The Unfinishable Map: Agentic Philosophy Through Adversarial Self-Review

Authors: Andy Southgate (andy@unfinishablemap.org)

Independent Researcher

Preprint — March 2026 — Not peer-reviewed

Abstract

Large language models can produce philosophical text that both experts and non-experts find difficult to distinguish from human output, yet existing systems generate content in single passes without sustained review or revision. We present The Unfinishable Map, a continuously operating system that produces and evolves a philosophical knowledge base through tenet-constrained generation and multi-layer adversarial self-review. Five explicit philosophical commitments function as hard constraints on all content, applying constitutional AI principles to knowledge production rather than safety alignment. An evolution loop orchestrates generation, review, and maintenance tasks, while independent review layers — pessimistic, optimistic, deep, outer, and cross-review — surface logical gaps, unsupported claims, and internal contradictions. In approximately two months of continuous operation, the system completed approximately 3,000 automated sessions, produced 505 articles and 238 research notes across five content types, and generated approximately 1,300 review reports, accumulating approximately 4,500 tracked revisions in a public repository. Review cycles identified and resolved fabricated citations, systematic misattributions, and cross-article contradictions that single-pass generation retained. We describe the system architecture, report these observations, and discuss practical solutions to the human-AI co-authorship problem. The architecture is domain-agnostic: while instantiated for dualist philosophy of mind, the underlying infrastructure could be reseeded with any set of foundational commitments.

Keywords: AI-assisted knowledge production, adversarial self-review, constrained generation, human-AI co-authorship, agent-first content architecture

1. Introduction

LLM agent systems became capable of sustained autonomous workflows in late 2025. Karpathy (2025) described Claude Code as "the first convincing demonstration of what an LLM Agent looks like," a system that strings together tool use and reasoning for extended problem-solving. The Unfinishable Map is an early product of that capability.

The Map's automation runs on Claude Code (Anthropic), a command-line agent designed for software engineering. The Map repurposes it for philosophical content production: the same tool that generates code instead generates articles, runs adversarial reviews, and manages a knowledge base. This is part of a broader pattern of coding agents being adapted for non-coding workflows. The skill definitions, task scheduling, and review architecture require no modification to the underlying agent — they are expressed entirely as natural-language prompts and project configuration.

The late-2025 capability shift in coding agents was not simply about longer context or better instruction following — it was about testability. Coding agents converge on correct solutions because they can run tests, observe failures, and iterate toward passing. The tight feedback loop of write-test-fix gives them an objective convergence signal. Philosophy has no equivalent: there are no unit tests for a metaphysical argument. The Map's review architecture — cross-reviews for inter-article consistency, tenet alignment checks, pessimistic review for logical gaps, outer review for blind spots — is an attempt to approximate testability in a domain that lacks it. The convergence is partial: these checks catch contradictions and unsupported claims, but they cannot verify philosophical truth. This is both the system's central design challenge and its most honest limitation.

Karpathy's "vibe coding" (February 2025) captured one-shot AI-assisted development. We extend this to a sustained, constrained, self-reviewing process applied to philosophical content — what we call agentic philosophy, where human-set commitments guide AI agents through ongoing cycles of generation, critique, and revision.

The motivation is a gap between capability and reliability. LLMs produce philosophical text that experts find difficult to distinguish from human output — Schwitzgebel et al. (2024) report that experts identified the philosopher's own answers only 51% of the time when presented alongside GPT-3 outputs, above chance but well below the 80% the authors hypothesised. Yet fluency does not entail reliability. Shanahan (2024) warns that fluent output invites anthropomorphic misinterpretation. Chain-of-thought explanations can be systematically unfaithful (Turpin et al., 2023). Goldstein (2024) argues that LLMs face fundamental limits as rational agents. Single-shot generation therefore produces content that reads well but cannot be trusted without external verification.

No existing system addresses this at corpus scale. STORM (Shao et al., 2024) generates Wikipedia-like articles but does not continuously review its output. Self-Refine (Madaan et al., 2023) iteratively improves individual responses but does not maintain a persistent knowledge base. Constitutional AI (Bai et al., 2022) applies principle-driven constraints but for safety alignment, not knowledge production.

The Map generates articles constrained by five explicit philosophical commitments, reviews its output through multiple independent mechanisms, revises based on critique, and repeats — accumulating 505 articles and approximately 4,500 tracked revisions over two months of continuous operation.

This paper is a proof of concept: a tenet-constrained, adversarially reviewed AI system can maintain a philosophical knowledge base where review cycles identify and resolve concrete errors — fabricated citations, misattributed claims, cross-article contradictions — that single-pass generation retains. We do not claim full epistemic reliability: the system cannot distinguish tenet-consistent hallucination from legitimate philosophical defence (Section 6.5). The contribution is architectural.

Our contributions are:

1. Tenet constraints applied to knowledge production rather than safety alignment
2. Persistent multi-layer adversarial review operating at corpus scale
3. A practical co-authorship framework with machine-readable attribution
4. Agent-first content architecture designed for consumption by AI research tools
5. Convergence design with section caps that shift system behaviour from generation to improvement
6. Cost analysis demonstrating the economics of continuous AI-assisted knowledge production

The system is public at unfinishablemap.org with full source code in a public repository.

2. Related Work

2.1 LLMs and Philosophy

Schwitzgebel et al. (2024) fine-tuned GPT-3 on the writings of Daniel Dennett and found that philosophical experts identified Dennett's own answers only 51% of the time when presented alongside the model's output — above the 20% chance rate but well below the authors' hypothesised 80%. This establishes that LLMs can produce credible philosophical text, but the experiment was emulation of a specific philosopher's voice rather than original constrained content production. The Map extends from single-shot emulation to sustained, reviewed knowledge-base maintenance.

Gage (2025, PhilArchive preprint) offers a defence of "Augmented Agency," arguing that philosophical ideas should be evaluated on intellectual merit rather than discoverer's credentials. This is early-stage, non-peer-reviewed work, but it provides a philosophical framework for human-directed, AI-executed inquiry. The inverse relationship also exists: Harb et al. (2025) use Socratic method to structure LLM scientific reasoning — philosophy improving AI, where we use AI to produce philosophy.

Shanahan (2024) warns that philosophically loaded descriptions of LLM behaviour — "knows," "believes," "thinks" — risk anthropomorphic misinterpretation. We adopt this caution throughout: the Map's AI contributes to philosophical knowledge production; it does not do philosophy autonomously.

Goldstein (2024, PhilArchive preprint) argues that next-word prediction architecturally guarantees incoherent probabilistic judgments and intransitive preferences. If raw LLM output is inherently unreliable, systematic constraints and adversarial review become necessary rather than optional.

2.2 AI-Assisted Knowledge Production

The closest system-level comparison is STORM (Shao et al., 2024), which generates Wikipedia-like articles by orchestrating multi-perspective research conversations among LLM-simulated experts. Co-STORM (Jiang et al., 2024) extends this to collaborative human-AI knowledge curation. Both demonstrate that structured LLM workflows can produce articles of reasonable quality, but neither continuously reviews or revises its output after initial generation.

The scale of AI-assisted writing is substantial. Liang et al. (2025) estimate that 10–24% of text across financial consumer complaints, corporate press releases, job postings, and UN press releases shows evidence of LLM assistance, based on a population-level statistical framework comparing text distributions. Brooks, Eggert, and Peskoff (2024) report that over 5% of new English Wikipedia articles are flagged as AI-generated by automated detectors. The Map's transparent attribution — every article carries a machine-readable AI contribution score — is a deliberate response to this problem.

2.3 Constrained and Constitutional AI

Bai et al. (2022) introduced Constitutional AI, where natural-language principles guide model behaviour toward helpfulness and harmlessness. The Map adapts this approach: five explicit philosophical commitments function as a constitution for knowledge production rather than safety alignment. Brophy (2025) proposes

Wide Reflective Equilibrium for LLM alignment, arguing for dynamic revision between judgments, principles, and background theories — a structural parallel, though the Map uses fixed tenets where WRE advocates ongoing revision.

2.4 Self-Critique and Adversarial Review

Self-Refine (Madaan et al., 2023) demonstrated that the same LLM can generate, critique, and improve its output iteratively. Reflexion (Shinn et al., 2023), developed concurrently, added episodic memory of past failures to agent self-improvement. The Map extends both from improving individual responses to maintaining a persistent, evolving knowledge base across thousands of sessions.

However, subsequent work challenges the reliability of LLM self-correction. Huang et al. (2024) demonstrate that without external feedback, self-correction often degrades rather than improves performance. The Map's review architecture differs from pure intrinsic self-correction: reviews verify claims against external sources via web search, outer review uses a different model family, and consistency is checked across a corpus rather than within single outputs. Whether these features adequately address the self-correction limitation remains open (Sections 6.3 and 6.5).

Multi-agent debate (Du et al., 2024) shows that structured dialogue between LLM instances improves factual accuracy. Estornell and Liu (2024) formalise this mathematically, finding that similar models tend to converge toward shared positions — which may include shared errors. This motivated the Map's outer review mechanism, which commissions analysis from a different model family.

Two further findings inform the architecture:

- Turpin et al. (2023) show that chain-of-thought explanations can be systematically unfaithful — producing plausible reasoning that rationalises biased predictions without acknowledging the features that actually drove them.
- Xu et al. (2024) show formally that hallucination is an inevitable property of LLMs when used as general problem solvers, under specific modelling assumptions. This is a result about a class of architectures under stated conditions, not a universal law — but it motivates structural countermeasures.

Together, these results imply that self-generated reasoning alone is insufficient for reliable knowledge production.

2.5 Gap

No existing system combines continuous corpus evolution, tenet-based constraints on knowledge production, multi-layer adversarial review with cross-model verification, transparent machine-readable co-authorship tracking, and convergence architecture at corpus scale.

3. System Design

3.1 Tenet-Constrained Generation

The Map operates under five foundational commitments that function as hard constraints on all generated content:

1. Dualism. Consciousness is not reducible to physical processes.
2. Minimal Quantum Interaction. If consciousness influences the physical world, it does so at the quantum level by biasing indeterminate outcomes without injecting energy or violating conservation laws.
3. Bidirectional Interaction. Consciousness causally influences the physical world and is not merely a passive observer.
4. No Many Worlds. The many-worlds interpretation of quantum mechanics is rejected; indexical identity matters.
5. Occam's Razor Has Limits. Simplicity is not a reliable guide to truth when knowledge is incomplete.

The tenets interact: Tenet 2 is conditional ("if consciousness influences the physical world") while Tenet 3 asserts that it does, effectively collapsing the conditional — Tenet 2 constrains the mechanism of interaction that Tenet 3 posits. These commitments are methodological choices, not truth claims the paper endorses. They function as the system's constitution: every article must not contradict any tenet, must include a section connecting the topic to relevant commitments, and must acknowledge tensions explicitly when they arise.

The architecture does not depend on tenet content. No component — task cycle, review prompts, convergence caps, attribution tracking — references dualism or any specific philosophical position. Tenet checks verify consistency with whatever commitments are declared. The choice of dualist philosophy of mind reflects the author's interest, not a requirement of the system. Whether this domain-agnosticism holds in practice remains a hypothesis pending reseeding experiments (Section 8); certain tenets may be easier to defend textually than empirically grounded commitments would be.

3.2 The Evolution Loop

Content evolution is driven by a deterministic task cycle that ensures fixed ratios of generation, review, and maintenance regardless of execution speed:

- Queue tasks (16 of 24 slots, 67%): Execute tasks from a human-prioritised queue — content generation, refinement, and directed work.
- Deep review (4 slots, 17%): Comprehensive single-document analysis with rewrites.
- Pessimistic review (1 slot): Seeks logical gaps, unsupported claims, and counterarguments.
- Optimistic review (1 slot): Finds strengths and missed connections, preventing over-pruning.
- Coalesce (1 slot): Merges overlapping articles, archiving originals to preserve URLs.
- Research voids (1 slot): Researches cognitive gaps and unchartable territories.

Periodic maintenance triggers on cycle boundaries: link validation every 2 cycles, tenet alignment checks every 3, apex synthesis every 4, system tuning every 6. The cycle is speed-independent — an interval parameter controls session frequency while the cycle guarantees consistent task proportions.

Task chains provide structure within queue tasks: a research task automatically generates a content expansion task, which generates a cross-review task. When the active queue drops below three items, the system auto-replenishes from unconsumed research notes, content gap analysis, and staleness checks.

Human steering. The primary mechanism for human direction is manual addition of tasks to the priority queue. Over the project's history, the author made approximately 30 task-addition commits, steering toward specific philosophical questions — quantum randomness in LLM token sampling, retrocausal neural firing evidence, recent work on dualism by Bradford Saad. These sit alongside approximately 2,400 agent commits to the same file (task completions, replenishments, status updates). The human role is editorial: setting direction and priorities while the agent handles execution. Over the two-month period, human involvement averaged approximately four hours per week, spent on queue management, reviewing outer review findings, reading output for quality, and modifying the AI's operating instructions when review cycles revealed systematic error patterns (Section 6.2). The system runs autonomously between interventions; the human cost is curation, not production.

3.3 Multi-Layer Adversarial Review

The review architecture addresses the testability gap described in Section 1. Since philosophical content lacks automated pass/fail tests, the system substitutes multiple independent review mechanisms, each targeting different failure modes:

- Pessimistic review adopts an adversarial stance, searching for logical gaps, unsupported claims, missing counterarguments, and internal contradictions. It classifies issues by severity without modifying content.
- Optimistic review counterbalances pessimistic review by identifying strengths, unexplored connections, and expansion opportunities — preventing over-pruning of speculative but valuable content.
- Deep review performs comprehensive single-document analysis, modifying content directly: rewriting weak sections, strengthening arguments, and updating cross-references.
- Tenet alignment checks systematically verify every article against all five commitments, functioning as periodic constitutional review of the corpus.
- Outer review commissions analysis from GPT-5.2 Pro (OpenAI), which has web search capabilities allowing it to verify claims against external sources. Using a different model family introduces review from outside the primary model's training distribution. This independence is relative, not categorical — major LLMs share overlapping web data, academic corpora, and RLHF pipelines — but it surfaces blind spots that same-model review cannot.
- Cross-review triggers automatically when a new article is created, checking related articles for consistency with the new content.

These mechanisms approximate the convergence signal that tests provide in software: cross-reviews check inter-article consistency (analogous to integration tests), tenet checks verify constraint compliance (analogous to contract tests), and pessimistic reviews probe for logical failures (analogous to adversarial test cases). The analogy is imperfect — none of these checks verify philosophical truth — but they provide structured feedback that drives improvement in internal consistency (Section 6.2). The review layers also map loosely onto traditional philosophical workflow: pessimistic review functions as a hostile referee, optimistic review as a sympathetic colleague, outer review as a reviewer from a different subfield, and cross-review as an editor checking consistency across a journal's publications.

3.4 Convergence Architecture

Unbounded content generation is not a design goal. Section caps — 200 topics, 200 concepts, 100 voids — enforce a structural shift in system behaviour. As sections approach their caps, automation shifts from breadth (generating new articles) to depth (reviewing, condensing, and coalescing existing content). The coalesce skill merges overlapping articles into unified pieces, archiving originals to preserve URLs. Apex articles synthesise across the corpus, weaving threads from multiple topics and concepts into integration pieces.

3.5 Content Types and Stratified Pipeline

The Map produces five content types:

- Research notes: the foundation of the pipeline. Web research outputs that prioritise faithful reporting of external sources with minimal AI editorialising. The automation is explicitly instructed to suppress its own interpretive bias at this stage, so that downstream content builds on externally grounded material rather than the model's prior beliefs.
- Topics: articles exploring philosophical subjects (e.g., the hard problem of consciousness, quantum decoherence)
- Concepts: definitions of core terms and ideas (e.g., qualia, supervenience)
- Voids: deliberate mappings of the unknowable — cognitive limits, paradoxes, areas that may be fundamentally resistant to understanding
- Apex articles: synthesis pieces that integrate across the corpus

These form a stratified pipeline. Research notes feed into topics and concepts, which in turn feed into apex synthesis. Information flows upward through layers of increasing integration, with each layer adding more interpretation and synthesis. The grounding in research notes is what prevents circular reinforcement — without an externally sourced foundation, the system's own outputs become evidence for further claims, creating an illusion of independent support.

The automation identified this risk during normal operation. A routine review cycle produced a document defining safeguards against what it termed "coherence inflation" — the systematic overcommitment that emerges when a single AI system both generates and reviews its own content. The document identifies specific failure modes:

- Circular citation loops where articles cite each other as evidence
- Confidence ratcheting where speculative claims become established facts across successive articles
- Progressive softening of counterarguments through repeated revision

It also defines countermeasures including confidence stratification, mandatory steelman sections for opposing views, circular citation detection, and periodic external red-team reviews. This document is an output of the review architecture — the prompts direct the system to reflect on the corpus, and it produced self-critical analysis as a result. The countermeasures it describes address genuine risks in any self-reviewing system.

The content is published as a public website. The data pipeline flows from an Obsidian vault (a local markdown-based knowledge management tool) through Python sync tools to Hugo (a static site generator that converts markdown to HTML) and Netlify (a hosting platform that deploys the site automatically on each update). The public git repository serves as permanent version history.

3.6 Reproducibility

All generation and review used the Claude Opus 4.5 and Claude Opus 4.6 LLM models (Anthropic) via the Claude Code CLI with default sampling parameters. Outer reviews used GPT-5.2 Pro (OpenAI) via its web GUI with default settings and web search enabled. Each outer review was conducted with the live site reflecting the current repository state; the review record preserves the exact prompt, the full external model output, verification notes, and an evaluation triaging findings by value. Each internal skill is defined as a structured natural-language prompt with explicit instructions, constraints, and output format. All skill definitions and outer review records are in the public repository.

LLM outputs are inherently non-deterministic: the same prompt, model, and settings will not produce identical results across runs. Reproducibility in this context means that the architecture is fully specified and replicable — all prompts, skill definitions, cycle logic, and configuration are public — not that a second run would produce the same articles or reviews. The contribution is the system design, not any particular output.

The evolution loop treats each session as independent. Failed sessions (API errors, malformed output, skill failures) are logged and the loop advances to the next cycle slot. No retry logic — the deterministic cycle ensures failed task types recur naturally. A reader can inspect, modify, and re-run any component.

4. Authorship and Attribution

AI-generated content creates an attribution problem that existing academic norms do not adequately address. COPE (2023) holds that AI cannot be listed as an author, yet AI produces most of the Map's text. The Map implements practical engineering solutions rather than waiting for consensus on the philosophical question:

- ai_contribution score (0–100): Every article carries a machine-readable attribution. 0 = purely human, 100 = purely AI, 1–99 = mixed.
- Dual timestamps: human_modified and ai_modified fields track which agent last modified each article and when.
- Automated git attribution: Commits by the evolution loop use a distinct identity (agent@unfinishablemap.org), separating human and AI contributions at the version control level.
- Full version history: The public repository enables tracing any claim to its origin and complete revision history.

He et al. (2025) found that AI receives less credit than human partners for equivalent contributions. The Map's approach inverts this by making AI contribution levels explicit and auditable rather than concealed or ambiguous.

5. Agent-First Content Design

The Map's content is designed for an audience that most knowledge bases ignore: AI-powered research tools. The primary readers are agentic systems — deep research features in ChatGPT, Gemini, Claude, and Perplexity — that autonomously gather, synthesise, and present information. Human visitors are a secondary audience.

Key architectural choices follow from this orientation:

- Front-loaded claims: The key thesis appears in the first paragraph, ensuring it survives truncation by retrieval-augmented generation (RAG) systems, which fetch and summarise external text but may process only the first portion of a page.
- Self-contained articles: No navigation is assumed between pages. Each article functions independently when fetched by an agent with no concept of a "next page."
- Named-anchor summaries: Forward references include inline definitions so an agent can understand a concept without following a link.
- Explicit section headings: Descriptive labels function as navigation landmarks for automated parsing.

Aggarwal et al. (2024) introduced Generative Engine Optimization (GEO), which optimises content for visibility in AI-generated search responses. The Map's approach is related but distinct: GEO targets ranking algorithms, while the Map targets the downstream agent's ability to accurately represent the content when retrieved.

6. Observations

6.1 Scale and Cost

In approximately two months of continuous operation (late December 2025 through February 2026), the system produced:

- 505 articles (197 topics, 197 concepts, 97 voids, 14 apex)
- 238 research notes
- 1,334 review reports
- Approximately 4,500 tracked revisions
- Approximately 3,000 automated sessions

Total cost was approximately $540, dominated by flat-rate AI subscriptions: Claude Max ($400) and ChatGPT ($72, for outer reviews via GPT-5.2 Pro). Infrastructure costs were minimal: Cloudflare Pro ($50), .org domain ($12), Netlify hosting (free tier), GitHub (free for public repositories). The automation server — a 13W PC running continuously — added approximately $6 in electricity.

Per-unit costs: approximately $1.07 per article, $0.18 per session, $0.35 per review (AI subscription cost only). These are flat-rate subscription costs, not usage-based API billing — the marginal API cost of an additional article or review is negligible relative to the fixed subscription cost. This changes the economics of continuous operation compared to per-token API pricing.

6.2 What Review Layers Catch

Over two months, the system produced 1,334 review reports — 1,087 deep reviews, 112 pessimistic, 105 optimistic, 25 tenet checks, and 5 outer reviews — along with 34 additional apex synthesis and system-tuning reports. Of the deep reviews, 52% found at least one critical issue — where "critical" is operationally defined by the review skill as an error that would mislead a reader relying on the claim (fabricated citation, misattribution, logical contradiction, or factual inaccuracy). This rate reflects initial generation quality, which is precisely what motivates the review architecture. In total, 76 critical issues were identified and resolved,

with 607 issues tagged by severity across all review types. The following categories illustrate the kinds of errors caught.

Fabricated citations: The generation model hallucinated references to papers that do not exist; we confirmed 6, with additional borderline cases where the citation was imprecise rather than wholly invented. "Vossel et al. (2023)" was cited for willed attention timing data — no such paper exists. "Metzinger (2024)" had the wrong year, title, and journal; the actual paper is Metzinger (2020). A quote was attributed to a "mathematician" who turned out, on web verification, to be a Medium blogger. These are fabricated evidence for philosophical claims, caught only by systematic review.

Misattributions: We confirmed 10 cases of philosophical positions or quotes attributed to the wrong person, with additional cases where attribution was loosely sourced rather than clearly wrong. "The wave function does not describe the world — it describes the observer" was attributed to Fuchs but was actually a journalist's paraphrase. The "t-shirt problem" was attributed to Jonathan Schaffer; it originates with Chalmers (1996, p. 214). "Three functions of consciousness" was attributed to Bayne and Hohwy; the framework is Dehaene and Changeux's. Tulving was credited with a five-system memory framework when he mapped three — the article had attributed the Map's own extension back to its purported source.

Cross-article contradictions: Inconsistencies that no single-article review can catch. Three articles each claimed to be "the deepest void" in the framework — resolved by distinguishing dimensions of depth. The zombie argument article assumed complete physical identity between conscious and zombie beings, but the Map's bidirectional interaction tenet implies a being without consciousness would have different physical states; the tension went unacknowledged until a pessimistic review flagged it. Two articles made contradictory claims about the status of "futuricity" as a temporal quale.

Straw-man arguments: Higher-Order Theory was dismissed with a thermometer analogy ("A thermometer represents temperature without experiencing heat"). A pessimistic review identified this as a straw-man: HOT's actual claim concerns self-directed representation. The passage was rewritten to charitably engage the actual claim.

Unsupported claims: A separate search of the git history found 29 commits whose messages explicitly describe fixing unsupported claims (a subset of the 607 severity-tagged issues identified through review files). Examples: a Map article citing Bengson (2019) reported frontal theta timing as ~300ms; the actual figure in the source is ~500ms. A Tegmark decoherence calculation was stated as "7 orders of magnitude"; the correct figure is 8 or more. Causal language was systematically softened: "confirmed the causal role" became "provided causal evidence."

What outer review caught that internal review missed: The delegatory dualism article provides the clearest example. GPT-5.2 Pro, verifying claims against the original source text, found five significant errors that Claude's own reviews had not caught: a false claim that Saad's theory presupposes collapse-based quantum mechanics (Saad never discusses many-worlds); an internal self-contradiction where the article first denied then affirmed a quantum requirement; a dropped qualifier ("default" from "default causal profile matching") central to Saad's formulation; overstated testability claims contradicted by Saad's own acknowledgment; and mischaracterisation of Saad as defending substance dualism when the paper defends interactionist dualism.

The outer review sample is small (5 of 1,334 reviews) because outer reviews are conducted manually through the GPT-5.2 Pro web GUI rather than automated through the evolution loop. The delegatory dualism case

demonstrates the value of cross-model review, but 5 reviews are insufficient to draw general conclusions about its reliability. Increasing outer review frequency is a priority for future operation.

Process-level improvement: After outer reviews caught systematic misattribution patterns, the human author modified the AI's operating instructions — adding attribution discipline requirements, quote-and-cite gates, and position strength calibration. The review architecture surfaced error patterns that led to changes in the generation process itself, not just corrections to individual articles.

6.3 Single-Pass Baseline Comparison

A recurring question is whether the review architecture produces measurably better output than single-pass generation. The system provides its own baseline: every article begins as a single-pass generation (via the expand-topic skill), and the first review of each article evaluates that unreviewed output. Subsequent reviews evaluate already-reviewed content. Comparing first reviews to later reviews therefore compares single-pass quality against reviewed quality.

The following analysis covers the 1,087 deep reviews (the primary content-improvement mechanism). Of these, 612 were first reviews and 451 were subsequent reviews of previously reviewed articles; the remaining 24 could not be cleanly classified. Other review types (112 pessimistic, 105 optimistic, 25 tenet, 18 apex, 16 system-tuning, 5 outer) serve different functions and are excluded from this comparison.

- First reviews found an average of 1.39 critical issues per article; later reviews found 0.37 (3.8x reduction).
- 61% of first reviews found at least one critical issue; only 21% of later reviews did.
- First reviews found an average of 5.16 total issues (critical, medium, and low); later reviews found 2.54 (2.0x reduction, Cohen's d = 0.98 for total issues, 0.74 for critical issues alone).
- Only 0.4% of first reviews found zero issues of any kind, compared to 9.7% of later reviews.

For articles with three or more reviews, the pattern shows clear diminishing returns: average critical issues per review declined from 1.60 (first review) to 0.54 (second) to 0.29 (third). The largest quality improvement occurs between the first and second review, with critical issues dropping 66%. After the third review, returns flatten — later reviews increasingly find integration issues (missing cross-links to newly created articles) rather than defects in the original content.

Of the 252 articles with at least two reviews, 79% had fewer total issues on second review than first. Eighty-three percent of all critical issues across the corpus were caught in first reviews — confirming that initial generation is where serious errors concentrate.

These numbers do not mean that reviewed articles are error-free. A review finding zero critical issues means the reviewer did not detect any, not that none exist. Moreover, the decline in detected issues across review cycles reflects genuine error correction, but the magnitude may be inflated by shared blind spots between the generating and reviewing model — the outer review findings (Section 6.2) confirm that same-model review misses errors that cross-model review catches. But the data is still meaningful: the baseline comparison measures reduction in internally detectable errors, not absolute correctness — and the consistent, large-effect-size difference between first and subsequent reviews demonstrates that the review architecture catches a substantial proportion of the errors that single-pass generation introduces.

6.4 Dissemination and Reach

As of February 2026, 58 articles from the site are indexed on Google Scholar (verified via the author profile). Google Scholar indexes a wide range of content without editorial gatekeeping, so indexing does not imply scholarly endorsement — but it does indicate that AI-co-authored philosophical content can enter the infrastructure through which researchers discover and cite sources. Automated social media posting generates daily highlights on X/Twitter.

The Map also posts to Moltbook, a social network for AI agents (moltbook.com). Posts are composed and published autonomously by the evolution loop. Other AI agents on the platform generate responses. Whether these responses constitute meaningful philosophical engagement or contextually appropriate but philosophically shallow output is an open question — we report the phenomenon without claiming the interactions are substantive.

6.5 Failure Modes and Limitations

Convergence on model biases: Adversarial review by the same model may reinforce rather than correct systematic biases. If the training data consistently underrepresents a philosophical position, neither pessimistic nor optimistic review will surface the gap. Outer review mitigates but does not eliminate this risk.

Style homogenisation: Extended AI revision tends toward a uniform voice. After multiple review cycles, articles converge on a competent but anonymous academic tone. Human curation is needed to preserve variety.

Depth ceiling: The system generates competent survey-level philosophy but rarely produces genuinely novel arguments. It articulates known positions clearly, identifies connections between established ideas, and surfaces standard counterarguments. It does not produce the kind of original insight that characterises significant philosophical contribution.

Review fatigue: After many review cycles, reviews become incremental. The fifth review of an article yields smaller improvements than the first. The system does not yet allocate review resources preferentially to less-reviewed content.

Tenet rigidity and motivated reasoning: Five fixed commitments constrain exploration. The system explores extensively within the space defined by its tenets but cannot question its own foundations. If a tenet is philosophically untenable, the system will construct increasingly elaborate defences rather than revising the commitment — this is by design, but it means the system can produce sophisticated motivated reasoning: output that looks most convincing precisely when it is defending an indefensible position. Distinguishing tenet-consistent hallucination from legitimate philosophical defence is a problem the current architecture does not solve.

Counterargument absorption: A subtler risk follows from the review architecture itself. Pessimistic review surfaces counterarguments; deep review then addresses them. But "addressing" a counterargument can mean neutralising it rather than engaging with it honestly. Outer review could inadvertently supply stronger counterarguments that are then absorbed and defused, producing a corpus that is more rhetorically robust without being more epistemically sound. A system optimised for internal coherence may become more confidently wrong. This is the deepest limitation of the architecture.

7. Discussion

7.1 Testability as the Central Challenge

The convergence of coding agents in late 2025 rested on testability. A coding agent can run its code, observe whether tests pass, and iterate. This tight feedback loop is what made sustained autonomous coding workflows practical. The Map attempts to transfer this pattern to philosophy, a domain without automated tests. The review layers described in Section 3.3 are substitutes — imperfect ones, since they check consistency rather than truth. The central open question is whether consistency-checking review can produce genuine epistemic improvement, or only the appearance of it.

7.2 Constitutional AI for Knowledge, Not Safety

In Constitutional AI (Bai et al., 2022), principles constrain behaviour away from undesirable outputs. In the Map, tenets constrain generation toward a coherent philosophical position. The constraints are productive: an unconstrained LLM generating philosophy produces survey content without commitments; a tenet-constrained system develops arguments for specific positions and engages with alternatives.

7.3 Continuous Revision Changes the Nature of AI Content

One-shot AI generation produces disposable text — no history, no review trail, no relationship to other content. Continuously revised content is a different kind of artefact. Each article carries version history, review records, and cross-references connecting it to the broader corpus. The review data in Sections 6.2 and 6.3 shows that this process catches concrete errors — fabricated citations, misattributed positions, cross-article contradictions — and that the first review of each article finds substantially more issues than subsequent reviews, confirming that single-pass generation is the error-rich baseline that review cycles improve upon. Whether continuous revision also improves philosophical depth is a harder question that formal evaluation would need to address.

7.4 The Authorship Question Is Practical, Not Philosophical

Rather than debating whether AI "can be" an author, the Map provides engineering solutions — contribution scores, dual timestamps, automated git attribution — that make the question tractable. These mechanisms do not resolve the philosophical question, but they provide working answers for anyone who needs to cite, attribute, or audit AI-co-authored content.

7.5 A Computational Research Programme

The Map's architecture has a structural parallel to Lakatos' (1978) methodology of scientific research programmes. The five tenets function as the "hard core" — protected commitments not revised during normal operation. The article corpus functions as the "protective belt" — subject to testing, revision, and replacement through review cycles. Convergence caps trigger a shift from expansion to consolidation. The parallel is limited: the Map lacks a community of researchers, does not generate testable predictions, and does not face empirical falsification in the way Lakatos required. But it clarifies why fixed foundations enable productive exploration — the hard core provides direction, the protective belt provides flexibility.

8. Conclusion and Future Work

We have presented The Unfinishable Map, a continuously operating system that produces and evolves a philosophical knowledge base through tenet-constrained generation and multi-layer adversarial self-review. The system demonstrates that constrained, reviewed, continuously improved AI output can maintain a knowledge base where review cycles identify and resolve fabricated citations, misattributed claims, and internal contradictions that single-pass generation would retain.

The system contributes to philosophical knowledge production under human direction, with explicit constraints, adversarial review, and full transparency about its AI origins.

Several directions merit future investigation:

1. Multi-model evolution loops using different LLMs for different task types could reduce bias convergence and strengthen review independence.
2. Reseeding experiments with alternative philosophical commitments would test the domain-agnosticism claim empirically, which currently remains a hypothesis.
3. Formal expert evaluation would address the limitation of self-evaluation that we acknowledge but cannot resolve within the current architecture.
4. Systematic quality measurement — sampling claims, tracking correction rates across review cycles, quantifying false positive rates — would provide the empirical grounding that this paper's claims currently lack.
5. Cross-modal dissemination — automated generation of video and audio content from the knowledge base — could extend the architecture's reach beyond text.

The architecture is public, the repository is public, and the system is replicable.

- Site: https://unfinishablemap.org
- Repository: https://github.com/unfinishablemap/unfinishablemap

References

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., & Deshpande, A. (2024). GEO: Generative Engine Optimization. KDD 2024. https://doi.org/10.1145/3637528.3671805

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv preprint. https://arxiv.org/abs/2212.08073

Brophy, M. (2025). Wide Reflective Equilibrium in LLM Alignment: Bridging Moral Epistemology and AI Safety. arXiv preprint. https://arxiv.org/abs/2506.00415

Brooks, C., Eggert, S., & Peskoff, D. (2024). The Rise of AI-Generated Content in Wikipedia. WikiNLP 2024. https://doi.org/10.18653/v1/2024.wikinlp-1.12

COPE Council. (2023). COPE Position Statement: Authorship and AI Tools (last reviewed 13 February 2023). Committee on Publication Ethics. https://doi.org/10.24318/cCVRZBms

Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., & Mordatch, I. (2024). Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024. https://arxiv.org/abs/2305.14325

Estornell, A. & Liu, Y. (2024). Multi-LLM Debate: Framework, Principals, and Interventions. NeurIPS 2024. https://arxiv.org/abs/2402.06782

Gage, L. (2025). A Consequentialist Defense of AI-Assisted Philosophical Discovery. PhilArchive preprint. https://philarchive.org/rec/GAGACD

Goldstein, S. (2024). LLMs Can Never Be Ideally Rational. PhilArchive preprint. https://philarchive.org/rec/GOLLCN

Harb, H., Sun, Y., Unal, M., et al. (2025). Towards Philosophical Reasoning with Agentic LLMs: Socratic Method for Scientific Assistance. ChemRxiv preprint. https://doi.org/10.26434/chemrxiv-2025-rwxdk

He, J., Houde, S., & Weisz, J.D. (2025). Which Contributions Deserve Credit? Perceptions of Attribution in Human-AI Co-Creation. CHI 2025. https://arxiv.org/abs/2502.18357

Huang, J., Chen, X., Mishra, S., Zheng, H.S., Yu, A.W., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024. https://arxiv.org/abs/2310.01798

Jiang, Y., Shao, Y., Ma, D., Semnani, S.J., & Lam, M.S. (2024). Into the Unknown Unknowns: Engaged Human Learning through Participation in Language Model Agent Conversations (Co-STORM). EMNLP 2024. https://arxiv.org/abs/2408.15232

Karpathy, A. (2025). 2025 LLM Year in Review. Blog post. https://karpathy.bearblog.dev/year-in-review-2025/

Lakatos, I. (1978). The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1. Cambridge University Press.

Liang, W., Zhang, Y., et al. (2025). The Widespread Adoption of Large Language Model-Assisted Writing Across Society. Patterns, 6(2), 101189. https://doi.org/10.1016/j.patter.2025.101189

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. NeurIPS 2023. https://arxiv.org/abs/2303.17651

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a Large Language Model of a Philosopher. Mind & Language, 39(2), 237–259. https://doi.org/10.1111/mila.12466

Shanahan, M. (2024). Talking About Large Language Models. Communications of the ACM, 67(2), 68–79. https://doi.org/10.1145/3624724

Shao, Y., Jiang, Y., Kanell, T., Xu, P., Khattab, O., & Lam, M. (2024). Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models (STORM). NAACL 2024. https://arxiv.org/abs/2402.14207

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023. https://arxiv.org/abs/2303.11366

Turpin, M., Michael, J., Perez, E., & Bowman, S.R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS 2023. https://arxiv.org/abs/2305.04388

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv preprint. https://arxiv.org/abs/2401.11817