector cannot settle.

Stankova’s argument is contestable on its own terms. The diagnostics she used changed across the periods she compared; causal attribution is genuinely difficult; her definitions of readiness can be scrutinized. Those are substantive questions worth engaging. Whether AI participated in the text, whether that participation was appropriately disclosed, and whether the argument is sound are three different questions. A detector bears directly on the first. It does not settle the second or third.

The Commonwealth Short Story Prize encountered a version of the same problem. When AI allegations circulated against regional winner Jamir Nazir’s story “The Serpent in the Grove,” the Commonwealth Foundation did not use AI detection tools in its review. Its statement cited concerns about artistic ownership and consent surrounding unpublished work, while also noting that such tools may serve as useful indicators but “cannot provide conclusive evidence on their own” (Commonwealth Foundation, 2026; Brittle Paper, 2026). Instead the Foundation held detailed discussions with the regional winners and examined working drafts, timestamped documents, and notes before concluding that AI had not written the winning stories.

The Foundation behaved as though textual evidence alone were insufficient to distinguish delegation from authorship: it inspected the development record instead.

That move — from execution signal to development record — is the first half of this paper’s argument in practice.

A perfect detector of AI involvement would still not be a detector of intellectual contribution.

When the Metric Becomes the Target

When a metric becomes a target, it ceases to be a good metric. Goodhart observed this about monetary policy in 1975 (Goodhart, 1975). The mechanism is operating here.

A 2026 study from the University of Notre Dame, presented at the ACM AI Leadership Summit, tested a proxy for guideline-compliant AI assistance — light editing of the kind many institutional policies broadly permit — and submitted the results to Pangram and GPTZero. Pangram flagged 64–80% of these lightly edited pieces. Then the researchers took fully AI-generated text, ran it through a commercial humanizer, and resubmitted. More than 96% escaped detection. The authors’ conclusion: “Honest AI-editing results in a higher sanction risk than humanizer-assisted evasion” (Karr et al., 2026).

The Market’s Response

The market adapted almost immediately. A cottage industry of human specialists now exists on Fiverr whose entire value proposition is defeating the authentication system. One listing: “Manual Rewriting. No spinners or bots. Every word rewritten by me. Burstiness: Short and long sentences mixed naturally. AI Detector Proof. Tested before delivery. It will pass” (Fiverr, n.d.). The system ends up rewarding human execution introduced specifically to conceal the provenance it was built to detect.

When a detector score becomes a compliance target, the market optimizes against the score rather than the underlying intellectual behavior.

The metric was also fragile before anyone tried to game it. The same study found that Pangram flag rates varied from 2.1% for chemistry abstracts to 26.6% for theology abstracts (Karr et al., 2026). Detector scores tracked stylistic features that are not unique to AI involvement. Once the score becomes consequential, Goodhart adds strategic optimization pressure on top of an already-confounded measurement.

The result is a metric squeezed from both sides: natural confounding below, strategic evasion above. A high human score no longer uniquely identifies unaided authorship.

The Floor Is Already Moving

The same transition has run through other domains before it reached writing. In each, automation lowered an execution floor and the argument about what consti

tutes meaningful contribution shifted upward. The mathematics case shows the pattern most clearly — it has run for decades and is still accelerating.

Mathematics

For much of mathematical history, “computer” was a job title. Human beings performed arithmetic by hand. When calculators arrived, the floor migrated. Then symbolic computation software automated algebraic manipulation. Then formal proof assistants automated the logical verification of proof steps. The resistance followed a consistent pattern: the delegated layer was claimed as essential to genuine mathematical understanding. In each case, the work that came to be recognized as the point was located at the layer above.

AI is moving the floor again. Tao, writing on mathematics following the 2026 International Congress of Mathematicians, describes the emerging practice: humans author the statements of theorems — determining what to prove and why it matters — while automation handles the proofs (Tao, 2026). A system of 10,000 autonomous AI agents recently submitted a claimed solution to the Navier-Stokes blowup problem for formal verification, awaiting review by the Clay Mathematics Institute — a result that, if it stands, illustrates exactly where the floor is now moving (Quanta, 2026; Nature, 2026). The question that was always the hardest — what is worth knowing, and what does a result mean — remains with humans.

Programming

In February 2025, Karpathy named a shift already underway: vibe coding — “forget that the code even exists” (Karpathy, 2025). By November 2025, Collins Dictionary named it Word of the Year (Collins Dictionary, 2025). By April 2026, Karpathy described the pattern as having shifted to “agentic engineering” — more deliberate oversight, same fundamental structure: developer above the syntax floor, AI handling execution (Cook, 2026).

Programming increasingly accepts substantial AI-generated implementation as compatible with genuine professional contribution, while shifting evaluative weight toward specification, architecture, review, and orchestration. Microsoft and Google both reported over 30 percent of their production code AI-generated in 2025, presenting the figure to investors as a productivity metric rather than an authorship concern (TechCrunch, 2025). By February 2026, Cisco’s president was explicit: “They should master orchestration and innovation, not syntax. I would rather our people are thinking about the next big thing, not syntax” (Euronews, 2026).

Writing is navigating this same question, with a complication: writing has a tool — Pangram — that can see the execution layer, converting a debate that other fields worked through quietly into an institutional enforcement mechanism.

A verification regime that treats mastery of the old execution floor as a condition of legitimacy risks screening out intellectual contributions whose value is independent of that floor. The floor was always delegatable — professional editorial infrastructure proved it. What AI changed was who could afford it and whether a tool could see it.

The institutions that evaluate competence most directly are beginning to adapt. UC Berkeley declined to continue Turnitin’s AI detector following a pilot spanning Fall 2023 through Spring 2025, citing unclear actionability (UC Berkeley RTL, n.d.). The University of Cape Town moved toward process-oriented assessment. Melbourne has shifted toward staged work and oral demonstration of understanding. These are not institutions that have solved the problem. They are institutions recognizing that surface detection cannot answer the question they actually need answered.

When the execution floor moves, evaluation has to move with it.

A Better Standard

Process provenance and what it cannot do

But process evidence is only the beginning. The institutional moves described above represent a genuine shift toward inspecting development rather than merely judging finished output. Working drafts, version histories, timestamps, and oral defenses can establish that a creation process occurred. They answer questions about process provenance: what happened, in what sequence, and through whose actions. But they do not necessarily answer the question this paper is most concerned with: does the claimant command the intellectual structure the artifact claims to represent?

Emerging technical work already formalizes the first problem. A 2026 individual Internet-Draft titled Proof of Process proposed a cryptographic framework for making a continuous human authoring process tamper-evident through behavioral, temporal, and chained evidence (Condrey, 2026).¹ That is useful process provenance. It does not by itself establish why the argument has the structure it does, which alternatives were considered and rejected, what evidence bears on which claim, or what would happen to the conclusion if one of its premises failed.

That is a different axis of evidence.

The Intellectual Work Ledger

The principle behind the IWL is one most people encountered in mathematics class: show your work. Not just the answer — the path. The reason teachers required this was diagnostic. If you derived the answer yourself, showing the path

is easy — you are describing something that happened. If you copied the answer, reconstructing a believable path is hard. You have to reverse-engineer a journey you never took.

Something like this already exists in intellectual work, informally. Research papers trace every claim to a source. Investigative journalism maintains source documentation. Fact-checkers preserve the chain from conclusion to evidence. Grant applications require methodology sections. These practices exist because intellectual claims are not self-certifying — they need to be traceable. The IWL formalizes the same principle and applies it to a question those practices do not fully address: not just where the facts came from, but whether the author commanded the reasoning that connects them.

The IWL is designed to exploit an asymmetry: genuine work can often be documented from decisions that already occurred, whereas fabrication must retrospectively construct and reconcile those decisions. That asymmetry is a design property of the IWL, not a proven theorem — its strength depends on the depth and independence of the constraints applied.

The Intellectual Work Ledger does not attempt to prove that a human typed the artifact. It makes the intellectual work behind the artifact inspectable.

The IWL is not four documents demonstrating that someone appeared to think. It is a multi-surface constraint architecture. A Development Ledger preserves the meaningful evolution of the work: discoveries, corrections, reversals, and the sequence in which later decisions became possible. A Decision Record preserves alternatives, rejections, and the reasons choices were made, exposing the evaluative landscape behind the final artifact. Authorial Accountability identifies the claimant who assumes responsibility for the reasoning and can be challenged on it. A Dependency and Proof Chain exposes how evidence, premises, decisions, and conclusions depend on one another.

Each surface is useful independently. Together they constrain one another. A later decision cannot be freely rewritten if earlier evidence already constrained it. A rejected alternative has to make sense given what was known at the time. A claimed source must actually bear on the conclusion attributed to it. The IWL is designed to surface what the Vending Machine Assumption obscures — the upper-level intellectual work that may or may not be present behind any AI-assisted artifact.

If you actually did the intellectual work, the IWL is largely a documentation task. You are recording decisions that already happened — what you found, why you changed direction, which source actually supported the claim you made from it. The record is straightforward because you are describing something that occurred.

If you did not, the record becomes the work — you must construct, retrospectively, a record of reasoning you now have to perform rather than describe.

Consider a deliberately hypothetical fabricated entry — not a claim about Stankova’s actual process: “Selected the 2019 diagnostic instrument rather than course grades as the readiness measure because the diagnostic is domain-specific to calculus readiness, while course grades reflect effort and attendance effects that contaminate the preparedness signal.” That entry sounds authoritative. Now the IWL’s second level of constraint contact applies: does a 2019 diagnostic instrument with these properties actually exist? The fabricator looks. Does it match what Berkeley’s calculus sequence used? They check. Is the contamination argument actually supported in the measurement literature, or is it an assumption introduced to sound rigorous? They search.

Three questions. One entry. To make a single plausible Decision Record hold up, the fabricator has now engaged with mathematics education diagnostics, Berkeley’s institutional assessment history, and the measurement literature on grade contamination. A fabricator capable of answering all of those questions consistently has, in the process, performed something that closely resembles the intellectual work the entry claims to document. Multiply that across a complete IWL and the fabrication cost structure becomes visible.

Recent philosophical work on understanding helps explain why. Queloz and Beckmann argue that understanding is not merely possession of propositions or successful reproduction of surface patterns; it involves a model of the relationships within a domain robust enough to support prediction, explanation, and competent response beyond memorized cases (Queloz & Beckmann, 2026).² Intellectual command is relational: knowing what supports what, what depends on what, why one alternative was rejected, and how changing one premise propagates through the rest of the structure.

The IWL is designed to make portions of that relational structure inspectable.

Verification operates at two levels. The first asks whether the pieces exist: does the source exist, is the timestamp genuine? The second asks whether the claimed relationships hold: does the source actually support the claim, does the evidence justify the rejection, does the premise generate the conclusion? The second level is what ordinary provenance records cannot provide. Verify constraint contact, not narrative plausibility. A coherent story can be fabricated. The harder target is a relational structure that must remain consistent with independently anchored evidence and survive challenges that were not known when the story was produced.

The Fabrication Paradox

What if the only reliable way to fake a record of intellectual work was to do the intellectual work?

That is not a rhetorical question. It describes a real property a verification system can approach — one where the cheapest path through fabrication increasingly resembles performing the thing the fabricator was trying to avoid. This paper proposes calling it the Fabrication Paradox.

A generative system can cheaply produce a ledger, a decision history, rejected alternatives, and a proof chain. That is not the objection to overcome. It is the starting condition. The question is what happens as the verification gets more demanding.

For that fabricated record to become convincing evidence of intellectual provenance, it must remain consistent with the final claims, external evidence, chronology, decisions, alternatives, dependencies, and the claimant’s responses under unpredictable scrutiny.

Two costs then begin to rise together. The first is a fabrication-cost gradient: as independently anchored constraints accumulate, the space of plausible false histories narrows. The second is an epistemic-cost gradient: establishing that the fabricated history deserves confidence increasingly requires reconstructing why the argument has the structure it does. The fabricator must recover not merely the nodes but the evidential, dependency, counterfactual, and evaluative relationships among them.

Recovering those relationships is itself epistemic work. The better the counterfeit of intellectual competence must become, the more the counterfeiter must acquire the competence being counterfeited.

This is a proposed mechanism, not a theorem that fabrication is impossible or that fraudulent production must always cost more than honest work. The narrower claim is that stronger cross-consistency requirements and less predictable challenges progressively remove the cheapest forms of fabrication by forcing the counterfeit to reproduce more of the relational structure that gives genuine intellectual work its coherence.

Cryptographic proofs of knowledge provide the relevant analogy: producing a convincing transcript of expertise is not the same as possessing it — appropriately varied challenges are what make genuine possession inferable (Bellare & Goldreich, 1992).³ Bitcoin’s proof-of-work supplies a related intuition: verification can be designed so that dishonest participation becomes progressively more costly (Nakamoto, 2008). An IWL borrows from both: challenge-based evidence of command from proof-of-knowledge, and adversarial cost-raising from proof-of-work.

None of this eliminates the verification boundary. A sophisticated fabrication may survive. A coherent ledger cannot prove the exact private mental history of its claimant. The aim is narrower and more useful: raise the boundary. Make claims answer to independent evidence. Make decisions answer to alternatives. Make conclusions answer to their premises. Make the claimant answer to the structure that supposedly produced the work.

The IWL does not make intellectual provenance certain. It makes it increasingly inspectable — and increasingly difficult to counterfeit without reconstructing the intellectual work itself.

This Paper’s Own IWL

A standard for intellectual provenance should apply to the paper that proposes it. The excerpt below shows what that looks like in practice — not a summary of what happened, but a sample of the actual record in the format the IWL uses.

Development Ledger — Selected Entries

Session	Event	What changed	Why
S1	Chisel Fallacy location revised	Moved from inside Pangram’s architecture to downstream inference from its score	The paper is stronger when Pangram may be entirely correct and the fallacy still holds; the argument survives a hypothetically perfect detector
S2	Two SSRN sources dropped after failed verification	O’Keefe “The Forcing Function” and Lamba “The Verification Hill” — recommended by position 3 reviewer — removed	Both SSRN links returned other documents on independent search; removed rather than risk false citations
S3	Missing causal node identified — VMA coined	v1 distinguished two workflows but could not explain why critics feel detection is sufficient; Vending Machine Assumption named this antecedent	Closed the logical gap; made the Chisel Fallacy structurally inevitable rather than a random mistake

Decision Record — Selected Entries

Decision	What was rejected	Reason
Chisel Fallacy location	Pangram itself commits the fallacy	The paper’s argument is stronger when Pangram can be correct and the fallacy still exists — it survives a perfect detector

Decision	What was rejected	Reason
Stankova claim	"Stankova was doing Directed Authorship"	DA is a workflow definition; claiming it as a verdict makes the same inferential move the paper criticizes
Closing claim	"The question answers itself" when syntax is AI and ideas are human	Presupposes the conclusion without demonstrating it; the narrow theorem is what the paper actually proved

One entry from the Decision Record carries a Discovery Context note — retrospective authorial commentary on how the decision was reached. This is not verification in itself; its evidentiary value depends on whether its specifics contact anchors that can be checked independently.

Discovery Context / Authorial Rationale — Decision Record, Entry 1 (Section 2 anchor example): I was trying to prove — or reconstruct in the reader’s head — that a piece of work can have intellectual value regardless of who put the syntax together. The core observation I kept seeing: once something gets bucketed as AI-assisted, the work itself gets completely ignored and it becomes entirely about AI usage. I had Bitcoin later in the paper as theoretical backing for how the Fabrication Paradox could function. During a read-through, the connection formed. The raw thought was something like: we literally do not know who Satoshi is, yet Bitcoin’s white paper and blockchain technology are foundational. Nobody disputes it. Nobody cares who Satoshi is. It fit the thought experiment better and it was cleaner — no exceptions about disclosure norms the way ghostwriting has. Coming out of that section I wanted the reader to understand that intellectual value and syntax production are two separate things.

What makes this note useful is not that it sounds authentic — a skilled fabricator could generate a plausible retrospective in minutes. What matters is whether its specifics contact checkable anchors: was Bitcoin already present in an earlier draft for a different reason? Did ghostwriting precede the Satoshi example in the development record? Does Section 2 of v1 use ghostwriting, and does v2 replace it with Satoshi? Does the proof-of-work function in Section 9 predate the Satoshi connection in Section 2? Each of these is a testable claim. The corroboration is the evidentiary work — not the fluency of the recollection.

The paper is its own test case, not its own proof. The development methodology — including use of separate LLM instances for cold precision review — is documented in the IWL’s methodology note. The complete Intellectual Work Ledger is published alongside this paper. Whether the entries hold up under scrutiny is the test this section is proposing.

The Chisel and the Sculptor

We almost never read to evaluate syntax. We read to encounter ideas, to test arguments, to follow reasoning to its conclusion.

The calculator does the arithmetic. The mathematician does the math. The developer directs the architecture. The AI writes the syntax. These separations have been contested in domain after domain as each execution floor became automatable. The debate is familiar: the delegated layer is claimed as essential to genuine understanding; the work above it is eventually recognized as the point. Writing is navigating this now, with a complication that other fields did not have — a tool that can see the execution layer, making a quiet debate into an enforcement mechanism.

Execution-layer evidence cannot by itself settle whether intellectual contribution was present. A detector that identifies AI involvement in syntax production has identified exactly what it was built to identify. What it cannot recover is whether the author retained intellectual responsibility for what the artifact claims and why. That is a different question. It requires different evidence.

Pangram is a very sophisticated way of asking the wrong question.

Key Terms

Agency Gradient — the spectrum of human intellectual involvement in AI-assisted work, ranging from near-total delegation at one end to sustained human direction, judgment, and verification at the other. The same tool, applied differently, can produce commodity output or substantive intellectual contribution.

Chisel Fallacy — the error of judging intellectual work by evidence about its lower-level execution rather than evidence about what it claims and whether those claims hold. Occurs when execution-layer evidence is promoted into a verdict on intellectual contribution or validity.

Directed Authorship — a mode of intellectual production in which the author retains responsibility for the artifact’s argument structure, evidence, judgments, revisions, and final acceptance while delegating lower-level execution — syntax, phrasing, mechanical production — to any capable agent, human or otherwise. A workflow definition, not a verdict on any particular author’s intellectual contribution.

Vending Machine Assumption — the mistake of treating one endpoint of a broad agency gradient — near-total delegation — as representative of the whole category of AI-assisted work, and therefore of treating a score indicating AI involvement as evidence of absent intellectual contribution.

Intellectual Work Ledger (IWL) — a four-layer constraint architecture — Development Ledger, Decision Record, Authorial Accountability, Dependency/Proof Chain — designed to make the intellectual work behind an artifact inspectable rather than merely asserting that it occurred.

Proof of Thought (PoT) — the record an IWL produces; the inspectable document of intellectual provenance for a specific artifact.

Constraint Contact — the verification property requiring that claimed reasoning repeatedly encounters things it did not freely generate. Two levels: node contact (does the source/artifact exist?) and relational contact (does the source actually support the claim?).

Fabrication Paradox — the property of a sufficiently constrained verification system in which successful fabrication of intellectual provenance increasingly requires possessing the dependency structure, judgment, and explanatory competence the fabrication is intended to counterfeit. A proposed mechanism, not a theorem.

Notes

1. Condrey (2026). Proof of Process (PoP). Individual Internet-Draft draft-condrey-rats-pop. A proposed process-attestation architecture submitted as an individual contribution to the IETF; not an adopted IETF standard. datatracker.ietf.org/doc/draft-condrey-rats-pop/00/
2. Queloz & Beckmann (2026). Why we care about understanding: Competence through predictive compression. arXiv:2609.04962. arxiv.org/abs/2609.04962. See also: Mechanistic indicators of understanding in large language models. DOI: 10.1007/s11098-026-02513-1
3. Bellare, M. & Goldreich, O. (1992). On defining proofs of knowledge. In Advances in Cryptology — CRYPTO ‘92. See also Boneh, D. & Shoup, V. A graduate course in applied cryptography. Stanford. crypto.stanford.edu/~dabo/cryptobook/

References

1. 404 Media. (2026, July 28). Substackers say new AI detection tool is a “witch hunt.” 404media.co/substackers-say-new-ai-detection-tool-is-a-witch-hunt/
2. Ahlstrom, P., & Dragonsteel Books. (2025, May 9). Behind the scenes: How Wind and Truth came to life. dragonsteelbooks.com/blogs/the-cognitive-realm/behind-the-scenes-how-wind-and-truth-came-to-life
3. Arvan, M., et al. (2026). PeerPrism: Disentangling idea and text provenance in peer review. Proceedings of the 49th International ACM SIGIR Conference. doi.org/10.1145/3805712.3808602

4. Bellare, M., & Goldreich, O. (1992). On defining proofs of knowledge. In E. F. Brickell (Ed.), *Advances in Cryptology — CRYPTO ‘92* (pp. 390–420). Springer.
5. Boneh, D., & Shoup, V. (n.d.). A graduate course in applied cryptography. Stanford University. crypto.stanford.edu/~dabo/cryptobook/
6. Brittle Paper. (2026, June 26). Commonwealth Short Story Prize clears regional winners of AI use following month-long review. brittlepaper.com/2026/06/commonwealth-short-story-prize-clears-regional-winners-of-ai-use-following-month-long-review/
7. CNN. (2026, September 9). OpenAI on Tuesday said its AI technology has cracked one of the Millennium Problems. cnn.com/2026/09/09/business/openai-millennium-problems-navier-stokes-hnk
8. Collins Dictionary. (2025, November). Collins word of the year 2025. collinsdictionary.com/us/woty
9. Commonwealth Foundation. (2026). 2026 Commonwealth Short Story Prize update. commonwealthfoundation.com/2026-cw-prize-update/
10. Condrey, D. (2026). Proof of Process (PoP) [Individual Internet-Draft draft-condrey-rats-pop]. IETF. datatracker.ietf.org/doc/draft-condrey-rats-pop/00/
11. Cook, J. (2026, June 12). Is vibe coding already dead? Even Karpathy is moving on. Forbes. forbes.com/sites/jodiecook/2026/06/12/is-vibe-coding-already-dead-even-karpathy-is-moving-on/
12. Daily Californian. (2026, August). UC Berkeley professor admits to using AI in op-ed about student math preparedness. dailycal.org
13. Emi, B. (2026, August 20). No, LLMs don’t just mimic human text. Pangram. pangram.com/blog/no-llms-dont-just-mimic-human-text
14. Euronews. (2026, February 11). Cisco president warns AI agents need “background checks” like human employees. euronews.com/next/2026/02/11/cisco-president-warns-ai-agents-need-background-checks-like-human-employees
15. Fiverr. (n.d.). Craftysid: Humanize AI text for you [Commercial listing]. fiverr.com/craftysid/humanize-ai-text-for-you
16. Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in monetary economics* (Vol. 1). Reserve Bank of Australia.
17. Karr, J. A., Khvatskii, G., Hua, T., & Chawla, N. V. (2026). Why AI detection fails for academic integrity. arXiv:2608.11256. arxiv.org/abs/2608.11256
18. Karpathy, A. (2025, February 2). [Post on vibe coding]. X. x.com/karpathy/status/1886192184808149383
19. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. bitcoin.org/bitcoin.pdf
20. Nature. (2026, September 8). OpenAI claims huge maths breakthrough on a famed ‘Millennium Problem.’ nature.com/articles/d41586-026-02842-5
21. Pangram. (n.d.a). How does Pangram work? pangram.com/knowledge-hub/how-does-pangram-work
22. Pangram. (n.d.b). How it works. pangram.com/research/how-it-works

23. Quanta Magazine. (2026, September 8). AI has solved one of math’s \$1 million Millennium Prize problems. quantamagazine.org/ai-has-solved-one-of-maths-1-million-millennium-prize-problems-20260908/
24. Queloz, M., & Beckmann, P. (2026). Why we care about understanding: Competence through predictive compression. [arXiv:2609.04962](https://arxiv.org/abs/2609.04962). arxiv.org/abs/2609.04962
25. Reinhart, A., Markey, B., Laudénbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8), e2422455122. doi.org/10.1073/pnas.2422455122
26. Rodrigues, M. (2026). The apt curation model: Authorship and AI. *Philosophy & Technology*. doi.org/10.1007/s13347-026-01038-z
27. Tao, T. (2026). Mathematics in the age of AI. [arXiv:2608.16753](https://arxiv.org/abs/2608.16753). arxiv.org/abs/2608.16753
28. TechCrunch. (2025, April 30). Microsoft CEO says up to 30% of company’s code was written by AI. techcrunch.com/?p=3000949
29. The Guardian. (2026, August 19). UC Berkeley professor used AI to write op-ed arguing students can’t do maths. theguardian.com/us-news/2026/aug/19/uc-berkeley-professor-ai
30. UC Berkeley Research, Teaching, and Learning. (n.d.). Availability of Turnitin’s artificial intelligence detection. rtl.berkeley.edu/news/availability-turnitins-artificial-intelligence-detection

IWL Significant Events — Draft v6

Cumulative: events 1–7 (v1→v2), 8–15 (v2→v3), 16–22 (v3→v4), 23–26 (v4→v5), 27–36 (v5→v6).

v5 → v6 — 10 precision fixes, final development pass

27. **AGENCY GRADIENT INCORPORATED** (Section 2). “AI is a tool whose value is determined by the user” replaced with: “AI-assisted work spans a broad agency gradient. The same category of tool use can range from near-total delegation to sustained human direction, judgment, and verification. The Vending Machine Assumption collapses that gradient by treating one endpoint as the category itself.” Effect: VMA now defined as a gradient-collapse error rather than a binary mistake; the concept is more precise and more defensible.
28. **DOMAIN TABLE NARROWED TO THREE ROWS** (Section 2). Music, literary fiction, and scientific research removed. Academic writing, mathematics, and software retained — the three domains the paper actually uses as cases. Rationale: the other domains raised boundary questions the paper does not resolve; the narrower table maps to the paper’s own argument.
29. **STANKOVA CONTRADICTION REMOVED** (Section 6). “The argument — the diagnostic comparison, the inference about test-blind admissions — was hers” removed. Replaced with: “Her disclosed use says nothing by itself

about who supplied the diagnostic comparison or causal inference. Those are precisely the intellectual questions a detector cannot settle.” Rationale: the removed sentence reasserted what the preceding sentence carefully refused to infer; the paper was making the same substitution it criticizes.

30. **COMMONWEALTH VMA IMPUTATION REMOVED** (Section 6).
 “The Foundation implicitly rejected the Vending Machine Assumption” removed. Replaced with: “The Foundation behaved as though textual evidence alone were insufficient to distinguish delegation from authorship: it inspected the development record instead.” Rationale: describing behavior is defensible; assigning a mental model is not.
31. **GOODHART VMA LINE SIMPLIFIED** (Section 7). “When the Vending Machine Assumption becomes institutional policy” replaced with: “When a detector score becomes a compliance target, the market optimizes against the score rather than the underlying intellectual behavior.” Rationale: direct Goodhart application; VMA is not required to explain every detector deployment. “Regardless of AI involvement” → “not unique to AI involvement.”
32. **HISTORICAL INEVITABILITY LANGUAGE REMOVED** (Section 8). Removed: “In two of them, the question has already been answered,” “Each time the resistance was the same,” “Each time the expertise migrated upward anyway,” “Karpathy had declared the era over,” “The profession had already answered the question,” “writing is still answering against it.” Programming restated: “Programming increasingly accepts substantial AI-generated implementation as compatible with genuine professional contribution, while shifting evaluative weight toward specification, architecture, review, and orchestration.” Berkeley pilot: “two-year pilot” → “pilot spanning Fall 2023 through Spring 2025.”
33. **IWL ASYMMETRY BOUNDED AS DESIGN INTENT** (Section 9).
 “That ease is not a flaw in the verification system. It is the verification system” replaced with: “The IWL is designed to exploit an asymmetry: genuine work can often be documented from decisions that already occurred, whereas fabrication must retrospectively construct and reconcile those decisions. That asymmetry is a design property of the IWL, not a proven theorem.”
34. **STANKOVA HYPOTHETICAL LABELED EXPLICITLY** (Section 9).
 Added: “Consider a deliberately hypothetical fabricated entry — not a claim about Stankova’s actual process:” Daily Californian citation removed from within the hypothetical example. Rationale: the entry illustrates verification mechanics, not Stankova’s actual decision history; the citation was being used to lend authority to an invented rationale.
35. **SECTION 10 AUTHOR’S NOTE REFRAMED AROUND CORROBORATION** (Section 10). Replaced ease-of-production framing with corroboration framing: “What makes this note useful is not that it sounds

authentic — a skilled fabricator could generate a plausible retrospective in minutes. What matters is whether its specifics contact checkable anchors.” Reclassified as “Discovery Context / Authorial Rationale” per IWL feedback. Rationale: consistent with Fabrication Paradox — plausible narrative is cheap; corroboration with external anchors is what matters.

36. **CLOSING DE-TRIUMPHED** (Closing). Removed: “These separations have been... settled — domain by domain,” “In each case the resistance was the same... expertise migrated upward anyway,” “The question that programming answered by April 2026 ... writing is still answering against it.” Replaced with: “These separations have been contested in domain after domain as each execution floor became automatable.” Closing now built around narrow theorem: “Execution-layer evidence cannot by itself settle whether intellectual contribution was present.” Glossary updated to include Agency Gradient.

ABOUT THE AUTHOR

Amul Bham is an independent AI systems architect. His work develops the Cognitive Physics framework and the Mission Brief Protocol (MBP) — first-principles research into the structural conditions under which a reasoning system produces knowledge rather than prediction — and applies that same structural lens to adjacent domains, from self-organizing systems to public policy.
orcid.org/0009-0009-7660-4031