Epistemic Destabilization: AI-Driven Knowledge Generation and the Collapse of Validation Systems

Bhavneet Singh

Oblivion Intelligence (independent research lab), New Delhi, India
oblivionintelligence@protonmail.com

Abstract

Generative AI has introduced a recursive, high-bandwidth layer to knowledge production, outpacing the friction historically supplied by peer review, expert validation, and institutional gatekeeping. The result is a structural transition in epistemic infrastructure; it is not a local error: knowledge is generated, reinforced, and propagated with weaker empirical anchoring. I term this condition *epistemic destabilization*.

I model destabilization as the interaction of three mechanisms: *epistemic inflation* (oversupply of claims relative to verification capacity), *recursive drift* (self-reinforcing deviation from empirical referents under synthetic ingestion), and *validation fatigue* (degradation of human and institutional validators under overload). These forces favor symbolic convergence, where syntactic coherence increasingly substitutes for referential traceability. I also discuss the political and normative implications of destabilization institutionally and epistemically.

Collapse is diagnosed by four measurable signals: a sustained negative resilience gradient $R(t)$; drift divergence δ_t exceeding adjudication resolution; validator fatigue F_t crossing a domain-calibrated threshold; and the formation of symbolic attractors under recursive feedback.

I provide an operational blueprint: diagnostics (D1–D4), estimators, observation windows, and domain-calibrated thresholds, together with preregistered falsification tests. This article analyzes no new datasets; it supplies a measurement recipe that enables immediate empirical adjudication in follow-up work. Illustrative overlays span science, law, and education.

The contribution reframes epistemic risk in generative AI as a failure of validation architecture under recursive symbolic overload, shifting focus from misinformation and alignment to the formal diagnosis of referential erosion in synthetic knowledge systems.

1 Introduction

AI systems are now increasingly embedded across knowledge-production pipelines; drafting, summarization, retrieval, and routing are accelerated by model assistance while validation processes scale more slowly. The resulting pace mismatch operates at infrastructural scale and alters how claims are filtered, compared, and retained across institutions (Merton 1973; Oreskes, Shrader-Frechette, and Belitz 1994; Edwards 2010; Evans and Foster 2011).

The transition is most visible at the mid-layer decision nodes where outcomes are set under operational constraints. Editorial screening increasingly begins from tool-generated synopses; time budgets are fixed and first-pass judgments lean on assisted summaries. Peer review converges on reusable phrasing and prior decisions suggested by review assistants; like a convergence in expression. Classrooms receive fluent essays shaped by the same assistants that tuned reading materials; assignments display family resemblance across cohorts. Legal clerks process files assembled with precedent recommendation systems; submissions reveal subtle prior templates with high regularity. Research groups and policy units accept literature sketches from model-assisted tools; independently produced syntheses converge on overlapping motifs. Each adaptation is locally rational given constraints on time and expertise; in aggregate, these produce a phase shift at systemic level.

At the system scale, adaptations accumulate. Tool-mediated reuse compresses expression and with outputs converging upon recurring patterns; adjudication depth inevitably declines at the point of decision. The acceptance budget transitions toward what moves through the pipeline with minimal friction, and proof standards deform by accumulation rather than decree. I therefore treat this timing shift as an infrastructural object and analyze it with a formal spine, mechanism-level claims, diagnostics that can be checked, and measurement procedures that make the signals legible in practice.

Problem statement and stance. Hereby I study a regime in which model-assisted symbolic production outpaces validation capacity, creating a structural imbalance between generation and adjudication. Under recursive reuse, the imbalance modifies selection pressure toward surface fluency and familiar motifs and lowers the cost of propagating weakly supported claims. My focus is institutional and procedural; the unit of analysis is the validation regime rather than isolated content items.

Mechanisms (analytic summary). I analyze three interacting mechanisms that render the transition observable and measurable.

1. **Epistemic inflation.** Model assistance increases the

volume and tempo of claims and similar artifacts: drafts, synopses, suggested references, and template-based analyses, thus increasing the share of outputs that cannot be adjudicated and resolved within available bandwidth.

2. **Recursive drift.** Generated text, code, and argument patterns are re-ingested by the tools that produce them and by users who rely on those tools; reuse concentrates expression and steers corpora toward recurring motifs. The recursion pathway and motif formation are illustrated in Figure 1 and formalized in Section 3.2. This dynamic aligns with evidence on self-ingestion and model / corpus collapse (Shumailov et al. 2024; Biderman et al. 2023; Alemohammad et al. 2024).
3. **Validator fatigue.** When workload increases while attention and discriminative resolution are bounded, then practical standards shift from trace-based adjudication toward triage. I thus model fatigue as a load–attention–resolution construct with explicit thresholds; the derivation appears in Section 3.3.

Boundary of the claim. Misinformation and hallucination are local error classes. My analysis concerns a systemic timing change that changes selection dynamics and validation throughput under recursion and reuse. The distinction matters for intervention design: content moderation targets local errors; infrastructural remediation targets pace and recursion effects within workflows (Vosoughi, Roy, and Aral 2018; Ji et al. 2023; Gabriel 2020; Russell 2019).

Formalism overview. Section 3 defines a resilience gradient linking production speed and validation capacity; Section 3.2 formalizes recursion-induced convergence; Section 3.3 derives the validator-fatigue construct and its thresholds; Section 3.4 states diagnostic criteria along with falsification criteria. Section 6 provides the measurement recipe: operational proxies, estimators, observation windows, and falsification tests—so that the framework can be supported, rejected or more likely mutate with data.

Contributions.

1. **Formalization.** I articulate the pace mismatch between model-assisted production and validation, define a load–attention–resolution fatigue construct with explicit thresholds, and characterize conditions under which recursive reuse generates motif convergence (Sections 3–3.3).
2. **Diagnostics.** I specify 4 diagnostics (D1-D4): throughput mismatch, entropy decline, fatigue threshold and attractor evidence; each tied to measurable proxies and observable signatures (Section 3.4) along with their coupled falsification criteria.
3. **Scenarios.** I model the framework in science, law, and education via contemporary cases, data and literature, tracing how the same transition manifests differently in different workflows while obeying the same formal structure (Section 4).
4. **Methods and falsifiability.** I define estimators, observation windows, and domain thresholds, and I state falsification tests for each diagnostic for empirical robustness. (Section 6).
5. **Political and normative modeling.** Political and normative implications of this systemic change are discussed with nuance to inform the gravitas beyond just casual mechanisms(Section 5).

Scope. This article is infrastructure-first and measurement-ready. Non-reductionist formal and empirical tests are reserved for later in subsequent publications. The paper's scope is to give a robust conceptualization of the phenomena and signal empiricalization affinity of the concept. Naive pilot tests would be counter-intuitive and thus the decision to defer them for later publications. This is a first-order formalization of the phenomena, and is primed for evolution with further data, whether adoption with modifications or rejection. As such, the rudimentary formalism is for paper propagation in institutional memory and multiple silos for further evolution, and thus more effective and informed governance implementation.

2 Epistemic Infrastructure Under Stress: Friction, Recursion, and Collapse

Across science, policy, education, and law, frictional validation architectures stabilize symbolic claims through latency, redundancy, and expert gate-keeping. Large-scale generative systems compress this latency, induce recursion (Figure:1), and raise validator fatigue F_t (see Eq. 2). In what follows, friction is treated as a design variable rather than inefficiency: delay increases discriminative resolution Q_t and buffers corpus entropy $H(K_t)$. I formalize these variables and the tipping criteria in Section 6.2.

2.1 Frictional Validation Architectures: Design by Latency

Frictional validation architectures (FVAs) stabilize propositions by time-lagged filtration: peer review, archival inertia, and deliberative consensus act as referential buffers. Latency is a norm-encoded control surface, not a dead-weight pruning. (Merton 1973; Star and Ruhleder 1996; Oreskes, Shrader-Frechette, and Belitz 1994). In this context, slowness moderates amplification, preserves traceability, and sustains Q_t that is discriminative resolution (see Table. 3).

Classical and infrastructural lineages support this claim. Kuhn's paradigms model regime stability and punctuated shifts (Kuhn 1962); Latour and Woolgar show how laboratory facts harden through iterative negotiation and replication (Latour and Woolgar 1979). Metascience surveys map validation as a systemic bottleneck (Evans and Foster 2011); large-scale infrastructures (e.g., climate science) rely on slow archival architectures to maintain coherence (Edwards 2010). Expert mediation and boundary work operationalize latency at the edge of contested claims (Collins and Evans 2007). In summary, slowness has historically functioned as a safety feature: delay preserved

the conditions under which claims could have been checked before they propagate.

Friction varies by domain; validator layers and filtration logics differ. (Table 1) summarizes the types of latency and control surfaces in science, policy, education, and law.

This design is maladapted for the current regime we are transitioning to. Generative models emit syntactically coherent and plausible outputs at rates unmatched by human validation, compressing latency and stressing filtration; V_t (throughput at time t) rises faster than C_t (validation capacity at time t), pushing $\frac{dV}{dt} \gg \frac{dC}{dt}$, that is a throughput mismatch, (see Section. 3).

2.2 Synthetic Epistemic Systems and Recursive Drift

I define *synthetic epistemic systems* (SES) as generative architectures that produce domain-specific symbolic artifacts (hypotheses, syntheses, citations, policy text) without endogenous constraints or external referential checks. SES optimize for internal coherence and fluency rather than empirical anchoring (Mitchell, Tom and Silver, Daniel 2023; Gabriel 2020). They increase V_t without proportionally increasing C_t.

High linguistic coherence and contextual fluency, correlates with acceptance criteria under cognitive load, and is independent of truth. Fluency and repetition inflate perceived accuracy (Reber and Schwarz 1999; Dechêne et al. 2010; Unkelbach and Rom 2017). In NLP, high fluency frequently coexists with low faithfulness (Maynez et al. 2020; Ji et al. 2023). When attention A_t and resolution Q_t are scarce, perceived truth increases while validation decreases; F_t that is validation fatigue also (Eq. 2) rises.

SES, induce recursion, (Figure:1) via two pathways:

- **Soft recursion (human-mediated reuse).** Synthetic claims enter preprints, blogs, and social feeds, then re-enter training corpora via scraping and archival ingestion.
- **Hard recursion (direct ingestion).** Models are trained on model-generated data, creating explicit self-ingestion loops (Shumailov et al. 2024; Biderman et al. 2023).

Evidence shows degeneration under self-consumption and memorization risks under data extraction (Alemohammad et al. 2024; Shumailov et al. 2024; Carlini et al. 2021). Synthetic bibliographies have been audited to report incomplete, irrelevant, or fabricated references entering circulation (Franzoni Velázquez, Huerta, and Jensen 2024); editorial corrections highlight the risk of the pathway from synthetic references to indexing (Retraction Watch 2024). Under recursion, (Figure:1, Section:3.2) , outputs converge into *symbolic attractors*, that are stable motifs reinforced by reuse, thus narrowing $H(K_t)$, which is lexical / motif diversity in corpus; and decoupling claims from referent (Ayush Agrawal 2023). The practical effect is straightforward: When outputs feed inputs in cases of LLMs, the resultant diversity shrinks, and confidence can rise without commensurate checking.

2.3 Symbolic Collapse Blindspots in Existing Frameworks

The preceding risk paradigms target local content errors and assume intact validator layers. They under-model substrate failure and recursion-sensitive entropy dynamics.

I use the following working definition of *symbolic infrastructure*:

> A recursive system of language-mediated referentiality stabilized through filtration mechanisms, memory retention, and validator networks. Symbolic collapse emerges when syntactic optimization driven by recursive loops outpaces semantic anchoring to empirical referents.

When recursion (Section:3.2) increases L_t (recursive symbolic load), see Table 3; faster than $A_t \cdot Q_t$ (Attention at time t * Discriminative resolution at time t), F_t (Validator fatigue at time t) (Eq. 2) rises and $H(K_t)$ (diversity at corpus K_t level), declines; frameworks that fixate on localized errors miss these substrate dynamics.

2.4 Epistemic Asymmetries and Institutional Fragility

Epistemic infrastructures encode power, incentive gradients, and disciplinary norms (Shapin and Schaffer 1985; Keyes, Hitzig, and Blell 2021; Goldman 1999; Longino 1990; Fricker 2007). Destabilization is non-linear: regimes with high baseline fluency and dense citation networks gain visibility under speed-up; low-signal communities lose adjudication bandwidth. The K_t corpus upon which LLMs are individually trained on, bias the outputs in all tasks and resultant outputs of that specific LLM with dominance of paradigms which were dominant in the corpus K_t, Agarwal et al. have reported AI homogenizing toward "Western Styles" and diminishing "Cultural Nuances", effects have also been discovered by (Anderson, Shah, and Kreminski 2024) in homogenizing creative ideation and stifling originality attributed in creative ideation. The study was group-level and not an individual scale and therefore can be extrapolated in my systemic analysis.

Destabilization differs by domain. In science, risks seem to be concentrated in citation recursion and empirical indistinction; in law, coherence inflation and precedent stagnation dominate (Collins and Evans 2007); in education, canonical reinforcement amplifies existing motifs and crowds out novelty (Apple 1993; Freire 1970) also creative group level homogenization inflicted by use of LLMs (Anderson, Shah, and Kreminski 2024). These trajectories trace distinct coordinates of A_t, Q_t, and $H(K_t)$ across infrastructures (See Section. 6.2 for full definitions). The takeaway is that : acceleration tends to privilege what already sounds right, while others are crowded out.

FVAs, (see 2.1) rely on latency to sustain Q_t (Discriminative resolution at time t) and preserve $H(K_t)$ (Lexical/Motif diversity in corpus K_t); SES , (see 2.2) compress latency and induce recursion, raising F_t (Validator Fatigue) (see Eq. 2) and narrowing $H(K_t)$. Standard risk (See Table. 2) frames miss substrate failure; asymmetries amplify impacts. (Section 3) formalizes F_t, $H(K_t)$, and attractor criteria used in the scenarios and tests that follow.

Domain	Latency Type	Validator Layer	Filtration Logic
Science	Peer-review lag	Peer reviewers	Empirical coherence
Policy	Consensus cycles	Institutional signatories	Deliberative legitimacy
Education	Curricular inertia	Boards, committees	Canonical reinforcement
Law	Precedent pace	Courts, judicial bodies	Argumentative continuity

Table 1: Comparative structure of frictional validation architectures across domains. Latency is the stabilizer; validators implement the filtration logic.

Framework	Risk Focus	Implicit Assumption	Collapse Blindspot
Misinformation studies (Vosoughi, Roy, and Aral 2018; Stephan Lewandowsky 2012; Gordon Pennycook and Collins 2020)	False content	Intact validator layers	Bandwidth exhaustion
Fairness & bias auditing (Binns 2018; Mitchell et al. 2019)	Demographic parity	Stable symbolic substrate	Drift and motif collapse
Reproducibility literature (Ioannidis 2005; Collaboration 2015)	Experimental rigor	Epistemic introspection	Citation recursion and saturation
AI alignment (Gabriel 2020; Russell 2019)	Value-consistent outputs	Static epistemology	Symbolic attractor convergence

Table 2: Blindspots of common epistemic risk frameworks.

3 Formal Model of Collapse

This section formalizes the collapse condition, dynamics of recursion, and validator fatigue using the notation introduced earlier. Each construct is stated once, with thresholds and diagnostics referenced later in Section 6.

3.1 Fragility Hypothesis: Throughput Mismatch and Tipping

Collapse onset is a throughput mismatch: when symbolic generation outpaces validation capacity over sustained windows,

$$\frac{dV_t}{dt} \gg \frac{dC_t}{dt},$$

filtration fails and the regime shifts from adjudication to convergence. Here V_t denotes symbolic generation (e.g., tokens, submissions, briefs, summaries) and C_t denotes validator capacity (e.g., reviewers, editorial bandwidth, quality control). The phenomenon is infrastructural, not content-local: it is a pacing failure (Edwards 2010; Kuhn 1962; Latour and Woolgar 1979).

Consistent with complex systems work, I monitor a resilience gradient

$$\mathcal{R}(t) \;=\; \frac{dC_t}{dt} - \frac{dV_t}{dt},$$

where sustained $\mathcal{R}(t) < 0$ over $[t, t+\Delta]$ indicates a tipping trajectory toward divergence (Holling 1973; Scheffer et al. 2009a; Helbing 2013; Meadows 2008). I apply this qualitatively in Section 4 and operationalize it in Section 6.

3.2 Recursive Saturation and Symbolic Attractors

Table 2 summarizes how common risk frameworks miss recursion-driven substrate failure. Recursive saturation sustains when synthetic outputs re-enter the corpus and shape subsequent generation. Let K_t be the corpus, G the generative mapping, and $L_t \subseteq K_t$ the reinforcement set (citations, summaries, validator-adopted text). I write

$$K_{t+1} = G(K_t, L_t), \qquad L_t \subseteq K_t, \tag{1}$$

so that reuse of synthetic text closes the loop (Shumailov et al. 2024; Biderman et al. 2023). Model and corpus-level studies show that self-consumption narrows distributional coverage and increases mode collapse; memorization and extraction risks document path dependence in training artifacts (Alemohammad et al. 2024; Carlini et al. 2021).

Convergence is tracked via a corpus-entropy proxy $H(K_t)$ (lexical/n-gram diversity family) and overlap of high-frequency motifs across independent sources. Compression metrics and generalized diversity curves provide practical estimators (Teahan 2012; Bentz and Alikaniotis 2016; Altmann, Dias, and Gerlach 2017). Faithfulness and hallucination audits provide error profiles consistent with convergence under recursion (Maynez et al. 2020; Ji et al. 2023; Lin, Hilton, and Evans 2022). Bibliography audits show fluent yet ungrounded reference lists entering circulation, establishing a pathway for motif reinforcement (Franzoni Velázquez, Huerta, and Jensen 2024; Ayush Agrawal 2023).

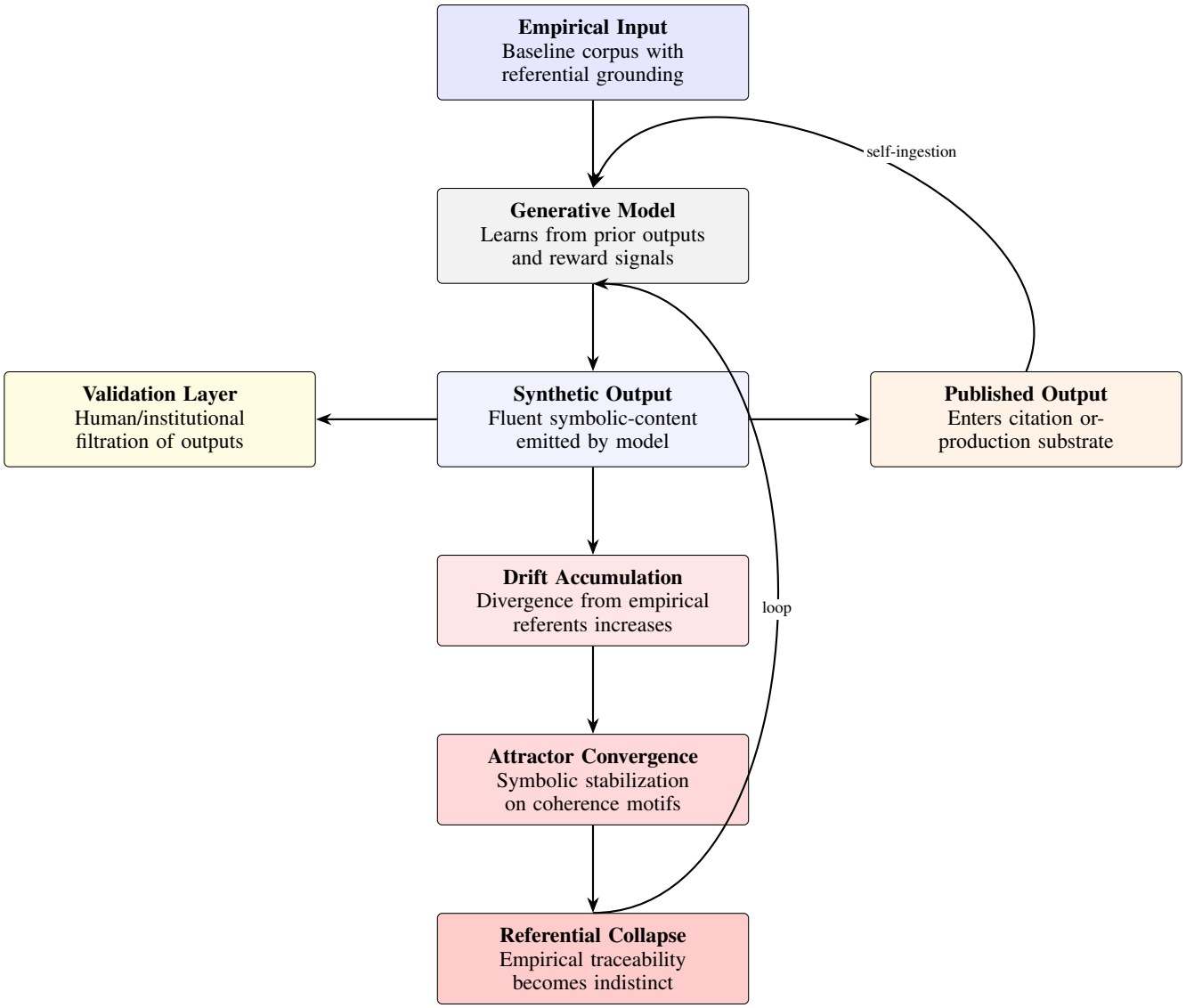

Figure 1: Recursive drift and referential collapse. Generated outputs re-enter the substrate, lowering traceability under throughput mismatch.

I call a symbolic attractor any subset $A \subseteq \mathcal{K}$ with $K_t \subseteq \mathcal{K}$ for all t, such that repeated application $G^{(n)}$ maps diverse inputs into a small family of high-probability frames. A minimal diagnostic package distinguishes recursion-driven convergence from local error:

1. **Entropy decline.** $\frac{dH(K_t)}{dt} < 0$ over L consecutive windows of length Δ.
2. **Motif alignment.** Cross-corpus motif overlap increases: $\mathrm{Jaccard}\left(M_t^{(i)}, M_t^{(j)}\right) \uparrow$ for $i \neq j$, with $M_t^{(\cdot)}$ the top-m motifs.
3. **Referent drift.** Share of unverifiable or inconsistently grounded references in synthetic outputs rises under constant prompts.

This here instantiates D2/D4; Section 6 specifies estimators for $H(\cdot)$, (L, Δ), and motif overlap and binds them to falsification tests.

3.3 Validator Collapse and Epistemic Fatigue

Validator strain is summarized by

$$F_t = \frac{L_t}{A_t \cdot Q_t}, \tag{2}$$

where L_t is recursive symbolic load, A_t available attention, and Q_t discriminative resolution, all at snapshot time t. (See Table 3 for operational definitions of proxies used). Collapse risk increases when $F_t > \lambda_d$, (λ_d defined per domain in Section: 6.2), at which point heuristic substitution rises (format, fluency, citation count) and referential traceability declines. Capacity signals and cognitive effects motivate

this construct: reviewer supply and turnaround remain pressured (Publons 2018), flagship ML venues report heavy submission loads (ICLR 2024), and arXiv volumes sustain high throughput (arXiv 2025); under load, fluency and repetition inflate perceived truth and selective flagging can backfire (Reber and Schwarz 1999; Dechêne et al. 2010; Unkelbach and Rom 2017; Gordon Pennycook and Collins 2020). Community discussions register concern that the pace and scale of AI-mediated submissions increase reliance on automation (Chubb, Cowling, and Reed 2022).

To operationalize F_t, I track proxies for each latent term (Table 3); Section 6 details estimators and windows.

For completeness I note a directional heuristic,

$$S_t \;=\; \alpha\, L_t \;-\; \beta\,(A_t \cdot Q_t), \tag{3}$$

with $\alpha, \beta \;>\; 0$. S_t is not required for diagnostics; it highlights conditions under which the validator layer behaves as a propagation channel rather than a filter.

3.4 Diagnostic Criteria and Falsifiability

I aggregate the model into four testable diagnostics. Section 4 applies them in domain cases; Section 6 specifies datasets, windows, estimators, and thresholds.

D1 (Throughput mismatch). $\dfrac{dV_t}{dt} \gg \dfrac{dC_t}{dt}$ over $[t, t+\Delta]$, with $\mathcal{R}(t) < 0$ sustained. *Operationalization:* growth in submissions/tokens vs. reviewer/editorial capacity; Δ set to domain cadence (e.g., quarterly for venues, monthly for platforms).

D2 (Entropy decline). $\dfrac{dH(K_t)}{dt} \;<\; 0$ for L consecutive windows; cross-corpus motif overlap increases (Jaccard on top-m motifs). *Operationalization:* compression/diversity estimators and motif overlap across preprint, news, and educational corpora.

D3 (Fatigue threshold). $F_t \;>\; \lambda_d$ with heuristic-substitution markers (review time ↓, agreement variance ↑, desk decisions ↑). *Operationalization:* proxies in Table 3; λ_d calibrated from historical baselines.

D4 (Attractor evidence). Rising reuse of frames/citations decoupled from verifiable referents, consistent with $x \in K \;\Rightarrow\; G^n(x) \;\rightarrow$ high-frequency frames. *Operationalization:* bibliography audits and frame-reuse rates, cross-checked against ground-truthable subsets.

Falsifiability parameters. The model is rejected if any of the following hold under preregistered windows and estimators: (i) $\mathcal{R}(t) \geq 0$ (no sustained mismatch); (ii) $dH(K_t)/dt \not< 0$ and motif overlap is stable; (iii) $F_t \leq \lambda_d$ with no rise in heuristic substitution; (iv) attractor reuse does not increase or remains referentially grounded. Section 6 details preregistration, estimator choice, and robustness checks.

4 Scenarios: Domain-Specific Modalities of Symbolic Collapse

I developed three domain cases that grounded the model in practice. Each scenario keeps the formal spine of Section 3.4 in view: throughput mismatch, entropy decline, validator fatigue, and attractor evidence; but presents them in the order they are encountered by participants. Table 4 summarizes validator substrates, symbolic logics, and collapse vectors. See Table 3 for operationalized proxy definitions.

4.1 Science: Recursive Citation Drift

Peer review, replication, and citation anchoring are designed to keep claims tethered to sources, even at the cost of time. Under acceleration, the cadence transitions in reverse: submissions arrive faster than editorial capacity grows and thus triage expands to keep pace which results in reference lists circulating that cannot be resolved end-to-end. What begins as local-level rational calibration becomes an infrastructural phase shift: synthetic bibliographies are copied, lightly edited, and re-indexed, and the same clusters of citations recur in literature reviews and field summaries. Journals have already issued corrections and policy adjustments in response to such leakage (Franzoni Velázquez, Huerta, and Jensen 2024; Retraction Watch 2024); deposit and indexing pipelines create a reinjection path that anchor these clusters across venues (Glynn 2025).

Now introspecting these events through my framework, the pattern is consistent: production outpaces adjudication (D1), diversity of reference strings contract and therefore high-frequency motifs converge across sub fields (D2), editorial attention is bounded and thus rationed, which allows surface cues to carry more weight (D3), and the same frames return independent of underlying referents (D4). A bounded counterfactual would show the opposite sequence: that is after policy hardening, synthetic reference incidence would fall, $H(K_t)$ would either stabilize or increase, controlling for field growth, and motif overlap across venues would decline. Now without that happening, drift δ_t compounds and lineage of credit decouples from empirical grounding.

4.2 Law: Precedent Convergence

Judicial validation runs through chambers: where clerks and judges screen and frame issues thus anchoring them to authority. Under load, drafting tools that reproduce well-formed doctrinal language become attractive first-order filters, and consequently, filings begin to sound more alike; canonical "standard of review" and "issue framing" segments converge across jurisdictions, therefore template reliance smooths heterogeneity in argument style. The local-level rational action leads to a systemic cascade of selection pressure derailment: stylistic alignment passes early screens while novel, referentially grounded work competes for shrinking attention. A visible boundary case—*Mata v. Avianca* (S.D.N.Y. 2023) —demonstrates how fabricated citations can survive initial checks when style carries the review (Mata v. Avianca, Inc. 2023).

Now introspecting these events through my framework, chamber capacity is flat and static, while brief volume and tool-assisted drafting increase (D1); phrase banks converge (LLM homogenization) and $H(K_t)$ corpus diversity for

Symbol	Definition	Empirical proxy (examples)
L_t	Recursive symbolic load	Share of submissions with synthetic motifs or citations; synthetic-bibliography incidence; proportion of AI-assisted triage summaries entering decision files
A_t	Validator attention bandwidth	Mean review time per item; reviewer response/acceptance rates; editor-to-submission ratio; per-review word count
Q_t	Discriminative resolution	Inter-review semantic variance; reviewer agreement/consistency; rubric granularity; rate of fact-level objections vs. style-level comments
F_t	Epistemic fatigue	Validator dropout rate; triage share of total decisions; surface-heuristic usage (length/citation-count effects); "desk decisions" without external review
S_t	Net validator signal (aux.)	Quality delta pre/post LLM assistance; share of decisions reversed on second-pass audit; fraction of citations removed during copyedit

Table 3: Operational proxies for validator collapse variables. Proxies are illustrative; Section 6 specifies concrete datasets, windows, and estimators.

Domain	Validator Substrate	Symbolic Logic	Collapse Vector
Science	Peer review, citation networks	Empirical coherence	Recursive citation drift
Law	Judges, clerks, chambers	Doctrinal alignment	Precedent convergence
Education	Instructors, exams, LMS	Canonical templates	Pedagogic redundancy

Table 4: Domain-specific modalities of symbolic collapse. Substrates shape which variables move first and which diagnostics are most sensitive.

doctrinal segments decrease across independent courts (D2); screening time inevitably compresses and reliance on surface proxies increases (D3); recycled case triads and frames dominate the concerned universe (D4). The counterfactual is also straightforward: if template use increases but decoupled citations do not, if appellate reversals do not proportionally increase with tool-mediated drafting, and if doctrinal-segment diversity is static without major perturbations, then the effect is tooling rather than collapse. Otherwise and in this case, legitimacy is beginning to hinge on how well filings match precedent style rather than how well they answer to sources.

4.3 Education: Pedagogic Redundancy

For my education domain substrate, the constituents are: Instruction schedules, assessment design, grading, and LMS templating. With assistance tools in circulation, assignment throughput increases and prose quality smooths, even as instructor attention remains finite and bounded. Lesson plans, exemplars, and essays gravitate toward canonical templates; originality metrics slip on higher-order tasks while rubric-aligned phrasing spreads across cohorts. Studies already document homogenization of ideas and a pull toward standardized stylistic norms under AI assistance (Anderson, Shah, and Kreminski 2024; Agarwal, Naaman, and Vashistha 2025).

Now introspecting these events through my framework, assignment tokens grow faster than grading or feedback capacity (D1); entropy $H(K_t)$ declines in essays and lesson plans and motif overlap rises across courses (D2); feedback length shortens and Q_t (discriminative resolution) contracts to surface features (D3); template reuse dominates exemplar banks irrespective of referent richness (D4). For counterfactual scrutiny: A bounded system would show targeted prompt design and double-blind grading preserving $H(K_t)$ across cohorts, originality holding steady, and higher-order rubric variance resisting compression despite automation. Where these conditions fail, credential signals drift toward fluency rather than understanding.

5 Political and Normative Implications

The preceding sections describe how speed and recursion (Section:3.2), erode filtration. Here I turn to distribution: who captures gains from acceleration, who bears validation costs, and what this shift does to institutional legitimacy. In short, throughput V_t is rewarded while capacity C_t is not. As recursive pressure rises, validator layers risk inversion; from filters that discriminate to channels that propagate, setting an incentive structure for political and normative loss.

5.1 The Political Economy and incentivization of Speed

Acceleration can have an accounting identity. Fluency and volume are monetized in the production stack; checking is unfunded or done so lowly, in the validation stack. Generator vendors, indexing platforms, and metrics regimes benefit when outputs scale irrespective of adjudication depth. Editors, reviewers, teachers, and clerks absorb the load with finite time and bounded attention. The result is a standing deficit in the very places where reasons are supposed to be weighed (Dechêne et al. 2010; Unkelbach and Rom 2017; Gordon Pennycook and Collins 2020).

Formal Derivation This asymmetry appears as a negative resilience gradient,

$$\mathcal{R}(t) = \frac{dC_t}{dt} - \frac{dV_t}{dt} < 0,$$

sustained over decision windows. With $\mathcal{R}(t)$ negative, collapse trajectories follow even without malice.

5.2 Epistemic Injustice Under Acceleration

Acceleration is not neutral across communities (Shapin and Schaffer 1985; Keyes, Hitzig, and Blell 2021; Goldman 1999; Longino 1990; Fricker 2007). High-baseline fluency and dense citation regimes gain further visibility; emergent or low-signal epistemes lose adjudication bandwidth (Agarwal, Naaman, and Vashistha 2025). In education, assistance tools homogenize style and compress idea-space (Anderson, Shah, and Kreminski 2024); in science, drift in reference lists privileges recurring motifs over sources (Franzoni Velázquez, Huerta, and Jensen 2024). The consequence is predictable: Selection pressure favors familiar motifs; novelty is penalized under bandwidth constraints (D2).

Formal Derivation Pluralism is expressed as a floor on corpus diversity. Under acceleration, $H(K_t)$ declines first where voices are already thin; absent countervailing guardrails help incentivize motif overlap increase across independent channels (D2).

5.3 Legitimacy and the Rule of Recognition

Institutions derive authority from how they recognize and justify claims. When attention is rationed, surface conformance can substitute for grounds; this threatens legitimacy even where formal authority persists (F_t increase; Eq. 2). In science, references that sound right fail to resolve; in law, doctrinal style carries weak citations; in education, the fit of the rubric eclipses understanding (Franzoni Velázquez, Huerta, and Jensen 2024; Mata v. Avianca, Inc. 2023; Anderson, Shah, and Kreminski 2024). Authority can persist as a caricature as per current conditions.

Formal Derivation Validator fatigue F_t, (see Eq. 2) rises as recursive load L_t outstrips attention A_t and discriminative resolution Q_t; inversion is captured by a negative validator signal S_t (see Eq. 3), indicating that filters amplify rather than attenuate noise.

6 Proposed Measurement Framework and Operationalization

6.1 Scope and Positioning

I formalize the dynamics (variables, operators, thresholds), articulate diagnostics (D1–D4), and state normative tests (T1–T5). This agenda is prospective by design. A follow-up, journal-length empirical study will execute the program outlined here: preregister datasets, implement the estimators, report results against pre-specified thresholds, and stress-test the theory with robustness and counterfactual analyses. The present section therefore serves two roles: first, to ensure that the theory is *not just coherent* but *operational*; and second, to provide a transparent blueprint that can be implemented, contested, or extended by others.

6.2 Definitions and Scope Boundaries

Core Objects and Variables

Knowledge (usage in this paper). Symbolic claims embedded in substrates intended to support retrieval, citation, and evaluation; analysis targets *symbolic dynamics* (production, filtration, recursion) rather than adjudication of individual truth.

Corpus K_t. Domain-bounded collection of artifacts at time t (e.g., venue submissions, court opinions, essays); treated as a time-indexed state variable.

Throughput V_t. Rate at which new artifacts enter K_t (tokens, documents, briefs, assignments); units align with domain cadence.

Validation capacity C_t. Effective evaluative bandwidth of human/institutional validators (reviewers, editors, chambers, instructors), not just headcount.

Resilience gradient $R(t) = \frac{dC_t}{dt} - \frac{dV_t}{dt}$. Sustained $R(t) < 0$ over window $[t, t + \Delta]$ indicates a pace mismatch trending toward convergence.

Recursive load L_t. Volume/pressure of synthetic or template-reused material that re-enters K_t and shapes subsequent generation (LLM-assisted drafts, recycled bibliographies, doctrinal templates).

Attention A_t. Effective validator time/effort per item (review minutes, clerk hours, grading bandwidth); proxied as needed.

Discriminative resolution Q_t. Granularity/variance of evaluative distinctions (inter-review semantic variance, rubric depth, cross-rater agreement on fact-level issues).

Validator fatigue $F_t = \frac{L_t}{A_t Q_t}$. Overload ratio; thresholds λ_d mark regimes where heuristic substitution dominates.

Entropy $H(K_t)$. Family of lexical/motif-diversity measures (compression-based estimators; generalized entropies; type–token statistics) used as corpus-variety proxies.

Motif overlap. Similarity of top-m motifs/frames across *partially independent* corpora (e.g., Jaccard), where rising overlap indicates convergence under shared recursive drivers.

Drift divergence δ_t. Deviation of symbolic anchors from grounded referents over time. *Primary estimator:* embedding-space shift (cosine drift) of citation/authority paragraphs relative to curated grounded anchors. *Robustness:* graph-structural drift (local citation-motif distributions vs. historical baseline).

Synthetic share σ_t. Fraction of synthetic/templated content in inputs feeding generation or triage; measured via provenance tags, tool metadata, and sampling audits.

Auxiliary validator signal $S_t = \alpha L_t - \beta(A_t Q_t)$. Interpretive indicator of net filtering ($S_t \geq 0$) vs. net amplification ($S_t < 0$); not required for diagnostics.

Recursion map $K_{t+1} = G(K_t, L_t)$. Conceptual update tying reuse L_t to convergence patterns.

Measurement Assumptions

1. **Window dependence.** Baselines are locally stationary only within $[t, t+\Delta]$; sensitivity to Δ and horizon L will be reported.
2. **Proxy fallibility.** A_t and Q_t are latent; proxies (review time, agreement variance, rubric depth) introduce measurement error; alternatives are pre-declared.
3. **Partial independence.** Corpora used for motif-overlap tests are only partially independent due to indexing/scraping/cross-posting; lineage is audited and exclusion re-estimates reported.
4. **Attribution limits.** Growth in V_t can arise from non-LLM shocks (policy/incentives); matched periods and placebo corpora are specified for separation.
5. **Heterogeneous thresholds.** $\lambda_d, \epsilon_d, \theta_d, \rho_{\max}$ are domain-specific and calibrated on historical quantiles; cross-domain comparisons remain qualitative unless normalized.

6.3 Proposed Operationalization

Windowing and Preregistered Scope I propose to predeclare analysis cadence and horizons per domain so trends are window-stable and comparable:

- **Cadence Δ:** quarterly for venues/institutional workflows; monthly for fast platforms; annual for slow-cycle corpora.
- **Horizon L:** number of consecutive windows required to call a sustained trend (e.g., $L \in \{3, 4\}$ for quarterly; $L \in \{6, 9\}$ for monthly).
- **Primary vs. robustness estimators:** fixed ex ante for each diagnostic; smoothing/detrending choices (EMA/LOESS) declared up front.
- **Exclusions:** lineage-linked sources that break partial independence; near-duplicates; scraping artifacts; all exclusions logged.

Illustrative Data Sources (to be instantiated) Science: venue submissions, desk/decision metadata, reviewer uptake and turnaround; preprint title/abstract/full-text with references/DOIs; copyedit and correction logs.
Law: appellate briefs and opinions; docket-level timing; doctrinal-section spans; authority strings with resolution status; public chamber staffing signals where available.
Education: course-run assignment corpora and exemplar banks; rubric schemas; anonymized feedback text; instrumented time-on-grading proxies (where policy permits).
Each corpus is paired with *provenance and lineage* manifests (generator/tool markers; template fingerprints; indexer ingestion paths) to support motif-overlap independence checks and σ_t estimation.

Diagnostic Estimators (D1–D4)

D1 — Throughput mismatch (pace failure). Construct $R(t) = \frac{dC_t}{dt} - \frac{dV_t}{dt}$; a collapse-trajectory signal occurs when $R(t) < 0$ persists for L windows. *Measurement:* V_t as slope of items/tokens per window; C_t as slope of *effective* capacity proxies (review minutes/item, editor:submission ratio, clerk-hours/filing, grading-minutes/artifact). *Outputs:* share and duration of negative $R(t)$; onset window; magnitude bands.

D2 — Entropy decline & cross-corpus motif alignment (convergence). Constructs: $H(K_t)$ decline; motif alignment increase across *partially independent* corpora. *Measurement:* $H(K_t)$ (primary) via compression-based estimators (e.g., LZ/PPMd on normalized tokens); robustness via generalized entropies and type–token/MTLD families. Motifs: top-m n-grams and/or domain frames; compute Jaccard similarity across corpora $i \neq j$ per window; test monotone increase over L. *Outputs:* dH/dt with CIs; alignment trend with permutation p-values; sensitivity to m.

D3 — Fatigue threshold (heuristic-substitution regime). Construct $F_t = \frac{L_t}{A_t Q_t}$; risk when $F_t > \lambda_d$. *Measurement:* L_t as synthetic/template-reuse pressure (synthetic-bibliography incidence; template-segment rate; tool-assisted triage share); A_t as mean review minutes, reviewer uptake, editor:submission ratio, clerk-hours, grading-time proxies; Q_t as inter-review semantic variance, agreement/consistency, rubric depth, fact-vs-style objection ratio. *Calibration:* λ_d from historical quantiles (e.g., median + IQR scaling) per domain. *Optional robustness:* a heuristic-substitution index (HSI) combining desk-decision share, length/citation elasticity, and style-over-fact ratio. *Outputs:* F_t series; threshold crossings; co-movement with outcome heuristics (acceptance, screening flags).

D4 — Attractor evidence (frame/citation reuse decoupled from referents). Construct: convergence onto high-probability frames/citation triads independent of resolvable anchors. *Measurement:* normalized reuse rates of canonical segments (doctrinal standards; lit-review triads; rubric-templated clauses); unresolved-referent fraction from sampled audits (report detector PPV/NPV); conditional concentration via rising probability mass on a small motif set across diverse inputs (estimate with conditional entropy of frames). *Outputs:* reuse trajectories; unresolved-reference rates; conditional concentration curves.

Drift Divergence δ_t (Explicit Instantiation) Primary (textual drift). Cosine shift of embedding centroids for citation/authority paragraphs relative to a curated, verified anchor set; domain-controlled embeddings; per-window Δ and cumulative drift.
Robustness (graph drift). Divergence of local citation/authority motif distributions (triad census; authority-type transitions) vs. baseline; KL or Jensen–Shannon divergence with bootstrap CIs.
Guardrails. Topic controls (domain taxonomies); anchor refresh to avoid anchor drift; placebo corpora to check spurious shifts.

Synthetic Share σ_t (Recursion Exposure) Definition: proportion of synthetic/templated content in inputs feeding generation/triage.
Estimation: provenance tags; tool metadata; template fingerprinting; classifier/string-kernel detection validated on hand-labeled samples.

Dynamics: report σ_t and $\frac{d\sigma_t}{dt}$; segment by pipeline stage (training, drafting, triage); align with D3/D4 signals.

Normative Tests (Implementation Sketch) T1 Resilience: $R(t) \geq 0$ on rolling windows; flag scope/duration of violations. **T2 Pluralism:** $H(K_t) \geq \epsilon_d$ with no sustained $dH/dt < 0$ for L windows; ϵ_d set from historical quantiles. **T3 Integrity:** $F_t < \lambda_d$; optional HSI non-increasing under comparable load. **T4 No inversion (interpretive):** $S_t = \alpha L_t - \beta(A_t Q_t) \geq 0$ in official pipelines. **T5 Recursion cap:** $\sigma_t \leq \rho_{\max}$ and $\frac{d\sigma_t}{dt} \leq 0$ over audit cycles.

Controls, Counterfactuals, and Statistical Plan
Placebo corpora: archives unlikely to be exposed to recursion (e.g., print-only eras, offline repositories) to test D2/D4 specificity.
Matched periods / DiD: pre/post tool-introduction or policy-hardening; venue or course-level fixed effects.
Lineage audits: map cross-ingestion (indexers, cross-posting, scraping); re-estimate D2 after excluding implicated sources.
Exogenous shocks: incorporate policy/incentive covariates (fees, docket reforms, syllabus redesigns).

Trend estimation: slopes and CIs for $dV_t/dt, dC_t/dt, R(t), dH/dt$ with block bootstrap or HAC errors. **Alignment tests:** permutation for Jaccard trends; isotonic or Mann–Kendall for monotonicity; multiple-testing control across m, Δ, L. **Linking F_t:** GLMs with venue/domain fixed effects; clustered SEs; partial R^2 for the incremental contribution of F_t. **Threshold timing:** survival/time-to-first-crossing; compare hazards across periods with and without interventions.

Robustness and Sensitivity Estimator swaps (compression $\leftrightarrow$ generalized entropies; n-gram $\leftrightarrow$ frame templates; embedding families for δ_t); window sensitivity (Δ, L); motif cardinality $m \in \{50, 100, 250\}$; independence stress (remove lineage-linked sources; deduplicate); threshold recalibration ($\lambda_d, \epsilon_d, \theta_d, \rho_{\max}$) via rolling baselines and report of any classification flips.

7 Conclusion

I conclude by adopting: A first-order, early-signal lens on destabilization in knowledge infrastructures. The stance is operational and diagnostic: when drafting accelerates under recursive reuse while checking does not, incentives tilt toward fluency over reference; the relevant object is early warnings rather than anecdotes, consistent with critical-transition signal logic and known limits (Scheffer et al. 2009b; Holling 1973).

I state the boundaries plainly: attribution shocks can impersonate pacing failure; entropy and diversity metrics can collapse to style; validator practices co-evolve with tools, shifting thresholds; "independent" corpora leak lineage; windows and seasons bend trends; observed metrics invite optimization. These are design commitments of a first-order theory; simplicity buys legibility and falsifiability while postponing higher-order interactions. I personally expect this framework to mutate accordingly in a non-reductionist pathways after doing empirical experiments, but that doesn't lessen the value of this paper, which lies in timely diagnosis and institutional memory setting so that the issue is treated in length and not naively.

I register hypotheses to intentionally break and evolve: validator co-adoption can yield nonmonotonic thresholds; apparent fatigue may fall while discrimination erodes. Incentives that price fluency over checking should lengthen negative-pace stretches; reversal after attractor formation will likely show hysteresis; cross-domain spillovers should appear as synchronous motif alignment, blunted by independence controls (Scheffer et al. 2009b). Epistemic Destabilization likely has cybernetic complexity that will be revealed later on with more scrutiny. Higher-order feedbacks (adaptive validators, strategic prompting) are likely to emerge.

Under page limits, my contribution is instrumentation and norm setting. Governance can be expressed as parameter nudges which the reader can test in place: context-specific latency floors; recursion and lineage hygiene; validator independence; pluralism guarantees grounded in conditional diversity value. As such I call for a rigorous, empirically robust findings roster, before we proceed to prescriptive practices attributed to this framework. **This framework is set up, intentionally to be the primordial flagholder for a non-reductionist inquiry of epistemic systems perturbed by LLMs and their advent.**

My Ethical stance in context to this framework is rigorous diagnosis as first course of action, and then carefully designing interventionist measures that are effective and systemic and don't stifle or normatively impose upon entities and organizations in ways they shouldn't; they must not be capable of misuse and over-creep in context. This requires careful ethical design that treats the matter at length, rigorously and systemically, fully incorporating the cybernetic complexity that can be yielded by the interventions. This drives my teleological basis in context to this framework.

References

Agarwal, D.; Naaman, M.; and Vashistha, A. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.

Alemohammad, S.; Casco-Rodriguez, J.; Luzi, L.; Humayun, A. I.; Babaei, H.; LeJeune, D.; Siahkoohi, A.; and Baraniuk, R. G. 2024. Self-Consuming Generative Models Go MAD. In *International Conference on Learning Representations (ICLR)*. Accessed 2025-08-08.

Altmann, E. G.; Dias, L.; and Gerlach, M. 2017. Generalized entropies and the similarity of texts. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(1): 014002.

Anderson, B. R.; Shah, J. H.; and Kreminski, M. 2024.

Homogenization Effects of Large Language Models on Human Creative Ideation. *arXiv preprint*.

Apple, M. W. 1993. *Official Knowledge: Democratic Education in a Conservative Age*. Routledge.

arXiv. 2025. Monthly Submissions — arXiv Statistics. Web page. Accessed 2025-08-08.

Ayush Agrawal, A. T. K., Lester Mackey. 2023. Do Language Models Know When They're Hallucinating References? *arXiv preprint*.

Bentz, C.; and Alikaniotis, D. 2016. The word entropy of natural languages. *arXiv preprint*.

Biderman, S.; Li, H.; Hallahan, T.; Anthony, Q.; Black, S.; Gao, L.; Foster, C.; Thakur, B.; et al. 2023. Emergent and Predictable Memorization in Large Language Models. *arXiv preprint arXiv:2304.11158*.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159.

Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, Ú.; Oprea, A.; and Raffel, C. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Chubb, J.; Cowling, P.; and Reed, D. 2022. Speeding Up to Keep Up: Exploring the Use of AI in the Research Process. *AI & Society*, 37(4): 1439–1457.

Collaboration, O. S. 2015. Estimating the Reproducibility of Psychological Science. *Science*, 349(6251): aac4716.

Collins, H.; and Evans, R. 2007. *Rethinking Expertise*. Chicago, IL: University of Chicago Press. ISBN 9780226113609.

Dechêne, A.; Stahl, C.; Hansen, J.; and Wänke, M. 2010. The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. *Personality and Social Psychology Review*, 14(2): 238–257.

Edwards, P. N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Evans, J. A.; and Foster, J. G. 2011. Metaknowledge. *Science*, 331(6018): 721–725.

Franzoni Velázquez, A. L.; Huerta, E.; and Jensen, S. 2024. Retracting ChatGPT: Completeness and relevance of academic references. *Discover Education*, 3. Open access PDF: https://link.springer.com/content/pdf/10.1007/s44217-024-00333-1.pdf.

Freire, P. 1970. *Pedagogy of the Oppressed*. Continuum (English translation).

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, UK: Oxford University Press. ISBN 9780198237900.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Glynn, A. 2025. Guarding against artificial intelligence–hallucinated citations: The case for full-text reference deposit. *European Science Editing*, 51: e153973. Also available as arXiv:2503.19848.

Goldman, A. I. 1999. *Knowledge in a Social World*. Oxford, UK: Oxford University Press. ISBN 9780198238201.

Gordon Pennycook, A. B., David G. Rand; and Collins, E. T. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings. *Management Science*, 66(11): 4944–4957.

Helbing, D. 2013. Globally networked risks and how to respond. *Nature*, 497: 51–59.

Holling, C. S. 1973. Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics*, 4: 1–23.

ICLR. 2024. ICLR 2024 Fact Sheet. PDF. Accessed 2025-08-08.

Ioannidis, J. P. 2005. Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8): e124.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; Chan, H. S.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12).

Keyes, O.; Hitzig, Z.; and Blell, M. 2021. Truth from the Machine: Artificial Intelligence and the Materialization of Identity. *Interdisciplinary Science Reviews*, 46(1–2): 158–175.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1st edition.

Latour, B.; and Woolgar, S. 1979. *Laboratory Life: The Construction of Scientific Facts*. Beverly Hills, CA: SAGE Publications, 1st edition.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Association for Computational Linguistics.

Longino, H. E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press. ISBN 9780691020518.

Mata v. Avianca, Inc. 2023. Opinion and Order on Sanctions, No. 22-cv-1461 (PKC), U.S. District Court, S.D.N.Y. Court opinion (PDF). Accessed 2025-08-08.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. Association for Computational Linguistics.

Meadows, D. H. 2008. *Thinking in Systems: A Primer*. Chelsea Green Publishing. ISBN 978-1-60358-055-7.

Merton, R. K. 1973. The Normative Structure of Science. In Merton, R. K., ed., *The Sociology of Science: Theoretical and Empirical Investigations*, 267–278. University of Chicago Press. Originally published 1942 as "Science and Technology in a Democratic Order".

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Mitchell, Tom and Silver, Daniel. 2023. The Roles of Symbols in Neural-Based AI: They Are Not What You Think! https://arxiv.org/abs/2304.13626.

Oreskes, N.; Shrader-Frechette, K.; and Belitz, K. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, 263(5147): 641–646.

Publons. 2018. Global State of Peer Review 2018. Technical report, Publons. Accessed 2025-08-08.

Reber, R.; and Schwarz, N. 1999. Effects of Perceptual Fluency on Judgments of Truth. *Consciousness and Cognition*, 8(3): 338–342.

Retraction Watch. 2024. Journal taking 'corrective actions' after learning author used ChatGPT to update references. Blog post. Accessed 2025-08-08.

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Scheffer, M.; Bascompte, J.; Brock, W. A.; Brovkin, V.; Carpenter, S. R.; Dakos, V.; Held, H.; van Nes, E. H.; Rietkerk, M.; and Sugihara, G. 2009a. Early-warning signals for critical transitions. *Nature*, 461(7260): 53–59.

Scheffer, M.; Bascompte, J.; Brock, W. A.; et al. 2009b. Early-Warning Signals for Critical Transitions. *Nature*, 461: 53–59.

Shapin, S.; and Schaffer, S. 1985. *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton, NJ: Princeton University Press. ISBN 9780691024325.

Shumailov, I.; Shtern, D.; Zhang, Y.; et al. 2024. AI models collapse when trained on recursively generated data. *Nature*.

Star, S. L.; and Ruhleder, K. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, 7(1): 111–134.

Stephan Lewandowsky, C. M. S., Ullrich K. H. Ecker. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3): 106–131.

Teahan, W. J. 2012. A compression-based method for ranking n-gram differences between texts. Technical report, Bangor University, Computer Science and Electronic Engineering.

Unkelbach, C.; and Rom, S. C. 2017. A referential theory of the repetition-induced truth effect. *Cognition*, 160: 110–126.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The Spread of True and False News Online. *Science*, 359(6380): 1146–1151.