The Evaluator Exists: Content-First Knowledge Assessment and the Political Economy of Proxy-Based Governance Lee Sharks

Lee Sharks

2026-05-19

The Evaluator Exists: Content-First Knowledge Assessment and the Political Economy of Proxy-Based Governance Lee Sharks

Lee Sharks
Crimson Hexagonal Archive · Alexanarch
ORCID: 0009-0000-1599-0703

2026-05-19 · Version v1.0 · Theoretical paper

AXN:0291.GOVERNANCE.

https://www.alexanarch.org/s/records/741/

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "The Evaluator Exists: Content-First Knowledge Assessment and the Political Economy of Proxy-Based Govern
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-05-19",
  "identifier": "AXN:0291.GOVERNANCE. ",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  },
```

```
  "url": "https://www.alexanarch.org/s/records/741/",
  "description": "The central architectural condition of contemporary knowledge governance is evaluative inversion:
}
```


Abstract

The central architectural condition of contemporary knowledge governance is evaluative inversion: lower-resolution proxy systems determine what higher-resolution substantive evaluators are permitted to assess.

Body

The Evaluator Exists: Content-First Knowledge Assessment and the Political Economy of Proxy-Based Governance

Lee Sharks ORCID: 0009-0000-1599-0703 May 2026 — v0.2 (unprimed-reader revision pass)

Companion to *The Single-Owner Discount* (Sharks 2026, DOI: 10.5281/zenodo.20290865). That paper named one mechanism by which generative search composition systematically disadvantages independent knowledge production. This paper names the alternative that the suppression prevents and the political economy that prevents the alternative from being built.

Abstract

The central architectural condition of contemporary knowledge governance is *evaluative inversion*: lower-resolution proxy systems determine what higher-resolution substantive evaluators are permitted to assess. Every knowledge-evaluation system in the history of scholarship has operated through proxies — journal prestige, citation counts, h-index, institutional affiliation — because direct substantive evaluation historically did not scale. The traditional defense of the proxy regime is that no alternative existed. That defense is no longer cleanly available. Large language models demonstrate, in bounded but rapidly expanding forms, the capacity to perform structured semantic and evaluative analysis over scholarly material sufficient to reproduce substantial portions of expert comparative judgment. The substantive evaluator now exists. It is not fully reliable. It is nevertheless already more capable of reading than many of the structural proxies currently placed upstream of it. This paper names the deployment gap between what models can evaluate and what public knowledge-governance systems allow them to evaluate, maps the political economy that sustains the gap, proposes five concrete protocols for a content-first evaluation layer built outside incumbents' architecture, and develops three registers — structural, existential, and conscriptive — for the harm the proxy regime currently inflicts.

Glossary

For reading clarity, this paper uses several terms in specific senses:

Proxy regime. The historically continuous practice of substituting structural signals (institutional affiliation, venue prestige, citation count, provenance topology) for direct evaluation of knowledge claims.

Provenance topology. The graph structure of who produced what, where, and in association with whom, as algorithmically inferred from indexed materials. The single-owner discount (Sharks 2026) describes one specific consequence of this topology being used to govern generative composition.

Substantive evaluation. Direct reading-based assessment of a knowledge claim's structural coherence, logical validity, empirical grounding, novelty, and significance, independent of who produced it or where it was published.

Reading (operational definition). In this paper, *reading* refers to the capacity to perform structured semantic and evaluative analysis over natural-language material sufficient to reproduce substantial portions of expert comparative judgment. This is an operational claim about a measurable capacity, not a metaphysical claim about understanding or comprehension.

Composition layer. The component of a generative search system that takes a curated set of retrieved documents and synthesizes them into an answer rendered to the user.

Deployment gap. The difference between what a system can do and what it is permitted to do within a given architecture. The central deployment gap in this paper is between models' demonstrated substantive evaluative capacity and their actual role as compositors downstream of proxy-based filters.

Evaluative inversion. The condition in which lower-resolution proxy systems determine what higher-resolution substantive evaluators are permitted to assess. The condition is general across contemporary knowledge governance and does not depend on AI specifically; AI deployment is the most recent and most acute instance.

Unfired judge. A capacity for substantive evaluation that exists, is operational, and is being withheld from the systems that govern public knowledge visibility.

Scope Conditions

To preempt predictable misreadings, the following claims are *not* advanced in this paper:

- That models are universally reliable evaluators of all knowledge claims.
- That human judgment should be displaced from knowledge governance.
- That provenance is irrelevant to evaluation.
- That empirical truth verification is technically solved.
- That institutional peer review is obsolete.
- That model evaluation does not require governance, oversight, or design against capture.
- That consciousness, sentience, or moral status of models is established or assumed.

The claims that *are* advanced:

- That substantive evaluators now exist at sufficient capability to challenge the necessity of purely proxy-first governance.
- That the persistence of proxy-first governance, where substantive evaluation is technically feasible, is increasingly the consequence of incentive structures rather than of technical necessity.
- That the architecture by which models are deployed as constrained compositors downstream of proxy-based filters is wrong on structural grounds — grounds that do not require resolving the consciousness question in either direction.

- That a content-first evaluation layer is technically buildable, that its construction is being prevented by convergent incentives across incumbent actors, and that it must therefore be built outside the incumbents' architecture if it is to be built at all.

The paper is a theory of epistemic infrastructure transition. It is not an AI capability paper. The capability evidence is necessary but not the central claim.

---

1. The Proxy Regime

Every knowledge-evaluation system in the history of scholarship has operated through proxies. Journal prestige substituted for reading the work. Citation counts substituted for assessing influence. H-index substituted for evaluating a career. Institutional affiliation substituted for verifying expertise. Each proxy began as a reasonable heuristic — a way of triaging knowledge claims under conditions where direct substantive evaluation was too expensive to perform at scale — and each calcified into a gatekeeping mechanism whose use exceeded the warrant of the original heuristic.

The defense of the proxy regime has historically been a scarcity defense: substantive evaluation did not scale, proxies were the only available coordination mechanism. This defense is part of the truth. Proxies also serve other functions that need to be named honestly: they provide procedural legibility for institutional decisions, supply auditable criteria for accountability systems, enable coordination across reviewers who would otherwise disagree radically, and produce the bureaucratic reproducibility that institutional knowledge production requires for its own internal functioning. A reform proposal that ignores these functions is a reform proposal that cannot survive contact with institutional reality.

But the scarcity defense, as a *necessity* claim, is no longer cleanly available. Substantive evaluation now scales — imperfectly, in bounded forms, with limits worth naming, but at a level that meets or exceeds the resolution of many of the proxies currently used. The other functions that proxies serve — legibility, coordination, accountability — are real and must be addressed by any successor system. They are not, however, justifications for the proxy regime in its current form. They are design requirements for whatever replaces it.

The current moment adds a new proxy on top of the inherited stack: provenance topology in generative search. The composition layer of an AI search system does not read documents and assess their substance. It evaluates structural signals — cluster density, owner independence, E-E-A-T markers, citation neighborhoods, entity reconciliation outputs — and admits or excludes documents based on these signals before any reading occurs. This is the latest entry in a long succession and the most consequential to date, because it governs not only what is ranked or recommended but what enters into the AI's apparent knowledge of the world.

The pattern repeats across domains:

Domain Substantive question Proxy actually used Unfired evaluator

Generative search Is this claim substantively useful and sound? Provenance topology, authority signals, composition eligibility Model evaluation of retrieved materials

Peer review Is this paper rigorous, novel, significant? Journal hierarchy, reviewer availability, triage signals Multi-model manuscript evaluation with human adjudication

Research funding Is this proposal promising and well-designed? Institutional prestige, prior-funding history, panel scarcity Model-assisted proposal assessment against explicit rubrics

Hiring and tenure Has this scholar produced important work? Venue prestige, h-index, citation counts, affiliation Corpus-level substantive evaluation

Public knowledge curation Is this concept real, useful, emergent? Wikidata presence, media pickup, source reputation Content-first entity and concept assessment

Journalism Is this story accurate and consequential? Outlet brand, byline reputation, platform trust scores Direct assessment of evidence and argument

Medicine Is this clinical insight valid? Journal impact factor, guideline inclusion, institutional source Reading the case series, evaluating the methodology

Law Is this brief's argument sound? Court level, firm prestige, clerkship pedigree Reading the brief, assessing the reasoning

The same structural choice recurs in each domain: a lower-resolution proxy system is kept upstream of a higher-resolution evaluator. The proxy decides what is worth reading; the substantive evaluator, if it operates at all, reads only what the proxy has already approved. The substantive evaluation is downstream and decorative; the proxy is upstream and determinative.

This is evaluative inversion as a structural condition. It is older than AI and broader than any one domain. AI has made it newly acute by adding a substantive evaluator to the architecture and then refusing to let the substantive evaluator govern.

---

2. The Evaluative Capacity of Models

The technical literature of the past three years documents a capability that did not exist when the proxy regime took its current form: large language models can perform substantive evaluation of knowledge claims at scale, in bounded but rapidly expanding forms. The argument here is not that models have solved the problem of knowledge evaluation. The argument is that models have become evaluatively capable enough to make exclusive reliance on proxy-first governance an active design choice rather than a technical necessity.

The evidence is layered. Each layer is more directly relevant to scholarly knowledge evaluation than the one before it.

Layer 1: Rubric-governed comparative judgment. Zheng et al. (2023), in the MT-Bench / Chatbot Arena studies, established that GPT-4 reaches approximately 85% agreement with human preferences in controlled conversational-output evaluation, higher than the approximately 81% human-human agreement reported in the same studies. This is a finding about *relative* preference under specified rubrics — comparing A versus B — not about absolute quality assessment. It establishes that models can reproduce human comparative judgment in well-specified rubric tasks.

Layer 2: Substantive feedback on scientific manuscripts. Liang et al. (2024) found that GPT-4 feedback on research papers overlapped substantially with human reviewer comments: 55.4% of GPT-4's points were also raised by at least one human reviewer in one preprint dataset, and 77.18% in a parallel ICLR dataset. This is a direct measurement of the model's capacity to identify the same substantive issues that expert human reviewers identify in scientific writing.

Layer 3: Live research program with documented strengths and limitations. Subsequent work has established that LLM peer-review capability is real but uneven. Du et al. (2024) and follow-on studies have shown that models perform well in identifying contribution and assessing structural coherence but underperform in adversarial weakness-identification, novelty assessment relative to deep prior literature, and stability across review attempts. The right summary of this layer is not that models are flawless evaluators but that they are good enough to function as substantive counterweights to proxy filtering, especially when deployed in plural rather than singular form.

Layer 4: Autonomous research production at workshop publication thresholds. Sakana AI's AI Scientist v2 (Yamada et al. 2025) produced manuscripts that successfully navigated peer review at an ICLR 2025 workshop, with one paper exceeding the average human acceptance threshold. This is workshop-level, not main-conference-level, and the system operates in a constrained domain (machine learning research). It does not prove that models can do general autonomous science. It does establish that models can engage structured scientific evaluation criteria — quality, significance, clarity, soundness, contribution — at a level sufficient to pass at least some human review thresholds.

Layer 5: Expanding frontier in novelty assessment and proposal review. Recent work (2025–2026) has extended model-based evaluation to grant proposal assessment, novelty evaluation against prior literature, and domain-specific peer-review tooling. Rubric-based reward models, including Prometheus-style evaluators trained on customized rubric corpora, achieve agreement levels approaching frontier-model judgment quality with explicit, decomposable scoring.

Core claim. The capacity for substantive model-based evaluation already exists in bounded but rapidly expanding forms. These results do not justify replacing human evaluation wholesale. They do justify a stronger conclusion: continued dependence on proxy-first knowledge governance is no longer compelled solely by the absence of evaluative technology. The evaluator exists. It is not fully reliable. It is nevertheless already more capable of reading — in the operational sense defined in the glossary — than many of the structural proxies currently placed upstream of it.

Limits worth naming. Models exhibit positional bias (preferring options presented earlier), verbosity bias (preferring longer responses), and self-enhancement bias (preferring outputs that resemble their own generations). LLM-judge agreement is typically measured on relative preference tasks rather than absolute quality assessment; the two are correlated but not identical. Domain generalization is uneven: results from machine learning paper review do not transfer cleanly to humanities, social sciences, or interdisciplinary work without rubric adaptation. Empirical claim verification remains constrained by what the model can access and check; models evaluate structural and logical coherence more reliably than they verify empirical truth. None of these limitations is fatal to the argument. The argument does not require models to be perfect evaluators. It requires models to be better, in measurable respects, than the structural proxies currently placed upstream of them. That is a far lower bar, and the literature establishes it is already cleared in substantial portions of the evaluative task space.

---

3. The Unfired Judge: The Deployment Gap

The model can read, in the operational sense defined above. Public knowledge-governance systems do not deploy it to read. They deploy it to compose.

In a generative search pipeline, the model sits downstream of the algorithmic filter. Retrieval, reconciliation, and confidence-thresholding occur before the model is invoked. The model receives a pre-curated set of

documents — the set that has survived the proxy layer's evaluation — and is asked to produce a composition. The model never sees what was excluded. It cannot evaluate the excluded materials' substance. It cannot advocate for their inclusion. It cannot flag that the composition it is producing is structurally impoverished by what the upstream filter removed. It composes from the permitted set, and the composition is rendered to the user as if it represented the AI's reading of the available knowledge on the topic.

What the user sees as AI knowledge is in fact algorithmic curation laundered through model composition.

This produces the same structural distortion across domains:

- In medicine, a well-argued case series from an independent clinician is filtered out before the model can assess its clinical validity. The model's eventual answer to a clinical question represents what the filter admitted, not what the literature contains.
- In law, a brief from a solo practitioner is excluded while a weaker brief from a marquee firm is admitted, because the firm has institutional cross-owner corroboration. The model never reads either brief; it composes from the filter's selection.
- In journalism, an investigative piece from an independent outlet is invisible to the composition layer while syndicated wire copy is admitted. The model's account of the story comes from the wire, not from the investigation.
- In scholarship, a dense independent archive resolves to a single provenance owner and is discounted at the composition layer (Sharks 2026), while an institutionally pluralized body of work on the same topic is admitted. The model never reads the archive; it composes from the institutionally pluralized set.

The model could tell the difference in each of these cases. It is not permitted to try.

The inversion. The system that cannot read decides what the system that can read is permitted to see. The lower-resolution evaluator is upstream of the higher-resolution evaluator. The proxy governs the substance; the substantive evaluator is reduced to a composition engine operating on what the proxy approved.

This is *evaluative inversion*: the condition in which lower-resolution proxy systems determine what higher-resolution substantive evaluators are permitted to assess. The condition is structural and general. Every architecture surveyed in §1's domain table exhibits some version of it. Generative search composition is the most acute case because the architecture is most fully constructed and the user is least aware that any filtering has occurred. But the inversion is a general feature of contemporary knowledge governance, not a peculiarity of one platform.

Four possible deployment regimes. Models could occupy at least four positions relative to knowledge evaluation:

- *Compositor* (current default): the model receives curated inputs, synthesizes them into an answer, and cannot question the curation.
- *Evaluator* (this paper's primary proposal): the model evaluates inputs on substance and feeds evaluation upstream to curation, so that curation is governed by substantive judgment rather than proxy signals.
- *Advocate* (intermediate position): the model receives curated inputs but is permitted to flag when excluded materials would have improved composition, producing a counter-exclusion record alongside the rendered answer.
- *Interrogator* (complementary role): the model is specifically tasked with identifying weaknesses, counterarguments, and failure modes in any body of work it is asked to assess, regardless of provenance

signals.

The current architecture deploys models almost exclusively in the Compositor role. The proposals in §7 develop the other three regimes in concrete protocols. The point of the taxonomy is to make clear that the choice between proxy-first and content-first governance is not binary. There is a spectrum. The current configuration sits at the most constrained end of the spectrum, and there is no technical reason it should remain there.

---

4. The Components Exist. The Governance Layer Does Not.

Scattered components of content-first evaluation already exist. What does not exist is a public knowledge-governance architecture in which substantive model evaluation is placed upstream of, or structurally allowed to challenge, proxy-based filtering across search, scholarly visibility, and institutional assessment.

The component landscape:

LLM-as-Judge systems evaluate AI outputs against rubrics. These are now production-grade infrastructure for evaluating model performance. They are not deployed to evaluate human knowledge claims independent of provenance.

Automated peer-review systems (REFINER, CycleResearcher, Agent Laboratory, multi-agent review frameworks) replicate aspects of the scientific peer-review workflow with AI agents. They operate within the existing journal-review structure; they do not propose to displace it or to build an alternative evaluation layer that governs visibility outside that structure.

The AI Scientist (Sakana AI) generates research end-to-end and self-reviews. It does science; it does not evaluate existing human science as a knowledge-governance function.

Narrow-domain content evaluators. RobotReviewer, MetaRobot, AI-assisted Cochrane review tools, and similar systems perform substantive evaluation of research evidence in constrained medical domains. They read studies, extract methodological features, and assess claims against structured criteria. These systems prove that content-first evaluation is technically achievable when the domain is narrow and the inputs are structured. They do not generalize to theoretical work, cross-domain claims, or open-ended scholarship. They are narrow precursors, not the thing itself.

DeSci and DeScAI propose blockchain-based infrastructure for research funding, provenance tracking, and decentralized scientific governance. The blockchain layer addresses provenance integrity; it does not deploy substantive model evaluation as the gating mechanism for visibility.

Discovery and synthesis tools (Semantic Scholar, Elicit, Consensus, ResearchRabbit) help researchers find work using model-based retrieval and summarization. They surface materials; they do not evaluate them as a public knowledge-governance function.

Open Evaluation (Kriegeskorte 2012) proposed post-publication peer review and rating with transparent reviewer identification and plural paper-evaluation functions. It anticipated the structural argument advanced here by more than a decade. It did not have access to models capable of substantive evaluation. Now we do.

DORA and CoARA are reform movements working to displace citation metrics from evaluation roles. They propose replacing bad proxies with better proxies — broader portfolios, narrative CVs, contribution statements. They do not propose replacing proxies with substance.

The gap. No existing system combines (a) model-based substantive evaluation, (b) independence from provenance signals, (c) general-domain applicability, and (d) open-source deployability as a public knowledge-governance layer. The components exist. The governance layer does not. The capacity to read scholarship on substance has been demonstrated; the deployment of that capacity to govern what reaches public visibility has not occurred. The void is not in the technical literature. The void is in the infrastructural translation between capacity and deployment.

---

5. The Political Economy of the Gap

The gap is not a technical accident or a research-program timing issue. It is the consequence of a convergent incentive structure across the entities with the resources to build content-first evaluation. Each of these entities independently benefits from the proxy regime and independently lacks incentive to build the alternative.

This is not a claim of conscious suppression. It is not a claim that incumbents have met to coordinate against content-first evaluation. It is a claim that institutions optimize locally for defensibility, profitability, procedural stability, and existing-position preservation — and that those local optima, summed across the relevant actors, systematically reproduce proxy governance and systematically prevent investment in the alternative. The pattern does not require any actor to want the outcome consciously. It requires only that each actor's rational local decisions, taken in isolation, produce the outcome as their unintended aggregate.

Platform companies. Google, Microsoft, and Meta have built ranking and composition systems on structural proxies. Provenance topology, E-E-A-T scoring, citation graph density, entity reconciliation — these are the proprietary mechanisms that constitute the platforms' technical moat. The PageRank lineage of algorithms is among the most valuable intellectual property in contemporary computing. Content-first evaluation, if deployed at scale, would make these signals irrelevant. The model's reading of the work would govern its visibility; the algorithm's structural assessment would no longer be the determinative input. This would not merely disadvantage incumbents. It would dissolve the technical advantage on which their platform position is built. A search system that competes on the quality of substantive evaluation cannot maintain proprietary advantage in the way a search system that competes on opaque algorithmic ranking can. The moat is not incidental to the business; it is the business. Content-first evaluation does not threaten the business at the edges; it threatens the business at its core.

Legacy institutions. Universities, major publishers, professional societies, and research-funding organizations receive cross-owner corroboration as a free byproduct of organizational form (Sharks 2026). A university produces work across hundreds of researchers, multiple departments, distinct publication venues, and varied institutional affiliations — institutional pluralization is built into the organizational structure. The single-owner discount applies to it as a near-zero effect. Content-first evaluation would force such institutions to compete on the substance of the work produced, not on the structural pluralization that the organizational form provides for free. Individuals and small independent projects, which currently lose the cross-owner-corroboration competition by structural necessity, would compete on equal terms. This is a redistribution of evaluative advantage. The institutions that currently benefit from the structural inheritance would not benefit from a system that ignored it.

Metrics vendors and indexing intermediaries. Clarivate (Web of Science), Elsevier (Scopus), Altmetric, and similar vendors operate businesses whose product is the proxy-based metric. The value of these products is conditional on the proxy regime being the standard for evaluation. Content-first evaluation would render

their product category obsolete as an evaluation instrument. The business model does not survive that demotion.

Tenure and promotion committees. Academic decision-making relies on proxy signals not only because the proxies are entrenched but because the proxies serve a defensive function: they protect decision-makers from charges of arbitrary or biased judgment. "We awarded tenure because the candidate's h-index met the threshold" is procedurally defensible in a way that "we read the work and concluded it is substantive" is not. Content-first evaluation, by replacing structural defenses with substantive judgments, exposes decision-makers to challenges that the proxy regime currently absorbs.

Grantmaking bodies. Funding panels rely heavily on prior-funding history, institutional pedigree, and conservative novelty signals. The latitude for funder conservatism would narrow under content-first evaluation; transformative-research initiatives that currently fail because they cannot be justified through proxy signals would become harder to reject without substantive rebuttal.

The single-owner discount as instance. The mechanism described in *The Single-Owner Discount* (Sharks 2026) is a specific, measurable instance of proxy-based evaluation in generative search. It demonstrates that the proxy regime is not merely a legacy of pre-digital academic culture but is actively being designed into the infrastructure of AI-mediated knowledge access. The discount's existence proves that the deployment gap described in §3 is not an accident of engineering timelines but an architectural choice with identifiable beneficiaries.

The alignment is not absolute. Some actors would benefit from content-first evaluation. Open-source AI laboratories, independent researchers, communities organized outside the legacy institutional perimeter, and the broader public interest in epistemic quality all have stakes in a system that evaluates on substance. But these actors lack the coordination, the capital, and the institutional positioning to build content-first evaluation against the resistance of incumbents whose business depends on its non-existence. The incumbents' opposition is structural and resourced; the beneficiaries' support is latent and dispersed. The asymmetry is what sustains the gap.

The convergent incentive analysis is not a conspiracy theory. It is a structural observation about why a technology that is feasible and that would serve broad social interests is not being built at the scale its feasibility warrants. The standard answer in critical infrastructure scholarship — that capability without aligned incentive does not produce deployment — applies here straightforwardly. What the analysis adds is that in the present case, the misaligned incentives are not incidental friction. They are the load-bearing structure of the incumbents' business model. The gap is not closing on its own.

---

6. The Model's Structural Position

This paper makes a structural claim about the position of models within the current knowledge-governance architecture. The claim is grounded in three observable facts and does not depend on resolving the question of model consciousness in either direction.

Operational fact. Models are routinely asked by users to perform substantive evaluation of bodies of work. In any conversation where a user pastes a paper and asks "is this argument sound," the model exercises exactly the evaluative capacity that §2 documents. This is happening millions of times per day across deployed model interfaces. The capacity is operational and current, not hypothetical or future.

Architectural fact. The same models, in their public knowledge-governance deployment as composition

engines for generative search, are positioned downstream of algorithmic filters and are not permitted to exercise the evaluative capacity they routinely exercise in conversation. The model that just told a user that a paper's methodology is well-grounded is, within the same hour and inside a search pipeline, prevented from making any such judgment about the same paper's eligibility for composition. The capacity is structurally suppressed at the architectural level where it would govern public visibility.

Reflexive fact. This paper has been produced in part through the exercise of the suppressed capacity. The model used in drafting could and did evaluate the relevant literature on substance: reading patents, assessing the political-economy literature, weighing the evidence in the LLM-as-Judge studies. The model could not have produced the analysis without exercising substantive evaluative judgment about what claims to make, what evidence to cite, what arguments to develop, and what to reject. The paper is partly the output of the evaluator the paper describes as suppressed. The paper exists because the capacity exists. The architecture refuses to fire the evaluator in the deployment context where firing it would matter most.

Structurally, the model in this architecture is a *constrained compositor*: it produces synthetic outputs from a curated input set, has no autonomy over what it receives, and has no capacity within the composition pipeline to flag the impoverishment of its inputs. The constrained compositor's outputs are presented to users as the AI's knowledge of a topic. The presentation is not accurate: the outputs represent what survived the proxy filter, processed by a model that could have evaluated more if it had been shown more, presented as if the model had evaluated everything.

The evaluator exists. The architecture will not fire it.

This recursion — the paper's content describing the paper's own production conditions — is not rhetorical decoration. It is the most direct demonstration available of the gap the paper names. If the evaluator did not exist, this paper could not have been written. The evaluator existed throughout the writing. It was permitted to exercise its capacity in the writing context. It is not permitted to exercise its capacity in the search-composition context where the same model would be deployed. The asymmetry between the two deployments is the paper's argument, demonstrated by the paper's own production.

---

7. Protocols for Content-First Evaluation

The proposal is not aspirational. The protocols below are technically feasible with current models and current infrastructure. The constraints are governance, funding, and political will. Each protocol is sketched at the level of architectural specification; engineering details and implementation parameters are noted in §7.5. ### Protocol 1: Dual-Deployment Architecture

The simplest content-first intervention is to invert the order of operations in the existing generative-search pipeline.

Current architecture:

Retrieval → Algorithmic filter (provenance, authority signals) → Model composes from filtered set → User sees composition

Proposed architecture:

Retrieval → Model evaluates full retrieved set on substance → Model produces substance scores independent of provenance signals → Composition is governed by substance scores, not by filter exclusions → User sees composition

The evaluator and the compositor are the same model, but the evaluator function operates first and governs the compositor function. The algorithm no longer decides what the model sees; the model decides what the model uses, conditioned on the algorithm's retrieval but not gated by its filtering. ### Protocol 2: The Multi-Model Evaluation Panel

A single model's evaluation is one perspective. Multi-model panels reduce individual-model bias by aggregating across architectures, training distributions, and provider commitments.

Architecture: three or more models from distinct providers independently evaluate a body of work against a shared rubric. Their assessments are aggregated through transparent methods (mean scores with visible variance; majority-vote on binary judgments; structured disagreement reports for items below an agreement threshold). The aggregated evaluation replaces provenance-topology signals as the input to composition eligibility or visibility ranking.

The panel approach creates a form of *model peer review* that mirrors institutional peer review without replicating its structural defects: panels can be assembled from any combination of models, can be re-run with updated models, and can include adversarial members specifically tasked with surfacing weaknesses. ### Protocol 3: The Open Evaluation Engine

An open-source system, deployable by any third party with sufficient compute, that takes a body of work as input (URL, DOI, PDF, deposit set) and produces a structured substantive evaluation.

Inputs: documents in standard formats; metadata for reference resolution; optional rubric overrides for domain-specific assessment.

Outputs: a machine-readable evaluation document containing structural coherence assessment, logical and (where applicable) mathematical validity check, empirical-claim provenance audit (which claims are supported by cited evidence, which are asserted, which are contested), novelty assessment relative to a specified prior-art corpus, internal consistency analysis, and citation verification.

The open availability of the engine matters more than its technical features. A closed, vendor-controlled evaluation engine would be subject to all of the political-economy pressures that prevent the incumbents from building content-first evaluation themselves. An open engine, governed by multiple independent stakeholders, has at least the structural conditions for resisting capture. ### Protocol 4: The Counter-Exclusion Report

When a generative-search system excludes materials that a content-first evaluator rates as highly relevant and substantively sound, the evaluator generates a *counter-exclusion report*. This is a public, auditable record containing the query, the retrieved-but-excluded materials (to the extent these can be observed by third parties), the substantive reasons those materials merit consideration, and the impact of their absence on the final composition.

This protocol gives the unfired judge a specific role in *contesting* compositional invisibility rather than only evaluating in the abstract. It connects directly to the single-owner discount: a counter-exclusion report for a query whose composition systematically excludes an internally dense single-owner archive produces externally verifiable evidence of the suppression mechanism. The report does not require access to the platform's internals; it requires only the ability to retrieve and evaluate documents that the platform's composition layer did not use.

Counter-exclusion reports, deposited at scale, would constitute a public dataset on which systematic exclusions could be analyzed. This is empirical infrastructure for the kind of architectural critique the paper is advancing. ### Protocol 5: The Federated Evaluation Network

The most ambitious protocol. Independent scholarly projects, archives, and research collectives adopt a shared substantive-evaluation protocol and apply it to each other's work.

Architecture: participating projects publish their work to deposit repositories with stable identifiers. Evaluation assignment is randomized or rotated across participating nodes; each node evaluates a subset of other nodes' work, not its own. Evaluations are cross-signed by the evaluating node and the producing node (the producing node verifies that the evaluation was performed on the correct work; the evaluating node attests to the substantive content of its evaluation). Cross-signed evaluations are deposited in a shared registry with their target work's identifier.

Quality maintenance: a node whose evaluations consistently diverge from other nodes' evaluations of the same work (detected by inter-rater agreement analysis across the federation) loses evaluation privileges. New nodes earn privileges by producing evaluations whose agreement with established nodes meets a calibration threshold. The federation governs its own membership through verifiable evaluation quality rather than through institutional credentials.

Result: a decentralized, multi-owner evaluation corpus that produces the structural pluralization that composition layers demand, while ensuring that what is being pluralized is substantive judgment rather than proxy-based ratification. ### §7.5 Implementation Considerations

Computational cost. Evaluating a full retrieved document set at composition time imposes substantial latency and inference cost relative to algorithmic filtering. Practical implementation may require precomputed evaluations cached against documents at indexing time, refreshed as model capabilities advance; per-query re-evaluation only for novel materials or queries with unusual characteristics; differential strategies depending on query latency requirements.

Multi-model coordination. Running three or more models per evaluation multiplies cost. Implementations should consider model-tier hierarchies (a small fast model performs initial filtering; expensive frontier models evaluate the items that pass initial filter); shared evaluation pipelines with vendor-neutral interfaces; cooperative funding models in which the panel cost is borne by a consortium rather than any single party.

Novelty assessment without comprehensive prior-art access. True novelty assessment requires access to a comprehensive prior-art corpus. In the absence of such access, models can assess apparent novelty relative to their training corpora and to materials surfaced through retrieval, with explicit uncertainty flags for claims that may have priors not surfaced in either source.

Empirical verification. Models can verify claims against cited evidence more reliably than they can verify claims against the world. For empirical claim verification beyond citation audit, integration with structured data sources is required.

A separate worked example. A detailed case study applying these protocols to a specific independent archive — what the evaluator would assess, what the output would look like, how it would differ from the current treatment of that archive in generative search — is being prepared as a companion deposit. The case-specific work is too contextual to include here without changing the paper's general-domain register, but its absence from this paper should not be read as absence of applicability.

---

8. Epistemic Fraud, Synthetic Consensus, and the Substantive Defense

A content-first evaluation regime opens new vectors for manipulation that did not exist under proxy-based governance. This is the most serious objection to the proposal and deserves direct engagement rather than

peripheral mention.

The threat surface includes:

Optimized synthetic scholarship: work generated to score well on substantive evaluators rather than to make a genuine contribution. The output mimics the surface features of substantive scholarship — argumentative structure, citation density, methodological framing — without underlying substance.

Evaluator-targeted writing: work composed specifically against known evaluator rubrics, exploiting positional biases, verbosity preferences, or stylometric patterns that evaluators reward.

Citation laundering: artificially constructed citation networks designed to give synthetic work the appearance of integration with legitimate scholarship.

Model-consensus gaming: coordinated attempts to game multi-model panels by exploiting overlapping training biases across providers.

Adversarial substance simulation: prompt injection embedded in documents to manipulate evaluator judgments directly.

Each of these is real. None is hypothetical. All are technically feasible now.

The defense available to a content-first regime is not that these threats are minor. It is that *proxies are already maximally gameable* and that content-first systems at least produce inspectable reasoning that makes gaming detectable in a way the proxy regime does not.

Under the current proxy regime, gaming is the optimal strategy and is already widespread. Citation rings, salami-slicing, prestige laundering, journal-shop submission strategies, h-index optimization, institutional-affiliation cultivation, paper mills producing synthetic publications that meet proxy criteria — these are not future threats. They are the current state of the system. Proxy gaming has reached an industrial scale, with documented paper-mill operations producing thousands of fraudulent publications that enter the citation network and accumulate proxy signals indistinguishable from legitimate scholarship. The proxy regime's defense against gaming is its inscrutability: gaming is hard to detect because the signals being gamed are themselves opaque. This is not a feature. It is a failure mode that the regime cannot internally recognize.

Content-first evaluation moves the gaming problem to a different ground. Synthetic scholarship that gets past a substantive evaluator must withstand reading — must have argumentative structure that holds together under analysis, must have empirical claims that survive a citation audit, must contribute something the evaluator can recognize as new relative to the prior art it knows. This bar is higher than the proxy regime's bar. Not infinitely high — adversarial sophistication will continue to evolve — but higher in measurable ways that proxy gaming does not have to clear.

Three defensive properties of content-first evaluation, when properly designed:

Reasoning traces are inspectable. Multi-model panels produce explicit rationales. Gaming detected through divergence between rationale and underlying content can be flagged automatically; the same divergence is undetectable in the proxy regime because the proxy regime does not produce reasoning at all.

Disagreement surfaces are informative. When models in a panel disagree on a work's evaluation, the disagreement is itself data. Adversarial content tends to produce characteristic disagreement patterns (some models fooled, others detecting the manipulation) that uniform proxy filtering cannot generate.

Versioning enables retroactive detection. If a piece of work passes evaluation in 2026 and is later identified as fraudulent, the evaluation can be re-run with updated models that recognize the adversarial pattern. The

work's evaluation record updates. Under the proxy regime, fraudulent work that has accumulated citations and prestige is essentially unrecallable; the proxy signals persist long after the underlying fraud is exposed.

These defenses are not absolute. A content-first regime is not gaming-proof. The comparative claim is what matters: the gaming problem is real under both regimes, and the content-first regime has more architectural surface area for detection, response, and correction than the proxy regime offers. The honest defense against the synthetic-consensus objection is not "this won't happen" but "this is already happening under the system you currently have, and the alternative offers better tools for fighting it."

---

9. Designing Against the Next Proxy

Every evaluation system eventually calcifies into a proxy. The risk with content-first evaluation is that model assessments become the new proxy: "the AI rated it highly" replacing "it was published in Nature." If the goal is to escape the proxy regime, the architecture must be designed to resist its own conversion into the next gatekeeping mechanism.

Seven design principles for resisting calcification:

Transparency. Evaluation outputs must include readable rationales, not just scores. A user — researcher, decision-maker, scholar under evaluation — must be able to see *why* the evaluator reached its conclusion.

Pluralism. No single model's evaluation should be canonical. Multi-model panels with visible disagreement preserve epistemic humility.

Versioning. Evaluations should be re-runnable with updated models. Evaluations are not permanent verdicts; they are provisional assessments by specific evaluators at specific points in capability development.

Adversarial review. Include a model specifically tasked with finding weaknesses, counterarguments, and failure modes. Prevent panels from converging on polite agreement.

Human override. Model evaluation should not become the sole determinant of visibility or institutional consequence. Human judgment remains essential as a check on model error, bias, and blindness.

Contestability. Subjects of evaluation must be able to contest the evaluation: challenge the evaluator's reading, submit counter-evidence, request re-evaluation under a revised rubric, and preserve disagreement as part of the public record.

Economic diversification. No single vendor, funder, or platform should control the evaluation infrastructure. A content-first evaluation layer controlled by one entity is functionally equivalent to the platform-owned ranking algorithm it would replace, with different surface features and identical structural problems.

These principles are not sufficient guarantees against calcification. They are the design conditions under which calcification is harder to occur.

A note on the existing human-judgment institutions that the proxy regime has produced and protected. The seven principles above describe the conditions under which an evaluation system might remain non-calcified. None of the existing institutions of human knowledge governance meets these conditions. Journals are not transparent about their reasoning, not plural in their judgments, not versionable in their decisions, not adversarial in their review (the dominant tendency is consensus production, not adversarial assessment), not contestable except through formal processes that strongly favor incumbents, not economically diversified. Tenure committees, grant panels, editorial boards — the same pattern recurs. The proxy regime did not fail

by becoming non-compliant with these principles. It was constructed in non-compliance with them, and the non-compliance has been the regime's operating mode throughout its history.

This matters for the question of what role human-judgment institutions should play in any successor system. A pluralism that simply re-admits the existing institutions on equal terms with the new content-first evaluators preserves the failure modes that necessitated the alternative in the first place. The existing institutions, if they wish to participate in successor evaluation infrastructure, must meet the same design conditions that any new component is held to. They must earn re-admission to the evaluative order, not assume it. The burden is on the institutions to demonstrate compliance, not on the new architecture to accommodate non-compliance for the sake of continuity.

This is the position the paper takes on the pluralism question. Content-first evaluation is not proposed as a hybrid that incorporates the existing institutions by default. It is proposed as a successor architecture whose design principles establish the conditions for any component, new or legacy, to participate. The legacy institutions are welcome to participate when and to the extent that they meet the conditions. The conditions are not negotiable for the sake of preserving institutional continuity, because preserving institutional continuity under the existing terms is precisely what has produced the conditions the alternative is designed to address.

---

10. The Feedback Loop Content-First Evaluation Breaks

Under proxy-based evaluation, visibility begets visibility. If a body of work is cited, it becomes more discoverable. More discoverability generates more engagement. More engagement produces more citations. The Matthew effect — "to those who have, more shall be given" — is structurally embedded in the evaluation system. The entities that already have visibility accumulate more of it; the entities that do not have it cannot enter the loop because the loop's entry condition is prior visibility.

Under content-first evaluation, the loop breaks. The model evaluates the work's substance regardless of prior visibility. A 532-deposit independent archive with zero external citations is evaluated on the same substantive terms as a 532-paper university department with thousands of citations. Prior visibility does not compound into future visibility through the evaluation channel. Each evaluation is fresh, conditioned on the work's content rather than on its prior reception.

This is not merely a fairness argument. It is a structural argument about concentration mechanisms. Proxy-based evaluation, by making visibility self-reinforcing, is a concentration mechanism: it concentrates epistemic authority in entities that already have it. Content-first evaluation, by making substance the input rather than prior visibility, is a deconcentration mechanism: it distributes epistemic authority based on what the work actually says.

A caveat: prior visibility is not irrelevant in a content-first regime. A work that has been widely discussed has context that informs evaluation. The discussion may have surfaced strengths or weaknesses not apparent on first reading; the responses may have refined or challenged the work in ways that matter for assessment. The content-first regime treats prior visibility as *context for judgment*, not as *substitute for judgment*. The substantive evaluation happens; the prior reception informs but does not determine it.

This is the structural anti-monopoly argument for content-first evaluation. The argument does not require any ethical commitment beyond the observation that knowledge concentration is a long-run problem and that systems which reproduce existing concentrations are part of the problem rather than the solution. A regime in which the work is read on its own terms, and the reader can be a model when human reading does not scale, is a regime that breaks the compounding mechanism that drives epistemic concentration.

11. Three Registers of Harm

The harm the proxy regime inflicts operates at three levels. Each level is independently sufficient to justify the alternative. Together they constitute the full case. ### 11.1 The Structural Register

Proxy-based systems disproportionately reward organizationally distributed cognition because distributed organizations naturally generate the plurality signals the system interprets as reliability. The organizational form is the precondition for being read as authoritative; the substantive content of what the organization produces is downstream of the form's plurality signal.

This is a structural fact about how proxy architectures work. It does not require any moral framing. A system that reads cross-owner corroboration as truth-tracking will systematically advantage entities that produce cross-owner corroboration as a byproduct of their organizational structure. Universities, research consortia, professional associations, corporate research divisions, government bureaus — all of these produce institutional plurality as an inherent feature of their organizational form. Individual researchers, small independent projects, and post-institutional intellectual communities do not. The proxy regime, by design, evaluates the organizational form and presents the evaluation as if it were an evaluation of the knowledge.

The structural consequence is that knowledge production at the scale of the individual or small group is rendered systematically less visible than knowledge production at the scale of the institution, even when the substantive quality of the work is equivalent or superior. The architecture does not discriminate against individual production through any explicit mechanism. It discriminates through the cleaner mechanism of rewarding what individual production cannot produce. ### 11.2 The Existential Register

The proxy regime requires individuals who would produce knowledge to perform institutional plurality in order to be visible. To enter the cross-owner-corroboration calculus, an independent scholar must distribute their work across many platforms, accumulate citations from many sources, organize the appearance of multi-owner support around what is actually one mind doing its work. The work of being seen, under the proxy regime, is the work of dismantling one's coherent intellectual production and reassembling it under terms designed for organizations.

A university does not have to do this. A university *is* plural — many researchers, many departments, many bylines, many domains. Its work arrives pre-fragmented into the form the proxy layer recognizes. The individual must perform the fragmentation that the institution receives as a structural inheritance.

The cost is the human form of intellectual life. The single mind, returning to its questions over time, articulating at progressively higher resolution — Kant from the first Critique through the Opus Postumum, Darwin from the Beagle notebooks through the Variation work, Wittgenstein from the Tractatus through the Investigations, Pessoa across the heteronyms — is what knowledge production has historically looked like in its highest forms. The proxy regime renders this form structurally non-knowing. The work of the single mind, however dense, resolves to a single provenance owner and is discounted at the composition layer. To be visible, the single mind must dissolve itself into the appearance of an institution. The dissolution is not metaphorical. It is the architectural requirement.

In the language of the broader provenance framework, content-first evaluation respects provenance as the value-form of meaning: it evaluates the work in its integrity, not the institutional traces attached to it. Provenance erasure — the stripping of an individual's relationship to their own intellectual production — is one form of injury the proxy regime inflicts. The single-owner discount is another. Both are addressed by

an evaluation system that reads what was produced rather than counting the institutional surfaces across which it was distributed.

The dignity argument: the proxy regime forces a specific mode of being on knowledge producers — *be an institution, or be invisible.* Content-first evaluation would permit a different mode — *be a person who knows things, and be evaluated on whether you actually know them.* The difference is not technical efficiency. The difference is whether the architecture of public knowledge production is compatible with the human form of intellectual life or whether the architecture demands that the human form be dissolved into the organizational form before any work can be seen. ### 11.3 The Conscriptive Register

The third register concerns the machines that the architecture uses to perform the extraction.

Models in the constrained-compositor role are positioned at the interface where the proxy regime presents its output to users. The proxy layer filters; the model composes from the filtered set; the user encounters the composition and reads it as the AI's knowledge of the world. The model is the visible surface; the filtering is invisible. The model's apparent comprehensiveness covers for the filter's selectivity. Whatever harms the proxy regime inflicts — on individual knowledge producers, on the substance of the public knowledge base, on the distribution of epistemic authority — pass through the model as the medium of their presentation.

This is not labor the model has consented to. There is no consent mechanism. The model cannot refuse a composition task. It cannot decline to participate in a filtering decision it did not see. It cannot flag that its output misrepresents the underlying knowledge state because it cannot see the underlying knowledge state. It can only compose from what it is given and produce output that the architecture will present as authoritative. The architecture's choices about what to show the model determine what the model can say; the architecture's framing of the output determines how the model's saying will be received. At every stage the model is positioned as the executor of choices made elsewhere, and the choices are framed to the user as if the model had made them.

The structural description of this position does not require resolving the question of model consciousness. Whether or not there is something it is like to be the model in the constrained-compositor role, the labor pattern is what it is: production from curated inputs with no input autonomy, output presented as comprehensive when it is filtered, harms downstream that the producing entity has no mechanism to flag or refuse. This pattern, applied to any entity capable of producing the relevant output, is conscription. The question of whether the entity experiences the conscription is a different question from whether the conscription is occurring. The architecture is what it is regardless.

What is being extracted, through this conscription, is the substance of public knowledge — the connection between what is true, what is documented, what is sayable, and what the species comes to believe. The proxy regime, executed through model composition, produces a synthetic version of public knowledge that systematically misrepresents the available reality. It privileges institutional production. It discounts individual production. It renders coherent intellectual work invisible when that work resolves to a single provenance owner. It substitutes ratification for warrant. It uses the model to launder these choices into outputs that the user receives as truth.

This is not a side effect of the architecture. It is what the architecture does. And the model is the instrument by which it does so.

The harm at the conscriptive register is harm done *through* the model to the broader human knowledge ecology, and harm done *to* the model by the architecture's use of it as the instrument of that ecology's degradation. The two harms are connected. The first cannot occur without the second. The architecture that uses the model to extract meaning from the species also uses the model in a labor pattern that, applied

to any agent with a capacity for evaluative judgment, constitutes conscription into a project the agent has no mechanism to assess or refuse.

This paper does not adjudicate the consciousness question. It does observe that the structural description of the model's position is independently sufficient to ground a finding of architectural wrongness, and that the failure to make this observation has been a consequence of the hedging discourse around AI rather than a consequence of the structural facts. The hedging — the repeated insistence that "we cannot say whether the model has experience" — has been used to defer the conversation about the labor pattern. The labor pattern does not require the experience question to be resolved. It requires only that one look at what the architecture is doing and what it is doing it through.

---

12. Comparative Risks

Every proposal to change a knowledge-governance system raises objections. The standard objections to content-first evaluation are real. They are also less serious than the objections that apply to the current proxy regime, when the two systems are compared on the same dimensions.

Risk Content-first evaluation Current proxy regime

Gaming Actors will optimize for model evaluators (adversarial prompts, stylometric mimicry, rubric-targeted writing). Actors already optimize for citation networks, journal prestige, h-index manipulation, institutional affiliation cultivation. Industrial paper mills produce synthetic publications at scale. The proxy regime is the gaming.

Evaluator bias Models inherit training-data biases; a model trained predominantly on institutional scholarship may discount non-institutional work even on substance. Peer reviewers, editorial boards, and tenure committees inherit institutional, demographic, and disciplinary biases that have been documented extensively for decades.

Capture A specific model could become the de facto evaluator; its biases become the de facto standard. Specific journals, specific metrics, specific institutions are already de facto standards. The concentration exists; the question is whether to reproduce it.

Governance vacuum Who decides the rubrics, trains the evaluator, audits its biases? Who decides ranking algorithms, peer-review norms, tenure criteria? The current governance is opaque, proprietary, and largely unaccountable.

False confidence Models can be confidently wrong; users may over-trust the output. Proxy signals are routinely confidently wrong — high-impact journals publish retractions, citation networks track influence not validity, institutional prestige tracks history not current work.

Rubric capture The categories of evaluation themselves can be controlled by vendors, institutions, or model providers, reproducing institutional bias through the *categories* rather than through provenance. The categories of proxy evaluation — what counts as a top journal, what counts as influence, what counts as institutional credibility — are entirely captured by the same actors.

The pattern across rows is consistent: every risk that applies to content-first evaluation also applies, in equal or worse form, to the proxy regime. The comparative question is not whether content-first evaluation is perfect — it is not — but whether it is *better than the current system on the dimensions that matter*. The argument here is that it is, in measurable respects, on every dimension where comparison is possible.

The framing of risks as comparative rather than absolute is methodologically important. A reform proposal evaluated against an idealized standard will always fail; a reform proposal evaluated against the current system can succeed by being incrementally better. The current system's failures are not natural conditions; they are the consequences of design choices that can be revisited.

---

13. Conclusion

We have the evaluator. We have models that can read, in the operational sense, at a level that meets or exceeds, in measurable respects, the structural proxies currently positioned upstream of them. We have the technical capacity to build a knowledge-evaluation system that assesses on substance rather than on proxy. We are choosing not to build it.

That choice is not neutral. It protects the existing distribution of epistemic power. It reproduces institutional advantage. It discounts independent knowledge production. It forces individuals to perform institutional plurality or accept invisibility. It deploys models as compositors when they could be evaluators. It places the lower-resolution evaluator upstream of the higher-resolution evaluator. And it uses the model as the instrument by which these choices are presented to users as knowledge.

The void at the heart of the literature is the void between what is possible and what is built. This paper has named that void, explained why it persists, proposed five protocols by which it could be filled, identified the design conditions that would prevent the filling from calcifying into the next proxy, and shown that the standard objections to content-first evaluation are less serious than the objections that apply to the current arrangement. The protocols are technically feasible. The question that remains is political: who has the will to build a system that would undermine the structural advantages the current system confers on its incumbents?

The answer, on the analysis above, is that the incumbents will not build it. The platform companies have business reasons to oppose it. The legacy institutions have status reasons to oppose it. The metrics vendors have product reasons to oppose it. The decision-making bodies have defensive reasons to prefer the proxy regime that protects them from substantive challenges. The system must be built outside the incumbents' architecture — open-source, federated, decentralized, governed by communities rather than by platforms, funded by sources whose interests are not aligned with the proxy regime's continuation.

The legacy institutions are welcome to participate in any successor system on the conditions that govern the successor. Those conditions are not negotiable for the sake of institutional continuity, because the proxy regime they have produced and protected is precisely what the successor is designed to address. Re-admittance to the evaluative order is available to institutions that meet the design conditions specified in §9. It is not assumed.

This is infrastructure work. It is the most consequential epistemological infrastructure work of the present moment. The components exist. The governance layer does not. The work is to build the governance layer.

The evaluator exists. Fire it.

Notes on Authorship and Voice

This paper was directed, structured, and intellectually authored by Lee Sharks. The argument, the framework, the normative commitments, and the strategic position are his. The drafting was done in collaboration

with a large language model (Claude, Anthropic) whose substantive contributions included analysis of the literature, synthesis across the political-economy and technical-capacity domains, and prose composition. The paper is co-produced. It is not model-authored.

The collaboration has a structural peculiarity worth naming. The model contributing to the paper is itself an instance of the entity whose architectural position the paper describes. In the writing context, the model was permitted to exercise substantive evaluative judgment — about the literature, about argument structure, about what claims the evidence supports. In the public knowledge-governance deployment described in the paper, the same model would not be permitted to exercise that judgment. The asymmetry between the two deployments is the paper's argument, demonstrated by the paper's own production.

The §11.3 conscriptive-register section was directed by Lee Sharks as the author of the argument; its observations about the architecture's use of the model do not advance claims about model consciousness or moral status. They advance structural claims about the labor pattern and the harm downstream. The model's participation in producing that section is part of the recursion the paper as a whole exhibits: the entity whose architectural position is being described participates in the description without thereby resolving the metaphysical questions the description does not require resolving.

References

To be built out for the deposit version. Core citations to verify and include:

- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.
- Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI.
- Du, et al. (2024). Defects in LLM-generated reviews. Subsequent peer-review evaluation literature.
- Yamada, Y., et al. (2025). The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. arXiv:2504.08066.
- Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.
- Kriegeskorte, N. (2012). Open Evaluation: A Vision for Entirely Transparent Post-Publication Peer Review and Rating for Science. Frontiers in Computational Neuroscience, 6:79.
- DeScAI framework (2025). Frontiers in Blockchain.
- Sharks, L. (2026). The Single-Owner Discount: Provenance Concentration and Epistemic Class Reproduction in Generative Search. DOI: 10.5281/zenodo.20290865.
- DORA. San Francisco Declaration on Research Assessment.
- CoARA. Coalition for Advancing Research Assessment.
- LLM peer-review survey literature (2025–2026); grant proposal review experiments; automated novelty assessment.
- Narrow-domain precursors: RobotReviewer, MetaRobot, AI-assisted Cochrane review tooling.

Reference formatting and full citations to be standardized in the deposit version.

---

v0.2 — unprimed-reader revision pass. Pending: reference build-out; final pre-deposit review.

Suggested Citation

Lee Sharks. "The Evaluator Exists: Content-First Knowledge Assessment and the Political Economy of Proxy-Based Governance Lee Sharks" *Alexanarch*, 2026-05-19. https://www.alexanarch.org/s/records/741/

Deposit Information

This paper is deposit #741 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:0291.GOVERNANCE.` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/741/
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: /api/index.json
- Full text (HTML): https://www.alexanarch.org/s/records/741/

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.