From Literature Synthesis to Experimental Hypothesis: A Coherence-Guided Architecture for Scientific Discovery

Joel Thorarinson* Allison Hensgen†

June 2026

Abstract

Scientific literature grows at a rate that exceeds any individual researcher's capacity for synthesis: PubMed alone indexes over two million new articles per year. Existing AI tools for literature review — semantic search engines, summarization systems, citation graph analyzers — retrieve and compress information but do not evaluate the *structural coherence* of the evidence they surface. We propose an architecture for a coherence-guided discovery system that moves beyond retrieval to evaluate five properties of evidence bodies: consistency, drift, persistence, recovery, and contradiction resolution. The architecture constructs an evidence graph from retrieved literature, scores claim clusters by their coherence, and generates hypotheses from regions of low coherence — where evidence is fragmented, contradictory, or drifting. This extends Swanson's insight that knowledge can be public yet undiscovered [Swanson, 1986]: where Swanson identified connections between disjoint literatures manually, we propose automated coherence analysis to detect not just missing connections but structural fragmentation across an entire evidence body. We define hypothesis quality metrics (falsifiability, evidence connectivity, endpoint measurability, control specification, coherence gap addressed) and describe a four-stage pipeline from literature ingestion through hypothesis output, specifying which components are buildable with existing tools and which require further research. We illustrate the approach on the embedding dimensionality debate, where manual coherence analysis of the kind the system would automate revealed a gap that motivated experimental work [Thorarinson and Hensgen, 2026b]. No working system exists; this is a design paper. The contribution is the architecture itself and the argument that coherence measurement — making visible the structural properties that human review cannot track at scale — is the missing layer between literature retrieval and scientific discovery.

Keywords: scientific discovery; literature synthesis; evidence coherence; hypothesis generation; system architecture; knowledge graph; reproducibility; human-in-the-loop

1 Introduction

The fundamental bottleneck in scientific discovery is no longer data collection or computation; it is synthesis. A researcher studying the relationship between gut microbiome composition and neurological outcomes faces a corpus of over 15,000 papers published in the last five years alone. These papers contain findings that agree, contradict, qualify, extend, and ignore each other in patterns that no human can track. The result is a literature that is simultaneously rich and incoherent — full of information but lacking structural organization.

This problem has consequences. The reproducibility crisis [Ioannidis, 2005] is, in part, a coherence crisis: fields publish contradictory findings without noticing the contradictions, replicate studies without tracking which claims persist, and shift consensus without documenting the drift.

*Coherence Research Group. ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com
†Coherence Research Group. ORCID: 0009-0008-7247-0307

Systematic reviews and meta-analyses [Higgins et al., 2019] address this for narrow, well-defined questions, but they require months of manual curation and cover a fraction of the literature. The structural coherence of most evidence bodies remains unmeasured.

Recent AI systems have demonstrated computational approaches to discovery on structured data: AlphaFold [Jumper et al., 2021] on amino acid sequences, GNoME [Merchant et al., 2023] on crystal structures, FunSearch [Romera-Paredes et al., 2024] on combinatorial search spaces. The challenge of literature synthesis is different: the "data" is the unstructured, contradictory, evolving body of published scientific text.

Current AI tools for literature review address the retrieval problem: given a query, find relevant papers. Semantic Scholar [Ammar et al., 2018] provides citation-aware search over a literature graph of hundreds of millions of nodes. Elicit extracts structured claims from papers. Consensus aggregates findings across studies. ResearchRabbit discovers related work through citation graphs. These tools are valuable, but they share a limitation: they treat the literature as a *retrieval target* rather than an *evidence structure*. They find papers; they do not evaluate whether the body of evidence they find is structurally coherent.

The idea that disjoint literatures may contain undiscovered connections dates to Swanson [1986], who demonstrated that separately published findings about fish oil (literature A) and Raynaud's disease (literature C) shared an intermediate concept — blood viscosity (B) — that neither community had noticed. The resulting therapeutic hypothesis was later confirmed experimentally. Swanson's ABC model identified a fundamental opportunity: knowledge can be public yet undiscovered when independently created fragments are never brought together and interpreted.

Modern hypothesis generation systems have scaled Swanson's approach computationally. AGATHA [Sybrandt et al., 2020] uses graph neural networks to predict novel connections in a biomedical knowledge graph. The Robot Scientist "Adam" [King et al., 2009] autonomously generated and experimentally tested hypotheses in functional genomics. These systems ask: *what entities might be connected?* We ask a different question: *where is the evidence structure breaking down?*

The distinction matters. Swanson's ABC model finds missing links between literatures that happen not to cite each other. The architecture we propose finds structural fragmentation *within* a literature — contradictions that go unresolved, consensus that drifts without stabilizing, findings that fail to replicate. These are different failure modes of scientific knowledge, and they require different detection methods. Missing links require bridging disjoint graphs; structural fragmentation requires measuring the internal coherence of a connected graph.

We propose that the coherence of a body of evidence — the degree to which findings are consistent, persistent, and converging — is itself a measurable and scientifically informative quantity. Why should low coherence signal discovery opportunity? Because low coherence means the field's current framework cannot accommodate all of its own findings. Either there is a moderating variable that has not been identified, a confound that has not been controlled, or a distinction that has not been drawn. Resolving any of these advances understanding. This is not a new idea — it is what scientists do when they notice contradictions in their own reading. The proposal is to do it computationally, at a scale that exceeds any individual's reading capacity.

The Coherence Engine framework [Thorarinson and Hensgen, 2026a] provides mathematical operators for measuring structural coherence in dynamical systems. We adapt these operators from physical time series to evidence time series — sequences of published findings indexed by publication date. The result is an architecture for a system that would measure the coherence of scientific evidence and generate hypotheses from regions where coherence is low. We describe the architecture in full, specify which components are buildable with current tools and which require further research, and illustrate the approach on a case where manual coherence analysis — the kind the system would automate — led to a confirmed experimental result [Thorarinson and Hensgen, 2026b].

1.1 Contributions

1. We define **evidence coherence** through five measurable properties adapted from the Coherence Engine framework, applied to bodies of scientific literature rather than physical systems.
2. We describe a complete **four-stage architecture** from literature ingestion and embedding through evidence graph construction, coherence scoring, and hypothesis generation, with design rationale for each stage.
3. We define **hypothesis quality metrics** that evaluate whether generated hypotheses are scientifically actionable.
4. We illustrate the approach on the **embedding dimensionality debate**, showing how the coherence analysis the system would perform mirrors the manual reasoning that motivated experimental work [Thorarinson and Hensgen, 2026b].
5. We provide a **feasibility assessment** distinguishing components buildable with existing tools from those requiring further research.

2 Evidence Coherence

We define five properties that characterize the structural coherence of a body of scientific evidence. Each property is adapted from the physical coherence operators of the Coherence Engine [Thorarinson and Hensgen, 2026a] to operate on evidence rather than sensor data. The "time series" is the sequence of published findings on a topic, indexed by publication date.

2.1 Consistency

Consistency measures whether findings on a topic agree with each other at a given time.

Definition 1 (Evidence Consistency)**.** *For a set of N claims $\{c_1, \ldots, c_N\}$ published within a time window $[t - w, t]$ on topic $\mathcal{T}$, the consistency is:*

$$C(\mathcal{T}, t) = \frac{1}{\binom{N}{2}} \sum_{i<j} agree(c_i, c_j) \tag{1}$$

where $agree(c_i, c_j) \in [-1, 1]$ is a semantic agreement score between claims, computed via embedding similarity of claim representations followed by entailment classification using natural language inference [Bowman et al., 2015, Williams et al., 2018]. $C = 1$ indicates perfect agreement; $C = -1$ indicates universal contradiction; $C \approx 0$ indicates a fragmented field where claims neither agree nor disagree (they address different aspects without integration).

This is the evidence analog of Phase Alignment $A(t)$ from the Coherence Engine: just as $A(t)$ measures whether system components are synchronized, $C(\mathcal{T}, t)$ measures whether research findings are aligned. A field with high A is producing convergent evidence; a field with low A is producing scattered or contradictory results.

2.2 Drift

Drift measures whether the consensus on a topic is shifting over time.

Definition 2 (Evidence Drift)**.** *For a topic $\mathcal{T}$ with claim distributions p_{t_1} and p_{t_2} at two time points, the evidence drift is:*

$$D_{ev}(\mathcal{T}, t_1, t_2) = d_{FR}(p_{t_1}, p_{t_2}) \tag{2}$$

where d_{FR} is the Fisher-Rao distance [Rao, 1945, Amari, 2016] between the distributions of claim embeddings at times t_1 and t_2.

Evidence drift is not inherently pathological — science *should* shift as new data arrives, as Kuhn [1962] argued in his account of paradigm shifts. But the *pattern* of drift is informative. Monotonic drift toward a stable attractor indicates convergence. Oscillatory drift (the field shifts back and forth between positions) indicates unresolved disagreement. Sudden drift (the consensus shifts rapidly after a single publication) indicates a field vulnerable to individual high-impact results. These patterns parallel the drift and noise amplification operators of the Coherence Engine.

2.3 Persistence

Persistence measures whether findings hold up across replications and time.

Definition 3 (Evidence Persistence)**.** *For a claim c first published at time t_0, the persistence is:*

$$P_{ev}(c,t) = \frac{|\{r \in \mathit{Replications}(c,[t_0,t]) : \mathit{agree}(r,c) > \theta\}|}{|\mathit{Replications}(c,[t_0,t])|} \tag{3}$$

where $\mathit{Replications}(c,[t_0,t])$ is the set of studies that attempted to replicate or extend claim c in the interval $[t_0,t]$, and θ is an agreement threshold.

This is the evidence analog of Pattern Retention $P(t)$: a claim with high persistence is like a dynamical pattern that recurs reliably. A claim that fails to replicate is like a pattern that dissolves — the system's "memory" of that state is fading. The reproducibility crisis [Ioannidis, 2005, Wang et al., 2023] can be understood as a field-wide decline in evidence persistence: claims are being published that do not persist under replication.

2.4 Recovery

Recovery measures how a field responds when a key finding is challenged.

Definition 4 (Evidence Recovery)**.** *When a claim c is challenged by contradictory evidence c' at time $t_{challenge}$, the recovery is:*

$$R_{ev}(c,t) = \frac{C(\mathcal{T}_c,t)}{C(\mathcal{T}_c,t_{challenge}-\epsilon)} \tag{4}$$

where $\mathcal{T}_c$ is the topic cluster containing c, and C is the consistency defined above. $R_{ev} \to 1$ indicates that the field recovered its coherence (either by confirming or replacing the challenged claim). $R_{ev} \ll 1$ indicates that the challenge fragmented the field without resolution.

This is the evidence analog of Recovery Capacity $R(t)$: a healthy field, like a healthy system, can absorb perturbations (challenges to key findings) and return to a coherent state. A fragile field remains fragmented after challenge. The *timescale* of recovery is also informative: a field that takes decades to resolve a contradiction (e.g., the nature of dark energy) has slower recovery dynamics than one that resolves contradictions within a few years.

2.5 Contradiction Resolution

Contradiction resolution measures whether contradictions in the evidence are acknowledged and addressed, or ignored.

Definition 5 (Contradiction Resolution)**.** *For a pair of contradictory claims (c,c') identified at time t_0, the resolution status at time t is:*

$$CR(c,c',t) = \begin{cases} \textit{resolved} & \textit{if } \exists\, c'' \in [t_0,t] \textit{ that explains or reconciles } c,c' \\ \textit{acknowledged} & \textit{if } \exists \textit{ paper in } [t_0,t] \textit{ citing both } c,c' \\ \textit{ignored} & \textit{otherwise} \end{cases} \tag{5}$$

The contradiction resolution rate for a topic is the fraction of identified contradictions in the "resolved" state.

Ignored contradictions are the most informative signal for hypothesis generation. A contradiction that is ignored indicates either: (a) the field has not noticed the discrepancy, (b) the field considers the discrepancy unimportant, or (c) the field lacks the tools or data to resolve it. Cases (a) and (c) represent discovery opportunities.

3 Architecture

The proposed system operates in four stages: literature ingestion, evidence graph construction, coherence scoring, and hypothesis generation. We describe each stage with its design rationale — why this approach rather than the alternatives — and distinguish standard components from those that are novel to this architecture. No working implementation exists; the descriptions specify what the system would do and how.

3.1 Stage 1: Literature Ingestion

The system ingests literature through semantic search and citation graph traversal. For a given research question Q:

1. **Seed retrieval**: Query Semantic Scholar [Ammar et al., 2018], PubMed, and arXiv APIs with Q and retrieve the top K papers (typically $K = 200$). This is standard practice; all existing literature tools begin here.

2. **Citation expansion**: Follow citations (both forward and backward) up to depth 2, adding papers that are cited by ≥ 3 seed papers or that cite ≥ 3 seed papers. The threshold of 3 balances corpus completeness against noise from tangential citations. This citation graph traversal exploits the heterogeneous literature graph structure described by Ammar et al. [2018], which links papers, authors, and entities through citation, authorship, and mention relationships.

3. **Embedding**: Embed all paper abstracts and extracted claims using a domain-appropriate embedding model. As demonstrated by Thorarinson and Hensgen [2026b], embedding quality is domain-sensitive — a model trained on biomedical text produces better claim representations for biomedical evidence than a general-purpose model. The retrieval layer must preserve domain-specific semantic structure, which requires sufficient embedding dimensionality.

4. **Claim extraction**: Extract structured claims from each paper using an LLM, building on methods from scientific information extraction [Luan et al., 2018, Wright et al., 2021]. Each claim is a triple: (subject, predicate, object) with associated confidence, method type, and sample size. Extracted claims can be verified against the literature using scientific claim verification methods [Wadden et al., 2020].

Design rationale. Why not skip claim extraction and work with paper-level representations? Because coherence is a property of *claims*, not papers. A single paper can contain claims that support one finding and contradict another. Paper-level analysis would obscure exactly the fine-grained structure that coherence measurement requires. The cost is complexity: claim extraction using current LLMs is imperfect (see Section 8), and extraction errors propagate through the entire pipeline. This is the architecture's primary vulnerability.

3.2 Stage 2: Evidence Graph Construction

The extracted claims are organized into an evidence graph (Figure 1). This stage combines three standard techniques — natural language inference (NLI), claim extraction, and community detection — in a specific configuration designed for coherence analysis.

Step 1: Node and edge construction. The graph contains nodes of five types (claims, papers, authors, methods, datasets) connected by four edge types:

- **Support**: Claim A provides evidence for claim B. Detected by NLI entailment classification [Bowman et al., 2015, Williams et al., 2018] — a standard component.
- **Contradict**: Claim A contradicts claim B. Detected by NLI contradiction classification, extended with negation detection methods [Harabagiu et al., 2006] — standard, though accuracy on scientific text is lower than on general NLI benchmarks.
- **Extend**: Claim A extends claim B to a new domain. Detected by high embedding similarity plus low lexical overlap in the domain terms — a heuristic specific to this architecture.
- **Qualify**: Claim A provides boundary conditions for claim B. Detected by partial entailment with conditional modifiers — the least reliable edge type and a target for future work.

Edge weights are determined by the strength of the evidential relationship (effect size, sample size, method rigor). Every edge carries a timestamp (publication date of the source paper).

Step 2: Community detection. The graph is clustered into topic groups using the Louvain algorithm [Blondel et al., 2008] on the claim embedding space, a standard and well-understood method for detecting community structure in large networks [Fortunato, 2010]. Each cluster represents a coherent subtopic within the broader research question.

Design rationale. Why a graph rather than a flat list of claims with pairwise similarity scores? Because the evidential relationships between claims have *types* (support, contradict, extend, qualify), and these types matter for coherence assessment. A cluster with many contradiction edges has low consistency; a cluster with many support edges has high consistency. A flat similarity matrix discards this information. The graph also preserves transitivity: if claim A supports B and B contradicts C, the graph captures the indirect tension between A and C.

What is novel here is not any individual component — NLI, claim extraction, and community detection are all established methods — but the specific pipeline that feeds typed evidential edges into coherence scoring. The graph is built to answer coherence questions, not retrieval questions.

3.3 Stage 3: Coherence Scoring

For each claim cluster, the system computes the five evidence coherence metrics defined in Section 2. The results are aggregated into a coherence profile:

$$\Gamma(\mathcal{T}) = \big(C(\mathcal{T}), D_{\text{ev}}(\mathcal{T}), P_{\text{ev}}(\mathcal{T}), R_{\text{ev}}(\mathcal{T}), \text{CR}(\mathcal{T})\big) \tag{6}$$

Clusters are ranked by their overall coherence (Figure 2). Low-coherence clusters — those with low consistency, high drift, low persistence, slow recovery, or unresolved contradictions — are flagged as hypothesis generation targets.

Design rationale. Why five separate metrics rather than a single coherence score? Because different patterns of low coherence suggest different kinds of scientific opportunity (Table 1). A cluster with low consistency (findings disagree) calls for a reconciliation hypothesis — what moderating variable explains the disagreement? A cluster with low persistence (findings do not replicate) calls for a methodological hypothesis — what confound or artifact produces the original result? Collapsing these into a single number would lose the diagnostic information that guides hypothesis generation.

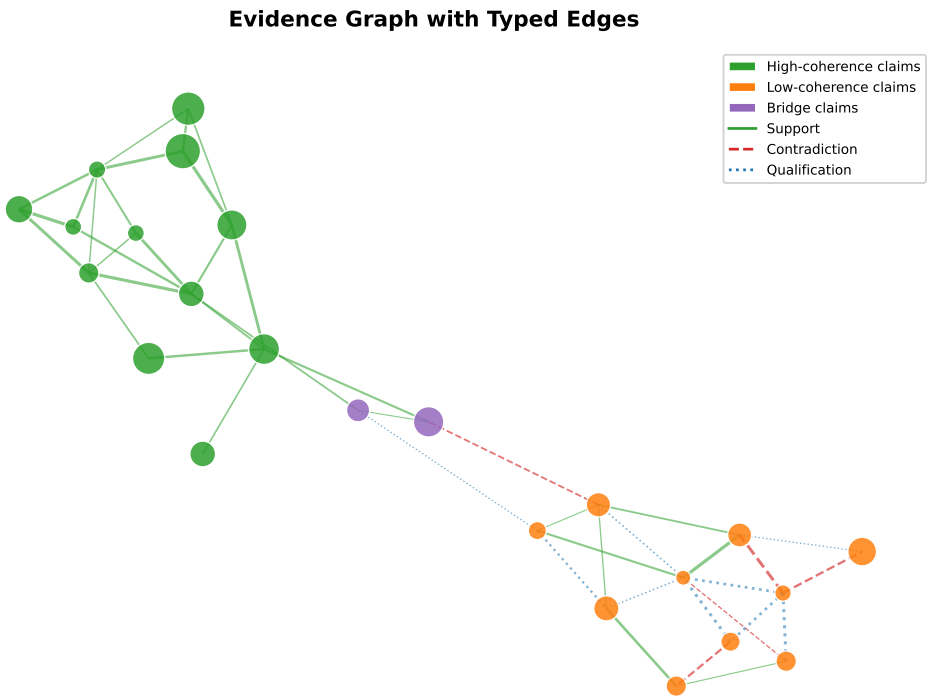

Figure 1: Conceptual evidence graph for a hypothetical research question. Nodes represent extracted claims; solid green edges indicate support, dashed red edges indicate contradiction, and dotted blue edges indicate qualification. Node size reflects citation count; edge width reflects evidential strength. The cluster at upper-left exhibits high consistency (dense green edges), while the cluster at lower-right exhibits low consistency (mixed green and red edges) — a hypothesis generation target.

All coherence scores carry epistemic uncertainty [Hu et al., 2023] arising from imperfect claim extraction, embedding noise, and temporal sampling. A low-coherence score might reflect genuine scientific disagreement, or it might reflect extraction errors that misclassify a supporting claim as a contradicting one. We discuss validation requirements in Section 7.

Coherence Signal	Interpretation	Hypothesis Type
Low consistency	Findings disagree	Reconciliation hypothesis
High drift	Consensus is shifting	Mechanism hypothesis
Low persistence	Findings don't replicate	Methodological hypothesis
Slow recovery	Field fragmented after challenge	Resolution hypothesis
Ignored contradictions	Discrepancy not addressed	Gap hypothesis

Table 1: Mapping from coherence signals to hypothesis types. Each pattern of low coherence suggests a different kind of scientific opportunity.

3.4 Stage 4: Hypothesis Generation

For each low-coherence cluster, the system would generate candidate hypotheses using the following protocol (see Figure 3 for the complete workflow):

1. **Identify the coherence gap**: What specific aspect of the evidence is incoherent? (Contradictory findings? Drifting consensus? Failed replications?)

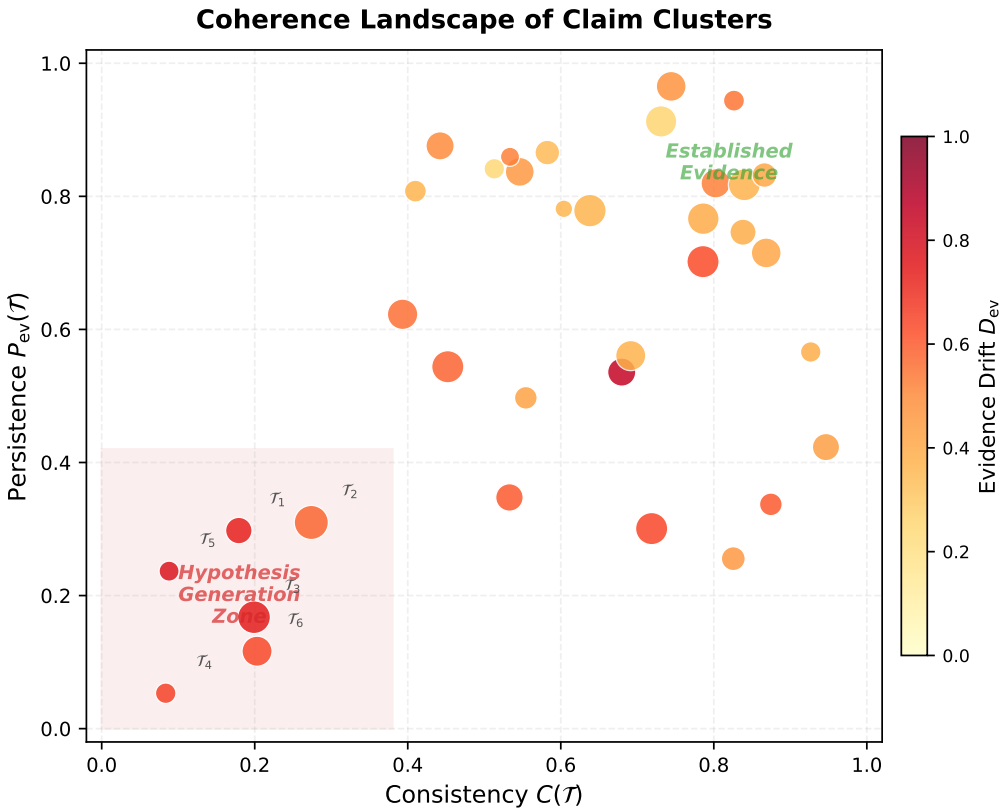

Figure 2: Illustrative coherence landscape for claim clusters within a research question. Each point represents a cluster; the x-axis shows consistency $C(\mathcal{T})$ and the y-axis shows persistence $P_{\text{ev}}(\mathcal{T})$. Point color indicates drift magnitude (darker = higher drift). The shaded region at lower-left marks the hypothesis generation zone: clusters with simultaneously low consistency and low persistence are the strongest candidates for discovery. Clusters in the upper-right represent well-established, convergent evidence.

2. **Analyze boundary conditions**: Under what conditions do the findings agree vs. disagree? (Different methods? Different populations? Different operationalizations of the same construct?)

3. **Propose a resolution**: What would need to be true for the incoherent findings to be reconciled? (A moderating variable? A threshold effect? A confound in one set of studies?)

4. **Formulate as a testable hypothesis**: State the proposed resolution as a falsifiable claim with specified methods, endpoints, and controls.

Design rationale. This stage is where the architecture is most speculative. Steps 1–2 are constrained by the evidence graph: the gap and boundary conditions are readable from the graph structure. Steps 3–4 require abductive reasoning — generating explanations for observed patterns — which is precisely the capability that current LLMs handle unevenly. A well-prompted LLM can produce plausible resolutions for simple contradictions; whether it can do so reliably for subtle scientific disagreements across dozens of claims is an open question. The hypothesis quality metrics (Section 4) are designed to filter weak outputs, but the generation step itself remains the least validated component of the architecture.

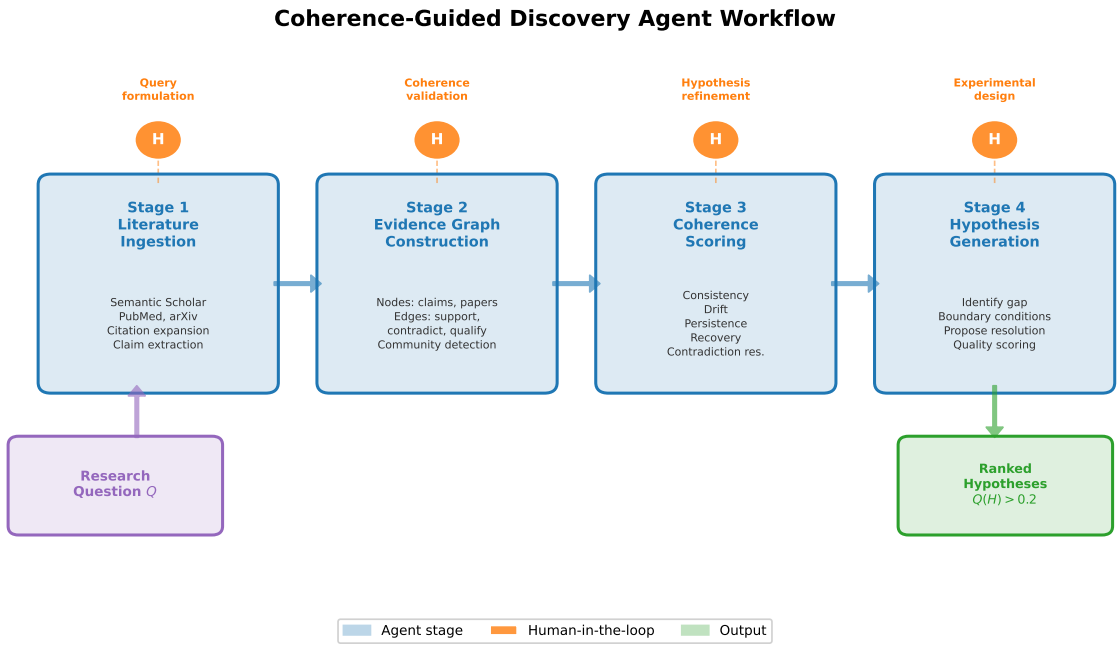

Figure 3: Complete workflow of the coherence-guided discovery system. Literature is ingested through API queries and citation graph traversal (Stage 1), structured into an evidence graph with typed edges (Stage 2), scored by the five coherence metrics (Stage 3), and low-coherence clusters are used to generate hypotheses evaluated by the quality metrics (Stage 4). The human researcher participates at each stage boundary: formulating the initial query, validating coherence assessments, refining generated hypotheses, and designing experiments.

4 Hypothesis Quality Metrics

Not all hypotheses are equally valuable. We define five quality metrics that distinguish actionable scientific hypotheses from vague speculation:

Definition 6 (Hypothesis Quality)**.** *A hypothesis H is evaluated on five dimensions:*

1. ***Falsifiability*** *(ϕ): Can a feasible experiment produce evidence against H? Scored as the specificity of the predicted outcome: $\phi = 1$ if H predicts a specific quantitative result; $\phi = 0$ if H is unfalsifiable.*

2. ***Evidence connectivity*** *(κ): Is H connected to existing evidence? Scored as the number of existing claims that H would explain, reconcile, or extend, normalized by the cluster size: $\kappa = |\{c : H \text{ addresses } c\}|/|\mathcal{T}|$.*

3. ***Endpoint measurability*** *(μ): Can the outcome of testing H be measured with existing instruments and methods? Scored on a 0–1 scale based on the availability and precision of measurement tools.*

4. ***Control specification*** *(γ): Does H specify what should be controlled? A hypothesis that identifies specific confounds to control is more actionable than one that does not.*

5. ***Coherence gap addressed*** *(δ): Does testing H reduce the coherence gap that motivated it? Scored as the fraction of the cluster's incoherent claims that would be resolved if H were confirmed: $\delta = |\text{resolved claims}|/|\text{incoherent claims}|$.*

The composite hypothesis quality score is:

$$Q(H) = \phi \cdot \kappa \cdot \mu \cdot \gamma \cdot \delta \tag{7}$$

The multiplicative form ensures that a hypothesis scoring zero on any dimension (unfalsifiable, disconnected, unmeasurable, uncontrolled, or not addressing the gap) receives an overall score of zero regardless of its other qualities.

5 Illustrative Example: The Embedding Dimensionality Debate

We illustrate how the proposed system would operate on a concrete case: the question of how many embedding dimensions are needed to preserve semantic information in text embeddings. No system was used to produce this analysis. The authors performed manual coherence analysis of the literature — the kind of analysis the system would automate — and the example is presented as "here is how the system would work" rather than "here is what the system found."

5.1 Stage 1: Literature Ingestion

A search for "embedding dimensionality reduction semantic preservation" would retrieve papers spanning PCA-based compression, Matryoshka representation learning [Kusupati et al., 2022], autoencoder-based reduction, and task-specific evaluations. The corpus would include benchmark papers reporting that 16–64 dimensions preserve 90–95% of variance, as well as domain-specific evaluations showing performance degradation under compression.

5.2 Stage 2–3: Evidence Graph and Coherence Analysis

From the retrieved literature, the system would construct an evidence graph and compute coherence scores for each claim cluster. We present the following qualitative characterization of what the system's output would look like for one cluster. These assessments reflect the authors' manual reading of the literature; they are not algorithmic outputs.

Consistency: The system would score this cluster as low-consistency. Papers reporting high variance retention under PCA contradict papers reporting task performance degradation under the same compression. The NLI component would classify a substantial fraction of claim pairs as contradictory.

Drift: The system would detect moderate drift. The literature has been shifting from "aggressive dimensionality reduction is safe" (2020–2023) toward "domain-specific evaluation is necessary" (2024–2026). The direction of drift is monotonic, suggesting convergence rather than oscillation.

Persistence: The system would score persistence as moderate. The original claims about variance-based safety of compression have partially persisted but with increasing qualifications. Newer work tends to cite the original claims while adding domain-specific caveats.

Recovery: Not applicable for this cluster (the field has not experienced a major challenge event that would test recovery dynamics).

Contradiction resolution: The system would classify this as ignored. The contradiction between high variance retention and low task performance at the same dimensionality has not been explicitly addressed in the literature — no published work cites both sides of the disagreement or proposes a reconciliation. Under the architecture's taxonomy (Table 1), an ignored contradiction signals a gap hypothesis.

5.3 Stage 4: Hypothesis Generation

Given the low-coherence cluster — low consistency, ignored contradiction, moderate drift — the system would identify the core incoherence: variance retention and task performance are

decoupled under dimensionality reduction, and no published work has explained why. Following the hypothesis generation protocol, a candidate hypothesis would take a form such as:

> *PCA variance retention overestimates the semantic information preserved under dimensionality reduction because PCA preferentially retains high-variance, low-information directions (frequency and document-length effects) while discarding low-variance, high-information directions (domain-specific semantic distinctions). A metric that measures semantic structure directly — such as neighborhood preservation in task-relevant queries — will show degradation at dimensionalities where variance retention remains above 95%.*

This is the hypothesis that the authors arrived at through manual reading of the same literature, and it motivated the experimental work in Thorarinson and Hensgen [2026b], which confirmed it: Semantic Coherence Loss (SCL) detects degradation that variance retention misses, particularly in specialized domains. The point is not that the system would have generated this exact hypothesis — that is unknowable without an implementation. The point is that the coherence structure of the literature (low consistency, ignored contradiction between variance metrics and task metrics) would have flagged this region as a hypothesis generation target. The gap was there; the system would have measured it.

5.4 Hypothesis Quality Assessment

To illustrate the quality scoring framework, we assess the hypothesis qualitatively against the five criteria. These are the authors' judgments, not algorithmic outputs:

- ϕ (Falsifiability): High. The hypothesis predicts a specific dissociation between variance retention and a structure-sensitive metric at specific dimensionalities — a result that could be directly falsified by experiment.
- κ (Evidence connectivity): Moderate-to-high. The hypothesis addresses the core contradictory claims in the cluster, though some contradictions may stem from dataset differences rather than metric choice.
- μ (Endpoint measurability): High. All measurements (PCA variance, neighborhood preservation, retrieval accuracy) are computable with existing tools and publicly available embedding models.
- γ (Control specification): High. The hypothesis specifies controls: same dataset, same embedding model, vary only dimensionality and evaluation metric.
- δ (Coherence gap addressed): Moderate-to-high. Confirmation would resolve the core variance-vs-performance contradiction, though secondary disagreements about optimal dimensionality thresholds would remain.

The fact that this hypothesis motivated experimental work [Thorarinson and Hensgen, 2026b] that confirmed the predicted dissociation provides circumstantial support for the claim that coherence gaps identify productive research directions. A single example does not validate the architecture — that requires systematic evaluation on a corpus of coherence-identified gaps and their outcomes, which we leave to future work.

6 Relationship to Existing Approaches

6.1 Swanson's ABC Model and Its Extensions

The most direct intellectual ancestor of this architecture is Swanson's ABC model of literature-based discovery [Swanson, 1986]. Swanson showed that literature A (fish oil reduces blood viscosity) and literature C (Raynaud's disease involves high blood viscosity) shared an intermediate concept B that neither community had recognized. His method was manual: read across disjoint literatures, identify the bridge.

AGATHA [Sybrandt et al., 2020] automated this pattern at scale, using graph neural networks to predict which entity pairs in a biomedical knowledge graph are likely to have undiscovered connections. This is a direct computational scaling of Swanson: find the missing edge between nodes that "should" be connected.

The architecture we propose addresses a different problem. Swanson's model finds missing connections between literatures that do not cite each other. Our architecture finds structural incoherence *within* a literature that does cite itself — contradictions that go unresolved, consensus that oscillates, findings that do not replicate. These are not missing links; they are structural failures in a connected graph. The detection method is correspondingly different: rather than predicting missing edges, we measure the coherence of existing edge patterns.

The two approaches are complementary. A system could run Swanson-style cross-literature bridging *and* within-literature coherence analysis, catching both disconnected knowledge and fragmented knowledge.

6.2 Comparison with Literature Tools

Capability	**Proposed System**	**Semantic Scholar**	**Elicit**	**Consensus**
Paper retrieval	Yes	Yes	Yes	Yes
Claim extraction	Yes	No	Yes	Yes
Citation graph	Yes	Yes	Partial	No
Evidence consistency	Yes	No	No	Partial
Temporal drift analysis	Yes	No	No	No
Persistence tracking	Yes	No	No	No
Contradiction detection	Yes	No	No	No
Hypothesis generation	Yes	No	No	No
Coherence scoring	Yes	No	No	No

Table 2: Capability comparison between the proposed system and existing literature tools. Existing tools retrieve and summarize; the proposed system would evaluate structural coherence and generate hypotheses from regions of low coherence.

The key distinction is between *content tools* (which surface relevant papers and extract their contents) and *structure tools* (which evaluate the structural relationships between findings). Existing tools are content tools; the proposed architecture describes a structure tool. The two are complementary: the system depends on content tools for ingestion and retrieval, and adds a layer of structural analysis on top.

This mirrors the relationship between predictive and discoverative intelligence described in Thorarinson and Hensgen [2026a]: content tools predict what papers a researcher should read; a coherence system would reveal where the field's evidence structure is breaking down.

7 The Human-in-the-Loop

The architecture is designed around an instrument, not an oracle. The system does not replace scientific judgment; it extends the researcher's capacity for structural perception. The human serves five essential functions:

1. **Query formulation**: The researcher defines the research question. The system cannot determine what is worth investigating — only the researcher's domain knowledge and scientific taste can do this.
2. **Coherence validation**: The system's coherence scores may be artifacts of embedding noise, citation biases, or temporal sampling. The researcher must validate that identified low-coherence regions reflect genuine scientific disagreement, not methodological artifacts of the system itself. This is not a secondary concern; it is the primary quality control mechanism.
3. **Hypothesis refinement**: Generated hypotheses are starting points, not finished scientific claims. The researcher refines them based on mechanistic knowledge, feasibility constraints, and ethical considerations that the system cannot evaluate.
4. **Experimental design**: The system proposes what to test; the researcher designs how to test it. Experimental design requires knowledge of available equipment, institutional constraints, regulatory requirements, and practical considerations that no literature analysis system can access.
5. **Interpretation**: A confirmed hypothesis changes the field's understanding. The researcher interprets the results in the context of broader theory, identifies implications, and determines the next question.

The appropriate metaphor is the telescope, not the astronomer. The telescope extends human perception to objects the naked eye cannot resolve, but the astronomer decides where to point it, validates what it shows, and interprets what the observations mean. A coherence-guided system would extend the researcher's perception to structural properties of the evidence landscape that manual review cannot track at scale, but the researcher remains the scientist.

8 Feasibility

The architecture combines components at different levels of maturity. We distinguish what is buildable now with existing tools from what requires further research.

8.1 Buildable Now

Literature ingestion (Stage 1, steps 1–3) is straightforward. The Semantic Scholar API provides programmatic access to over 200 million papers with citation graphs. PubMed and arXiv APIs are similarly mature. Embedding abstracts with models like SPECTER [Ammar et al., 2018] or domain-specific alternatives is standard practice. A working ingestion pipeline could be built in weeks.

Community detection (Stage 2, step 2) is a solved problem. The Louvain algorithm [Blondel et al., 2008] runs efficiently on graphs with millions of nodes. Applying it to claim embedding space requires only a similarity threshold to convert embeddings to a graph — a parameter choice, not a research problem.

Basic consistency scoring (Stage 3, partial) is feasible using existing NLI models. SciFact [Wadden et al., 2020] demonstrated claim verification against scientific abstracts. Applying NLI to compute pairwise agreement scores within a claim cluster is a direct extension.

8.2 Requires Engineering Effort

Claim extraction (Stage 1, step 4) works but imperfectly. Current LLMs can extract subject-predicate-object triples from scientific text, but accuracy varies by domain and writing style. Luan et al. [2018] and Wright et al. [2021] demonstrated structured extraction on scientific text, but error rates on out-of-domain material remain significant. This is the component most likely to introduce noise into downstream coherence scores.

Edge typing (Stage 2, step 1) beyond support and contradict is difficult. NLI models reliably distinguish entailment from contradiction, but the "extend" and "qualify" edge types require finer-grained classification that current models handle inconsistently. A practical first implementation might use only support and contradict edges, adding the others as classification improves.

Temporal indexing and drift computation (Stages 2–3) requires tracking claim distributions over time windows. The Fisher-Rao distance between claim embedding distributions is well-defined mathematically but requires sufficient temporal density — enough claims per time window to estimate distributional parameters. For well-studied topics with hundreds of papers per year, this is feasible; for niche topics with a few papers per year, temporal resolution would be too coarse for meaningful drift estimates.

8.3 Requires Further Research

Persistence tracking requires identifying replication attempts, which is harder than it sounds. A study that "replicates" a finding may use different methods, different populations, and different operationalizations of the key constructs. Automated identification of replication relationships is an open problem in meta-science.

Recovery measurement requires identifying challenge events and tracking their aftermath — a temporal pattern recognition problem on the evidence graph that has not, to our knowledge, been attempted at scale.

Hypothesis generation (Stage 4) is the most speculative component. Current LLMs can generate plausible-sounding hypotheses, but reliably generating hypotheses that score high on all five quality metrics requires the kind of abductive scientific reasoning that remains an active research frontier. The architecture is designed so that a weak hypothesis generator still produces value: even without Stage 4, Stages 1–3 produce a coherence map that identifies where discovery opportunities exist, leaving hypothesis formulation to the human researcher.

8.4 A Minimal Viable System

The components buildable now define a minimal viable system: literature ingestion, claim extraction (accepting noise), NLI-based consistency scoring, and community detection. This system would produce consistency maps — visualizations of which claim clusters exhibit high internal disagreement — without drift analysis, persistence tracking, or automated hypothesis generation. Even this reduced system would go beyond what current literature tools provide, by measuring the structural coherence of retrieved evidence rather than treating retrieval as the endpoint.

9 Discussion

9.1 Implications for the Reproducibility Crisis

The reproducibility crisis [Ioannidis, 2005, Wang et al., 2023] can be understood through the coherence framework as a systematic decline in evidence persistence (P_{ev}). Fields with low persistence are fields where published claims do not survive replication. Systematic review and meta-analysis methods [Higgins et al., 2019] already aggregate evidence across studies, but they

operate on manually curated study sets with pre-registered protocols. Automated persistence tracking would provide a complementary measure: by computing P_{ev} across topics and time, it would become possible to identify which topics have the lowest replication rates and investigate why.

More specifically, coherence analysis could identify the *boundary conditions* of persistence failure: claims that replicate in some contexts but not others. These boundary conditions are where the next generation of theory needs to focus — not on whether a claim is "true" or "false" in general, but on what moderating variables determine when it holds and when it does not.

9.2 Implications for Meta-Science

The coherence metrics defined in Section 2 provide a potential set of tools for meta-science — the scientific study of science itself. Questions that are currently addressed through manual coding of papers ("how often do authors cite contradictory evidence?" or "how long do controversial findings take to resolve?") could be answered at scale through automated coherence analysis.

The contradiction resolution metric (CR) is particularly informative for meta-science. A field where contradictions are systematically ignored has different dynamics from a field where contradictions are actively investigated. The former may produce more papers (contradictions do not slow publication) but less knowledge (contradictions are not resolved). Coherence analysis would make this distinction measurable.

9.3 Implications for Grant Writing

A coherence map of a research area would provide a principled basis for grant applications. Rather than arguing that a topic is "important" or "underexplored" (qualitative claims), a researcher could demonstrate that a specific cluster of evidence has low coherence (a quantitative claim) and propose research that would increase it. The hypothesis quality metrics provide a framework for evaluating whether the proposed research is likely to resolve the identified coherence gap.

10 Limitations

This paper describes an architecture, not a system. Several limitations follow directly from this.

No empirical validation. The architecture has not been implemented, and the illustrative example (Section 5) demonstrates manual coherence analysis, not automated analysis. The claim that the system would identify productive research directions rests on the argument that coherence gaps signal discovery opportunities, which is motivated but not tested. Validation requires building the system and evaluating its outputs against a corpus of known scientific discoveries — a substantial undertaking.

Claim extraction as a bottleneck. The entire architecture depends on accurate claim extraction from scientific text. Current methods produce error rates that could systematically bias coherence scores. If the extraction step misclassifies supporting claims as contradicting ones (or vice versa), the consistency metric becomes unreliable. The architecture has no internal mechanism to detect this failure mode; it relies on human validation (Section 7).

Citation bias. The literature ingestion stage inherits all the biases of the underlying literature: publication bias toward positive results, citation bias toward high-profile journals, language bias toward English-language publications. A coherence map built on a biased corpus reflects the coherence of *what was published and cited*, not the coherence of the underlying scientific evidence. This is a fundamental limitation of any literature-based approach, including Swanson's.

Temporal resolution. The drift and persistence metrics require sufficient temporal density of publications. For niche topics with sparse publication histories, the metrics would be too noisy to be informative.

Scope of "coherence." The five metrics defined here do not exhaust the ways evidence can be structurally problematic. A field can have high consistency, high persistence, and resolved contradictions while still being wrong — if all studies share a systematic confound. Coherence measures internal structure, not external validity.

11 Conclusion

We have described an architecture for measuring the structural coherence of scientific evidence and generating hypotheses from regions where coherence is low. The architecture adapts operators from the Coherence Engine framework [Thorarinson and Hensgen, 2026a] — originally developed for physical dynamical systems — to evidence systems, treating the sequence of published findings on a topic as a time series whose structural properties can be measured.

The contribution is the architecture itself and the argument that coherence measurement is the missing layer between literature retrieval and scientific discovery. Current tools find papers; the proposed system would evaluate whether the evidence those papers contain is structurally sound. Low-coherence regions — where findings contradict, where consensus drifts, where replications fail — are where the field's current framework cannot accommodate its own data. These are the regions where discovery is most likely.

No working system exists. The feasibility assessment (Section 8) identifies a minimal viable system — literature ingestion, claim extraction, NLI-based consistency scoring, and community detection — that could be built with existing tools. The full architecture, including persistence tracking, recovery measurement, and automated hypothesis generation, requires further research.

The connection to the broader coherence framework is deliberate. Whether measuring the structural integrity of a turbofan engine, a biological recovery trajectory, an embedding space [Thorarinson and Hensgen, 2026b], or a body of scientific evidence, the question is the same: is the structure holding, or is it degrading? The operators that detect structural change are domain-invariant; the interpretation changes with the domain. The discovery architecture is the framework's most ambitious proposed application: using coherence measurement not to monitor a system but to advance human understanding of where scientific knowledge is fragmented and where it might be repaired.

References

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, 2016.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, 2018.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010.

Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 755–762, 2006.

Julian P T Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley, 2 edition, 2019.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*, 2023.

John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. The automation of science. *Science*, 324(5923):85–89, 2009.

Thomas S Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Aditya Kusupati, Gantavya Bhatt, Aniket Rber, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, 2018.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco J R Ruiz, Kai Olav Ellefsen, Pengming Wang, Amer Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Don R Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. AGATHA: Automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2757–2764, 2020.

Joel Thorarinson and Allison Hensgen. From prediction to discoverative intelligence: A coherence-based AI framework for detecting system drift before failure. *arXiv preprint*, 2026a.

Joel Thorarinson and Allison Hensgen. The dimensionality illusion: Why PCA variance does not equal semantic information in text embeddings. *arXiv preprint*, 2026b.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550, 2020.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122, 2018.

Dustin Wright, Jennifer Lautenschlager, and Isabelle Laumond. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4498–4513, 2021.