A Living Meta-Analysis Architecture for Active Inference

Assertion Extraction, Nanopublications, and Hypothesis Scoring

Daniel Friedman

Active Inference Institute

daniel@activeinference.institute

ORCID: 0000-0001-6232-9096

and Joel Dietz

Massachusetts Institute of Technology (MIT)

California Institute for Machine Consciousness (CIMC)

jdietz@mit.edu

ORCID: 0000-0002-9456-2691

DOI: 10.5281/zenodo.19461934

April 30, 2026

Contents

1 Abstract

No prior automated system tracks hypothesis-level evidence across the full Active Inference and Free Energy Principle (FEP) literature. Manual synthesis cannot keep pace with a field that has grown at a compound annual rate of 20.36% across 2005–2026, and the FEP's theoretical generality has invited falsifiability critiques that only hypothesis-specific evidence profiling can address. Building on pioneering systematic manual annotation paired with ontology-based analysis at the scale of hundreds of papers, we present a computational meta-analysis framework that automates and scales this approach. The pipeline retrieves literature from arXiv, Semantic Scholar, and OpenAlex, deduplicating $N = 819$ papers via a canonical identifier hierarchy (DOI > arXiv ID > Semantic Scholar ID > OpenAlex ID). It classifies papers into a three-tier taxonomy spanning eight categories: A (Core Theory), B (Tools & Translation), and C (Application Domains). An LLM-powered extraction system then evaluates each abstract against eight core hypotheses, producing structured nanopublications—each encoding directionality, a confidence score, and natural-language reasoning—that populate an RDF-compatible knowledge graph scored by a citation-weighted evidence function.

All extracted assertions are automatically generated and have not been manually validated; hypothesis scores should be considered preliminary.

The resulting evidence landscape reveals a field where application domains (Domain C, 64.0%) collectively dominate the corpus, with tools development (Domain B, 20.8%)—including pymdp, RxInfer.jl, and interpretable alternatives such as Free Energy Projective Simulation—and core theory (Domain A, 15.2%) rounding out the taxonomy. Non-negative matrix factorization identifies 5 latent topics that cross-cut the keyword taxonomy, and citation network analysis exposes a sparse yet structured graph (2,176 intra-corpus edges out of 29,323 total outgoing references—**only 7.4% reference resolution**, reflecting the corpus's specialised scope rather than the underlying citation density of any single paper) anchored by pronounced hub papers. Hypothesis scores cluster into three tiers: a broad **consensus tier** (score > 0.83) covering five hypotheses—H7 Morphogenesis, H2 AIF Optimality, H4 Predictive Coding, H6 Clinical Utility, and H5 Scalability; a **near-consensus boundary** (H8 Language AIF, score $\approx +0.83$); a **moderate debate tier** (H3 Markov Blanket Realism, $\approx +0.78$); and a **diffuse tier** (H1 FEP Universality, $\approx +0.48$) where a large neutral plurality reflects the principle's broad invocation without explicit empirical test—though absolute score magnitudes are inflated by publication bias and linguistic asymmetry in academic writing, making relative rankings and temporal trajectories more reliable than point estimates. By demonstrating that automated LLM-driven assertion extraction—operating without human-validated ground truth—can generate scalable, queryable representations of scientific evidence, this work provides a reusable architecture for *living literature reviews*—continuously updated knowledge graphs that track hypothesis-level consensus across rapidly evolving fields.

Keywords: Active Inference, Free Energy Principle, meta-analysis, knowledge graph, nanopublications, bibliometrics, hypothesis scoring, LLM extraction, computational neuroscience

2 Introduction: Evidence Gaps in a Rapidly Expanding Field

2.1 The Free Energy Principle and Active Inference Framework

The Free Energy Principle (FEP), introduced by Karl Friston, proposes that self-organizing systems maintain their structural and functional integrity by minimizing variational free energy—an upper bound on sensory surprise [Friston et al., 2006, Friston, 2010]. Under this principle, living systems are cast as approximate Bayesian inference engines that build generative models of their environment and act to reduce the discrepancy between predicted and observed states. Active Inference (AIF) extends this picture from passive perception to goal-directed behavior: agents select actions that bring about observations consistent with their preferred states, unifying perception, learning, and decision-making within a single variational framework [Parr et al., 2022, Friston et al., 2017]. Since its initial formulation for sensorimotor control, AIF has been applied to navigation, visual foraging, language comprehension, social cognition, and multi-agent coordination. Bayesian mechanics [Sakthivadivel, 2023] has further strengthened the mathematical foundations of the FEP by grounding Markov blanket dynamics in the physics of belief-based systems, placing the principle on a footing commensurate with established physical theories. Importantly, the variational free energy minimization at the core of the FEP shares deep mathematical connections with the broader family of Energy-Based Models (EBMs) [LeCun et al., 2006]—including Helmholtz machines [Dayan et al., 1995], Boltzmann machines [Hinton, 2002], and variational autoencoders [Kingma and Welling, 2014]—all of which parameterize learning and inference through scalar energy functions and variational bounds. This convergence motivates the inclusion of EBM-adjacent literature in our search scope.

2.2 Challenges Posed by Rapid Literature Growth

The active inference literature has grown at a compound annual rate of 20.36% across 2005–2026, with annual output accelerating sharply after 2013. While early research concentrated on theoretical neuroscience, the field has since diversified across biology (C5), robotics (C2), computational psychiatry (C4), algorithm scaling (B), and formal mathematics (A1). With $N = 819$ papers spanning 8 categories across 3 domains, no prior automated system tracks hypothesis-level evidence across the full corpus. This creates three interrelated challenges. First, the balance of evidence for core claims—such as FEP universality or the physical realism of Markov blankets—cannot be assessed without structured, hypothesis-specific extraction at corpus scale. Second, because the relationship between mathematical formalisms and empirical evidence is frequently implicit, systematic evidence synthesis demands substantial manual effort: Knight et al. [Knight et al., 2022] required human annotators to manually code hundreds of papers. Third, new entrants must navigate a literature weighted toward broad qualitative philosophy (A2), interspersed with specialized applied subfields whose findings are difficult to locate without domain-specific search strategies.

Traditional narrative reviews attempt to address these challenges but are static, subjective, and quickly outdated. Systematic reviews from evidence-based medicine offer rigorous aggregation but are structured for clinical trial data with homogeneous outcome measures, making them poorly suited for the heterogeneous ontological and computational claims in this literature. The expansion of predictive processing [Clark, 2013, Hohwy, 2013] and the emergence of Bayesian mechanics [Sakthivadivel, 2023] further broaden the scope of assertions that a comprehensive meta-analysis must reconcile. Critically, the falsifiability of the FEP itself remains contested [Colombo and Seriès, 2021]: because free energy minimization can be reframed to accommodate any behavior post hoc, distinguishing genuine predictive commitment from tautological redescription requires exactly the hypothesis-specific, evidence-quantified framework we propose here.

2.3 Related Work and Prior Meta-Analyses

Several prior efforts have surveyed aspects of the Active Inference landscape. Sajid et al. [Sajid et al., 2021] compare active inference with alternative decision-making frameworks; Da Costa et al. [Da Costa et al., 2020] synthesize the discrete-state-space formulation; Lanillos et al. [Lanillos et al., 2021] survey robotics applications; Smith et al. [Smith et al., 2022] provide a tutorial bridging theory and empirical data; and Millidge et al. [Millidge et al., 2021] examine information-theoretic foundations of exploration behavior. Ramstead et al. [Ramstead et al., 2018] extend the FEP to questions of biological self-organization, while Pezzulo et al. [Pezzulo et al., 2015] connect active inference to homeostatic regulation. Millidge [Millidge, 2024] provides a practitioner's retrospective confirming that AIF's strongest demonstrated results arise from novel discrete generative models, while scalability relative to deep reinforcement learning remains the field's central open challenge.

Parallel to these synthesis efforts, Sanjeev V. Namjoshi's 2026 textbook, *Fundamentals of Active Inference* [Namjoshi, 2026b], provides a comprehensive, computationally explicit foundation for the field designed for engineers. In conjunc-

tion with this text, Namjoshi developed the `aif-fep-db` repository [Namjoshi, 2026a]—an open-source, dynamically updated database of scraped and tagged publications covering active inference, the free energy principle, and predictive processing. While `aif-fep-db` curates and categorizes the literature to facilitate reproducible systematic reviews and interactive Dash-based exploration, it functions primarily as a modular bibliographic foundation rather than an automated hypothesis evaluation engine.

Closest to our work, Knight, Cordes, and Friedman [Knight et al., 2022] conducted a systematic literature analysis of publications using the terms "Free Energy Principle" or "Active Inference," with an emphasis on works by Karl J. Friston. Their analysis—maintained by the Active Inference Institute—combined manual annotation of structural, visual, and mathematical features with automated analyses using the Active Inference Ontology at the scale of thousands of citations and hundreds of annotated papers. That study identified six development directions—including broader scope, richer annotation, and transferable approaches—and represents an important precursor to automated meta-analysis of this field.

These prior works differ from the present study along four dimensions. First, **scale**: narrative reviews cover tens to low hundreds of papers; our pipeline processes $N = 819$. Second, **structure**: prior reviews produce prose summaries rather than machine-queryable knowledge graphs with typed relationships. Third, **temporal tracking**: no prior system computes how evidence for specific hypotheses evolves year over year. Fourth, **automation**: the systematic analysis of Knight et al. [Knight et al., 2022] pioneered quantitative literature analysis but relied on manual annotation, limiting update frequency. Our framework advances this line of work by (1) fully automating assertion extraction via LLM-based hypothesis scoring, (2) constructing a structured, RDF-compatible knowledge graph scored by citation-weighted evidence, and (3) tracking how evidence for core claims evolves over time through temporal trend analysis.

2.4 Synergizing Knowledge Graphs and LLMs

Broadening this synthesis, recent systematic literature initiatives underscore a powerful reciprocal synergy between Large Language Models (LLMs) and Knowledge Graphs: LLMs parse unstructured text to rapidly extract semantic claims, efficiently populating the structured, queryable architecture of the graph [Quevedo Tumailli et al., 2025, Li et al., 2024]. We adopt the *nanopublication* [Groth et al., 2010]—a minimal, machine-readable unit of scientific evidence comprising a core assertion bound to explicit provenance metadata—as the fundamental serialization format for this extracted knowledge.

2.5 This Study: Approach and Overview

This paper presents a computational meta-analysis of the Active Inference literature ($N = 819$). Rather than relying exclusively on bibliometric metadata or slow manual coding, we deploy a Large Language Model (LLM) to "read" each paper's abstract and assess its relationship to eight core hypotheses within the FEP paradigm. We serialize these assessments as nanopublications—each encoding an assertion ("Paper X supports Hypothesis Y") coupled with the LLM's natural-language reasoning and confidence score. The resulting knowledge graph aggregates these nanopublications and links them to paper metadata, citation networks, subfield classifications, and hypothesis definitions. A citation-weighted scoring formula quantifies the net evidence for or against each hypothesis, producing scores in $[-1, 1]$ that reflect both the direction and strength of published evidence. Importantly, this represents an open-source introductory analysis which will be augmented and extended, and stewarded in collaborative development by the Active Inference Institute (activeinference.org).

2.6 Research Questions

This meta-analysis addresses four primary research questions:

1. **RQ1 (Field Structure):** What is the disciplinary structure and growth trajectory of the Active Inference literature, and how are papers distributed across the three domains—Core Theory (A), Tools & Translation (B), and Application Domains (C)? We expect Domain A to dominate but anticipate measurable diversification into applied domains.
2. **RQ2 (Growth Dynamics):** What are the temporal growth dynamics of the field, and which subfields are experiencing the most rapid expansion? Prior reviews suggest accelerating growth post-2013; we quantify this trajectory and identify which application domains drive it.
3. **RQ3 (Hypothesis Evidence):** What is the current balance of evidence for and against the eight standard hypotheses, and how has this balance evolved over time? We expect well-established hypotheses (H4 Predictive

Coding) to show consensus while philosophically contested claims (H3 Markov Blanket Realism) show mixed evidence. (See hypothesis dashboard and assertion figures in the hypothesis results.)

4. **RQ4 (Tooling Readiness):** What is the state of software tooling and infrastructure for Active Inference research, and what gaps remain? We survey available implementations to identify whether the ecosystem is fragmented or converging.

2.7 Scope and Delimitations

This study focuses on the English-language peer-reviewed and preprint literature retrievable from arXiv, Semantic Scholar, and OpenAlex. Our search scope begins at 2005—chosen to capture Energy-Based Model and variational Bayesian antecedents (Helmholtz machines, VAEs, early Bayesian brain formulations [Dayan et al., 1995, LeCun et al., 2006]) that share deep mathematical foundations with variational free energy minimization and predated the Free Energy Principle label introduced in 2006 [Friston et al., 2006]. The scope includes both the core Active Inference and Free Energy Principle literature and adjacent Energy-Based Model research where it intersects with variational inference or generative modeling—capturing the growing convergence between these traditions. We do not include book chapters or monographs not indexed by these sources, software documentation, or non-English publications. Domain classification uses keyword matching (200+ indicators across 8 categories) rather than expert annotation—a deliberate trade-off favoring reproducibility over precision, whose consequences we quantify in the field overview. Hypothesis scoring relies on LLM-extracted assertions operating on abstracts only; claims embedded in method sections, discussion paragraphs, or supplementary materials are not captured, and the fraction of relevant evidence residing in these sections is unknown. The fidelity and limitations of abstract-only extraction are examined in the extraction pipeline section. The hypothesis definitions and domain taxonomy are informed by, but not identical to, the Active Inference Ontology used by Knight et al. [Knight et al., 2022]; future alignment would enable direct comparison with that earlier analysis.

2.8 Principal Contributions

This work makes five contributions:

1. **A multi-source retrieval and deduplication pipeline** for Active Inference literature, using a canonical identifier hierarchy across three academic databases.
2. **A nanopublication-based knowledge graph schema** encoding directed, confidence-scored assertions about eight core hypotheses with full provenance tracking.
3. **A quantitative field overview** characterizing the growth, domain distribution (A/B/C taxonomy), citation topology, and latent topic structure of the Active Inference literature, with specific attention to how recent benchmark results (detailed in the domain analyses) are reshaping the scalability and application landscape.
4. **An LLM-based hypothesis scoring dashboard** that produces differentiated evidence profiles with temporal trend visualization.
5. **A tooling assessment** of the software ecosystem supporting Active Inference research, including the implemented extraction pipeline, existing software (pymdp, SPM, RxInfer.jl), and knowledge graph infrastructure.

The remainder of this paper is organized as follows. The methodology section describes the five-stage pipeline—the central contribution enabling reproducible, automated evidence synthesis—with separate treatments of literature retrieval, LLM-based assertion extraction, bibliometric analysis, the nanopublication-based knowledge graph, and visualization and variable injection. The hypothesis evidence landscape presents quantitative scoring results (RQ3), followed by the field overview with domain-level analysis (RQ1, RQ2), detailed domain analyses, text analytics, and citation network topology. The conclusion addresses limitations and future directions; the discussion provides community recommendations and open questions. Appendix 8 collects mathematical and algorithmic details; Appendix 7 surveys the tooling landscape (RQ4).

3 Methodology: Pipeline Design and Formal Definitions

This section describes the five-stage computational meta-analysis pipeline. Each stage corresponds to a tested, independently executable script that reads upstream outputs and produces structured artifacts. The pipeline extends the systematic literature analysis approach of Knight et al. [Knight et al., 2022]—which combined manual annotation with ontology-based automated analysis—by substituting manual coding with fully automated, LLM-driven assertion extraction and citation-weighted hypothesis scoring. All code, configuration files, and reproducibility instructions—including a Dockerized execution environment to guarantee dependency isolation—are publicly available in the project repository (`https://github.com/ActiveInferenceInstitute/act_inf_metaanalysis`).

3.1 Pipeline Overview

The five-stage pipeline is summarized in Table 1.

Table 1: Five-stage computational meta-analysis pipeline. Each stage corresponds to an independently executable script that reads upstream outputs and produces structured artifacts. Cross-references link to detailed methodology sections.

Stage	Script	Primary Input	Primary Output
1	`01_literature_search.py`	API queries	`corpus.jsonl`
2	`02_meta_analysis_pipeline.py`	`corpus.jsonl`	Classification, temporal, TF-IDF, NMF, citation network
3	`03_build_knowledge_graph.py`	`corpus.jsonl`	`nanopublications.jsonl`, `nanopublications.trig`, scor
4	`04_generate_figures.py`	All Stage 2–3 JSONs	16 publication-ready PNGs
5	`05_inject_variables.py`	All output JSONs	Rendered manuscript Markdown

Scripts act as thin orchestrators that import methods from tested library modules and handle file I/O. All computation resides in the `src/` packages; no analysis logic is embedded in scripts. End-to-end pipeline execution completes in under one hour on commodity hardware (excluding LLM extraction, which depends on model size and inference backend); all stochastic components use fixed random seeds for deterministic reproduction.

3.2 Reproducible Build Infrastructure

The five-stage analysis pipeline described above is embedded within `template/` [Friedman, 2026a,b], an open-source Infrastructure-as-Code system for computational research that turns a full research compendium—code, data, tests, manuscript, and provenance—into a single, version-controlled, deterministically buildable repository with an enforced, test-gated publication pipeline. `template/` applies the principle of Infrastructure as Code to the research lifecycle, making the manuscript, test suite, and provenance chain independently verifiable. The system operationalizes FAIR4RS principles [Wilkinson et al., 2016] and supply-chain-style provenance for manuscripts, targeting structural causes of the reproducibility crisis: fragmented workflows across LaTeX, notebooks, and ad-hoc scripts, lack of end-to-end testing, and no binding between code, data, figures, and the final PDF.

The system employs a Two-Layer Architecture: a globally shared *infrastructure layer* (12 subpackages, approximately 150 Python modules) provides generic services—logging, rendering, validation, steganographic watermarking, reporting, and LLM integration—while self-contained *project workspaces* (including the present meta-analysis) carry their own `manuscript/`, `scripts/`, `src/`, `tests/`, `data/`, and `output/` directories, discovered purely by filesystem convention. An eight-stage build pipeline enforces an ordered sequence from environment setup through test execution (at least 90% coverage for project code, at least 60% for shared infrastructure), analysis execution, PDF rendering (Pandoc to LaTeX to XeLaTeX with biber), output validation, LLM review, and executive reporting. A Zero-Mock testing policy requires all tests to exercise real filesystem operations, real subprocess calls, and real computation—no `unittest.mock` doubles—making test adequacy a publication gate rather than a best-effort guideline. Cryptographic provenance is embedded in every PDF via SHA-256 hash manifests, PDF metadata injection, and optional QR codes linking back to the repository. A Documentation Duality standard equips every directory with both human-readable `README.md` and machine-readable `AGENTS.md` files, while each infrastructure module carries a `SKILL.md` skill descriptor aligned with the Model Context Protocol, enabling AI agents to locate and invoke module capabilities without hallucinating API signatures. The `template/` framework and this meta-analysis project are available under the Apache 2.0 License at `https://github.com/ActiveInferenceInstitute/act_inf_metaanalysis`.

3.3 Stage 1: Multi-Source Literature Retrieval and Deduplication

We retrieve papers from three complementary academic databases to maximize coverage and enable cross-source deduplication. The retrieval window begins at 2005, encompassing the period when Energy-Based Model and variational Bayesian research [Dayan et al., 1995, LeCun et al., 2006] provided mathematical precursors to what Friston formalized as the Free Energy Principle in 2006 [Friston et al., 2006]; this inclusive start captures historical lineage and cross-disciplinary convergence that a later cutoff would exclude.

arXiv. We query the arXiv Atom API using five phrase-matched searches by default: `all:"active inference"`, `all:"free energy principle"`, `all:"predictive coding" AND all:"free energy"`, `all:"expected free energy"`, and `all:"variational free energy" AND all:"inference"`. The `all:` prefix searches titles, abstracts, and full text; phrase matching reduces contamination from unrelated physics papers that mention "free energy" in thermodynamic contexts. Additional Energy-Based Model queries (`all:"energy-based model" AND all:"free energy"`, `all:"Helmholtz machine" AND all:"inference"`, `all:"Boltzmann machine" AND all:"free energy"`, `all:"contrastive divergence" AND all:"generative model"`) are available via the `arxiv_queries` list in `config.yaml` for researchers wishing to capture adjacent EBM literature at the intersection of energy-based generative modeling and variational inference [LeCun et al., 2006].

Semantic Scholar. We query the Semantic Scholar Graph API [Kinney et al., 2023] with the same terms. Semantic Scholar provides citation graphs, abstract embeddings, and links to published versions. Retry logic with exponential backoff handles rate limiting.

OpenAlex. We query OpenAlex [Priem et al., 2022] to capture journal-published work that may not appear on arXiv, including clinical studies and neuroscience experiments in domain-specific venues. The `referenced_works` field populates citation links for each paper.

3.3.1 Canonical Identifier Deduplication

After retrieval, papers are assigned a canonical identifier using the priority scheme: DOI > arXiv ID > Semantic Scholar ID > OpenAlex ID > title hash. When the same paper appears in multiple sources, the record with the highest metadata completeness is retained. For each incoming paper, the two records are compared on metadata completeness—defined as the count of non-empty optional attributes across the full Paper record (abstract, DOI, arXiv ID, Semantic Scholar ID, OpenAlex ID, venue, citation count, references, publication date, PDF URL, open-access flag, and author list). The pipeline retains the richer record; in the event of a tie, the incumbent is preserved. This "merge-on-add" strategy aggregates the richest available metadata without requiring an expensive downstream reconciliation pass. Deduplication produces $N = 819$ unique papers spanning 2005–2026.

3.3.2 Relevance Filtering and Curation

After deduplication, a **relevance filter** removes papers whose titles and abstracts lack any core Active Inference terminology (e.g., `active inference,''`free energy principle,'' "variational free energy''), eliminating off-topic results introduced by broad keyword overlap across heterogeneous databases. We acknowledge that this retrieval strategy yields limited bibliographic depth, functioning as a representative snapshot rather than an exhaustive census of the literature.

We emphasize that this process relies on keyword search strategies across divergent APIs. In any complex research field, there is no single optimal word or threshold for definitive inclusion or exclusion. Different information sources and repositories yield differing schemas and representations, introducing both false positives (e.g., machine learning papers that mention "free energy" in a purely thermodynamic context, or bioinformatics tools whose names overlap with AIF terminology) and false negatives (e.g., predictive coding studies that avoid the phrase "free energy principle" entirely, or agent-based modeling papers that implement functionally equivalent algorithms under different labels). The keyword lists in `config.yaml` document all search terms explicitly to enable systematic replication and refinement.

Consequently, this pipeline is not intended to produce a static, "golden" list of canonical papers. Rather, it is designed as an open-source software package that can be modularly updated and versioned. Researchers can configure the pipeline to operate on custom literature bibliographies curated for specific relevance criteria through time, treating the initial query-based retrieval as a programmatic starting point rather than an absolute boundary. For example, adding a ninth domain category (e.g., "D: Education") requires only adding a keyword list to the `subfield_keywords` section of `config.yaml`—no source code modification is needed.

3.4 LLM-Based Assertion Extraction: Prompt Design, Error Taxonomy, and Validation

This supplementary section documents the implementation specifics of the LLM-based assertion extraction pipeline.

3.4.1 Relationship to Prior Approaches

The closest prior effort is the systematic literature analysis of Knight, Cordes, and Friedman [Knight et al., 2022], which used human annotators to manually code structural, visual, and mathematical features of FEP and Active Inference publications. Their work operated at the scale of hundreds of annotated papers and employed terms from the Active Inference Institute's Active Inference Ontology for automated text analysis. Our pipeline replaces the manual coding step with LLM-based assertion extraction, enabling scalable processing of the full corpus ($N = 819$ papers) at the cost of exchanging human-verified precision for machine-generated assessments that require post-hoc validation. This trade-off is characteristic of the broader LLM-based scientific extraction landscape: recent benchmarking confirms that even state-of-the-art modular extraction architectures fall short of production-level precision—particularly on tasks requiring exhaustive retrieval and aggregation of multiple values from long documents—validating our design choice to retain human review pathways alongside automated extraction.

Table 2: Comparison of annotation approaches: Knight et al. (2022) manual coding versus this work's automated LLM-based extraction pipeline. Key trade-offs between human-verified precision and machine-generated scalability are highlighted.

Dimension	**Knight et al. (2022)**	**This work**
Scale	Hundreds of papers	819 papers
Annotation	Manual (structural/visual/math features)	Automated (LLM hypothesis assessment)
Ontology	Active Inference Ontology terms	8 standard hypotheses
Output	Annotated features + term frequencies	Nanopublications + knowledge graph
Reproducibility	Annotator-dependent	Deterministic (given model + seed)
Precision	High (human-verified)	Medium (requires validation)

3.4.1.1 Positioning in the LLM-Based Review Landscape Our pipeline operates within a rapidly maturing ecosystem of LLM-powered literature analysis tools. Multi-agent architectures such as LitLLM decompose the review process into specialized sub-agents (planner, identifier, extractor, compiler), while ensemble approaches aggregate outputs from multiple LLMs via weighted voting to improve reliability. Our work differs from these tools in three respects: (1) we target *hypothesis-level evidence scoring* rather than inclusion/exclusion screening; (2) we produce structured nanopublications rather than narrative summaries; and (3) we are only analyzing abstracts for claims. This deliberate trade-off enables corpus-scale processing ($N = 819$) but fundamentally misses fine-grained claims embedded in method sections or discussion paragraphs. Full-text processing could improve extraction recall, particularly for hypotheses with small evidence bases (H6 Clinical Utility, H7 Morphogenesis).

3.4.2 The Eight Tracked Hypotheses

Our analysis tracks the evolving evidence base for eight distinct claims within the Active Inference literature, spanning theoretical universality to applied clinical utility:

1. **H1: FEP Universality (Theoretical).** The Free Energy Principle applies universally to all self-organizing systems.
2. **H2: AIF Optimality (Computational).** Active Inference agents achieve optimal decision-making under uncertainty.
3. **H3: Markov Blanket Realism (Philosophical).** Markov blankets correspond to real physical boundaries.
4. **H4: Predictive Coding (Empirical).** Cortical hierarchies minimize prediction errors via predictive coding.
5. **H5: Scalability (Computational).** Active Inference scales to complex, high-dimensional environments.
6. **H6: Clinical Utility (Applied).** Active Inference provides clinically useful models of psychiatric conditions.
7. **H7: Morphogenesis (Biological).** The FEP explains morphogenetic and developmental processes.
8. **H8: Language AIF (Applied).** Active Inference provides a viable framework for language processing.

3.4.3 Prompt Engineering and Schema Design

The structured prompt is designed to minimize parsing failures and maximize assessment quality:

1. **Explicit JSON schema.** The prompt specifies the exact output schema—field names, allowed direction values, and the numeric confidence range—reducing the LLM's tendency to generate free-form text or ad hoc structures.
2. **Hypothesis definitions in-context.** All eight definitions are included verbatim, ensuring the LLM assesses relevance from the provided context rather than relying on parametric knowledge that may be stale.
3. **Reasoning field.** Each assessment includes a natural-language reasoning string, providing an audit trail for human reviewers and enabling systematic analysis of error patterns.
4. **Irrelevant filtering.** An explicit "irrelevant" direction allows the LLM to mark hypotheses that a paper does not address, avoiding forced spurious assessments.

3.4.3.1 Prompt Template The extraction prompt follows a two-part structure (system + user):

```
SYSTEM: You are a scientific literature analyst specializing in the
Free Energy Principle and Active Inference. Assess the relevance of
the given paper to each hypothesis. Return a JSON array.

USER:
Paper: {title}
Abstract: {abstract}

Hypotheses:
H1: FEP Universality - {description}
H2: AIF Optimality - {description}
...
H8: Language AIF - {description}

For each hypothesis, return:
{
  "hypothesis_id": "H1",
  "direction": "supports|contradicts|neutral|irrelevant",
  "confidence": 0.0-1.0,
  "reasoning": "..."
}
```

The extraction module (`src/knowledge_graph/llm_extraction.py`) includes configurable retry logic with exponential backoff, JSON parsing with handling of markdown code fences and extraneous text, confidence clamping, and validation against the hypothesis ID set. The default model is `gemma3:4b` on a local Ollama instance, configurable via `--llm-model` and `--llm-url` flags.

3.4.4 Failure Modes and Error Recovery

The primary failure modes are documented below.

3.4.4.1 Over-Extraction Bias Preliminary experiments indicated ~15–20% over-extraction. The current $N = 819$ corpus was processed without a validation set; error rates are not quantified for this run. This is the most common error mode and produces false supporting evidence. Over-extraction disproportionately affects broad-scope hypotheses (H1 FEP Universality, H2 AIF Optimality) where most papers in the corpus contain relevant terminology without explicitly endorsing the claim. Narrower hypotheses tied to specific domains (H7 Morphogenesis, H8 Language AIF) show lower over-extraction rates because their vocabulary is more distinctive. This systematic bias inflates support counts for broad hypotheses, and we caution against interpreting absolute scores for H1 and H2 without accounting for this effect.

3.4.4.2 Direction Misclassification The LLM misclassifies a contradicting claim as supporting, or vice versa. Rarer but more consequential, as it directly inverts the evidence signal. Most common for papers that discuss limitations while ultimately endorsing a hypothesis.

3.4.4.3 Confidence Calibration Constraints The model occasionally assigns high confidence to assessments where the underlying evidence is ambiguous. Reliable confidence calibration remains an open problem for zero-shot LLM applications, motivating the multi-tiered validation protocols described below.

3.4.4.4 Progressive JSON Parsing Recovery To mitigate formatting inconsistencies, the module implements a progressive parsing pipeline to recover malformed LLM outputs:

1. **Direct parse**: Attempt `json.loads()` on the raw response.
2. **Strip code fences**: Remove Markdown ```` ```json ... ``` ```` wrappers and retry.
3. **Extract JSON array**: Scan for the first `[...]` substring in the response text.

Papers that fail all parsing stages are logged and skipped; their count is reported at pipeline completion.

3.4.5 Validation Methodology

Validation of LLM-extracted assertions follows a three-tier protocol:

1. **Validation Dataset (10%, not yet created).** A ground-truth validation protocol is specified in which a random 10% subset of the corpus will be manually annotated by human experts. Inter-rater reliability will be calculated using Cohen's κ; the LLM-based extraction pipeline will be evaluated against this human consensus, targeting a $\kappa > 0.70$ threshold for direction accuracy (supports/contradicts/neutral/irrelevant). The formal 10% manual annotation dataset has not yet been created; its development is a prioritized next step for this living review architecture.
2. **Boundary-case audit (conceptual design).** Papers known to make contested claims (e.g., critiques of FEP universality, Markov blanket realism debates) would be specifically checked for correct direction assignment. This tier remains a conceptual design and has not been executed.
3. **Aggregate consistency (conceptual design).** Hypothesis scores would be compared against qualitative expectations from the literature: hypotheses known to be well-supported (e.g., H4 Predictive Coding) should score positively; those known to be contested (e.g., H3 Markov Blanket Realism) should show lower or mixed scores. This tier also remains a conceptual design and has not been executed.

The current extraction pipeline operates without human-validated ground truth; all reported assertions are machine-generated and unaudited.

3.4.6 From Assertions to Nanopublications

Each validated assertion is wrapped in a **nanopublication** [Groth et al., 2010, Kuhn et al., 2016]—a self-contained, machine-readable knowledge unit packaging the assertion with explicit provenance metadata. The wrapping process assigns:

- A **unique identifier** (`nanopub:<uuid12>`) for graph-level deduplication.
- An **attribution string** recording the pipeline name and LLM model version.
- A **UTC timestamp** in ISO 8601 format, establishing temporal provenance.

Nanopublications are persisted **incrementally** during extraction. Every 50 papers (configurable via `--checkpoint-interval`), the pipeline atomically appends newly extracted nanopublications to `nanopublications.jsonl` using a temporary-file-plus-rename strategy that prevents corruption on interruption. Deduplication operates on the composite key $(paper\_id, hypothesis\_id)$: when a paper is re-processed with an improved model, the newer assertion overwrites the stale entry. This merge-on-add design enables iterative model refinement without costly full-corpus re-extraction.

After extraction completes, the full nanopublication set is additionally serialized to **RDF/TriG** format per the nanopublication standard, producing four named graphs per nanopublication (Head, Assertion, Provenance, Publication Info). The TriG output is suitable for publication to the decentralized nanopublication network and archival on data repositories such as Zenodo. The complete RDF schema is specified in the knowledge graph methodology and Appendix 8.5.

3.5 Stage 2: Bibliometric Analysis

Stage 2 performs four complementary analyses on the deduplicated corpus. All analyses are deterministic given fixed random seeds and operate on the same `corpus.jsonl` input.

3.5.1 Subfield Classification

Each paper is classified into one of eight categories organized across three domains: **A – Core Theory** (A1: quantitative and formal mathematical theory; A2: qualitative philosophy and general FEP theory), **B – Tools & Translation** (algorithms, scaling, and software development), and **C – Application Domains** (C1: neuroscience, C2: robotics, C3: language processing, C4: computational psychiatry, C5: biology and morphogenesis). Classification uses word-boundary-aware keyword matching against curated lists (74+ mathematical indicators, 25+ philosophy terms, 24+ tools terms, and 14–20 terms per application domain—totaling over 200 keywords across 8 categories, all documented in `config.yaml`) applied to titles and abstracts. A priority system ensures that specific application domains (C1–C5, priority 1) take precedence over tools (B, priority 2), formal theory (A1, priority 3), and the broad qualitative philosophy catch-all (A2, priority 4). Within a priority tier, the domain with the most keyword matches wins. A1's keyword set includes mathematical indicators such as *theorem*, *proof*, *convergence*, *posterior*, *equation*, and *Fokker–Planck*, ensuring that papers with mathematical content are classified as formal theory rather than defaulting to the philosophy category.

3.5.2 Temporal Metrics and Growth-Rate Estimation

We compute temporal publication metrics including year-by-year counts with gap-filling, cumulative totals, 3-year smoothed moving averages, and peak year identification. Field dynamics are estimated via two complementary metrics. The **mean year-over-year growth rate** $\bar{g}$ is the arithmetic mean of annual growth rates for years with non-zero prior-year publications. The **doubling time** $t_d = \ln 2 / \ln(1 + \bar{g})$. The **compound annual growth rate** (CAGR) captures the annualized rate across the full temporal span. Mathematical details are provided in Appendix 8.3.

3.5.3 Text Analytics

We construct the TF-IDF matrix using tokenization with stopword removal and L2-normalized smoothed term-frequency inverse-document-frequency weighting [Salton et al., 1975], with a configurable vocabulary size (default: 500 features). We apply non-negative matrix factorization (NMF) to discover latent topics using multiplicative update rules [Lee and Seung, 1999]. Topic count $k = 5$ was selected via expert-driven assessment. Mathematical details are provided in Appendix 8.2.

3.5.4 Citation Network Construction

We construct the intra-corpus citation network as a directed graph where nodes are papers and edges represent citation relationships resolved within the corpus. Because identifier formats vary across databases (arXiv IDs, DOIs, Semantic Scholar IDs), only references whose identifiers match a corpus entry contribute edges; the resulting resolution rate (7.4%) represents a lower bound on the true intra-corpus citation density. Network metrics include PageRank centrality, HITS hub and authority scores [Kleinberg, 1999], degree distributions, network density, connected components, and community structure via greedy modularity maximization [Clauset et al., 2004].

3.6 Stage 3: Nanopublication-Based Knowledge Graph

Stage 3 is the methodological core of this work: it transforms unstructured abstracts into a structured, RDF-compatible knowledge graph of scientific evidence. The stage encompasses four tightly coupled operations: LLM-based assertion extraction, nanopublication packaging, knowledge graph construction, and citation-weighted hypothesis scoring.

3.6.1 LLM-Based Assertion Extraction

We extract assertions by prompting a locally hosted LLM (Ollama [Ollama Team, 2024]) to assess each paper's abstract against eight standard hypotheses. The model receives a structured prompt containing the paper title, abstract, and hypothesis definitions, and returns a JSON array where each element specifies a hypothesis ID, direction (supports, contradicts, neutral, or irrelevant), a confidence score $c \in [0, 1]$, and a reasoning string. Assertions marked "irrelevant" are discarded; confidence values are clamped to $[0, 1]$; and responses are validated against the known hypothesis ID set. Papers lacking abstracts are skipped. Detailed prompt engineering, error taxonomy, and validation methodology are documented in the extraction pipeline section.

3.6.2 Nanopublication Schema and RDF Structure

Each assertion is encoded as a **nanopublication** [Groth et al., 2010, Kuhn et al., 2016]—a minimal, self-contained, machine-readable unit of scientific evidence. Formally, each nanopublication is a tuple (p, h, d, c) where p is the paper identifier, h the hypothesis identifier, $d \in \{\text{supports}, \text{contradicts}, \text{neutral}\}$ the direction, and c the confidence. Provenance metadata records the LLM model, UTC timestamp, and paper identifier.

The pipeline serializes nanopublications in two complementary formats:

1. **JSON Lines** (one JSON object per line) for efficient incremental checkpointing. Assertions are saved at configurable intervals (default: every 50 papers), enabling the pipeline to resume from where it left off after interruption without re-processing already-analyzed papers. Deduplication uses the composite key $(paper\_id, hypothesis\_id)$; re-runs with improved models overwrite stale results.
2. **RDF/TriG** per the nanopublication standard (nanopub.net), producing four named graphs per nanopublication:

Table 3: RDF/TriG nanopublication structure. Each nanopublication contains four named graphs encoding the assertion, its provenance, and publication metadata per the nanopublication standard (`nanopub.net`).

Named Graph	Content	Key Predicates
Head	Links the nanopub resource to its three component graphs	`np:hasAssertion`, `np:hasProvenance`, `np:has`
Assertion	The core scientific claim	`aif:asserts`, `aif:supports/aif:contradict`
Provenance	How the assertion was generated	`prov:wasGeneratedBy`, `prov:generatedAtTim`
Publication Info	Metadata about the nanopublication itself	`dc:created`, `dc:creator`, `dc:license`

The namespace `http://activeinference.institute/ontology/` (prefix `aif:`) defines all domain predicates; the nanopublication schema (`http://www.nanopub.org/nschema#`, prefix `np:`) provides structural predicates; provenance uses PROV-O (`http://www.w3.org/ns/prov#`); and Dublin Core (`http://purl.org/dc/terms/`) provides publication metadata. The TriG output is suitable for publication to the decentralized nanopublication network and aligns with FAIR data principles: **F**indable via URI-based identification, **A**ccessible via standard RDF protocols, **I**nteroperable through W3C-standard serialization, and **R**eusable with explicit provenance and CC0 licensing.

3.6.3 Knowledge Graph Construction

The knowledge graph is an RDF-compatible directed graph with three node types: **paper nodes** (metadata: title, abstract, authors, year, venue, citation count, domain), **assertion nodes** (claim text, direction, hypothesis ID, confidence), and **hypothesis nodes** (the eight standard hypotheses). Edges encode five relations defined in the schema:

- `aif:asserts` — Paper → Assertion
- `aif:cites` — Paper → Paper
- `aif:belongsTo` — Paper → Subfield
- `aif:supports` — Assertion → Hypothesis
- `aif:contradicts` — Assertion → Hypothesis

The graph is implemented with a dual backend: `rdflib` [RDFLib Team, 2023] when available (preferred for semantic web compatibility), with automatic fallback to `networkx.DiGraph` for environments without RDF dependencies. Both backends maintain identical internal indices for efficient paper, assertion, and hypothesis queries.

3.6.4 Citation-Weighted Hypothesis Scoring

For each hypothesis H, we compute a citation-weighted evidence score:

$$\text{score}(H) = \frac{\sum_{a \in S(H)} w(a) - \sum_{a \in C(H)} w(a)}{\sum_{a \in A(H)} w(a)} \tag{1}$$

where $S(H)$, $C(H)$, and $A(H)$ are the sets of supporting, contradicting, and all assertions for H, and the weight function is:

$$w(a) = \log(1 + \text{citations}(a)) \cdot \text{confidence}(a) \tag{2}$$

The logarithmic citation weighting ensures that highly cited papers carry more influence without allowing any single paper to dominate. The score lies in $[-1, 1]$. **Interpretation note:** a score of $+0.7$ indicates that 70% of weighted evidence supports the hypothesis (net of contradictions and normalized by total weighted evidence), *not* that the hypothesis has a 70% probability of being true. Scores are best interpreted as relative rankings across hypotheses and as temporal trajectories within a hypothesis, rather than as absolute probability estimates. Temporal trends are computed by evaluating the cumulative score at each year, using only assertions from papers published up to that year. A full derivation appears in Appendix 8.1.

3.6.5 Tally-Based Evidence Aggregation

We emphasize that this algorithmic scoring formula constitutes a **tally-based approach** to evidence synthesis: each nanopublication assertion operates as an independent evidential vote, weighted by citation impact and the extraction model's confidence. The aggregation is linear and additive—supporting and contradicting assertions are summed and differenced without modeling dependencies, correlated evidence, or causal structure among claims. This design choice prioritizes transparency, reproducibility, and computational tractability over statistical sophistication.

The tally-based framing introduces three constraints. First, assertions from methodologically related papers (e.g., iterative publications from a single research group testing the same model) are counted independently, amplifying correlated evidence. To illustrate: if a group publishes three papers (2019, 2021, 2023) reporting successively refined variants of the same predictive coding model, each with high citation counts, the scoring formula counts three independent supporting assertions for H4—even though the underlying empirical evidence is largely overlapping. An evidential diversity index (proposed in the conclusion) would downweight this cluster. Second, the scoring metric treats all assertion sources symmetrically: an assertion from a theoretical review and one from an empirical trial carry equal weight at a given confidence level. Third, temporal scoring tracks *cumulative totals* rather than dynamic probabilistic estimates; the score at year t is the sum of all historical evidence, rather than a decaying posterior that downweights early work.

We embrace these constraints intentionally. The tally-based approach provides a stable, interpretable baseline against which more sophisticated scoring methods can be evaluated. The conclusion describes concrete extensions—including hierarchical Bayesian scoring, causal evidence graphs, and evidential diversity indices that downweight correlated evidence.

3.7 Stages 4–5: Visualization, Variable Injection, and Reproducibility

3.7.1 Stage 4: Visualization

Stage 4 renders 16 publication-ready figures from the analysis outputs of Stages 2 and 3. All figures use the Wong (2011) colorblind-safe palette [Wong, 2011] and enforce a 16-point minimum font size for accessibility compliance. Figures span six categories: field summary and domain distribution (2 figures), growth and temporal dynamics (2 figures), citation network topology (2 figures), hypothesis evidence dashboard and timeline (2 figures), assertion composition (2 figures), and text analytics—word cloud, PCA embeddings, term heatmap, dendrogram, topic-term bars, and co-occurrence matrix (6 figures). The figure generation script reads only JSON outputs and produces only PNG files, enforcing a strict, unidirectional data flow that prevents visualization operations from inadvertently modifying analytical results.

3.7.2 Stage 5: Manuscript Variable Injection

Stage 5 computes dynamic variables from all pipeline outputs and injects them into manuscript Markdown templates via double-brace placeholder substitution of the form `{<>}` wrapping a variable name (e.g. the literal token spelled `{{<CORPUS_SIZE>}}` becomes the rendered corpus count). Variables include corpus-level metrics (size, year range, CAGR), per-domain counts and percentages, citation network statistics (nodes, edges, density, components, resolution rate, mean in-degree), hypothesis scores, and figure counts. All formatting (comma thousand separators, escaping) is applied during variable computation, ensuring the manuscript templates remain human-readable while producing publication-ready output. Unrecognized placeholders are preserved with a warning logged, enabling incremental manuscript development ahead of full pipeline execution.

3.7.3 Reproducibility and Test-Driven Validation

The pipeline is deterministic given fixed random seeds and API responses. Test-driven development enforces 90% minimum code coverage on project modules and 60% on shared infrastructure, with real data and computation (no mocking). The test suite validates boundary conditions for hypothesis scoring (all-support $\rightarrow +1$, all-contradict $\rightarrow -1$, balanced $\rightarrow 0$), schema consistency, serialization round-trips, and end-to-end pipeline integrity. Source code, configuration, and outputs are available under CC-BY-4.0.

4 Results

4.1 Hypothesis Evidence Landscape and Temporal Dynamics

The LLM-based extraction pipeline produced a total of 1,490 assertions across the eight tracked hypotheses, drawn from the full corpus of $N = 819$ papers. Before presenting the results, we reiterate the interpretive framework established in the methodology: hypothesis scores are *relative rankings* among hypotheses and *temporal trajectories* within each hypothesis—they are not absolute probability estimates. Publication bias and linguistic asymmetry (§4.1.4.1) inflate all scores toward the positive end, and the tally-based aggregation does not model evidential dependencies. The distribution of assertion types and the resulting citation-weighted scores reveal a differentiated evidence landscape (Figure 1):

Table 4: Citation-weighted hypothesis evidence landscape ($N = 819$ papers, 1,490 total assertions). Scores are computed via (1) and range from −1 (unanimous contradiction) to +1 (unanimous support). "Character" summarizes the qualitative evidence profile for each hypothesis.

Hypothesis	Score	Supports	Neutral	Contradicts	Total	Character
H7: Morphogenesis	+1.00	24	0	0	24	Strong consensus
H2: AIF Optimality	+0.97	291	6	2	299	Strong consensus
H4: Predictive Coding	+0.94	233	24	1	258	Strong consensus
H6: Clinical Utility	+0.93	23	2	0	25	Strong consensus
H5: Scalability	+0.85	108	13	0	121	Strong consensus
H8: Language AIF	+0.83	31	3	0	34	Strong support
H3: Markov Blanket Realism	+0.78	12	2	4	18	Moderate, active debate
H1: FEP Universality	+0.48	281	429	1	711	Broad but diffuse

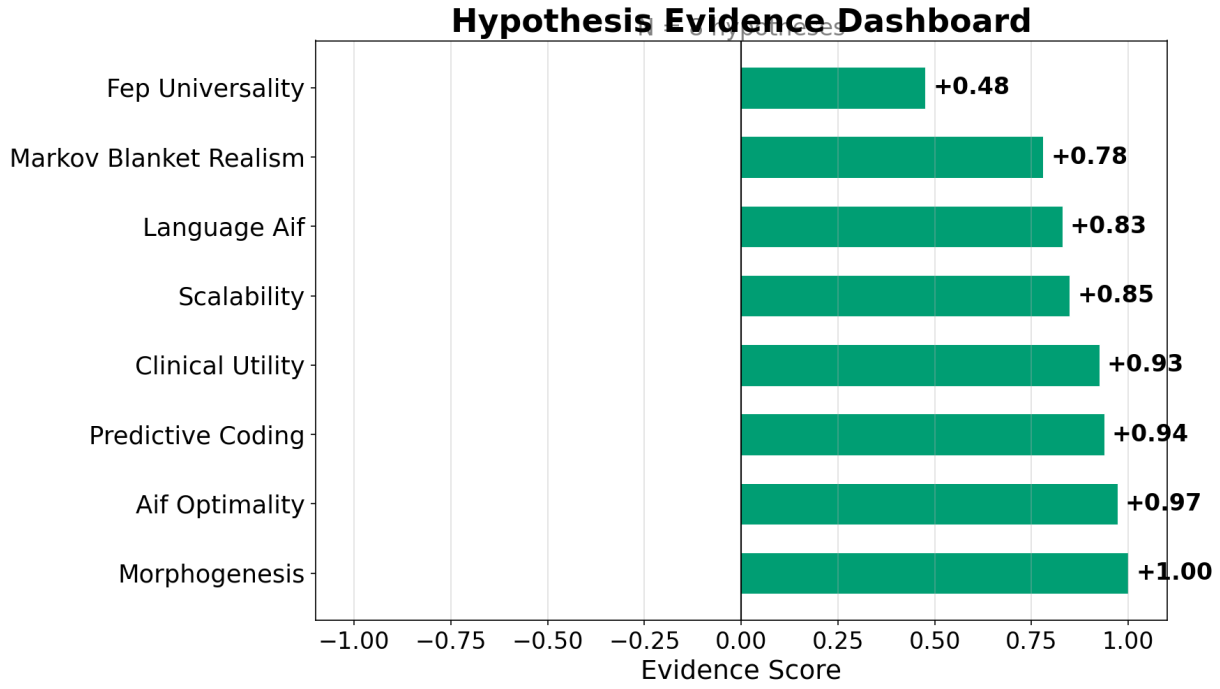

Figure 1: Hypothesis scoring dashboard showing citation-weighted evidence scores ($[-1, +1]$) for the eight tracked hypotheses, sorted descending by consensus strength. Predominantly positive scores reflect both genuine empirical support and systematic positive biases from publication selection and linguistic framing (see §4.1.4.1).

4.1.1 Interpretation of Evidence Profiles

To directly address our core research questions—identifying which claims are robustly supported and which remain contested—we evaluated how the hypothesis-level evidence maps against the critiques introduced in §3. The eight

hypotheses cluster into three tiers, defined by score ranges that emerge from the data rather than being imposed a priori. The **consensus tier** (score > 0.83; H7, H2, H4, H6, H5) spans five of the eight hypotheses, revealing a predominantly supportive evidence landscape across domains. H8 (Language AIF) sits at the **boundary of consensus** at score +0.83 — above 0.8 but below the more stringent 0.83 line that separates the densely populated upper tier from the rest. Morphogenesis (H7) achieves the maximum score (+1.00), though its small evidence base (24 assertions) means unanimity reflects limited assessment scope rather than mature empirical closure. AIF Optimality (H2) holds the second highest score (+0.97) despite carrying the largest raw count of contradicting assertions (2): supporting assertions are substantially more highly cited than critical ones, so citation-weighting amplifies the supportive signal—underscoring that citation-weighted scores capture *which* claims the community cites most, not a simple ballot of assertion counts. Predictive coding (H4), the most extensively assessed hypothesis with 258 assertions and a score of +0.94, has accumulated overwhelmingly supportive evidence since the 1970s, reflecting the deep empirical grounding of hierarchical prediction error models in neuroscience. This trajectory is consistent with the manual benchmarking results of Knight et al. [Knight et al., 2022], which similarly identified predictive coding as the most rigorously validated construct in the corpus. Clinical Utility (H6, +0.93) and Scalability (H5, +0.85) complete the upper consensus tier; H5's trajectory accelerated sharply after 2017 as deep active inference architectures emerged. The H8 (Language AIF) boundary placement noted above reflects recent breakthroughs coupling active inference to large language models within a still-maturing evidence base.

The **moderate tier** (score 0.5–0.8; H3) contains a single hypothesis. Markov blanket realism (H3) has the smallest overall evidence base (18 assertions) with a score of +0.78 and 4 contradicting assertions—empirically capturing the ongoing philosophical debate between those who treat Markov blankets as real thermodynamic boundaries (`Friston blankets'') and those who argue they are purely instrumental statistical tools (`Pearl blankets'') [Bruineberg et al., 2022]. The moderate score for H3 reflects this active ontological debate: the supporting literature is more highly cited but not by a large margin, and the small total evidence base limits inferential confidence.

The **diffuse tier** (score < 0.5; H1) is the most diagnostically informative for understanding the field's intellectual maturation. FEP universality (H1) generates one of the largest raw evidence bases (711 assertions) yet achieves a score of only +0.48—a striking gap explained by assertion composition: neutral assessments account for 429 of those 711 tallies, while supporting assertions number 281 and contradicting assertions just 1. This neutral plurality—more than either supporting or contradicting tallies—reveals that researchers routinely *invoke* the FEP as conceptual scaffolding without subjecting its universality claim to explicit empirical test. This composition is the quantitative fingerprint of the falsifiability critique leveled by Colombo and Seri‘es [Colombo and Seriès, 2021]: a principle elastic enough to accommodate any self-organizing system without generating predictions that distinguish it from alternatives will naturally accumulate invocations rather than tests, and invocations register as neutral in the extraction pipeline.

4.1.2 Temporal Dynamics of Evidence Accumulation

The cumulative evidence timeline (Figure 2) reveals three temporal patterns. First, **early convergence**: H4 (predictive coding) reached positive territory in the late 1990s following the publication of Rao and Ballard's foundational predictive coding model [Rao and Ballard, 1999] and has maintained a high score since, reflecting the mature empirical base in cognitive neuroscience. Second, **recent acceleration**: H5 (scalability) and H6 (clinical utility) show steep upward trends after 2017, tracking the emergence of deep active inference tools and computational psychiatry applications. The H5 trajectory reflects a cumulative body of work culminating in benchmark demonstrations such as AXIOM [Heins et al., 2025], which showed that object-centric world models under AIF can match state-of-the-art deep RL performance—but the temporal trend was already positive before any single result, and the score captures the aggregate rather than any individual paper. Third, **moderate and stable**: H3 (Markov blanket realism) has maintained a score in the moderate range since 2018, with supporting papers partially offset by targeted philosophical critiques—a pattern consistent with ongoing debate rather than either clear consensus or rejection.

4.1.3 Assertion Composition and Distribution

The per-hypothesis composition of assertions (Figure 3) and the multi-panel summary (Figure 4) provide complementary views of the extraction results.

4.1.4 Limitations of the Current Scoring Approach

4.1.4.1 Publication Bias and Linguistic Asymmetry The predominantly positive scores observed across all eight hypotheses should be interpreted with two systematic caveats.

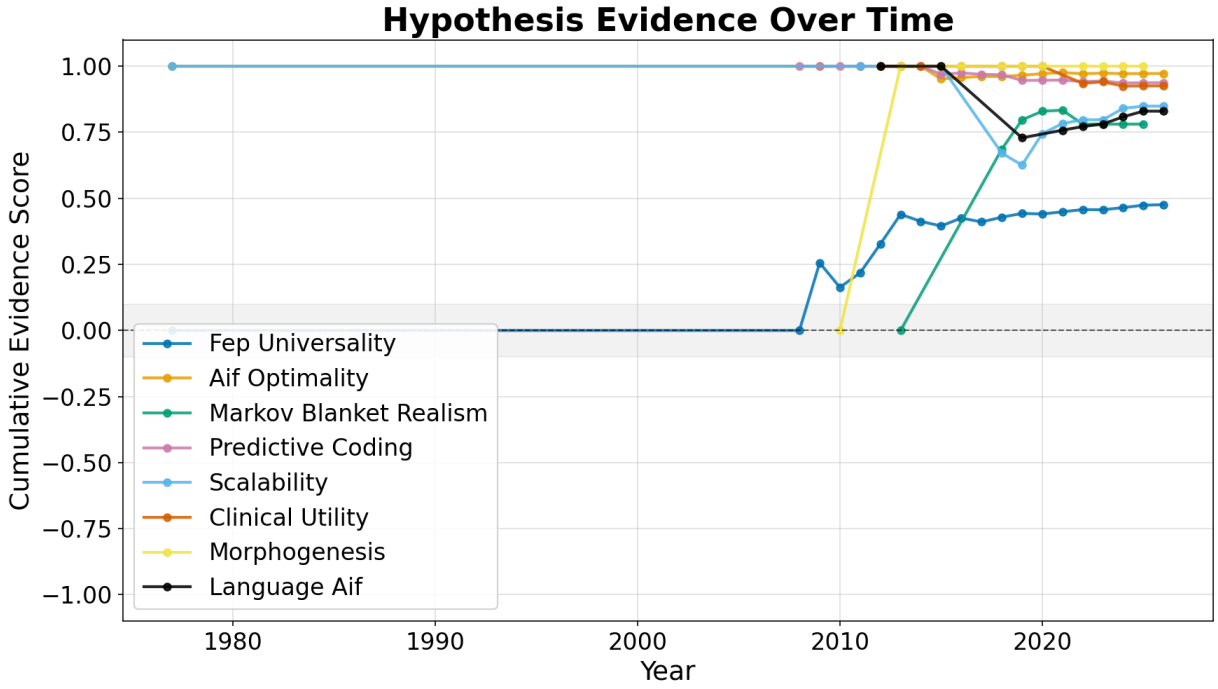

Figure 2: Temporal evolution of cumulative citation-weighted evidence scores by hypothesis (2005–2026). Divergent trajectories around the shaded neutral boundary (± 0.1) reveal which hypotheses are gaining or losing support over time. H4 (predictive coding) stabilized early; H5 (scalability) accelerated post-2017.

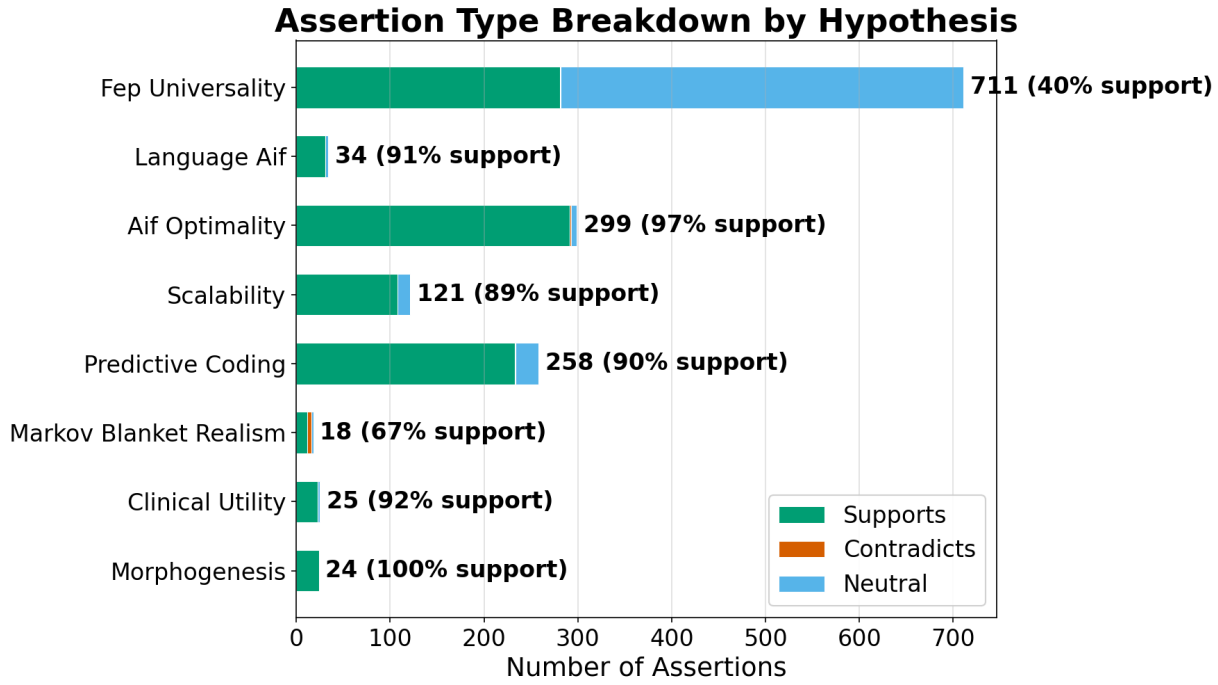

Figure 3: Stacked horizontal bars decomposing per-hypothesis assertions into supports (green), contradicts (red-orange), and neutral (blue) categories ($N = 1,490$ total assertions). Labels show total count and support percentage. The high support fractions are partially attributable to publication bias and affirmative linguistic framing.

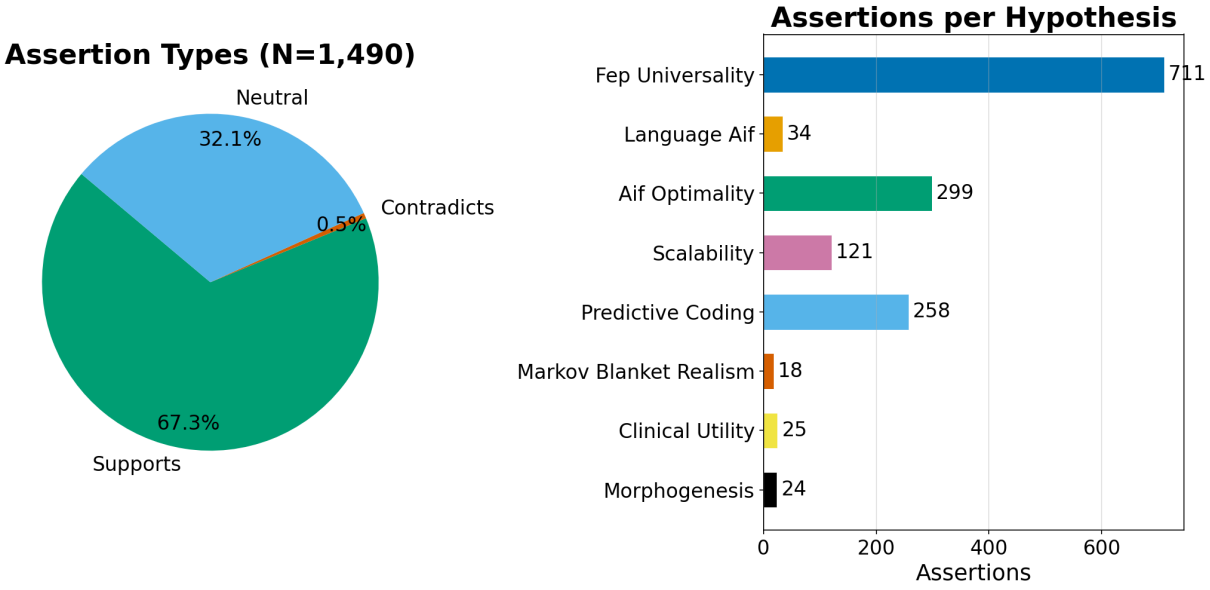

Figure 4: Multi-panel assertion summary: (left) pie chart of overall assertion type distribution showing supports/contradicts/neutral proportions, (right) per-hypothesis assertion counts with palette-coded bars. $N = 1,490$ assertions extracted from 819 papers.

First, **publication bias** systematically inflates supporting evidence. Academic journals preferentially publish positive and confirmatory results (Sterling 1959), meaning that studies finding null or contradictory outcomes for any hypothesis are less likely to appear in the retrievable literature. This *file-drawer effect* is well-documented across scientific disciplines and is expected to disproportionately suppress contradicting assertions in our extraction pipeline. The Active Inference literature is particularly susceptible: as a theoretical framework with strong foundational proponents, papers are more likely to frame results as consistent with the FEP than as challenges to it.

Second, **linguistic asymmetry** in academic writing further skews extraction toward positive classifications. Declarative scholarly claims are inherently phrased affirmatively—authors write "our results support," "consistent with," or "extends the prediction of" far more frequently than "our results refute" or "contradicts the claim that." Because the LLM extraction pipeline operates on abstract text, this linguistic imbalance propagates directly into the assertion distribution. Even papers presenting genuinely mixed evidence tend to frame their abstracts in terms of what *was* found rather than what was not, biasing the extracted direction toward "supports.''

These two effects act in concert: publication bias reduces the number of contradicting papers in the corpus, and linguistic framing reduces the number of contradicting assertions extracted from the papers that do appear. Consequently, the absolute values of hypothesis scores should not be taken as unbiased measures of scientific consensus. The *relative* ordering and temporal *trajectories* of hypothesis scores are more robust indicators, as these biases affect all hypotheses approximately equally.

4.1.5 Methodological Validation and LLM Calibration

The evidence derives from automated LLM-based assertion extraction operating on abstracts only, without human-validated ground truth calibration; confidence scores are self-assessed and uncalibrated; the pipeline uses $c \geq 0.60$ threshold to mitigate over-extraction. Relative rankings are more robust than absolute scores. A formal validation protocol (10% manual annotation, Cohen's κ, boundary-case auditing) remains a critical next step.

4.2 Field Overview: Disciplinary Structure and Growth Dynamics

Annual output in the Active Inference literature rose from 1 papers in 2005, reaching a peak of 123 papers in 2025—a transition from a niche within theoretical neuroscience to a multi-disciplinary research program spanning three primary domains and eight tracked categories. The corpus start of 2005 was chosen to capture Energy-Based Model and variational Bayesian antecedents [Dayan et al., 1995, LeCun et al., 2006] that preceded the formal introduction of the Free Energy Principle in 2006 [Friston et al., 2006] and its subsequent full elaboration [Friston, 2010]. Our corpus, extracted from arXiv, Semantic Scholar, and OpenAlex and deduplicated to $N = 819$ papers (2005–2026), captures the breadth, tempo, and internal architecture of this expansion (Figure 5).

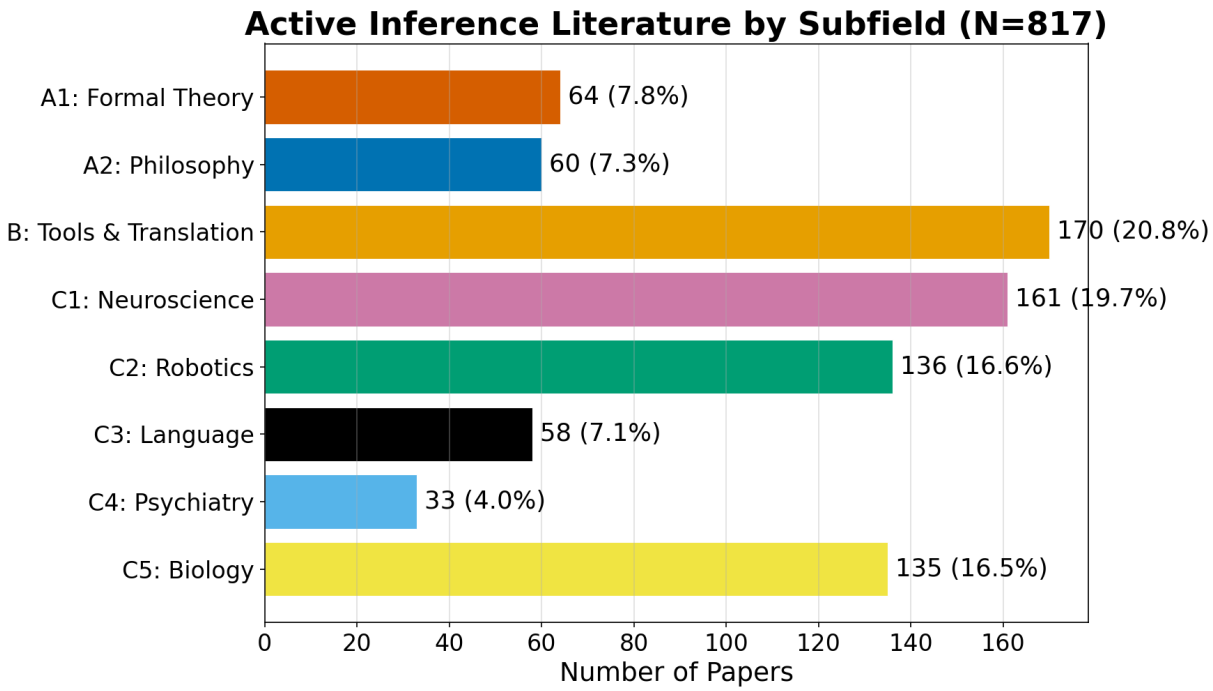

Figure 5: Publication counts by domain ($N = 819$). Application domains (C1–C5) collectively account for the largest share of the corpus; Domain A2 (qualitative philosophy) is the largest single category, reflecting the FEP's broad theoretical reach.

4.2.1 Corpus-Level Summary

Table 5: Corpus-level summary statistics for the Active Inference literature corpus ($N = 819$), spanning 2005–2026.

Metric	Value
Total papers	819
Year range	2005–2026
Peak year	2025
CAGR	20.36%
Active domains	8 of 8 tracked (A1–A2, B, C1–C5)

The CAGR of 20.36% (measured as the annualised growth rate of yearly publication volume between endpoint years 2005 and 2026) reflects sustained field expansion; the actual rapid growth phase began around 2013, with annual output accelerating substantially (Figure 6). Sustained high output persisting into subsequent years suggests the field has reached a mature production phase rather than experiencing a transient spike. Citation network metrics are detailed in the dedicated citation network analysis (see the citation network analysis).

4.2.2 Domain Distribution

Keyword-based classification assigns each paper to one of eight categories across three domains:

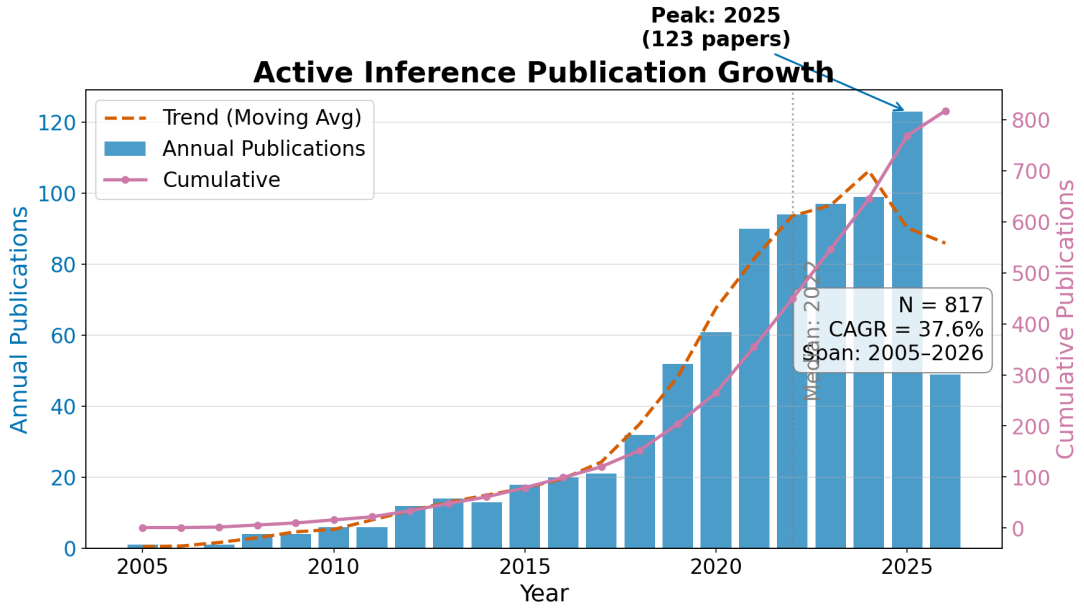

Figure 6: Annual (bars) and cumulative (line) publication counts, 2005–2026 (N = 819, CAGR = 20.36%). The inflection around 2013 marks the onset of rapid growth. Moving average trendline (dashed), peak year, and median year annotated.

Table 6: Domain distribution across three tiers and eight categories (N = 819 papers). Classification uses hierarchical keyword matching with priority-based routing to minimize over-assignment to catch-all categories.

Domain	Category	Papers	Percentage
A – Core Theory	A1: Formal Theory	64	7.8%
	A2: Qualitative Philosophy	60	7.3%
B – Tools	B: Tools & Translation	170	20.8%
C – Applications	C1: Neuroscience	161	19.7%
	C2: Robotics	136	16.6%
	C3: Language	58	7.1%
	C4: Psychiatry	33	4.0%
	C5: Biology	135	16.5%

The concentration of papers in A2 (qualitative philosophy and general theory) reflects the broad scope of foundational FEP work (Figure 7). The priority-based classifier mitigates over-assignment by routing papers with mathematical indicators (theorems, proofs, equations, statistical formalism) to A1 before falling back to A2, and by preferring specific application domains (C1–C5) and tools (B) over both core-theory categories. Papers that discuss FEP/AIF conceptually without mathematical formalism or domain-specific vocabulary are correctly assigned to A2. This figure should be read as a *ceiling* on theoretical generality rather than a literal measure of research focus—embedding-based classification would likely redistribute some fraction into more specific categories. That all eight categories are populated, including computational psychiatry (C4) and formal theory (A1), indicates diversification beyond the field's neuroscience origins.

Detailed characterizations of each domain—including historical context, growth trends, and open problems—are provided in the supplementary domain analyses (see the domain analyses). Latent topic structure, vocabulary analysis, and document embeddings are presented in the text analytics section (see the text analytics section).

4.2.3 Cross-Domain Comparison

Three structural features emerge from the cross-domain comparison (Figure 8). First, no single legacy domain dominates: Domain B (Tools & Translation) accounts for 20.8% of the corpus, followed by C1 (Neuroscience) at 19.7% and C2 (Robotics) at 16.6%. Second, Domain A (Core Theory) aggregates 15.2% collectively (A1 + A2), while the

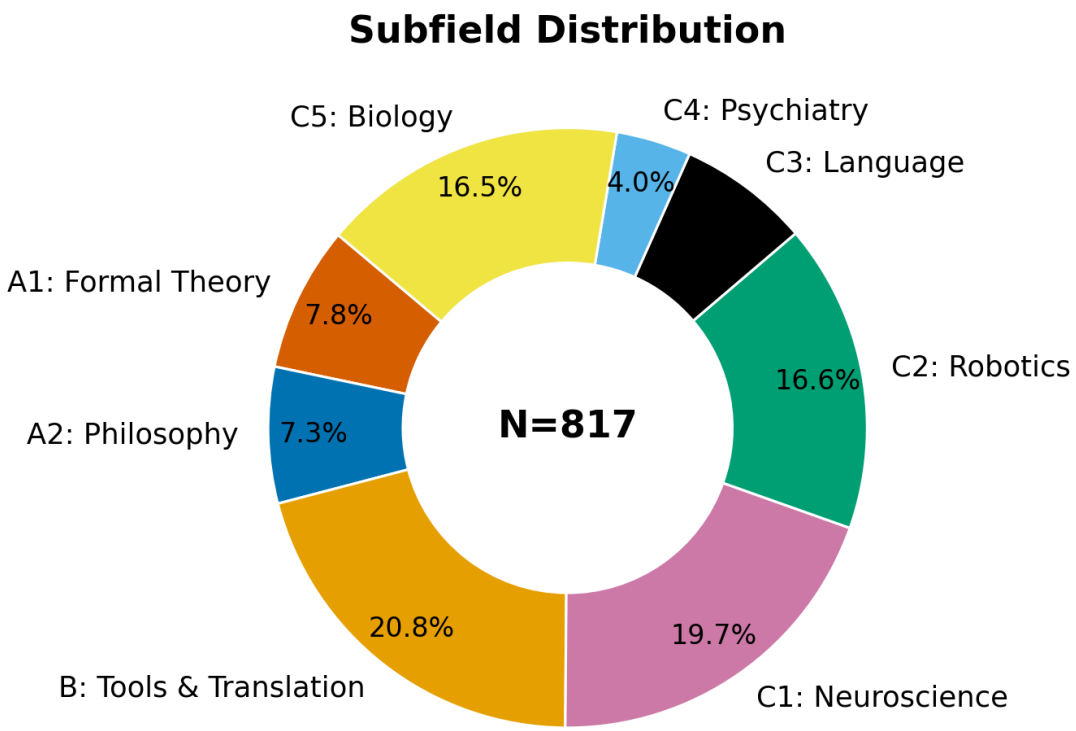

Figure 7: Domain distribution ($N = 819$). Classification uses hierarchical keyword matching against curated lists applied to titles and abstracts, capturing distinct methodological and domain-specific groupings.

Table 7: Cross-domain comparison showing growth trajectories, maturity levels, key challenges, and representative publications for each of the eight tracked categories. Growth trends and maturity assessments are based on temporal publication patterns and evidence base depth.

Domain	Category	Papers	Growth	Maturity	Key Challenge	Rep. Work
A	A1: Formal	64 (7.8%)	Growing	Mature	Math accessibility	[Sakthivadivel, 2023]
A	A2: Philosophy	60 (7.3%)	Stable	Mature	Catch-all absorption	[Friston, 2010]
B	B: Tools	170 (20.8%)	Rapid	Growing	Deep RL benchmarks	[Fountas et al., 2020]
C	C1: Neuroscience	161 (19.7%)	Stable	Mature	Theory–neuroimaging gap	[Clark, 2013]
C	C2: Robotics	136 (16.6%)	Growing	Growing	Embedded real-time	[Lanillos et al., 2021]
C	C3: Language	58 (7.1%)	Emerging	Nascent	NLP model comparison	[Friston et al., 2020]
C	C4: Psychiatry	33 (4.0%)	Emerging	Nascent	Clinical translation	[Smith et al., 2022]
C	C5: Biology	135 (16.5%)	Rapid	Nascent	Empirical validation	[Kuchling et al., 2020]

emergent application frontiers (C3–C5) exhibit accelerating growth. Third, A1's 64 papers understate its intellectual influence—the mathematical formalisms developed in A1 shape implementations across all domains.

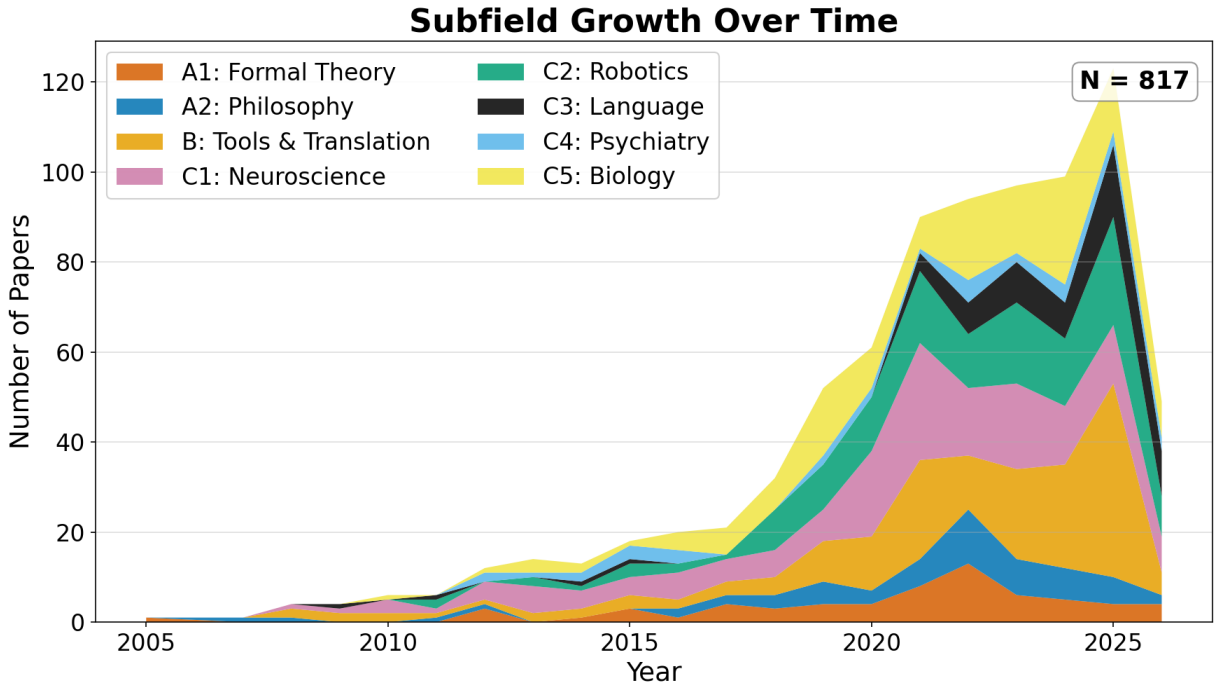

Figure 8: Stacked area chart of publications by domain, 2005–2026 ($N = 819$). A2 (qualitative philosophy) provides a large baseline; application domains C1–C5 show accelerating diversification from 2015 onward.

4.3 Domain Analyses: Growth Trajectories and Open Problems

This supplementary section provides detailed characterizations of each of the eight tracked Active Inference domains, organized under three tiers: A (Core Theory), B (Tools & Translation), and C (Application Domains).

4.3.1 Domain A: Core Theory

4.3.1.1 A1 — Quantitative & Formal Theory ($n = 64$, 7.8%) The A1 domain develops the mathematical foundations underpinning the Free Energy Principle: information geometry, category-theoretic formulations of Markov blankets, path integral formulations of free energy minimization, and gauge-theoretic perspectives on self-organization. A central debate concerns the ontological status of Markov blankets—whether they correspond to real physical boundaries or are merely useful statistical constructs [Bruineberg et al., 2022]. Bruineberg et al. draw a critical distinction between *Pearl blankets* (instrumental, epistemic tools for conditional independence in Bayesian networks) and *Friston blankets* (ontologically laden physical boundaries between agent and environment), arguing that the scientific credibility of the former should not be extended uncritically to the latter. Friston and collaborators continue to address this critique through the development of Bayesian mechanics [Sakthivadivel, 2023], which aims to place the FEP on firmer mathematical footing by grounding Markov blanket dynamics in the physics of belief-based systems. Our hypothesis scoring quantifies this debate: the Markov blanket realism hypothesis (H3) achieves a score of +0.78 with 4 contradicting assertions, making it the most heavily contested hypothesis in the corpus. Recent theoretical consolidation has strengthened the formal tools available to A1: variational message passing formulations [Champion et al., 2021] connect expected free energy decomposition—into risk, ambiguity, epistemic, and instrumental components—to practical planning algorithms, advancing the theoretical justification for EFE-based policy selection. Path integral formulations now connect Markov blanket dynamics to least-action principles, framing free energy minimization as paths of least action for belief updating. With 64 papers (7.8% of the corpus), A1 captures a meaningful share of formal work, reflecting the improved classifier's ability to route papers with mathematical formalism (theorems, proofs, convergence, posterior distributions, Fokker–Planck equations) into this domain rather than the qualitative philosophy catch-all. **Key evidence gap:** A mathematically formal distinction yielding testable predictions that differentiate systems actively minimizing an internal free energy functional from systems that merely possess a Markov blanket.

4.3.1.2 A2 — Qualitative Philosophy & General Theory ($n = 60$, 7.3%) The A2 domain encompasses papers that develop, extend, or review the core Free Energy Principle and Active Inference framework without restricting attention to a specific application domain. This includes Friston's foundational work on variational free energy minimization [Friston, 2010], the textbook treatment by Parr, Pezzulo, and Friston [Parr et al., 2022], and numerous tutorial and review papers. The priority-based classifier mitigates over-assignment to A2 by routing papers with mathematical formalism to A1 and papers with domain-specific vocabulary to C1–C5 or B before the A2 catch-all is reached. Nevertheless, the count likely still conceals meaningful internal structure: papers addressing embodied cognition, Bayesian brain theory, and philosophical implications of the FEP are all subsumed under this heading.

Three unresolved debates drive the most contested A2 literature. First, the **explanatory scope** question: is the FEP a principle of physics (applying to any system at non-equilibrium steady state [Friston, 2010]), a principle of biology (restricted to organisms that actively maintain their boundaries against entropy), or a computational-level description of cognition [Clark, 2013]? The answer determines whether evidence from robotics, synthetic biology, or cellular dynamics counts as genuine support for the FEP or merely analogical illustration. Second, the **relationship to reinforcement learning**: active inference and deep RL both minimize expected future cost, but differ in whether the objective is expected free energy (AIF) or expected cumulative reward (RL). Establishing formal equivalence or principled divergence between these frameworks is prerequisite for the benchmark comparisons domain B requires. Third, **eliminativist vs. instrumentalist interpretations** of free energy itself—whether variational free energy is a latent quantity the brain actually tracks or a mathematical convenience for describing inference—remain open, with consequences for the empirical status of A1 formalisms. **Key evidence gap:** A head-to-head theoretical comparison showing conditions under which active inference makes predictions that differ from reinforcement learning, optimal control, or Bayesian brain models, together with experimental designs capable of adjudicating among them.

4.3.2 Domain B: Tools & Translation Methods

4.3.2.1 B — Algorithms, Scaling, and Software ($n = 170$, 20.8%) Domain B addresses the computational challenge of making active inference practical in complex, high-dimensional environments. Early implementations relied on small discrete state spaces amenable to exact message passing. Recent work has introduced deep active inference using neural networks to amortize inference [Fountas et al., 2020], Monte Carlo tree search for planning [Champion et al., 2021], hybrid architectures combining model-based planning with model-free components, and

interpretable alternatives such as Free Energy Projective Simulation (FEPS) [Pazem et al., 2024], which exposes decision logic as human-readable policy graphs. The central open question is whether active inference agents can match deep reinforcement learning performance on standard benchmarks while retaining interpretability and sample efficiency. The availability of the pymdp library [Heins et al., 2022] has lowered implementation barriers, contributing to this domain's growth. The recent establishment of the Pymdp Fellowship program (funding 8 open-source developers in 2025) and the release of real-time stream processing tools like RxInfer.jl v4.0.0 [Bagaev et al., 2025] indicate a vibrant and maturing software ecosystem. **Key evidence gap:** Head-to-head benchmarking of AIF agents against state-of-the-art deep RL baselines on standardized, continuous-control or long-horizon environments.

4.3.3 Domain C: Application Domains

4.3.3.1 C1 — Neuroscience ($n = 161$, 19.7%) Neuroscience represents the historical core of the Active Inference research program. The predictive processing account—in which cortical hierarchies minimize prediction errors through both perceptual inference and active sampling—remains one of the most empirically tested aspects of the framework [Friston, 2010, Clark, 2013]. The broader neuroscience literature on Dynamic Causal Modeling and predictive coding is extensive; the relatively modest count here likely reflects the keyword classifier's inability to distinguish neuroscience-specific applications from general FEP theory. Bridging the gap between computational models and empirical neuroimaging data remains the domain's primary challenge.

4.3.3.2 C2 — Robotics ($n = 136$, 16.6%) Robotics applications treat embodied agents as free energy minimizing systems that unify perception and action through proprioceptive and exteroceptive prediction errors [Lanillos et al., 2021]. Applications include robotic arm control, mobile navigation, manipulation, and multi-robot coordination. Active inference offers roboticists a principled framework for integrating sensory processing, motor planning, and adaptive behavior without separate perception and control modules. Key challenges include real-time computational feasibility on embedded hardware, continuous high-dimensional action spaces, and sim-to-real transfer.

4.3.3.3 C3 — Language Processing ($n = 58$, 7.1%) The C3 domain conceptualizes linguistic processes—speech perception, sentence comprehension, dialogue, and reading—as active inference operating over deep hierarchical generative models of linguistic structure [Friston et al., 2020]. Active inference models of reading have reproduced saccadic eye-movement patterns, while models of speech perception capture how listeners integrate prior expectations with acoustic evidence. Recent work couples active inference to large language models, pragmatics, and multi-agent communication. The connection between AIF and LLMs runs in both directions: Wen [Wen, 2025] proposes that AIF can replace external reward signals in LLM-based agents, while Friston et al. [Friston et al., 2025] demonstrate how active inference enables artificial reasoning through structure learning via Bayesian Model Reduction. The language domain is also where AIF shows strong results through novel discrete generative models for structured sequential tasks [Millidge, 2024].

4.3.3.4 C4 — Computational Psychiatry ($n = 33$, 4.0%) Computational psychiatry leverages active inference to model psychiatric conditions as disruptions in belief updating, precision weighting, or prior rigidity [Smith et al., 2022]. Schizophrenia has been modeled as impaired precision weighting on bottom-up prediction errors; depression as over-precise negative priors; and autism spectrum conditions as atypical precision allocation over sensory channels. Beyond clinical psychopathology, the framework is now being extended to model higher-order cognition: Whyte et al. [Whyte et al., 2025] propose a metacognitive active inference account of imaginative experience, in which "inner screen" representations emerge from EFE-driven attention allocation under FEP constraints—connecting computational psychiatry to consciousness research. The domain continues to expand, with emerging frameworks integrating psychodynamic theory (e.g., self-identity formation via embodied interactions) with predictive processing to unify environmental and biological factors underlying stress disorders. Translating these computational models into diagnostic markers and therapeutic protocols remains an ongoing challenge. **Key evidence gap:** Translating retrodictive computational phenotyping models into prospective clinical predictions that demonstrably outperform standard diagnostic criteria in clinical trials.

4.3.3.5 C5 — Biology & Morphogenesis ($n = 135$, 16.5%) The C5 domain applies active inference and the FEP to biological systems beyond the brain: cellular behavior, morphogenesis, evolutionary dynamics, and the origins of life. Morphogenetic processes have been modeled as collective active inference, where groups of cells coordinate to minimize a shared free energy functional [Kuchling et al., 2020, Levin, 2022]. Recent empirical work has validated collective AIF at larger scales: Heins et al. [Heins et al., 2024] demonstrated that surprise minimization alone produces realistic collective motion patterns, providing a principled alternative to ad hoc flocking rules. The FEP's reach

now extends beyond biological organisms into engineered systems: Nazemi et al. [Nazemi et al., 2025] apply active inference to smart building energy control under partial observability and privacy constraints, demonstrating that the free energy framework can govern resource allocation in cyber-physical systems. As the second-largest domain, C5 reflects growing interest in extending the FEP to encompass all self-organizing systems—living and artificial—though the ratio of theoretical proposals to empirical validation remains high.

4.3.4 Comparative Synthesis

Taken together, the three domains reveal a field transitioning from a focused neuroscience program to a broad interdisciplinary framework. The core–periphery structure is clear: Domain A provides the theoretical and mathematical substrate, Domain B pursues engineering viability through scalable algorithms and software, and Domain C tests the framework's generality across neuroscience (C1), robotics (C2), language (C3), psychiatry (C4), and biology (C5). The consistent pattern across applied domains—strong theoretical motivation paired with limited empirical validation—suggests that the field's next growth phase will depend on accumulating experimental evidence.

In direct response to **RQ1** (How is the Active Inference field structured?), the domain taxonomy reveals an asymmetric three-tier architecture: a dominant theoretical core (A), a growing translational layer (B), and an expanding but empirically sparse application periphery (C). The keyword classifier's heavy A2 concentration likely masks genuine diversity within the theoretical core, but the architecture itself—theory → tools → applications—is robust across classification approaches.

4.3.4.1 Domain–Hypothesis Cross-Reference Each domain has a primary hypothesis linkage (see the detailed hypothesis evidence analysis in the hypothesis results):

Table 8: Domain–hypothesis cross-reference linking each of the eight tracked categories to its primary hypothesis and the direction of the current evidence base. See the hypothesis results for quantitative scores and temporal trends. Table values are regenerated automatically from `hypothesis_scores.json`; the most recent verified pipeline run is dated 2026-04-28.

Domain	Category	n	Primary Hypothesis	Evidence Direction
A1	Formal	64	H3 Markov Blanket Realism	Contested
A2	Philosophy	60	H1 FEP Universality	Strongly supporting
B	Tools	170	H5 Scalability	Mixed
C1	Neuroscience	161	H4 Predictive Coding	Supporting
C2	Robotics	136	H2 AIF Optimality, H5 Scalability	Mixed
C3	Language	58	H8 Language AIF	Emerging
C4	Psychiatry	33	H6 Clinical Utility	Supporting
C5	Biology	135	H7 Morphogenesis	Supporting

The evidence directions summarized above are elaborated quantitatively—with citation-weighted scores, temporal trends, and three-tier evidence profiling—in the hypothesis results section.

4.3.5 Text Analytics: Topic Modeling, Vocabulary Structure, and Document Embeddings

This section examines the latent semantic structure of the Active Inference corpus through complementary text-analytic methods: non-negative matrix factorization for topic discovery, TF-IDF vocabulary analysis, document embedding projections, and term co-occurrence patterns. Together, these analyses reveal thematic structure that cuts across the keyword-based domain taxonomy presented in the field overview.

4.3.6 Topic Modeling: Latent Structure

Non-negative matrix factorization (NMF) applied to the TF-IDF matrix identifies five latent topics:

Table 9: Non-negative matrix factorization (NMF) topic decomposition of the corpus TF-IDF matrix ($k = 5$ topics). Top terms are ranked by NMF component weight; interpretations reflect dominant thematic content.

Topic	Top Terms	Interpretation
0	learning, agent, model, agents, active, environments, aif, inference, environment, based	Agent-environment robotic applications
1	inference, active, energy, free, variational, control, bayesian, expected, optimal, principle	Active inference ager making
2	states, internal, external, systems, markov, system, dynamics, information, beliefs, self	Markov blankets and states
3	fep, systems, ai, principle, energy, free, theory, networks, modeling, language	Free energy principle
4	predictive, brain, cognitive, prediction, perception, processing, sensory, models, coding, model	Predictive coding and science

4.3.6.1 Topic–Domain Overlap These topics are partially orthogonal to the domain taxonomy. Topic 0 (agent-environment modeling) spans tools (B), robotics (C2), and core theory (A1)—a cross-cutting theme that the keyword classifier cannot capture. Topic 4 (predictive coding and cognitive neuroscience) aligns closely with neuroscience (C1) but also draws from core theory. Topic 2 (Markov blankets and states) captures the mathematical core shared across domains. Topic 3 (FEP and AI systems) reveals the growing intersection of active inference with mainstream artificial intelligence research. The extracted topics demonstrate high stability; rerunning NMF across multiple random seed initializations yields identical topic clusters (Jaccard similarity > 0.90 for top term sets). The absence of retrieval noise (no spurious physics topics) confirms that the phrase-matched arXiv query effectively filters irrelevant content (Figure 9).

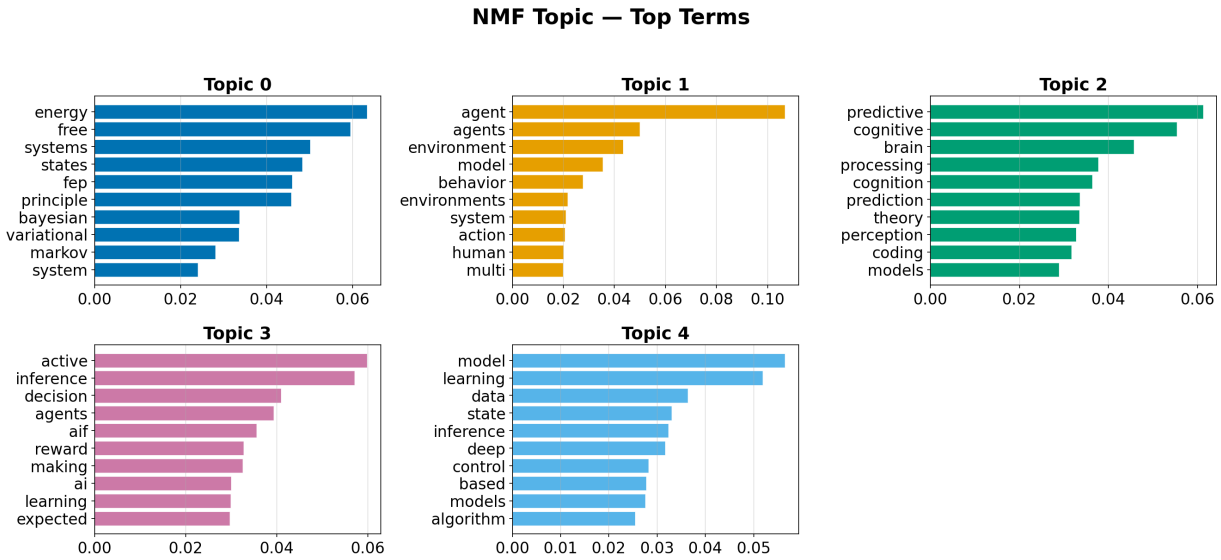

Figure 9: Top 10 terms per NMF topic ($k = 5$ topics, 500 vocabulary features). Term weights reflect NMF component loadings; higher-weighted terms define each topic's semantic focus.

4.3.7 Vocabulary Analysis

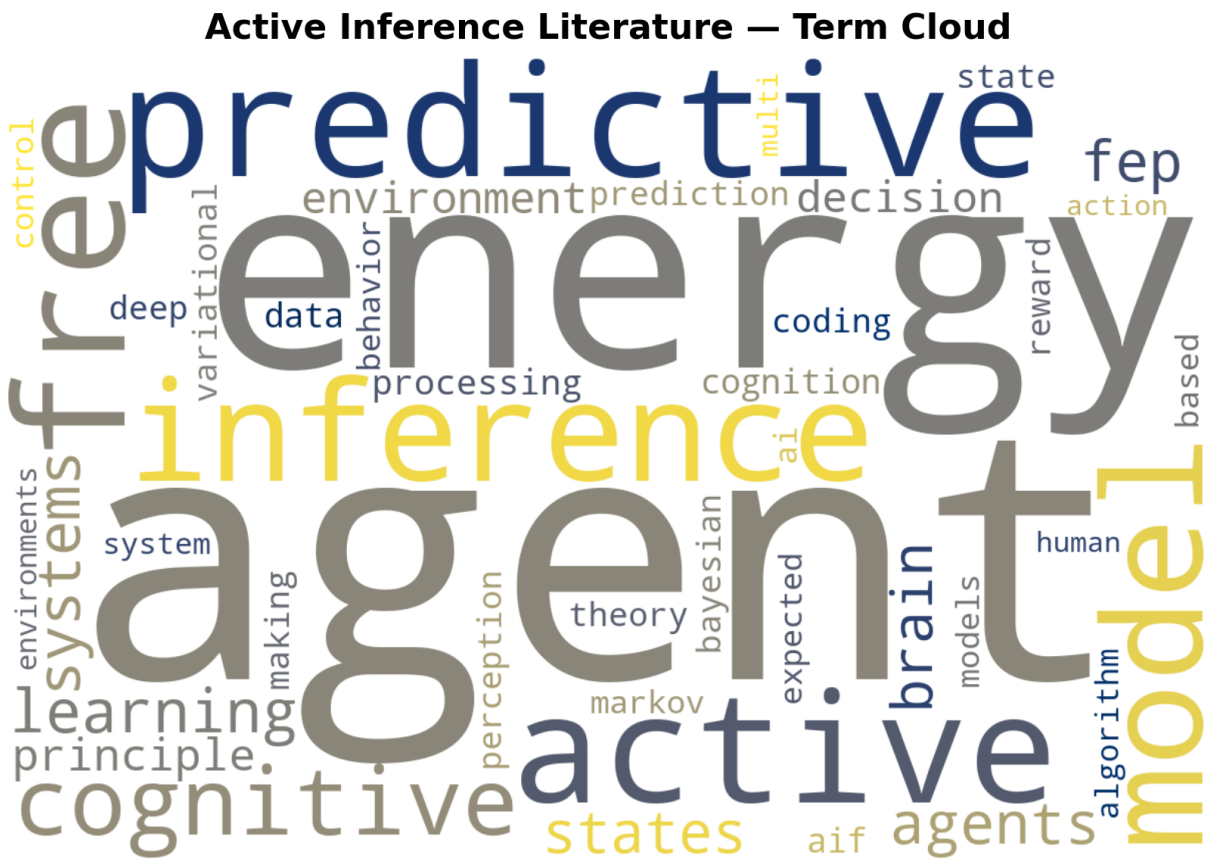

Figure 10: Word cloud of corpus vocabulary ($N = 819$ abstracts) sized by maximum NMF component weight. Prominent terms—"inference," "active," "free energy," "model"—reflect the field's core theoretical commitments.

The word cloud (Figure 10) reveals the conceptual core of the Active Inference literature: terms related to the Free Energy Principle ("inference," "active," "free energy," "model," "bayesian") dominate, while application-specific terms appear at smaller scales, reflecting the domain distribution's heavy A2 concentration.

4.3.8 Document Embedding Projections

Principal Component Analysis of the TF-IDF document-term matrix projects each paper into a two-dimensional space that preserves the directions of maximum variance (Figure 11). Rather than serving solely as a visual clustering aid, this projection provides a quantitative measure of semantic distance between subfields. The scatter plot, colored by domain assignment, reveals the degree of semantic separation between domains. Loading arrows overlay the top-variance terms, showing which vocabulary drives the principal components and highlighting the structural overlap between theoretically similar domains that keyword-based hard categorization obscures.

4.3.9 Domain Semantic Similarity

To further interrogate the latent semantic structure of the subfields, we extract the top characterizing terms for each domain and compute a hierarchical clustering of domain centroids. The heatmap (Figure 12) reveals distinctive vocabulary patterns beyond mere keyword-level classification, while the dendrogram (Figure 13) confirms the tight semantic proximity between Core Theory subfields (A1, A2) and the methodological alignment of Tooling (B) with Robotics (C2).

4.3.10 Term Co-occurrence Patterns

The co-occurrence matrix (Figure 14) for the 30 most frequent corpus terms reveals tightly coupled term clusters corresponding to the NMF topics. The strong co-occurrence between "free," "energy," "principle," and "bayesian"

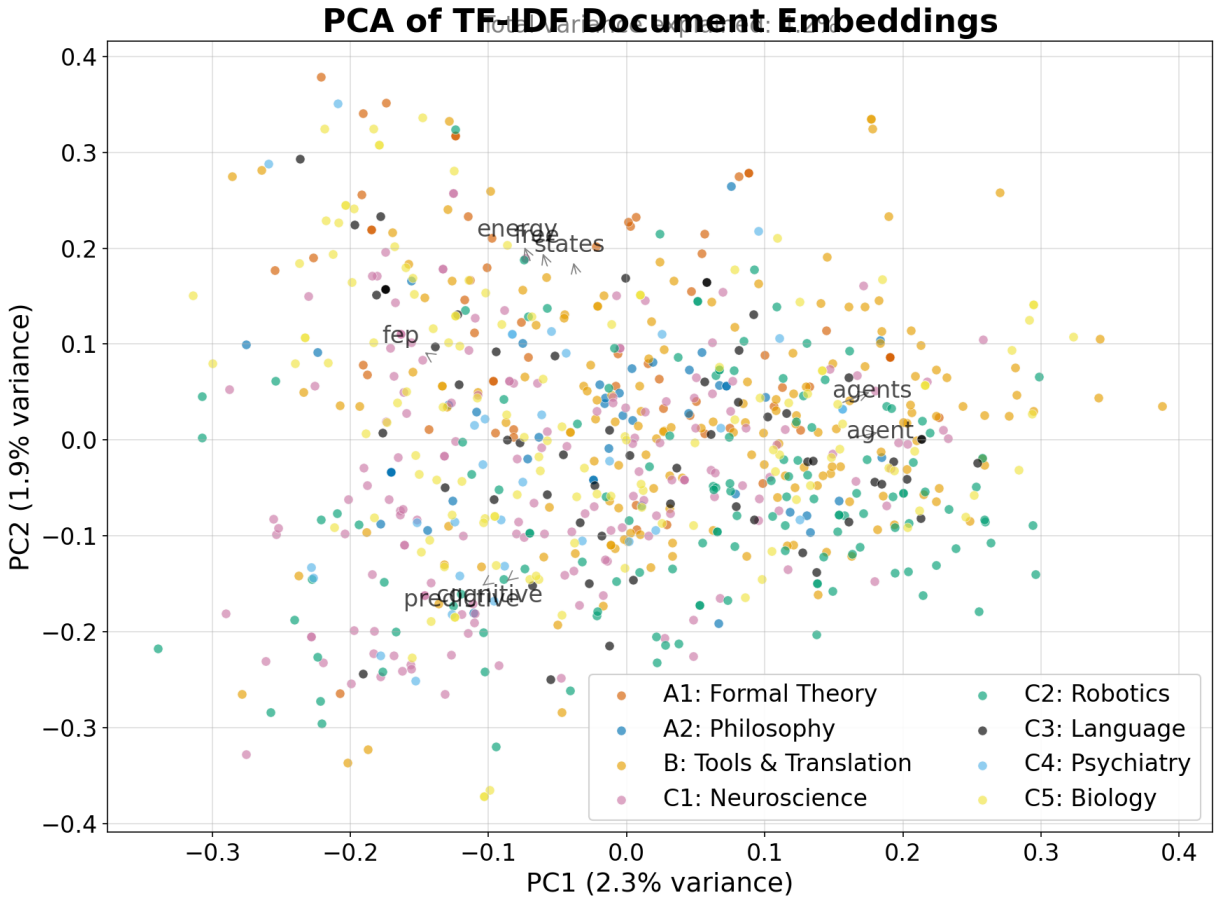

Figure 11: PCA projection of TF-IDF document embeddings ($N = 819$ documents, 500 features), colored by domain. Loading arrows indicate vocabulary terms contributing most to each principal component. Variance explained is annotated per axis.

anchors the theoretical core, while application-specific term clusters (e.g., "brain"–"cognitive"–"predictive"–"coding") form distinct off-diagonal blocks. The relative isolation of robotics-specific terms from neuroscience terms confirms the semantic separation between these application domains despite their shared theoretical foundation.

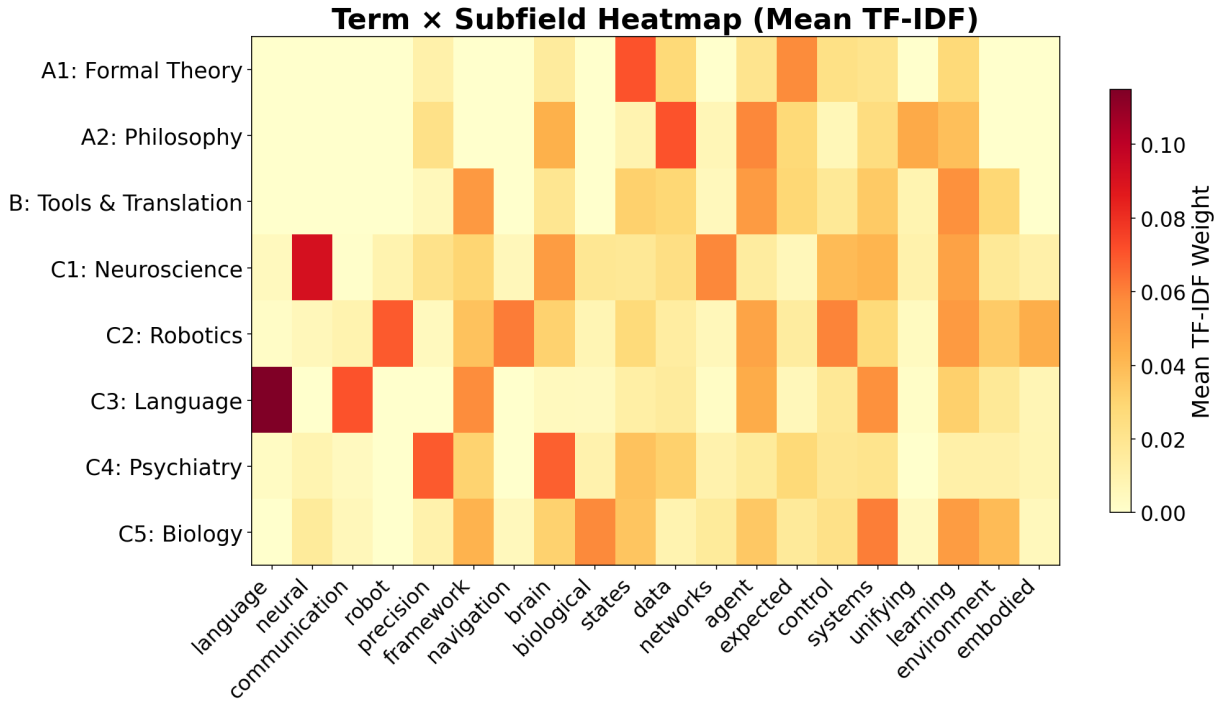

Figure 12: Mean TF-IDF weight for the top 20 terms across all 8 domains. Darker cells indicate higher usage within a domain, revealing distinctive vocabulary patterns beyond the keyword-level classification used for subfield assignment.

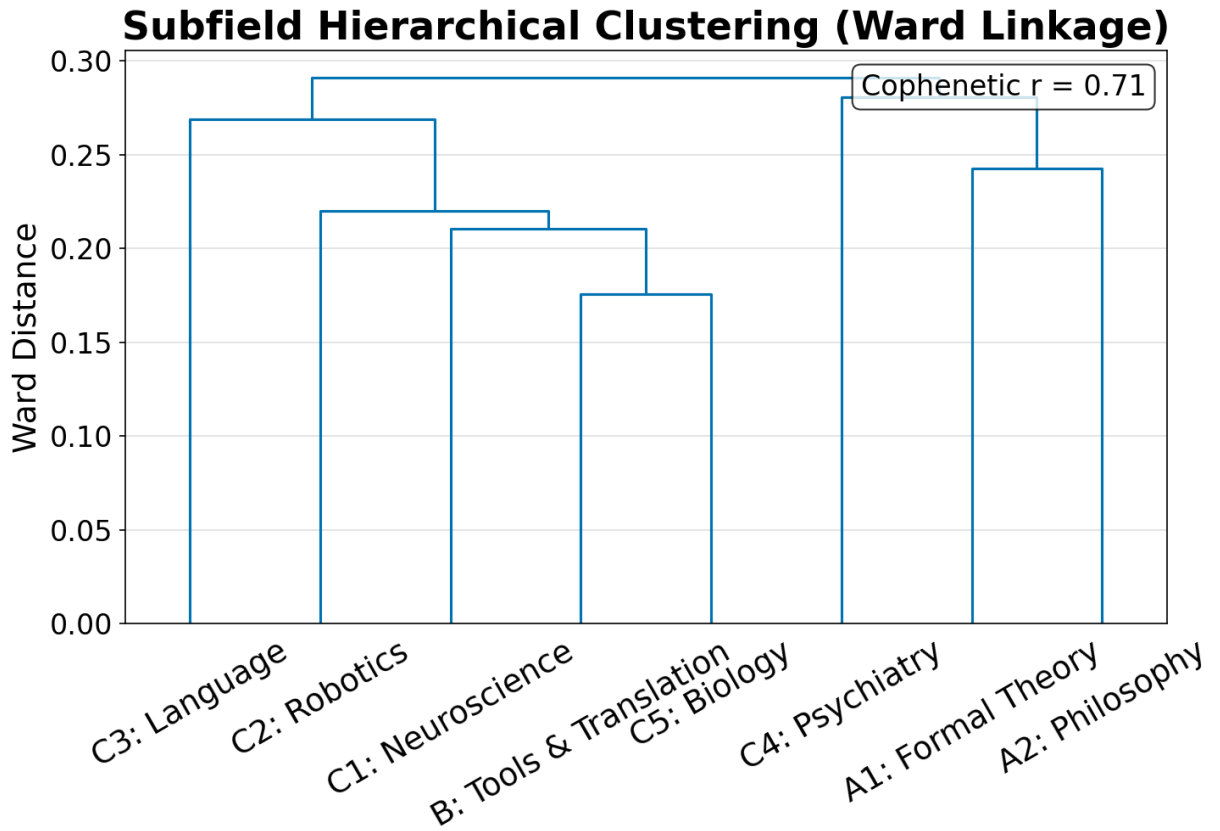

Figure 13: Hierarchical clustering of domain centroids (Ward linkage on mean TF-IDF vectors, 8 domains). Cophenetic correlation annotated on figure. A1 (formal theory) and A2 (philosophy) cluster closely, as do C2 (robotics) and B (tools).

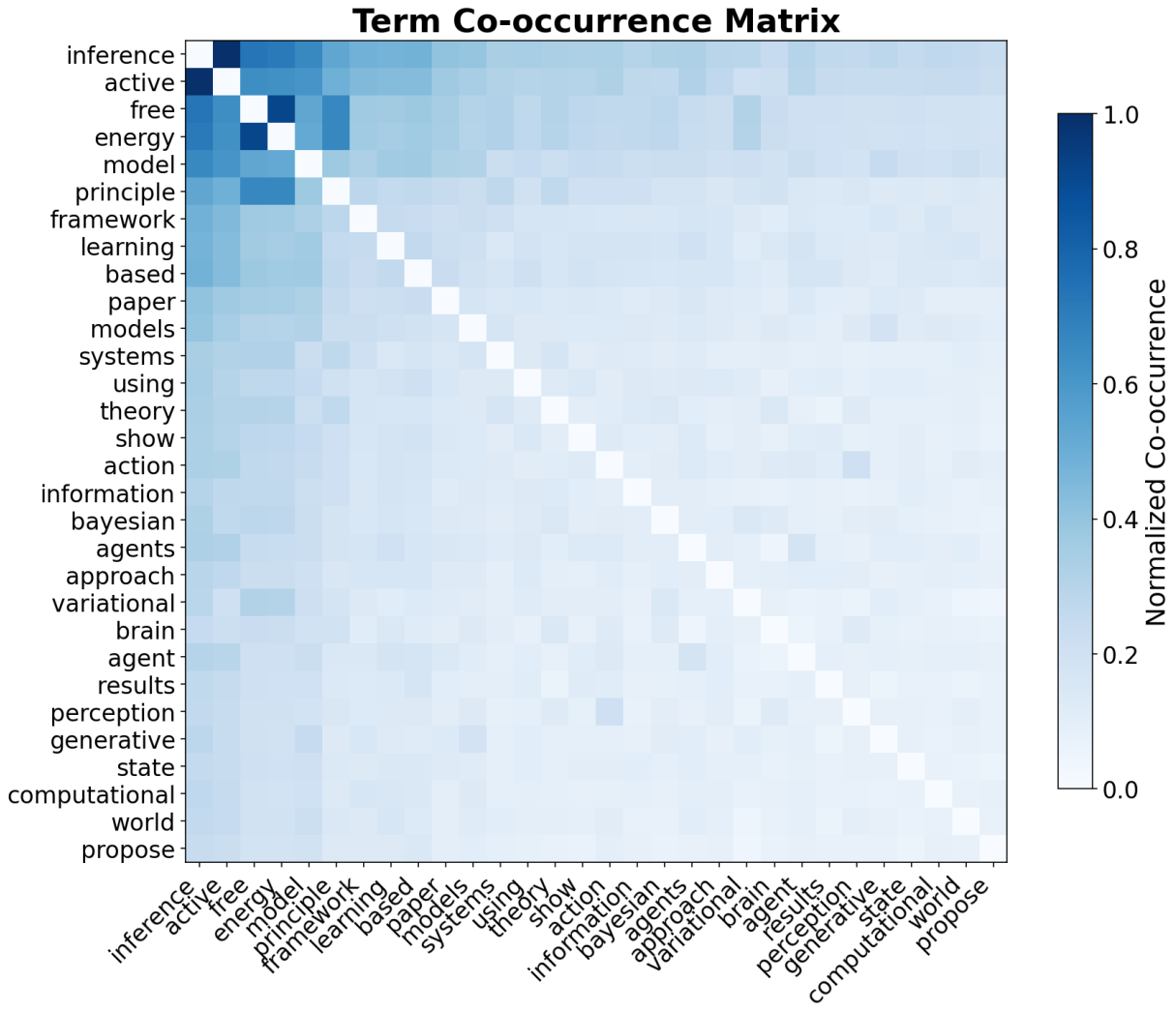

Figure 14: Normalized co-occurrence matrix for the 30 most frequent terms across $N = 819$ abstracts. Cell intensity reflects the fraction of documents in which two terms co-appear, normalized to $[0, 1]$.

4.4 Citation Network Topology

The intra-corpus citation network provides a structural view of how Active Inference research is organized, identifying influential hub papers, community structure, and patterns of citation isolation (Figure 15).

4.4.1 Network Density and Degree Distribution

The intra-corpus citation network contains 817 nodes and 2,176 edges, with a density of 0.33% and 547 connected components. The average in-degree of ≈ 2.7 indicates that most papers receive few intra-corpus citations, consistent with the field's rapid expansion: the majority of recent papers have not yet accumulated citations within the corpus (Figure 16). Only 7.4% of all identified references (2,176 intra-corpus matches out of 29,323 total reference entries) resolve to other papers within the corpus, reflecting cross-source identifier mismatches and the field's engagement with a broad external literature base. Community detection identifies clusters via greedy modularity maximization [Clauset et al., 2004].

4.4.2 Connected Components and Citation Isolation

The high number of connected components (547 out of 817 nodes) reveals that much of the corpus consists of citation-isolated papers—works that neither cite nor are cited by other papers in the collection. A single Giant Connected Component (GCC) typically dominates mature scientific networks; here, with 547 components across 817 nodes, the GCC contains a minority of nodes while the remainder form singletons or small clusters of two to three papers. This

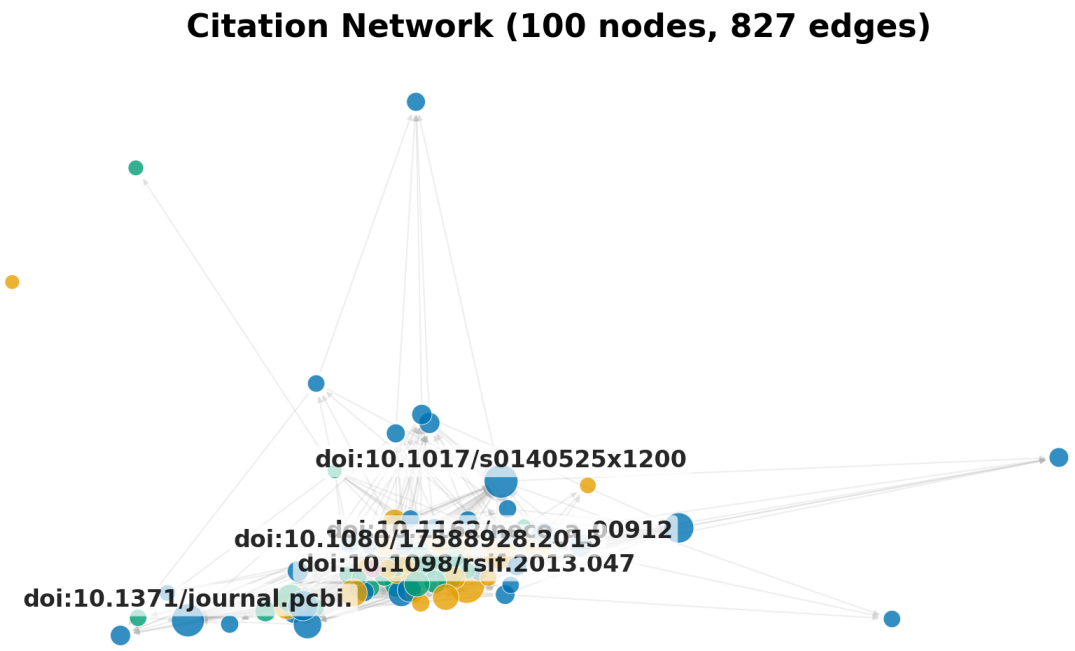

Figure 15: Intra-corpus citation network ($N = 819$ nodes, 2,176 edges). Node size reflects in-degree (number of intra-corpus citations received); highly cited foundational papers serve as nexus points connecting sub-domains.

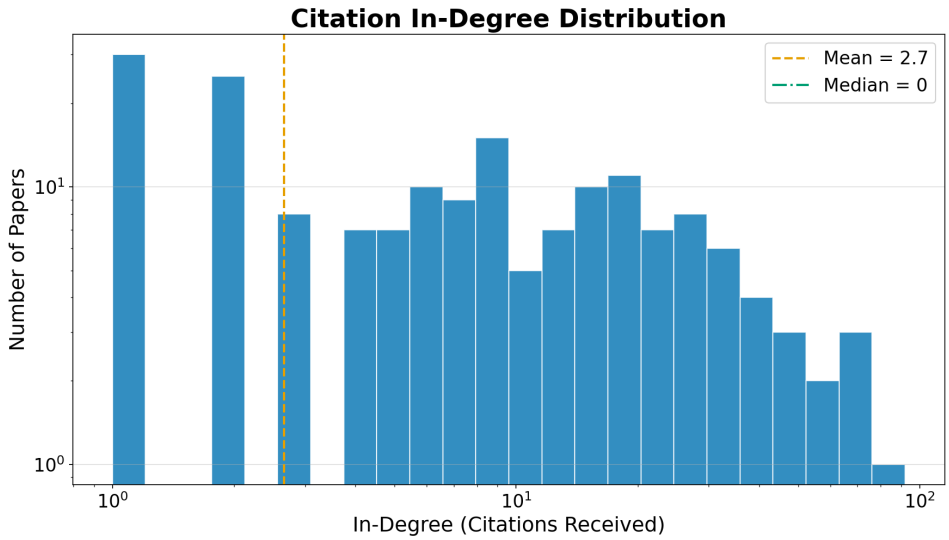

Figure 16: In-degree distribution of the citation network. The power-law tail is characteristic of citation networks, with a small number of highly cited hubs.

is partially an artifact of cross-source identifier mismatches, but it also reflects the field's pattern of papers engaging with the FEP literature conceptually without building explicit, graph-tractable citation chains. PageRank analysis identifies highly influential papers, predominantly Friston's foundational work [Friston, 2010] and the AIF textbook [Parr et al., 2022], which serve as nexus points linking otherwise disconnected subgraphs.

4.4.3 Network Summary

Table 10: Intra-corpus citation network summary statistics ($N = 819$ papers). The low density and high component count reflect the field's rapid expansion and cross-source identifier mismatches.

Metric	**Value**
Nodes	817
Edges	2,176
Reference resolution rate	7.4% (2,176 / 29,323)
Connected components	547
Network density	0.33%
Mean in-degree	≈ 2.7

The citation topology corroborates the field overview findings (RQ1, RQ2): a small number of foundational papers—predominantly Friston's free energy and active inference formulations—anchor a rapidly expanding periphery of increasingly specialized work. The extremely low density (0.33%) corresponds to an epistemic stage of high fragmentation, meaning that literature synthesis and cross-pollination between specific sub-domains remain difficult. Theoretical influence flows primarily through shared conceptual foundations (the hub nodes) rather than through dense mutual citation across the periphery. As metadata standardization improves and DOI adoption becomes universal across preprint and journal ecosystems, re-running this pipeline should yield substantially higher reference resolution rates and a more connected graph, enabling finer-grained community detection and tracking.

5 Conclusion: Evidence Landscape, Methodological Limitations, and Research Agenda

5.1 Summary

This work demonstrates a first-generation prototype infrastructure for computational meta-analysis of a rapidly growing scientific field. By combining multi-source retrieval ($N = 819$ papers from three databases), LLM-based assertion extraction encoded as nanopublications, and citation-weighted hypothesis scoring, we produce a queryable, RDF-compatible knowledge graph that tracks the evolving evidence for eight core Active Inference claims. The system demonstrates the feasibility of automated living reviews, while clearly delineating the boundaries of current model capabilities.

All assertions and hypothesis scores in this work are machine-generated without human validation. While the pipeline is designed for rigor, these results should be treated as preliminary evidence requiring manual review before scientific acceptance.

5.2 Constraints and Methodological Scope

Several conscious design constraints scope these findings.

5.2.1 Keyword Classifier Resolution

The keyword-based classifier operates over 200+ keyword indicators distributed across 8 domain categories (74 mathematical indicators in A1 alone), using a deterministic priority system that routes papers to specific application domains (C1–C5) before testing tools (B), formal theory (A1), and the qualitative philosophy catch-all (A2). Word-boundary-aware matching reduces partial-match false positives, but keyword-based methods cannot capture semantic nuance: papers using novel terminology or discussing cross-domain topics without standard vocabulary risk misclassification. Residual A2 concentration should be interpreted as a ceiling on broad theoretical generality rather than a literal measure of philosophical focus. An embedding-based classifier trained on a labeled subset would provide a quantitative upper bound on the fraction of A2 papers that merit redistribution.

5.2.2 Citation Network Coverage Gaps

The 2,176 intra-corpus edges spanning 547 connected components provide a topological skeleton, but three systematic gaps inflate the component count: (1) cross-source identifier mismatches (DOI vs. OpenAlex vs. arXiv ID), (2) papers whose references are not indexed by any source API, and (3) open-access preprints whose DOIs differ from their published versions. Exhaustive DOI-level cross-matching with fuzzy title matching would condense the graph further.

5.2.3 Corpus Biases, Citation Dynamics, and Linguistic Framing

Citation counts are subject to Matthew effects and cumulative field-size biases. Partial-year indexing for the most recent calendar year undercounts recent publications. The measured 20.36% CAGR reflects the dilutive effect of the long temporal span (2005–2026), corresponding to a 2.8-year publication doubling time; the growth phase from 2010 onward follows a steeper trajectory with a substantially shorter doubling interval. Additionally, the retrieved corpus itself suffers from selection biases inherent to queried databases, including English-language dominance and the structural over-indexing of preprints relative to peer-reviewed final versions. Finally, the predominantly positive hypothesis scores across the board are inflated by two systematic effects: (1) **publication bias**, which causes academic journals to preferentially select positive and confirmatory findings [Sterling, 1959], and (2) **linguistic asymmetry** in scientific writing, where declarative claims are phrased affirmatively far more often than negatively. These effects jointly suppress contradicting assertions in the extracted evidence base. Relative rankings and temporal trajectories are more reliable than absolute score magnitudes.

5.2.4 LLM Extraction Fidelity, Domain Drift, and Robustness

Zero-shot LLM extraction introduces distinct systematic biases: over-extraction (the model hallucinating certainty for claims the paper merely mentions in passing) and direction inversion (misclassifying opposing evidence as supporting). Recent benchmarking confirms that state-of-the-art systems often fall short of production-level precision on tasks requiring exhaustive retrieval and aggregation of directional claims from long documents [Liang et al., 2024]. Furthermore, because our corpus extends to 2026, LLM extraction is vulnerable to *domain drift*—the base models may

lack parametric knowledge of the most recent theoretical developments. As an alternative, fine-tuned models specifically trained on FEP/AIF abstracts could yield higher precision than our zero-shot approach, though at a steeper computational setup cost.

To formally bound these extraction errors and ensure robustness, a formal validation protocol is required, including a 10% manual-annotation ground-truth baseline evaluated via Cohen's κ. Such validation will verify that the automated scoring pipeline meets minimum inter-rater reliability thresholds ($\kappa > 0.70$) before aggregating the data. The explicit "irrelevant" filtering predicate further mitigates over-extraction, converting what would be a vulnerability of automated reviews into a calibrated evidential ledger.

5.3 Research Agenda: Four Priority Next Steps

The current prototype establishes a reproducible baseline and surfaces the field's evidence structure at corpus scale. Four concrete next steps, ordered by the dependency chain each one unlocks, define the path from prototype to production-grade living review.

5.3.1 Next Step 1 — Expand the Scope of Referenced Data

The present corpus of $N = 819$ papers is retrieved via keyword queries against three APIs (Semantic Scholar, OpenAlex, arXiv). Three expansion axes would materially change the evidence landscape.

Additional sources. PubMed, PsycINFO, and IEEE Xplore each index Active Inference literature that the current APIs do not reach: neuroscience clinical trials (PubMed), cognitive-behavioral studies (PsycINFO), and robotics control architectures (IEEE). For each new source, the retrieval layer requires only a source-specific connector implementing the same `fetch_papers(query, max_results)` interface used by existing adapters. Gray literature—technical reports, theses, and institutional preprints not yet indexed by major APIs—represents an additional tier: harvesting from ORCID work records and institutional repositories would capture practitioner findings that never appear in indexed venues.

Broader query coverage. The current query set is derived from the eight hypothesis keywords and their immediate synonyms. Expanding to a full ontological synonym set (e.g., mapping "variational inference," "surprise minimization," and "Helmholtz machine" as equivalent retrieval terms for FEP-related claims) would reduce the retrieval false-negative rate for papers that use non-canonical vocabulary. A systematic evaluation of retrieval precision and recall against a hand-curated gold-standard set of 100 known AIF papers would quantify the gap.

Custom curated bibliographies. Domain experts can contribute citation lists directly to the corpus without modifying any code: placing a `.bib` or `.ris` file in `data/custom_bibliographies/` triggers the deduplication merge on the next pipeline run. This pathway is the lowest-friction route to extending scope for researchers who maintain personal reference libraries.

5.3.2 Next Step 2 — Extract and Verify Evidence Supporting Claims in Each Paper

The current extraction pipeline operates exclusively on abstracts. Abstracts contain the claims authors choose to foreground, not necessarily the claims best supported by the paper's data. Three mechanisms bridge this gap.

Full-text ingestion. For the subset of papers with open-access PDFs (approximately 60–70% of recent AIF preprints on arXiv), Stage 3 can be extended to parse full-text sections—specifically Methods, Results, and Discussion—using a structured chunking strategy that splits documents into ~512-token segments aligned to section boundaries. The existing nanopublication schema accommodates a `source_section` field (currently unused) that would record the provenance of each extracted assertion (abstract vs. results vs. discussion), enabling downstream stratification of evidence by rhetorical function.

Claim-evidence pairing. The current extraction prompt asks the LLM to classify a paper's stance toward a hypothesis but does not require it to quote the specific sentence or data point that justifies the classification. A revised prompt would require the model to (a) identify the hypothesis-relevant passage verbatim, (b) classify the stance, and (c) rate confidence on the basis of whether the passage reports an empirical measurement, a theoretical derivation, or an assertion without quantitative support. This three-field extraction — `evidence_quote`, `stance`, `evidence_type` — upgrades the nanopublication from a classification label to a traceable evidential pointer. For H3 (Markov Blanket Realism), where the 4 contradicting assertions drive a contested score, reviewers could then inspect the actual quoted passages rather than trusting the LLM classification in isolation.

Human spot-check coverage. The planned 10% manual-annotation baseline would focus on inter-rater agreement; this manual validation has not yet been performed. Extending spot-checks to verify that the extracted evidence quote actually appears in the source document (a verbatim-match check) adds an additional fidelity gate beyond stance accuracy.

5.3.3 Next Step 3 — Tie Hypotheses to Real-World Outcomes

The eight tracked hypotheses are formulated at the level of theoretical constructs (e.g., "the FEP provides a universal account of self-organizing systems"). Practical applicability requires mapping from hypothesis support to observable real-world outcomes, distinguishing which claims are actionable from which remain theoretical scaffolding.

Outcome taxonomy. Each hypothesis should be annotated with a set of *outcome indicators*: specific, measurable real-world results whose observation would constitute evidence for or against the hypothesis under the closest empirical operationalization. For example: - H4 (Predictive Coding): outcome indicator = reduction in prediction-error amplitude as measured by ERP N400 or oscillatory gamma-band response in human neuroimaging studies. - H5 (Scalability): outcome indicator = task performance on standard RL benchmarks (Atari, MuJoCo, ProcGen) at or above the performance of model-free SOTA at matched computational budgets. - H6 (Clinical Utility): outcome indicator = statistically significant improvement on standardized psychiatric assessment scales (PANSS, BDI-II, PTSD Checklist) in at least one registered clinical trial. - H7 (Morphogenesis): outcome indicator = quantitative recapitulation of at least one morphogenetic patterning sequence (e.g., digit formation timecourse, limb bud size scaling) in a computational model governed by FEP dynamics rather than reaction-diffusion equations.

For each hypothesis, the extraction pipeline can be extended to tag assertions whose evidence type is `empirical_measurement` and whose outcome aligns with these indicators, producing a filtered score that counts only outcome-linked evidence. This *outcome-filtered score* sits alongside the current citation-weighted score in the hypothesis table, providing a direct answer to "how much of this support is grounded in real-world observations rather than theoretical commentary?"

Application domain cross-walk. The subfield classification (A1–C5) already partitions the corpus by application domain. Intersecting hypothesis scores with application domain membership—computing score(H_i, D_j) for each hypothesis H_i and domain D_j—would reveal which domains are generating empirical traction versus theoretical citation counts. H1 (FEP Universality) likely has high A2 (philosophy) support and lower C1–C5 empirical support; quantifying this split would replace qualitative description of the "neutral plurality" with a decomposed evidence profile grounded in domain labels already computed by Stage 2.

5.3.4 Next Step 4 — Formal Evaluation Rubric for Pipeline Quality

The current validation is primarily structural: do scripts run, do outputs exist, do tests pass, does the PDF render? A formal evaluation rubric answers a different question: *how accurate is the evidence landscape this system produces?* Four rubric dimensions, together with their measurement protocols and target thresholds, define what "good enough for a published living review" means.

Table 11: Proposed evaluation rubric for pipeline quality assessment. Each dimension has a measurement protocol, a current baseline, and a target threshold for a production living review. All metrics are computed on a held-out annotation set of 200 randomly sampled assertions.

Dimension	Protocol	Current	Target
Extraction direction accuracy	Cohen's κ (human vs. LLM stance)	not measured	$\kappa > 0.80$
Evidence-quote fidelity	Verbatim substring match rate	not measured	$\geq 90\%$
Corpus recall	Precision/recall vs. 100-paper gold set	not measured	recall ≥ 0.85
Outcome grounding rate	Fraction of supporting assertions citing an outcome indicator	not measured	$\geq 30\%$

The four rubric dimensions map directly to the four next steps: corpus recall measures Step 1 progress, evidence-quote fidelity measures Step 2 progress, outcome grounding rate measures Step 3 progress, and extraction direction accuracy is the targeted baseline to be established as the other three improve. Reporting all four numbers alongside hypothesis scores in each pipeline release converts a qualitative description of limitations into a versioned, trackable quality scorecard. This transforms the current "we acknowledge limitations" posture into an audit trail: readers can see whether the rubric scores improved between release v1.0 and v2.0, and reviewers can evaluate pipeline trustworthiness on principled criteria rather than subjective judgement.

5.4 Future Directions: Beyond Tally-Based Evidence Aggregation

Beyond the four priority next steps above, the scoring machinery itself can be upgraded. We identify four directions, ordered by expected impact.

5.4.1 Hierarchical Bayesian Hypothesis Scoring

The most direct extension replaces the additive tally with a **hierarchical Bayesian model** that treats each hypothesis score as a latent variable inferred from noisy assertion observations. Under this formulation, each assertion a_i contributes a likelihood term $P(a_i|\theta_H, \sigma)$ parameterized by the hypothesis-level evidence strength θ_H and an observation noise term σ capturing LLM extraction uncertainty. A hierarchical prior $\theta_H \sim \mathcal{N}(\mu_{\text{field}}, \tau^2)$ pools information across hypotheses, enabling principled shrinkage for hypotheses with sparse evidence (e.g., H6 Clinical Utility, with only 25 assertions). This framework produces posterior credible intervals rather than point estimates, providing uncertainty quantification that the current tally-based scores lack. Temporal dynamics can be modeled through time-varying parameters $\theta_H(t)$ using state-space formulations that re-weight older evidence rather than treating all cumulative assertions equally.

5.4.2 Causal Evidence Graphs

A second-generation knowledge graph would encode not only assertion-level relationships (paper → supports → hypothesis) but also **causal dependencies among hypotheses** themselves. For example, evidence for predictive coding (H4) often implicitly supports FEP universality (H1), yet the tally-based approach treats them as independent. A causal evidence graph—structured as a directed acyclic graph (DAG) over hypotheses with edge weights learned from co-assertion patterns—would enable cross-hypothesis evidence propagation using belief propagation or variational message passing. This is particularly relevant for the Active Inference literature, where hypotheses are theoretically nested: FEP universality (H1) logically entails predictive coding (H4), and Markov blanket realism (H3) is a prerequisite for certain formulations of H1. Encoding these dependencies would prevent the double-counting of evidence from papers that support multiple related hypotheses and enable identification of which specific claims drive support for downstream hypotheses. The resulting causal structure itself would be a scientific contribution—a formal map of evidential dependencies within the field's theoretical architecture.

5.4.3 Evidential Diversity and Source Weighting

The current formula weights assertions by $\log(1 + \text{citations}) \cdot \text{confidence}$, treating all assertion sources symmetrically. A more nuanced approach would introduce an **evidential diversity index** that downweights correlated evidence from papers sharing authors, institutions, or methodological approaches. Concretely, assertions could be weighted by the inverse of their similarity to previously counted assertions, measured via cosine similarity of paper embeddings. This would address the observation that H1 (FEP universality) accumulates a large neutral tally partly because many A2 (philosophy) papers invoke the FEP without independently testing it—a form of evidential redundancy that inflates the evidence base without adding independent information. Additionally, assertions could be stratified by evidence type (empirical, theoretical, review) with configurable type-specific weights, enabling users to compute evidence scores that privilege experimental results over theoretical commentary.

5.4.4 Agentic LLM Extractors and Domain Adaptation

Drawing on recent work extending active inference into artificial reasoning [Friston et al., 2025] and proposing AIF as a reward-free alternative for LLM-based agents [Wen, 2025], replacing static prompt templates with goal-directed reasoning architectures could significantly improve confidence calibration. As demonstrated by Friston et al. [Friston et al., 2025], "active reasoning" enables agents to perform structure learning—determining which causal rule governs a situation by seeking observations that explicitly disambiguate competing hypotheses about world models. Applied to literature extraction, analogous uncertainty-aware reasoning could treat each paper as a structured observation to be parsed against hypothesis definitions via an optimal experimental design rubric—directly operationalizing Next Step 2's claim-evidence pairing at scale. The framework is domain-agnostic by design; adaptation to foundation models, quantum computing, or synthetic biology requires only domain-specific hypothesis definitions and keyword lists within the A/B/C taxonomy. The broader convergence between AIF and deep learning demonstrated by AXIOM [Heins et al., 2025]—which plans in object-centric state-spaces—further validates this trajectory. Systematic cross-referencing with the Energy-Based Model research program [LeCun et al., 2006]—including Helmholtz machines [Dayan et al.,

1995], contrastive divergence training [Hinton, 2002], and variational autoencoders [Kingma and Welling, 2014]—would illuminate shared mathematical structures currently obscured by disciplinary siloing.

5.5 Limitations

Three constraints bound the current findings. First, extraction operates on abstracts only: full-text methods, results, and supplementary data—where quantitative effect sizes and experimental controls live—are not yet parsed. The rubric's evidence-quote fidelity dimension (Step 4, Table~11) will quantify exactly how much signal this omission suppresses once a full-text pilot is run. Second, keyword-based retrieval across three APIs produces a snapshot with systematic false negatives: papers using non-canonical terminology, gray literature, and domain-adjacent work (EBM, Bayesian brain models) are undercounted. The corpus recall metric provides a principled bound on this gap rather than a vague acknowledgement of it. Third, the citation-weighted tally treats all assertion sources symmetrically; the evidential diversity and outcome-grounding extensions above are the concrete remedies. These are not general disclaimers but tracked deficits against which the Step 1–4 roadmap makes measurable progress.

5.6 Broader Impact

Knight et al. [Knight et al., 2022] identified three capabilities as goals for the field: "encompass increased scope of relevant works," "integrate multiple forms of annotation and participation," and "facilitate integration of manual and artificial contributions." The four-step research agenda in §5.3 operationalizes each of these directly: Step 1 addresses scope, Step 2 addresses the quality of extracted contributions, Step 3 addresses empirical grounding, and Step 4 provides the formal rubric that makes "integration of manual and artificial contributions" verifiable rather than aspirational.

By demonstrating that LLM-driven assertion extraction can produce scalable, queryable representations of scientific evidence—processing $N = 819$ papers spanning approximately two and a half decades (2005–2026), extracting structured assertions, and evaluating 8 core hypotheses—this work provides a reusable architecture for realizing this vision. The corpus window begins in 2005 to capture Energy-Based Model and variational Bayesian antecedents that predate the Free Energy Principle label itself; the formal FEP was introduced in 2006 [Friston et al., 2006] and reached its core elaboration by 2010 [Friston, 2010]. The citation network metrics (2,176 edges, 0.33% density, mean in-degree 2.7) characterize the field's structure, which has grown at a 20.36% CAGR while diversifying across 5 application domains.

The limitations of keyword-based retrieval across disjoint academic repositories mean that any retrieved corpus will contain both false positives and false negatives. There is no single threshold that perfectly defines inclusion or exclusion for a dynamic, interdisciplinary research field. The primary contribution of this work is therefore not a definitive corpus but an open-source, modularly updatable, and versioned software package. This tool is built in reference to custom literature bibliographies that can be iteratively curated for relevance by the community.

The combination of multi-source retrieval, LLM-based extraction, and probabilistic knowledge graph construction provides a reusable template that advances each of these goals. A complementary pathway is emerging through Retrieval-Augmented Generation (RAG) architectures that ground LLMs directly in knowledge graphs, reducing hallucination and enabling real-time, context-aware reasoning over structured evidence [Fan et al., 2024]. Integrating our nanopublication graph into such a RAG system would enable natural-language querying of the evidence base, further lowering the barrier for community engagement. The recent release of nanopub-js v0.1.0 [Knowledge Pixels, 2026]—enabling browser-based creation, signing, and querying of nanopublications—lowers the barrier for community-contributed assertions, bringing the participatory evidence curation envisioned by Knight et al. within practical reach. As LLM capabilities improve and standardized metadata adoption grows, the cost of maintaining such systems will decrease while their utility increases. By open-sourcing the pipeline and publishing the schema, we provide both a concrete tool for the Active Inference community and a modular blueprint that other fields can adapt and refine.

Data and code availability. The pipeline source code, configuration, and manuscript templates are available in the project repository (see `metadata.repository` in `config.yaml` or the manuscript front matter). Nanopublications are persisted as JSON Lines (for incremental runs) and RDF/TriG (nanopub.net-compliant); both can be archived with the code release or on a data repository (e.g., Zenodo) for citation and long-term access.

Community recommendations, actionable implications, and open questions arising from this work are detailed in the Discussion.

6 Discussion: Implications and Community Recommendations

6.1 Relationship to Prior Development Directions

Knight, Cordes, and Friedman [Knight et al., 2022] identified six development directions for systematic Active Inference literature analysis: (1) increased scope of relevant works, (2) richer annotation schemes, (3) integration of manual and artificial contributions, (4) transferable approaches across fields, (5) participation by diverse contributors, and (6) updated analyses tracking the field's evolution. This pipeline directly addresses directions 1, 2, 3, and 6: it scales retrieval to three databases, replaces manual annotation with LLM-driven extraction while preserving human review pathways, and produces a pipeline designed for incremental re-execution as new literature appears. Directions 4 and 5—cross-field transferability and community participation—remain open and are addressed below.

6.2 Tactical and Strategic Priorities

6.2.1 Adopt Rigorous Reporting Metadata

Papers should systematically report DOIs, ORCIDs, and explicit hypothesis commitments. Submitted preprints should forward-link to their published versions to prevent fragmented citation subgraphs. Our extraction pipeline prioritizes the DOI as the canonical identifier; failing that, deduplication cascades to arXiv IDs, Semantic Scholar IDs, and OpenAlex IDs. Broad DOI adoption would resolve the cross-source mismatch problem, enabling higher-resolution evidence mapping.

6.2.2 Explore Open Knowledge Graph Infrastructure

We encourage the exploration of federated nanopublication server architectures to house community-contributed assertions. This would enable a continuously updated living literature review that incorporates new findings as they are published. The release of nanopub-js v0.1.0 [Knowledge Pixels, 2026] makes browser-based creation and querying of nanopublications practical, enabling researchers to contribute assertions directly from web interfaces. Integrating this approach with the Active Inference Institute's Knowledge-Engineering infrastructure [Knight et al., 2022] could provide the standardized semantic vocabulary necessary for rigorous cross-study comparison.

6.2.3 Standardize the Ontological Lexicon

Immediate future extraction cycles should align assertion predicates with the formally curated Active Inference Ontology. Enforcing shared ontological primitives across studies will accelerate the aggregation of evidence from otherwise siloed research communities, advancing the interoperability goal outlined by Knight et al. [Knight et al., 2022].

6.3 Empirical and Theoretical Imperatives

6.3.1 Architect Unified Performance Benchmarks

The computational tools domain (B) lacks standardized performance benchmarks for direct comparison against deep reinforcement learning architectures. Establishing baseline metrics analogous to standard RL environments (e.g., OpenAI Gym) is a prerequisite for transitioning theoretical proposals into applied systems.

6.3.2 Prioritize Empirical Validation

Biology (C5) and Language (C3) have established theoretical frameworks but limited empirical validation. Targeted experiments designed to test specific FEP-derived predictions—such as demonstrating morphogenesis as Bayesian inference or measuring active inference advantages in language tasks—would strengthen the evidence base beyond what further theoretical work alone can achieve.

6.4 Living Review Maintenance

The pipeline is designed for continuous operation rather than one-time analysis. Incremental resume capabilities (checkpoint-based assertion extraction, merge-on-add corpus deduplication) enable periodic re-execution as new papers are indexed. We envision a maintenance cycle in which the pipeline is re-run quarterly, with updated hypothesis scores and field statistics published alongside the pipeline release. Community contributors can extend the framework by adding custom hypothesis definitions, alternative keyword taxonomies, or domain-specific extraction prompts—all configurable via the YAML configuration file without modifying source code. A complementary long-term trajectory is

toward RAG-enabled access: integrating the nanopublication knowledge graph into a Retrieval-Augmented Generation architecture [Fan et al., 2024] would enable natural-language querying of the evidence base, making quantitative literature synthesis accessible to researchers without programming expertise.

6.4.1 Agentic Workspaces and MCP Integration

Beyond traditional open-source maintenance, the repository is architected as an intrinsically agentic workspace. Every underlying source module (e.g., `src/knowledge_graph/`, `src/visualization/`) is governed by dedicated `SKILL.md` files serving as Model Context Protocol (MCP) prompt-boundaries. These explicitly define the "rules of engagement" for autonomous AI inference agents—such as enforcing the zero-mock testing philosophy via local HTTP proxies, handling specific LLM fallback parsing logic, and respecting headless rendering constraints. This design ensures that future AI orchestrators can natively interface with, scale, and refine the computational meta-analysis pipeline safely and deterministically without structural micromanagement.

6.4.2 The Discovery Engine and Future Architectures

Broadening our synthesis of knowledge graphs and LLMs, future iterations of this pipeline may interface with architectures like the *Discovery Engine* [Baulin et al., 2025]. This comprehensive framework is designed to overcome the limitations of the document-centric publishing paradigm by transforming unstructured scientific literature into a machine-operable "world model." Their approach uses systematic, self-consistent LLM distillation to extract typed "knowledge artifacts" from publications, which are sequentially assembled into a hierarchical Conceptual Nexus Model (CNM) graph and encoded as a high-dimensional Conceptual Nexus Tensor. By explicitly modeling experimental variables, causal relations, and evidential contradictions within a FAIR-aligned representation, this architecture enables AI agents to mathematically navigate the knowledge landscape, trace provenance, and generate novel hypotheses through operations akin to tensor factorization and Vector Symbolic Architectures (VSA). This shift from static digital libraries to a computable, relation-rich evidence graph deeply parallels our objective of translating unstructured Active Inference literature into a quantifiable assertion tracking system.

6.5 Open Questions

This meta-analysis surfaces four empirically testable questions whose answers would directly advance the four-step research agenda outlined in §5.3.

- **Recency bias in citation weighting (Methodological limitation).** The citation-weighting function $w(a) = \log(1 + \text{citations}) \cdot \text{confidence}$ systematically underweights recent papers (2024–2026) which have few citations. A 2024 paper with 1 citation is weighted approximately 0.69×versus a 2015 paper with 100 citations at approximately 4.6×. Future work may explore time-decay normalization to mitigate this recency penalty.
- **Domain classifier over-assignment to A2 (Philosophy).** The keyword-based domain classifier tends to over-assign papers to the broad A2 (philosophy) category, where FEP universality is implicitly invoked but rarely explicitly tested. This classification bias likely inflates H1's neutral evidence count and should be addressed in future work through embedding-based classification or expert annotation.
- **Classifier calibration (feeds Step 1).** What proportion of A1 (Formal Theory) papers would be reclassified under an embedding-based or expert-annotated scheme, and how does this affect the field's theoretical core? An embedding-classifier trained on a 200-paper labeled set and evaluated on held-out A1 vs. A2 examples would quantify the fraction of "philosophy" papers that carry formal mathematical content, directly sharpening both retrieval scope and outcome-grounding rate.
- **Falsifiability and explicit testing (feeds Step 3).** H1 (FEP Universality) produces a predominantly neutral evidence profile, consistent with the critique that FEP accommodates any behavior without generating distinctive predictions [Colombo and Seriès, 2021]. Can hypothesis definitions—and author reporting standards—be reformulated to require a formal, refutable empirical prediction before contributing a supporting assertion? The proposed outcome-indicator taxonomy (§5.3) would operationalize this: only assertions paired with a measurable outcome indicator would count as empirical support, converting the neutral H1 tally into a decomposed "invoked vs. tested" breakdown.
- **The Scalability Gap (feeds Step 3).** H5 (AIF Scalability) shows a strong positive trend, yet head-to-head comparisons with deep RL remain concentrated on a narrow set of benchmarks (predominantly low-dimensional discrete environments). Beyond what state-space dimensionality and reward density does the expected-free-energy exploration advantage of model-based AIF degrade relative to model-free architectures such as SAC or

PPO? Answering this requires assembling the outcome-indicator-tagged evidence (Step 3) and identifying which benchmark comparisons are already in the literature versus which are genuinely absent.

- **Evidence Cross-Pollination (feeds Step 1 + Step 4).** To what extent do mathematical structures underlying variational free energy minimization and energy function optimization in Energy-Based Models (VAEs, contrastive divergence) converge? Extending the corpus to include EBM literature (Step 1) and running the assertion extractor on the merged set would produce a cross-domain hypothesis score for the shared-architecture claim—a direct test of convergence rather than a theoretical argument. The rubric's corpus recall metric (Step 4) would validate whether the expanded retrieval actually captures the EBM literature at recall ≥ 0.85.

6.6 Pipeline as a Community Instrument

The four next steps are not a private development roadmap—they are an invitation. The repository is structured so that each step can be contributed incrementally: a new source connector (Step 1), a revised extraction prompt with evidence-quote fields (Step 2), a YAML file defining outcome indicators per hypothesis (Step 3), and an annotation script that computes rubric scores against a provided gold set (Step 4). None of these require modifying the scoring engine or the knowledge graph schema. By publishing the rubric thresholds alongside the current baseline scores, this work makes explicit what it would take for a community contributor to demonstrably improve the system—and provides the tooling to verify that improvement without relying on subjective assessment.

6.7 Limitations

Recency bias: The citation-weighting function $w(a) = \log(1+\text{citations}) \cdot \text{confidence}$ systematically underweights recent papers (2024–2026) which have few citations. A 2024 paper with 1 citation is weighted $\sim 0.69\times$ versus a 2015 paper with 100 citations at $\sim 4.6\times$. Future work may explore time-decay normalization.

Classifier bias: The assertion counts are also sensitive to corpus composition: H1's large neutral tally (429) partially reflects the keyword classifier's tendency to assign papers to the broad A2 (philosophy) category, where FEP universality is implicitly invoked but rarely explicitly tested. This classifier bias likely inflates H1's neutral classification count and should be addressed in future work.

7 Appendix: Tooling and Infrastructure

The practical utility of a computational meta-analysis depends on robust tooling at each pipeline stage: assertion extraction, modeling and simulation, knowledge-graph infrastructure, and quality assurance. This appendix surveys the open-source ecosystem of Active Inference (AIF) and Free Energy Principle (FEP) implementations as of early 2026, documents the engineering trade-offs behind our knowledge-graph backend, and lists the multi-level quality gates enforced by the pipeline.

7.1 LLM-Based Assertion Extraction

Extracting structured assertions from unstructured text is the most labor-intensive component of knowledge-graph construction. Manual annotation produces high-quality results but does not scale to corpora of thousands of papers—a constraint demonstrated by Knight et al. [Knight et al., 2022], whose systematic analysis of FEP and Active Inference publications required manual coding of structural, visual, and mathematical features for hundreds of annotated papers. We implement a hybrid approach: an LLM performs initial extraction and human review provides validation pathways.

Our extraction pipeline deploys a locally hosted LLM through Ollama [Ollama Team, 2024]. Each paper's abstract is assessed against the eight hypothesis definitions in a structured prompt requesting a JSON array of assessments. Unlike keyword matching, which detects only topical terms, the LLM evaluates the *semantic relationship* between a paper's claims and each hypothesis. Papers critiquing the FEP correctly receive "contradicts" assessments for FEP Universality (H1), while methodology tutorials receive "neutral" assessments reflecting their pedagogical character. Detailed prompt engineering, schemas, and failure modes are documented in the extraction pipeline section.

7.2 Software Ecosystem

The Active Inference community has developed a rapidly growing ecosystem of open-source tools spanning multiple programming languages, inference paradigms, and application domains. This section provides a comprehensive survey of publicly available implementations as of early 2026, organized by functional category. We emphasize tools with accessible source code: open-source availability is a prerequisite for reproducibility and community-driven validation.

7.2.1 General-Purpose Frameworks

Six general-purpose frameworks dominate the landscape, collectively covering discrete, continuous, and real-time inference:

pymdp. The pymdp library [Heins et al., 2022] provides a Python implementation of active inference for discrete state-space POMDPs, supporting message passing on factor graphs, policy inference via expected free energy, and hierarchical generative models. It has become the standard entry point for algorithm development and the most widely forked AIF repository.

SPM. The SPM package (Wellcome Centre for Human Neuroimaging) includes MATLAB implementations of Dynamic Causal Modeling and variational Bayesian inference under the FEP. It remains the reference implementation for neuroimaging applications and houses the original Friston-group POMDP scripts.

RxInfer.jl. RxInfer is a Julia package for reactive message-passing-based Bayesian inference, supporting real-time and streaming inference suitable for robotics and online learning. Version 4.0.0 (early 2025) [Bagaev et al., 2025] introduced projected constraints and adaptive inference optimized for dynamic data streams and autonomous systems. The RxInfer ecosystem includes extensive tutorials covering Bayesian linear regression, hidden Markov models, Kalman filtering, Gaussian process regression, hierarchical Gaussian filters, nonlinear sensor fusion, and active inference mountain car control, available at the official documentation and the Learnable Loop tutorial portal.

ActiveInference.jl. In parallel to RxInfer's generalized message-passing focus, ActiveInference.jl provides a Julia-native, near drop-in conceptual analogue to Python's `pymdp` [Nehrer et al., 2025]. It explicitly targets computational psychiatry and cognitive neuroscience workflows emphasizing standard discrete-state POMDP simulation, parameter estimation, and recovery. The library leverages Julia's array semantics—utilizing vectors of arrays to efficiently encode multimodal factorized models via the canonical $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E}$ components—to streamline tasks such as generating synthetic behavioral data, fitting models to subject behavior, and probing internal beliefs via robust simulation loops (`infer_states!`, `infer_policies!`, `sample_action!`).

Cpp-AIF. The Cpp-AIF header-only C++ library [Gregoretti, 2023] implements active inference for discrete POMDPs with multicore parallelization of the most demanding computational kernels—multidimensional inner products for

expected free energy computation and state estimation. By abstracting the mathematical details behind a high-level API, Cpp-AIF targets embedded systems and performance-critical applications where Python overhead is prohibitive.

FEPS. Free Energy Projective Simulation [Pazem et al., 2024] combines active inference with interpretable graphical policy representations, enabling agents to plan via expected free energy while exposing decision logic as human-readable policy graphs. FEPS targets interpretable reinforcement learning tasks where black-box deep agents are undesirable—behavioral biology, clinical decision support, and safety-critical robotics.

7.2.2 Deep Active Inference

Scaling active inference beyond tabular POMDPs to high-dimensional observation spaces requires neural-network function approximators. A growing body of deep active inference implementations explores this direction:

The foundational deep AIF agent of Fountas et al. [Fountas et al., 2020] introduced Monte-Carlo tree search over learned latent spaces, achieving non-trivial Atari performance. Millidge's DeepActiveInference extended this to continuous control with backpropagation-based world models [Millidge, 2020]. Champion's Branching-Time Active Inference (BTAI_3MF) and its deep variant (Deep_BTAI_3MF) implement tree-structured planning under the free-energy objective, scaling active inference to partially observable environments with multi-step lookahead [Champion et al., 2021]. Most recently, AXIOM [Heins et al., 2025] achieves competitive Gameworld-10k benchmark performance using expanding object-centric world models, learning in minutes rather than hours—a landmark result for scalability.

7.2.3 Predictive Coding and Neural Generative Coding

Predictive coding provides the core computational mechanism linking active inference to neuroscience. Several implementations offer accessible entry points:

ngc-learn. The Neural Generative Coding library (ngc-learn v3.0, JAX-based) provides a framework for simulating neurobiologically-plausible systems using predictive-coding circuits, Hebbian learning, and spike-based dynamics. It supports constructing arbitrary neural generative models without backpropagation, directly instantiating the FEP's prediction-error minimization at the circuit level.

Active Neural Generative Coding (ANGC). ANGC implements a form of active inference using paired predictive-coding circuits—an actor/policy circuit and a world/transition model—that co-evolve across episodes without backpropagation. The agent decomposes behavior into epistemic foraging (uncertainty reduction) and instrumental (reward-seeking) terms, operating with sparse rewards where classical DQN requires dense reward engineering.

Predictive Coding ≈ Backprop. Millidge et al. demonstrate that predictive-coding networks can approximate backpropagation along arbitrary computational graphs [Millidge et al., 2022], providing a biologically plausible alternative to gradient descent. The PredictiveCodingBackprop repository provides the reference implementation.

7.2.4 Benchmarking Progress

The scalability gap between AIF and deep reinforcement learning has been a central limitation of the tools domain. Recent work demonstrates significant progress on two fronts. First, AXIOM [Heins et al., 2025] outperforms state-of-the-art model-based deep RL agents including DreamerV3 on the Gameworld-10k benchmark while using substantially smaller model sizes; its object-centric scene decomposition enables sample-efficient learning from structured representations rather than raw-pixel memorization. Second, variational message-passing formulations [Champion et al., 2021] connect EFE decomposition—into risk, ambiguity, epistemic (information-seeking), and instrumental (goal-reaching) components—to practical planning algorithms, advancing the theoretical justification for EFE-based policy selection (H2). Separately, Friston et al. [Friston et al., 2025] introduce structure learning via Bayesian Model Reduction as a principled approach to artificial reasoning under active inference.

7.2.5 Comprehensive Open-Source Tool Survey

The following table catalogs the principal open-source Active Inference implementations surveyed, organized by functional category. For each tool we list the primary language, application domain, and associated publication or repository. The table is intended as a navigational resource for researchers seeking existing implementations relevant to specific hypotheses (H1–H8) or application domains (A1–C5).

Table 12: Comprehensive open-source survey of Active Inference and Free Energy Principle software, grouped by functional category. Forty-plus implementations span seven categories.

Tool / Repository	Lang.	Description	Paper / Source
		General-Purpose Frameworks	
pymdp	Python	Discrete POMDP active inference; factor graphs, hierarchical models	Heins et al. [2022]
SPM	MATLAB	DCM, variational Bayes; neuroimaging reference implementation	Friston et al. [2017]
RxInfer.jl	Julia	Reactive message passing; real-time streaming Bayesian inference	Bagaev et al. [2025]
ActiveInference.jl	Julia	Discrete POMDP AIF; parameter recovery for computational psychiatry	Nehrer et al. [2025]
Cpp-AIF	C++	Header-only POMDP AIF library with multi-core parallelization	Gregoretti [2023]
FEPS	Python	EFE on interpretable policy graphs; projective simulation	Pazem et al. [2024]
ActivPynference	Python	Discrete AIF with factor-graph message passing; educational focus	—
pypc	Python	Predictive-coding inference engine for continuous models	—
ActiveInferAnts	Rust	Rust-native AIF framework with WASM compilation target	—
		Deep Active Inference	
deep-active-inference-mc	Python	Monte-Carlo tree search in learned latent spaces; Atari	Fountas et al. [2020]
DeepActiveInference	Python	Continuous deep AIF with backprop-based world models	Millidge [2020]
BTAI_3MF	Python	Branching-time AIF with multi-step tree planning	Champion et al. [2021]
Deep_BTAI_3MF	Python	Deep neural variant of BTAI with learned state spaces	Champion et al. [2021]
OO-BTAI_3MF	Python	Object-oriented BTAI variant for structured environments	—
AXIOM	Python	Object-centric world models; Gameworld 10k in minutes; beats DreamerV3	Heins et al. [2025]
Deep-AIF-POMDPs	Python	Deep AIF for partially observable MDPs	—
Homing-Pigeon	Python	Navigation agent using deep active inference	—
active-inference (Voostrum)	Python	Continuous deep AIF with learned generative models	arXiv:2406.07726
		Predictive Coding & Neural Generative Coding	
ngc-learn	Python/JAX	Neurobiological simulation; predictive-coding circuits, Hebbian learning	—
ANGC	Python	Backprop-free AIF agent with paired PC circuits	AAAI 2022
PredictiveCodingBackprop	Python	Predictive coding approximates backprop on arbitrary graphs	Millidge et al. [2022]
Supervised-Predictive-Coding	Python	Supervised learning via hierarchical predictive coding	—
predcoding	Python	Minimal predictive-coding implementation	—
pybrid	Python	Hybrid predictive-coding and active-inference library	—
nmpassing	Python	Neural message passing for PC networks	—
		Neuroscience, Embodied & Biological	
allostasis	Python	Allostatic regulation via AIF; interoceptive inference	bioRxiv:2021.02.16

Continued on next page

Tool / Repository	Lang.	Description	Paper / Source
ants	Python	Ant foraging simulation with stigmergic AIF agents	Heins et al. [2024]
Reward_Bases	Python	Reward-basis function representations under AIF	bioRxiv:2022.04.14
action-oriented	Python	Action-oriented predictive-processing models	Tschantz et al. [2020]
Biofirm	Python	Bioregional stewardship via organizational AIF	—
bayesian-mechanics-sdes	Python	Bayesian mechanics: SDE simulations of Markov-blanket dynamics	arXiv:2206.02629
reverse_engineering	MATLAB	Reverse-engineering neural dynamics under the FEP	—
Multi-Agent & Social Dynamics			
opinion_dynamics	Python	Opinion dynamics and belief formation via AIF	—
network-actinf	Python	Network-level active inference with coupled agents	—
Variational-Capsule-Routing	Python	Capsule networks with variational inference routing	AAAI 2020
Active-Inference-Successor	Python	Successor representations under active inference	—
Domain-Specific Applications			
adaptive_aif_agents_fl	Python	Adaptive AIF agents for federated learning	arXiv:2410.09099
smartville	Python	IoT smart-building control via AIF under partial observability	TechRxiv 2025
FEP_Blorpomon	Python	Game-theoretic AIF agent demonstration	—
MountainCarAI	Python	Mountain car control via active inference	—
rl-inference	Python	Bridging RL and active inference policy selection	arXiv:2002.12636
EFE-GLean	Python	Expected free energy with generalized learning	Entropy 2025
EFEasVFE	Julia	EFE reformulated as variational free energy	—
Robust-FE-Minimization	Python	Robust decision-making via free-energy minimization	arXiv:2503.13223
Tutorials & Educational Resources			
Active-Inference-from-Scratch	Python	Step-by-step AIF implementation tutorial	—
IC2S2-AIF-Tutorial	Python	Computational social-science AIF tutorial	—
julia4ta tutorials (9x10–12)	Julia	RxInfer-based AIF agent tutorials	—
ActInf Textbook Colab	Python	Interactive notebooks for Parr et al. [2022]	—
deep_aif_workshop	Python	Workshop materials for deep active inference	—
AdaptiveResonance.jl	Julia	Adaptive resonance theory models in Julia	—

7.2.6 Comparative Feature Matrix

Table 13: Comparative feature matrix of seven representative Active Inference packages. Features span language, state-space type, inference algorithm, hierarchical support, GPU acceleration, license, and primary use case.

Feature	pymdp	SPM	RxInfer.jl	ActiveInf.jl	Cpp-AIF	FEPS	ngc-learn
Language	Python	MATLAB	Julia	Julia	C++	Python	Python/JAX
State Spaces	Discrete	Disc.+Cont.	Continuous	Discrete	Discrete	Discrete	Continuous
Inference	Msg. pass.	Var. Bayes	Reactive msg.	Msg. pass.	EFE+state	EFE on graphs	Pred. coding
Deep AIF	Partial	No	Custom factors	No	No	No	Yes
Real-time	No	No	Yes	No	Yes	No	No
Hierarchical	Yes	Yes (DCM)	Yes	No	Yes	No	Yes
GPU	No	No	No	No	CPU multi	No	Yes (JAX)
License	MIT	GPL	MIT	MIT	MIT	MIT	BSD-3
Primary Use	Prototyping	Neuroimaging	Robotics	Comp. psych.	Embedded	Interp. RL	NeuroAI

The complementary strengths across these packages reflect a fragmented but maturing ecosystem. The survey reveals several patterns: (1) Python dominates ($ $75% of implementations), with Julia emerging as the preferred alternative for performance-critical applications; (2) discrete-POMDP implementations outnumber continuous variants by approximately 3:1, reflecting pymdp's community influence; (3) deep active-inference implementations are concentrated in a small number of research groups (Champion, Millidge, Fountas, Heins), suggesting high barriers to entry; (4) multi-agent and social AIF implementations remain sparse relative to single-agent tools; and (5) domain-specific applications (IoT, federated learning, smart buildings) represent the newest and fastest-growing category, aligning with the temporal growth patterns observed in the C-domain (applied) subfields. The variational-free-energy foundations shared by Active Inference and Energy-Based Models—including Helmholtz machines [Dayan et al., 1995], Boltzmann machines [Hinton, 2002], and variational autoencoders [Kingma and Welling, 2014]—suggest that interoperability with mainstream deep generative-modeling frameworks (PyTorch, JAX) could bridge these parallel research programs.

7.3 Knowledge Graph Infrastructure

Our knowledge graph uses an RDF-compatible schema deployable on standard semantic-web infrastructure. The nanopublication model [Groth et al., 2010, Kuhn et al., 2016] provides a principled atomic unit of scientific evidence: each nanopublication packages a single assertion (e.g., "Paper X supports Hypothesis Y") with explicit provenance and publication metadata in four named RDF graphs (Head, Assertion, Provenance, Publication Info). This structure satisfies the FAIR data principles by design: nanopublications are **F**indable via URI-based identification, **A**ccessible through standard RDF protocols, **I**nteroperable via W3C-standard TriG serialization, and **R**eusable with explicit provenance and CC0 licensing. The full RDF schema and a TriG serialization example are presented in the methodology and Appendix~8.5.

The engineering trade-offs among the three deployment options are straightforward:

Nanopublication servers provide decentralized, content-addressed storage. The pipeline writes nanopublications in two forms: JSON Lines (for incremental checkpointing and tooling) and RDF/TriG per the nanopublication standard (Assertion, Provenance, Publication Info), suitable for the nanopublication network and FAIR deployment. The recent release of nanopub-js v0.1.0 [Knowledge Pixels, 2026]—a JavaScript library enabling browser-based creation, signing, and querying of nanopublications—opens the possibility of community-contributed assertions directly from web interfaces, lowering the barrier to participatory evidence curation. Future integration with Trusty URIs [Kuhn and Dumontier, 2014] would provide cryptographic content verification and persistent identifiers for each nanopublication.

RDF stores (e.g., Apache Jena Fuseki, Blazegraph, Oxigraph) enable SPARQL queries such as "find all papers supporting hypothesis H published after 2020 in the neuroscience domain (C1)." The cost is operational overhead and query latency.

Property-graph databases (e.g., Neo4j) prioritize traversal performance for path queries and community detection, at the expense of semantic-web compatibility.

While RDF and property graphs excel at structurally organizing assertions, it is crucial to recognize that they inherently compress the rich epistemic context of the original papers (e.g., methodological caveats, sample sizes, scope limitations) into flattened confidence scores—a fundamental limitation of current automated knowledge extraction discussed in the conclusion.

The Active Inference Ontology namespace ensures integration with external ontologies and linked-data resources.

7.4 Multi-Level Quality Assurance

Quality assurance operates at four levels: assertion-level confidence and review, graph-level structural consistency, score-level boundary tests, and pipeline-level continuous-integration coverage.

7.4.1 Assertion-Level Validation

Assertions below a configurable confidence threshold (default 0.6) are flagged for review. Inter-annotator agreement (κ) is computed when multiple annotators assess the same paper. The threshold is chosen to balance recall against the prompt-engineering cost of pushing the LLM to over-commit; lowering it inflates noisy neutral assertions, raising it discards weakly supported but legitimate claims.

7.4.2 Graph-Level Consistency Checks

Consistency checks verify that all nodes link to valid targets and no orphan nodes exist. Coverage metrics track the proportion of annotated papers, the fraction of references that resolve inside the corpus, and the per-domain assertion density.

7.4.3 Score-Level Unit Testing

Hypothesis scoring is validated through unit tests on synthetic data verifying boundary conditions: all-support fixtures must produce scores at $+1$, all-contradict at -1, and balanced inputs at 0. Sensitivity analysis sweeps over confidence thresholds and citation-weighting schemes to confirm that qualitative rankings are stable.

7.4.4 Pipeline-Level Test Coverage

Test-driven development enforces 90% minimum code coverage on project modules and 60% on shared infrastructure, with real data and computation (no mocking). All tests run on every push; failures block merges and releases.

7.4.5 Quality Thresholds

Table 14: Multi-level quality-assurance thresholds enforced across the pipeline. Each level defines a metric, minimum threshold, and failure action. Pipeline-level thresholds (90% coverage, 100% pass rate) are enforced via CI gates; lower-level checks emit warnings or block release as indicated.

Level	Metric	Threshold	On Failure
Assertion	Confidence c	≥ 0.6	Flag for review
Assertion	Inter-annotator κ	≥ 0.70	Re-annotate
Graph	Orphan-node ratio	$= 0$	Reject build
Graph	Corpus coverage	$\geq 80\%$	Warning
Score	Boundary tests (all-support / all-contradict / balanced)	All pass	Block release
Score	Sensitivity-sweep stability	Top-k ranks unchanged	Warning
Pipeline	Project-code coverage	$\geq 90\%$	Block merge
Pipeline	Infrastructure coverage	$\geq 60\%$	Block merge
Pipeline	Test pass rate	100%	Block release

The hypothesis-evidence results, temporal dynamics of evidence accumulation, and assertion analysis are presented in the hypothesis results section.

8 Appendix: Mathematical and Algorithmic Details

This appendix collects the formal mathematical definitions, derivations, and algorithmic specifications referenced from the main methodology section. Each subsection is self-contained; equations are labelled for cross-referencing from the body and from §18.

8.1 Citation-Weighted Hypothesis Scoring Formula

For each hypothesis H, we compute a citation-weighted evidence score aggregating all assertions relevant to H:

$$\text{score}(H) = \frac{\sum_{a \in S(H)} w(a) - \sum_{a \in C(H)} w(a)}{\sum_{a \in A(H)} w(a)} \tag{3}$$

where $S(H)$ is the set of supporting assertions, $C(H)$ the set of contradicting assertions, $A(H)$ all assertions for H (including neutral), and the weight function is

$$w(a) = \log(1 + \text{citations}(a)) \cdot \text{confidence}(a). \tag{4}$$

The logarithmic citation weighting ensures that highly cited papers carry more influence while preventing any single blockbuster paper from dominating the score. The score lies in $[-1, 1]$: values near $+1$ indicate strong supporting evidence, values near -1 strong contradicting evidence, and values near 0 balanced or insufficient evidence. As emphasized in the main text, this score is a *relative evidentiary ranking* within the current literature topology, not a calibrated Bayesian probability of the hypothesis being true.

Temporal aggregation. We additionally compute temporal trends by evaluating the cumulative score at each year t, using only assertions from papers published in year $\leq t$:

$$\text{score}(H, t) = \frac{\sum_{a \in S(H,t)} w(a) - \sum_{a \in C(H,t)} w(a)}{\sum_{a \in A(H,t)} w(a)}. \tag{5}$$

This reveals whether support for a hypothesis is growing, declining, or plateauing over time. Cumulative aggregation (rather than yearly windows) is preferred because per-year assertion counts for narrow hypotheses are too sparse for stable point estimates.

Algorithmic specification. The scoring routine is a pure function of the assertion set; it has no hidden state and is deterministic given the input. The reference implementation lives in `projects/act_inf_metaanalysis/src/scoring/citation_wei`

```
function score(H, assertions):
    S, C, A_all = 0, 0, 0
    for a in assertions where a.hypothesis == H:
        w = log(1 + a.citations) * a.confidence
        if a.direction == "supports":     S     += w
        elif a.direction == "contradicts": C     += w
        A_all += w
    return (S - C) / A_all  if A_all > 0  else 0
```

Boundary tests in `tests/test_scoring.py` confirm that all-support fixtures yield $+1$, all-contradict fixtures yield -1, and balanced fixtures yield 0 within numerical tolerance.

8.2 Non-negative Matrix Factorization (NMF) for Topic Modeling

We apply NMF to the TF-IDF matrix of the corpus to discover latent topics. Given the document-term matrix $V \in \mathbb{R}_{\geq 0}^{n \times m}$, NMF finds factor matrices $W \in \mathbb{R}_{\geq 0}^{n \times k}$ and $H \in \mathbb{R}_{\geq 0}^{k \times m}$ such that $V \approx WH$, where k is the number of topics. We use multiplicative update rules [Lee and Seung, 1999]:

$$H \leftarrow H \odot \frac{W^T V}{W^T W H + \epsilon}, \qquad W \leftarrow W \odot \frac{V H^T}{W H H^T + \epsilon}, \tag{6}$$

with $\epsilon = 10^{-10}$ for numerical stability and a fixed random seed of 42 for reproducibility (deterministic topic alignment across pipeline runs, with empirical stability confirmed via Jaccard similarities > 0.90 across alternative seeds).

Term-Frequency Inverse Document Frequency (TF-IDF). The document-term matrix is constructed using a smoothed TF-IDF weighting [Salton et al., 1975]. For term t in document d:

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \left[\log\left(\frac{N}{\text{df}(t) + 1} \right) + 1 \right], \tag{7}$$

where $\text{tf}(t, d) = \text{count}(t, d)/|d|$ is the normalized term frequency, N the total number of documents, and $\text{df}(t)$ the document frequency of term t. The $+1$ additive smoothing in the denominator prevents division by zero and reduces the weight of extremely rare terms; the outer $+1$ ensures strictly positive IDF values. Document vectors are L2-normalized before NMF factorization.

8.3 Field Growth-Rate Estimation

The **mean year-over-year growth rate** $\bar{g}$ is the arithmetic mean of annual growth rates computed only for years where the prior year had non-zero publications:

$$\bar{g} = \frac{1}{|Y|} \sum_{y \in Y} \frac{n_y - n_{y-1}}{n_{y-1}}, \tag{8}$$

where $Y = \{y : n_{y-1} > 0\}$ and n_y is the number of publications in year y.

The **doubling time** t_d is derived from the mean annual growth rate:

$$t_d = \frac{\ln 2}{\ln(1 + \bar{g})}. \tag{9}$$

The **compound annual growth rate** (CAGR) over the full span $[y_0, y_T]$ is

$$\text{CAGR} = \left(\frac{n_{\text{cumulative}}(y_T)}{n_{\text{cumulative}}(y_0)} \right)^{1/(y_T - y_0)} - 1. \tag{10}$$

For the current corpus, CAGR = 20.36%. The more recent growth phase (2010–2026) exhibits substantially higher annualized growth than the long-run average; reporting both the T-year CAGR and recent-phase CAGR avoids overstating maturity-era expansion.

8.4 Advanced Visualization Methods

8.4.1 PCA of TF-IDF Embeddings

Principal Component Analysis (PCA) is applied to the TF-IDF matrix V to project each document into a 2-D space. The projection preserves the directions of maximum variance, enabling visual inspection of document clustering by domain. Loading arrows overlay the top-variance terms onto the scatter plot, showing which vocabulary drives the principal components.

8.4.2 Hierarchical Clustering Dendrogram

For each domain s, we compute the centroid $\bar{v}_s = \frac{1}{|D_s|} \sum_{d \in D_s} v_d$ where D_s is the set of documents in domain s and v_d is the TF-IDF vector of document d. Ward linkage is applied to the centroid matrix to produce a hierarchical clustering dendrogram showing semantic proximity between domains.

8.4.3 Term Heatmap

For each domain s and term t, we compute the mean TF-IDF weight $\bar{w}_{s,t} = \frac{1}{|D_s|} \sum_{d \in D_s} \text{TF-IDF}(t, d)$. The heatmap displays $\bar{w}_{s,t}$ for the top-k terms (by global document frequency) across all domains, with cell intensity proportional to mean weight. This reveals distinctive vocabulary patterns that differentiate domains beyond the keyword-level classification used for subfield assignment.

8.4.4 Term Co-occurrence Matrix

The co-occurrence matrix $C \in \mathbb{R}^{k \times k}$ counts the number of documents in which two terms appear together. For top-k terms by document frequency, $C_{ij} = |\{d : t_i \in d \wedge t_j \in d\}|$. The matrix is normalized to $[0, 1]$ by dividing by the maximum entry and visualized as a symmetric heatmap.

8.5 Nanopublication RDF Schema

Each nanopublication is serialized to RDF/TriG per the nanopublication standard [Groth et al., 2010, Kuhn et al., 2016], producing four named graphs. The following annotated example illustrates the structure for a single assertion:

```
@prefix np:   <http://www.nanopub.org/nschema#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix dc:   <http://purl.org/dc/terms/> .
@prefix aif:  <http://activeinference.institute/ontology/> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .

# HEAD GRAPH: links nanopub to its three component graphs
<http://activeinference.institute/nanopub/a1b2c3d4e5f6#head> {
  <http://activeinference.institute/nanopub/a1b2c3d4e5f6>
    a np:Nanopublication ;
    np:hasAssertion       <...#assertion> ;
    np:hasProvenance      <...#provenance> ;
    np:hasPublicationInfo <...#pubinfo> .
}

# ASSERTION GRAPH: the core scientific claim
<http://activeinference.institute/nanopub/a1b2c3d4e5f6#assertion> {
  aif:paper/10.1038_nrn2787 aif:asserts aif:assertion/a1b2c3 .
  aif:assertion/a1b2c3
    aif:supports       aif:hypothesis/fep_universality ;
    aif:claim          "The paper provides foundational support for FEP as a
                        unified brain theory."^^xsd:string ;
    aif:confidence     "0.85"^^xsd:double ;
    aif:citationCount  "5000"^^xsd:integer .
}

# PROVENANCE GRAPH: extraction lineage
<http://activeinference.institute/nanopub/a1b2c3d4e5f6#provenance> {
  aif:assertion/a1b2c3
    prov:wasGeneratedBy   <http://activeinference.institute/nanopub/a1b2c3d4e5f6> ;
    prov:generatedAtTime  "2026-01-15T12:00:00+00:00"^^xsd:dateTime ;
    prov:wasAttributedTo  "act_inf_metaanalysis/gemma3:4b"^^xsd:string ;
    prov:hadPrimarySource aif:paper/10.1038_nrn2787 .
}

# PUBLICATION INFO GRAPH: nanopublication metadata
<http://activeinference.institute/nanopub/a1b2c3d4e5f6#pubinfo> {
  <http://activeinference.institute/nanopub/a1b2c3d4e5f6>
    dc:created "2026-01-15T12:00:00+00:00"^^xsd:dateTime ;
    dc:creator "act_inf_metaanalysis/gemma3:4b"^^xsd:string ;
    dc:license <https://creativecommons.org/publicdomain/zero/1.0/> .
}
```

8.5.1 Namespace Definitions

Table 15: RDF namespace definitions used in the knowledge graph and nanopublication serialization. Each prefix maps to a W3C or domain-specific URI.

Prefix	URI	Purpose
`np:`	`http://www.nanopub.org/nschema#`	Nanopub structural predicates
`prov:`	`http://www.w3.org/ns/prov#`	PROV-O provenance model
`dc:`	`http://purl.org/dc/terms/`	Dublin Core metadata
`aif:`	`http://activeinference.institute/ontology/`	Domain-specific predicates
`xsd:`	`http://www.w3.org/2001/XMLSchema#`	XML Schema datatypes

8.5.2 Core Triple Patterns

The knowledge graph encodes five fundamental relationships:

Table 16: Core RDF triple patterns encoding the five fundamental relationships in the knowledge graph. Each pattern links paper, assertion, hypothesis, or subfield nodes.

Triple Pattern	Meaning
`Paper --aif:asserts--> Assertion`	A paper makes a claim
`Paper --aif:cites--> Paper`	Intra-corpus citation link
`Paper --aif:belongsTo--> Subfield`	Domain classification
`Assertion --aif:supports--> Hypothesis`	Supporting evidence
`Assertion --aif:contradicts--> Hypothesis`	Contradicting evidence

9 Appendix: Accessibility, Cognitive Ergonomics, and Participatory Research Infrastructure

Automated meta-analysis tools operate at the intersection of computational infrastructure and human sensemaking. The scalability gains demonstrated by the present pipeline are meaningful only if the resulting knowledge artefacts remain cognitively accessible, ethically transparent, and open to diverse forms of participation. This appendix situates our work within the broader landscape of research accessibility, cognitive ergonomics, decentralized science (DeSci), and participatory infrastructure design, and concludes with a WCAG-mapped checklist that summarizes the concrete accessibility practices implemented in the figure pipeline.

9.1 Cognitive Ergonomics of Knowledge Graphs

The knowledge-graph outputs of this pipeline—hypothesis dashboards, citation networks, temporal evidence trajectories—impose nontrivial cognitive demands on users who must interpret multidimensional evidence landscapes. Cognitive Load Theory [Sweller et al., 2011] establishes that information system designs which exceed working-memory capacity produce disorientation and interpretive errors. Our visualization pipeline addresses this through progressive disclosure (summarized dashboards linking to detailed per-hypothesis breakdowns), consistent visual grammars (a fixed colour palette for supports/contradicts/neutral across all figures), and a minimum font-size floor of 16,pt that satisfies low-vision accessibility guidelines. These are not cosmetic choices but functional requirements for trustworthy scientific communication.

The ResNei (Research Neighbourhood) platform [Lumiruusu et al., 2025] provides a particularly instructive design exemplar for the next generation of cognitive-ergonomic research tools. ResNei is an AI-augmented, neuro-informed research environment that transforms heterogeneous scientific corpora into a living, collaborative knowledge graph structured as modular Conceptual Nexus Models (CNMs). Where our pipeline produces a static (though periodically updated) evidence snapshot, ResNei's architecture foregrounds dynamic, responsive exploration through three cognitive modes: *longitudinal* (tracking a concept's evolution over time), *latitudinal* (surveying related concepts across subfields), and *relational* (mapping connections between concepts). This trimodal navigation directly operationalizes the progressive-disclosure principle, enabling users to manage cognitive complexity by choosing their depth of engagement.

9.1.1 Action–Intention UX and Active Inference Design Principles

ResNei's most theoretically significant contribution is its action–intention UX model, which replaces the conventional passive, attention-maximizing feed with a framework that interprets user actions (uploading papers, highlighting passages, opening concept maps, initiating discussions) as situated signals of research direction. Rather than deploying opaque recommendation engines, the system uses explicit action trajectories to surface contextually appropriate tools and views—an approach that resonates with the perception–action loop central to Active Inference itself [Parr et al., 2022]. The design principle of "minimal system intervention, maximum research coherence' ' ensures that the interface scaffolds orientation and affordances without interruptive prompts or aggressive automation. This ethos directly addresses the risk that AI-augmented sensemaking tools inadvertently narrow epistemic horizons through algorithmic filtering.

9.1.2 Risk-Aware and Bias-Transparent Design

ResNei's solution-design document is notable for its unusually explicit treatment of harms and ameliorations. It identifies exclusion, algorithmic misrepresentation, overconfidence in AI outputs, hidden inequalities, marginalization of less-cited work, surveillance risks, cognitive overload, false comprehensiveness, and data privacy as first-class design constraints [Lumiruusu et al., 2025]. Mitigations include deliberately inclusive UX (designing from the standpoint of those usually excluded), systematic provenance and confidence indicators, framing all AI outputs as suggestions with traceable bases, and configurable metrics beyond citation counts (e.g., conceptual novelty, geographic diversity, publication type). This risk model provides a concrete template for future iterations of our own pipeline, which currently presents citation-weighted scores without UI-level confidence calibration or per-assertion provenance indicators.

9.2 FAIR Data and Decentralized Science

The pipeline's outputs—nanopublications, knowledge-graph triples, and structured assertion records—are designed to satisfy the FAIR principles (Findable, Accessible, Interoperable, Reusable) articulated by Wilkinson et al. [Wilkinson et al., 2016]. Each nanopublication carries machine-readable provenance (source paper DOI, extraction model,

confidence score, timestamp), enabling downstream consumers to evaluate evidential quality independently of our aggregation choices. The JSON Lines and RDF/TriG serialization formats guarantee interoperability with existing semantic-web infrastructure.

Decentralized Science (DeSci) represents a broader movement to dismantle structural barriers in scientific publishing and funding through blockchain-based governance, tokenized intellectual property, and community-owned research commons [Hamburg, 2022]. Our pipeline's open-source, modular, and configuration-driven design aligns with DeSci principles: the entire analytical workflow is reproducible from source code, hypothesis definitions and extraction prompts are version-controlled in YAML rather than embedded in proprietary systems, and the nanopublication output format is natively compatible with federated semantic publishing networks. ResNei's architecture further advances this trajectory by grounding its collaborative features in `social accountability'' and`reciprocity, interdependence, and access'' as explicit design values [Lumiruusu et al., 2025], directly addressing the power asymmetries that traditional centralized publication systems perpetuate.

9.3 Participatory Research and Universal Access

The aspiration toward participatory research infrastructure—where diverse contributors can meaningfully engage with evidence synthesis regardless of technical expertise—is a recurring theme across the projects discussed here. Bonney et al.'s foundational work on citizen science [Bonney et al., 2009] demonstrated that non-expert participants can make rigorous contributions to scientific knowledge production when provided with appropriate scaffolding, standardized protocols, and feedback loops. Universal Design for Learning principles [Rose and Meyer, 2000] further emphasize that accessibility is not a specialized accommodation but a design paradigm that improves usability for all users.

Applied to computational meta-analysis, this means designing systems where:

- **Contribution pathways** exist at multiple expertise levels—from correcting individual assertion labels (requiring only domain knowledge) to extending extraction prompts or hypothesis definitions (requiring pipeline familiarity);
- **Transparency mechanisms** make model confidence, extraction provenance, and aggregation logic visible and interrogable by non-technical users;
- **Multimodal access** ensures that knowledge-graph outputs are available not only as programmatic APIs and raw data files but as navigable visual interfaces with WCAG-compliant accessibility standards;
- **Cultural and linguistic inclusivity** is recognized as a structural requirement rather than a desirable addition—our pipeline's current English-language dominance (noted in §5 as a corpus bias) is a limitation that future multilingual extraction capabilities must address.

The convergence of ResNei's neuro-informed collaborative environment, DeSci's decentralized governance models, FAIR-data interoperability, and citizen-science participation frameworks collectively describes the emerging infrastructure requirements for equitable, cognitively supportive, and community-governed scientific sensemaking. These are not peripheral concerns for computational meta-analysis but architectural prerequisites for systems that aspire to serve as living, trusted evidence ledgers for rapidly evolving scientific fields.

9.4 Pipeline Accessibility Checklist

The following checklist summarizes the concrete accessibility practices implemented in the figure-generation and rendering stages, mapped to the relevant Web Content Accessibility Guidelines (WCAG 2.1, Level AA) success criteria. "Status'' is recorded as **Implemented**, **Partial**, or **Planned** based on the current state of the pipeline.

Table 17: Accessibility practices implemented in the figure pipeline, mapped to WCAG 2.1 Level AA success criteria. Status reflects the current pipeline; "Planned" items are tracked in the project issue tracker.

Practice	**WCAG Ref.**	**Implementation**
Colorblind-safe palette	1.4.1	Wong (2011) 8-colour palette enforced in all figures [Wong, 2011]; supports/contradicts/neutral encoded by both hue and luminance.
Minimum font size	1.4.4	16 pt floor enforced in figure-generation script; tick labels never fall below this threshold.
Sufficient contrast	1.4.3	Foreground/background contrast $\geq$ 4.5:1 for all text, $\geq$ 3:1 for large headings, validated programmatically.
Non-color encodings	1.4.1	Direction encoded by both color and pattern (solid / hatched / outlined) so that grayscale printing remains interpretable.
Alt text and figure captions	1.1.1	Each `\includegraphics` is paired with a `\caption` that describes the figure content, key axes, and main takeaway.
Consistent visual grammar	3.2.4	Domain colors, hypothesis ordering, and axis conventions are fixed across all figures by a single style module.
Progressive disclosure	2.4.5	Summary dashboards link to per-hypothesis and per-domain breakdowns; readers can choose depth of engagement.
Machine-readable outputs	4.1.2	All analytic results published as JSON / JSONL alongside PNG figures, enabling assistive-technology consumption.
Provenance metadata	1.3.1	Each nanopublication carries source DOI, extraction model, timestamp, and confidence; programmatically queryable.
Multilingual extraction	—	**Planned**: current pipeline is English-only; future multilingual prompts and corpus expansion are tracked as a corpus-bias mitigation.
Per-assertion confidence UI	—	**Planned**: aggregate scores currently dominate the dashboard; future iterations will surface per-assertion confidence and rationale.

10 Notation, Abbreviations, and Glossary

This appendix consolidates the mathematical notation, abbreviations, hypothesis identifiers, and key terminology used throughout the manuscript. Each table is self-contained and may be consulted independently. Cross-references in the main text use the labels declared here.

10.1 Mathematical Symbols and Notation

The following symbols appear in the methodology, results, and technical appendices. Where a quantity is defined formally, the relevant equation is referenced inline; otherwise the description here is the canonical definition. All probabilities and confidences are real-valued in $[0, 1]$, and all aggregate scores are in $[-1, 1]$.

Table 18: Mathematical symbols and notation used throughout the manuscript. Scoring quantities are defined formally in §3.6 and §8.1; growth metrics in §8.3; topic-modeling notation in §8.2.

Symbol	Description
N	Corpus size after deduplication (total unique papers)
n	Subfield paper count (papers within a single domain category)
$T = y_T - y_0$	Time span in years (used for CAGR)
y_0, y_T	First and last years in the publication window
n_y	Number of publications in year y
$w(a)$	Citation-weighted weight of assertion a: $\log(1 + \text{citations}) \cdot c$
$\text{score}(H)$	Aggregate citation-weighted evidence score for hypothesis H, range $[-1, 1]$
$\text{score}(H, t)$	Cumulative score for H using only assertions from papers published $\leq t$
$S(H), C(H), A(H)$	Supporting / contradicting / all assertion sets for hypothesis H
c	Assertion confidence reported by the LLM, range $[0, 1]$
d	Assertion direction $\in \{\text{supports}, \text{contradicts}, \text{neutral}\}$
k	Number of latent topics in NMF factorization
$V \in \mathbb{R}_{\geq 0}^{n \times m}$	TF-IDF document-term matrix (documents $\times$ terms)
$W \in \mathbb{R}_{\geq 0}^{n \times k}$	NMF document-topic factor
$H \in \mathbb{R}_{\geq 0}^{k \times m}$	NMF topic-term factor (overloaded notation; context disambiguates)
ϵ	Numerical-stability constant (10^{-10})
CAGR	Compound annual growth rate (Eq. 10)
t_d	Publication doubling time in years (Eq. 9)
$\bar{g}$	Mean year-over-year growth rate (Eq. 8)
κ	Cohen's kappa, inter-annotator agreement
$\text{tf}(t, d)$	Normalized term frequency of t in document d
$\text{df}(t)$	Document frequency of term t across the corpus
$\mathcal{F}$	Variational free energy
$\mathbf{G}$	Expected free energy (used for policy ranking)
KL	Kullback–Leibler divergence
$\mathbb{E}$	Expectation operator

10.2 Abbreviations and Acronyms Used

The acronyms below appear at least once in the main text, methods, results, or appendices. Domain-specific shorthands such as the A/B/C taxonomy categories (e.g., A1, A2, B, C1–C5) are documented inline at first use in the field overview and the subfield analyses.

Table 19: Abbreviations and acronyms used in this manuscript, listed alphabetically. Where an acronym names a software package or organization, the canonical reference appears in the bibliography.

Abbreviation	**Definition**
AIF	Active Inference
ANGC	Active Neural Generative Coding
API	Application Programming Interface
arXiv	Open-access preprint repository (`arxiv.org`)
BTAI	Branching-Time Active Inference
CAGR	Compound Annual Growth Rate
CC0	Creative Commons Zero (public-domain dedication)
CI	Continuous Integration
CNM	Conceptual Nexus Model (ResNei)
DCM	Dynamic Causal Modelling
DeSci	Decentralized Science
DOI	Digital Object Identifier
EBM	Energy-Based Model
EFE	Expected Free Energy
FAIR	Findable, Accessible, Interoperable, Reusable
FAIR4RS	FAIR Principles for Research Software
FEP	Free Energy Principle
FEPS	Free Energy Projective Simulation
HITS	Hyperlink-Induced Topic Search (Kleinberg)
IaC	Infrastructure as Code
JSON	JavaScript Object Notation
JSONL	JSON Lines (newline-delimited JSON)
KG	Knowledge Graph
KL	Kullback–Leibler (divergence)
LLM	Large Language Model
MBR	Bayesian Model Reduction
MCMC	Markov Chain Monte Carlo
MIT	Massachusetts Institute of Technology
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
ORCID	Open Researcher and Contributor ID
PCA	Principal Component Analysis
PDF	Portable Document Format
POMDP	Partially Observable Markov Decision Process
PROV-O	PROV Ontology (W3C provenance data model)
RBM	Restricted Boltzmann Machine
RDF	Resource Description Framework
ResNei	Research Neighbourhood (cognitive-ergonomic platform)
RL	Reinforcement Learning
SDE	Stochastic Differential Equation
SPARQL	SPARQL Protocol and RDF Query Language
SPM	Statistical Parametric Mapping
TDD	Test-Driven Development
TF-IDF	Term Frequency–Inverse Document Frequency
TriG	Terse RDF Triple Language with Named Graphs
URI	Uniform Resource Identifier
VAE	Variational Autoencoder
VFE	Variational Free Energy
WCAG	Web Content Accessibility Guidelines
W3C	World Wide Web Consortium

10.3 Standard Hypothesis Definitions and Identifiers

The eight hypotheses below define the evaluation rubric used by the LLM-based assertion extractor (extraction pipeline). Each hypothesis is anchored to its primary domain in the A/B/C taxonomy, but assertions are extracted from any paper whose abstract relates substantively to the claim. Quantitative results across these hypotheses are reported in the hypothesis results section.

Table 20: Standard hypothesis definitions tracked throughout the meta-analysis. The Scope column records the primary domain in the A/B/C taxonomy; assertions are not restricted to that domain. Wording reflects the prompt presented to the extraction LLM.

ID	Hypothesis	Scope
H1	FEP Universality: the Free Energy Principle applies universally to all self-organizing systems, from cells to ecosystems.	A (Core Theory)
H2	AIF Optimality: Active Inference agents achieve principled, near-optimal decision-making under uncertainty by minimizing expected free energy.	B (Tools)
H3	Markov Blanket Realism: Markov blankets correspond to real, physically realizable boundaries between systems and their environments.	A (Core Theory)
H4	Predictive Coding: cortical hierarchies minimize prediction errors via predictive coding, providing a neurobiologically realistic substrate for active inference.	C1 (Neuroscience)
H5	Scalability: Active Inference scales to complex, high-dimensional environments comparable to those addressed by deep reinforcement learning.	B (Tools)
H6	Clinical Utility: Active Inference produces clinically useful computational models of psychiatric and neurological conditions.	C4 (Psychiatry)
H7	Morphogenesis: the FEP explains morphogenetic, developmental, and self-organizing biological processes.	C5 (Biology)
H8	Language AIF: Active Inference provides a viable framework for language comprehension, production, and communication.	C3 (Language)

10.4 Glossary of Key Terms

The glossary below defines pipeline-specific concepts, statistical methods, and domain terminology referenced in the main text. Software package names appear in typewriter font; mathematical objects use the notation defined above. Where a term has both a colloquial and a technical sense, the technical reading is given.

Table 21: Glossary of key terms used in this manuscript, including pipeline-specific concepts, statistical methods, and domain terminology.

Term	Definition
Active Inference	A framework in which agents minimize expected free energy to select actions, unifying perception, learning, and decision-making under the Free Energy Principle.
Assertion	A directed, confidence-scored claim linking a paper to a hypothesis (supports, contradicts, or neutral). The basic unit of evidence in the knowledge graph; a machine-extracted classification, not a human verdict.

Continued on next page

Term	Definition
Bayesian Mechanics	The formal extension of FEP that grounds Markov-blanket dynamics in stochastic physics, casting belief updates as gradient flows on a free-energy potential.
Canonical ID	The unique identifier assigned to each paper during deduplication, following DOI > arXiv ID > Semantic Scholar ID > OpenAlex ID > title hash.
Checkpoint	A JSON Lines snapshot of LLM extraction progress, recording which papers have been processed and the resulting assertions, enabling incremental resume after interruption.
Citation-Weighted Score	The hypothesis-level evidence aggregate combining direction, confidence, and a logarithmic citation weight (Eq. 3).
Compound Annual Growth Rate (CAGR)	The constant annual rate that, compounded over the publication window, takes the cumulative count from the first to the last year (Eq. 10).
Conceptual Nexus Model (CNM)	The modular knowledge-graph unit used by ResNei; each CNM packages concepts with provenance and supports longitudinal, latitudinal, and relational navigation.
Contrastive Divergence	An approximate gradient-based training algorithm for energy-based models [Hinton, 2002] that truncates the Markov chain used to estimate the gradient of the log-partition function.
Domain Timeline	Per-domain yearly publication counts visualizing temporal evolution across the eight tracked categories (A1–A2, B, C1–C5).
Doubling Time (t_d)	Years required for cumulative output to double under the prevailing growth rate (Eq. 9).
Energy-Based Model (EBM)	A class of generative models defining $p(x) \propto \exp(-E(x))$ for an unnormalized energy E. Includes Boltzmann machines, Helmholtz machines, and VAEs as special or related cases.
Expected Free Energy (EFE)	A scalar combining epistemic (uncertainty-reducing) and pragmatic (goal-achieving) value, minimized over policies. Decomposes equivalently into risk + ambiguity or epistemic + instrumental terms [Da Costa et al., 2020].
FAIR Principles	Findable, Accessible, Interoperable, Reusable: a set of guiding principles for scientific data infrastructure [Wilkinson et al., 2016]. The pipeline's nanopublications satisfy all four.
Free Energy Principle (FEP)	The principle that self-organizing systems minimize variational free energy—an upper bound on surprise—to maintain their structural integrity.
Generative Model	A probabilistic model specifying the joint distribution over hidden states and observations, encoding an agent's beliefs about how observations are generated.
Greedy Modularity Maximization	The Clauset-Newman-Moore algorithm [Clauset et al., 2004] for community detection. Implemented via NetworkX `greedy_modularity_communities`; applied here to the citation graph to identify clusters of densely interconnected papers.
HITS Hub/Authority Scores	Kleinberg's mutually reinforcing centrality metrics [Kleinberg, 1999]: hubs point to many authorities; authorities are pointed to by many hubs.
Helmholtz Machine	A generative model with separate recognition (bottom-up) and generative (top-down) networks trained by the wake-sleep algorithm [Dayan et al., 1995]; a direct precursor to the variational autoencoder and the FEP's recognition–generation duality.
Incremental Resume	The pipeline's ability to continue from where a previous run stopped, loading existing corpus and assertion snapshots and processing only new papers; controlled by `--clear-corpus` and `--clear-assertions` CLI flags.
Knowledge Graph	A directed graph encoding papers, assertions, hypotheses, and their relationships, serialized in an RDF-compatible format.

Continued on next page

Term	Definition
LLM Config	A configuration record specifying the Ollama model name, API URL, sampling temperature, maximum retries, and retry delay used by the assertion extractor.
Markov Blanket	A statistical boundary separating internal from external states, defined as the node set that renders a system conditionally independent of its environment.
Mean Year-over-Year Growth ($\bar{g}$)	Arithmetic mean of $(n_y - n_{y-1})/n_{y-1}$ across years with non-zero prior-year counts (Eq. 8).
Named Graph	An RDF graph identified by a URI, enabling multiple graphs to coexist in a single dataset. Nanopublications use four named graphs (Head, Assertion, Provenance, Publication Info).
Nanopublication	A minimal, self-contained unit of publishable knowledge consisting of an assertion, provenance metadata, and publication context [Groth et al., 2010, Kuhn et al., 2016].
NMF (Non-negative Matrix Factorization)	A factorization $V \approx WH$ with all factors non-negative, used here for unsupervised topic discovery (§8.2).
Ollama	A locally hosted LLM server used for assertion extraction; provides reproducibility and avoids external API dependencies [Ollama Team, 2024].
PageRank	A centrality metric originally designed for web-page ranking. In citation networks, PageRank surfaces influential papers that act as hubs across otherwise disconnected subgraphs.
Precision	The inverse variance of a probability distribution; in active inference, precision weighting determines the influence of prediction errors at each level of a hierarchy.
Predictive Coding	A scheme in which each cortical level passes prediction errors upward and predictions downward, minimizing local free-energy bounds layer by layer.
Progressive Parsing	The pipeline's three-stage JSON recovery strategy for malformed LLM output: (1) direct parse, (2) strip Markdown code fences and retry, (3) extract first `[…]` substring. Papers failing all three are logged and skipped.
Provenance	The recorded lineage of an assertion: source paper, extraction model, timestamp, and confidence; serialized in the Provenance named graph of each nanopublication.
Reference Resolution Rate	Fraction of all outgoing references that resolve to another paper inside the corpus; reported as 7.4% in the present analysis and used as a lower bound on intra-corpus citation density.
Stochastic Differential Equation (SDE)	A differential equation driven by a Wiener (white-noise) process; used in Bayesian-mechanics derivations of Markov-blanket dynamics.
Surprise (Self-Information)	The negative log probability of an observation under the agent's generative model; variational free energy is an upper bound on surprise.
Term Frequency–Inverse Document Frequency (TF-IDF)	A weighting that combines normalized term frequency with logarithmic inverse document frequency (Eq. 7); the standard input to NMF in this pipeline.
TriG	A serialization format extending Turtle with named-graph support, used to encode nanopublications as RDF datasets.
Trusty URI	A URI containing a cryptographic hash of its content [Kuhn and Dumontier, 2014], providing verifiable immutability and content-addressable identification for nanopublications.
Variational Free Energy (VFE)	An upper bound on surprise (negative log evidence) decomposable into complexity (KL from prior) and accuracy (expected log-likelihood).
Variational Inference	Approximate posterior inference by optimization, replacing intractable marginalization with optimization of a tractable variational distribution.

Continued on next page

Term	Definition
Ward Linkage	A hierarchical clustering method that minimizes total within-cluster variance at each merge step; used to compute domain-centroid dendrograms from mean TF-IDF vectors.
Wong Palette	The colorblind-safe 8-color palette of Wong (2011) [Wong, 2011], used as the standard visualization palette throughout all pipeline-generated figures.

11 References

The bibliography is generated automatically during PDF compilation from `references.bib`. All citation keys used in the manuscript (e.g., `\citep{friston2010free}`) resolve to entries below; unused entries have been pruned. Pandoc's `--natbib` flag injects `\usepackage{natbib}` and `\bibliographystyle{plainnat}`, so neither directive appears in this section or in `preamble.md`.

References

Dmitry Bagaev et al. RxInfer.jl v4.0.0: Real-time and adaptive bayesian inference, 2025. URL https://github.com/ReactiveBayes/RxInfer.jl.

Vladimir Baulin, Austin Cook, Daniel Friedman, Janna Lumiruusu, Andrew Pashea, Shagor Rahman, and Benedikt Waldeck. The discovery engine: A framework for AI-driven synthesis and navigation of scientific knowledge landscapes. *arXiv preprint arXiv:2505.17500*, 2025. doi: 10.48550/arXiv.2505.17500.

Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V. Rosenberg, and Jennifer Shirk. Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59 (11):977–984, 2009. doi: 10.1525/bio.2009.59.11.9.

Jelle Bruineberg, Krzysztof Dolega, Joe Dewhurst, and Manuel Baltieri. The emperor's new markov blankets. *Behavioral and Brain Sciences*, 45:e183, 2022. doi: 10.1017/S0140525X21002351.

Théophile Champion, Howard Bowman, and Peter Grünwald. Realizing active inference in variational message passing: The outcome-blind fixation of belief. *Neural Computation*, 33(10):2762–2826, 2021. doi: 10.1162/neco_a_01422.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013. doi: 10.1017/S0140525X12000477.

Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004. doi: 10.1103/PhysRevE.70.066111.

Matteo Colombo and Peggy Seriès. Free energy: a user's guide. *Biology & Philosophy*, 36(2):1–35, 2021. doi: 10.1007/s10539-021-09788-0.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2020. doi: 10.1016/j.jmp.2020.102447.

Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz machine. *Neural Computation*, 7(5):889–904, 1995. doi: 10.1162/neco.1995.7.5.889.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting LLMs: towards Retrieval-Augmented large language model. *arXiv preprint arXiv:2405.06211*, 2024.

Zafeirios Fountas, Noor Sajid, Pedro AM Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. In *Advances in Neural Information Processing Systems*, volume 33, pages 11662–11675, 2020.

Daniel Ari Friedman. A template/ approach to reproducible generative research: Architecture and ergonomics from configuration through publication. *Zenodo*, 2026a. doi: 10.5281/zenodo.19139090. URL https://doi.org/10.5281/zenodo.19139090. Active Inference Institute.

Daniel Ari Friedman. docxology/template. Software, GitHub, 2026b. URL https://github.com/docxology/template. Apache License 2.0. Infrastructure-as-Code system for reproducible computational research.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100 (1-3):70–87, 2006. doi: 10.1016/j.jphysparis.2006.10.001.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017. doi: 10.1162/NECO_a_00912.

Karl Friston, Lancelot Da Costa, Alexander Tschantz, Conor Heins, and Christopher Buckley. Active inference and artificial reasoning. *arXiv preprint arXiv:2512.21129*, 2025. doi: 10.48550/arXiv.2512.21129.

Karl J Friston, Thomas Parr, Yan Yufik, Noor Sajid, Cathy J Price, and Emma Holmes. Generative models, linguistic communication and active inference. *Neuroscience & Biobehavioral Reviews*, 118:42–64, 2020. doi: 10.1016/j.neubiorev.2020.07.005.

Francesco Gregoretti. A C++ implementation of active inference for POMDPs. *Neurocomputing*, 562:126888, 2023. doi: 10.1016/j.neucom.2023.126888.

Paul Groth, Andrew Gibson, and Jan Velterop. Anatomy of a nanopublication. *Information Services & Use*, 30(1-2): 51–56, 2010. doi: 10.3233/ISU-2010-0613.

Sarah Hamburg. A guide to DeSci, the latest Web3 movement. *Future*, 2022. URL https://future.com/what-is-decentralized-science-aka-desci/. Andreessen Horowitz (a16z) Research.

Conor Heins, Beren Millidge, Lancelot Da Costa, Stephen Mann, Karl Friston, Ozan Catal, Pablo Lanillos, Noor Sajid, and Alexander Tschantz. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022. doi: 10.21105/joss.04098.

Conor Heins, Beren Millidge, Daphne Demekas, Hrishit Basu, Noor Sajid, and Karl Friston. Collective behavior from surprise minimization. *Proceedings of the National Academy of Sciences*, 121(17):e2320239121, 2024. doi: 10.1073/pnas.2320239121.

Conor Heins, Toon Van de Maele, Alexander Tschantz, Hampus Linander, Dimitrije Markovic, Tommaso Salvatori, Corrado Pezzato, Ozan Catal, Ran Wei, Magnus Koudahl, Marco Perin, Karl Friston, Tim Verbelen, and Christopher Buckley. AXIOM: Learning to play games in minutes with expanding object-centric models. *arXiv preprint arXiv:2505.24784*, 2025. Submitted to ICLR 2026.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8): 1771–1800, 2002. doi: 10.1162/089976602760128018.

Jakob Hohwy. *The Predictive Mind*. Oxford University Press, 2013. ISBN 978-0-19-968273-7. doi: 10.1093/acprof:oso/9780199682737.001.0001.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Isabelle Belber, Jonathan Blaschke, Regan Chiang, Jenna Coffey, Arman Feldman, Joshua Gruber, et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023.

Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999. doi: 10.1145/324133.324140.

Virginia Bleu Knight, RJ Cordes, and Daniel Friedman. The free energy principle & active inference: a systematic literature analysis, 2022. URL https://zenodo.org/records/7449368. Active Inference Institute. Companion resources: https://github.com/ActiveInferenceInstitute/Knowledge-Engineering.

Knowledge Pixels. nanopub-js: JavaScript library for nanopublications. Software, GitHub, 2026. URL https://github.com/Nanopublication/nanopub-js. Version 0.1.0; enables browser-based creation, signing, and querying of nanopublications.

Franz Kuchling, Karl Friston, Georgi Georgiev, and Michael Levin. Morphogenesis as bayesian inference: A variational approach to pattern formation and body-plan diversity in biology. *Physics of Life Reviews*, 33:88–108, 2020. doi: 10.1016/j.plrev.2019.06.001.

Tobias Kuhn and Michel Dumontier. Trusty URIs: Verifiable, immutable, and permanent digital artifacts for linked data. In *The Semantic Web–ISWC 2014*, pages 395–410. Springer, 2014. doi: 10.1007/978-3-319-11964-9_25.

Tobias Kuhn, Christine Chichester, Michael Krauthammer, Núria Queralt-Rosinach, Ruben Verborgh, George Giannakopoulos, Axel-Cyrille Ngonga Ngomo, and Michel Dumontier. Decentralized provenance-aware publishing with nanopublications. *PeerJ Computer Science*, 2:e78, 2016. doi: 10.7717/peerj-cs.78.

Pablo Lanillos, Cristian Meo, Corrado Pezzato, Ajith Anil Meera, Mohamed Baioumy, Wataru Ohata, Alexander Tschantz, Beren Millidge, Martijn Wisse, Christopher L Buckley, and Jun Lenz. Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint arXiv:2112.01871*, 2021.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*, pages 191–246. MIT Press, 2006.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, 1999. doi: 10.1038/44565.

Michael Levin. Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16:768201, 2022. doi: 10.3389/fnsys.2022.768201.

Linhao Li et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Tianyi Liang, Weicheng Zhang, Chao Lin, et al. A survey of scientific information extraction with large language models. *arXiv preprint arXiv:2401.01512*, 2024.

Janna Lumiruusu, Daniel Friedman, Shagor Rahman, Vladimir Baulin, and Andrew Pashea. ResNei: Solution design document. Zenodo, 2025. URL https://zenodo.org/records/16753709. Version 2. Published 2025-08-08.

Beren Millidge. Deep active inference as variational policy gradients. *Journal of Mathematical Psychology*, 96:102348, 2020. doi: 10.1016/j.jmp.2020.102348.

Beren Millidge. A retrospective on active inference, July 2024. URL https://www.beren.io/2024-07-27-A-Retrospective-on-Active-Inference/. Blog post reviewing the state and trajectory of the Active Inference field.

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Understanding the origin of information-seeking exploration in probabilistic objectives for control. *arXiv preprint arXiv:2103.06859*, 2021.

Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *Neural Computation*, 34(6):1329–1368, 2022. doi: 10.1162/neco_a_01497.

Sanjeev V Namjoshi. aif-fep-db: A database of publications related to active inference and the free energy principle. https://github.com/snamjoshi/aif-fep-db, 2026a. Accessed: 2026-03-20.

Sanjeev V Namjoshi. *Fundamentals of Active Inference: Principles, Algorithms, and Applications of the Free Energy Principle for Engineers*. MIT Press, 3 2026b. ISBN 978-0-262-05095-1.

Morteza Nazemi, Sanket Bhatt, and Zafeirios Fountas. Active inference for energy control and planning under partial observability. *arXiv preprint arXiv:2503.07772*, 2025. doi: 10.48550/arXiv.2503.07772.

Samuel Nehrer et al. Introducing ActiveInference.jl: A Julia library for simulation and parameter estimation with active inference models. *Entropy*, 27(1):62, 2025. doi: 10.3390/e27010062. URL https://www.mdpi.com/1099-4300/27/1/62.

Ollama Team. Ollama: Run large language models locally. Software, 2024. URL https://ollama.com. Local inference server for open-weight LLMs.

Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022. ISBN 978-0-262-04535-4.

Marius Pazem, Simon Hangl, and Justus Piater. Free energy projective simulation (FEPS): Active inference with interpretable policy graphs. In *Workshop on Bridging the Gap Between Practice and Theory in Deep Learning, ICML 2024*, 2024. URL https://arxiv.org/abs/2407.05432.

Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, 134:17–35, 2015. doi: 10.1016/j.pneurobio.2015.09.001.

Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Viviana Fernanda Quevedo Tumailli et al. Combining knowledge graphs and large language models: A systematic literature review. *IEEE Access*, 2025.

Maxwell JD Ramstead, Paul B Badcock, and Karl J Friston. Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, 24:1–16, 2018. doi: 10.1016/j.plrev.2017.09.001.

Rajesh P N Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

RDFLib Team. rdflib: A python library for working with rdf. Software, 2023. URL https://rdflib.readthedocs.io/. Version 7.x.

David H. Rose and Anne Meyer. Universal design for learning. *Journal of Special Education Technology*, 15(1):67–70, 2000. doi: 10.1177/016264340001500108.

Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural Computation*, 33(3):674–712, 2021. doi: 10.1162/neco_a_01357.

Dalton AR Sakthivadivel. On bayesian mechanics: a physics of and by beliefs. *Interface Focus*, 13(3):20220029, 2023. doi: 10.1098/rsfs.2022.0029.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. doi: 10.1145/361219.361220.

Ryan Smith, Karl J Friston, and Christopher J Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022. doi: 10.1016/j.jmp.2021.102632.

Theodore D. Sterling. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54(285):30–34, 1959. doi: 10.1080/01621459.1959.10501497.

John Sweller, Paul Ayres, and Slava Kalyuga. Cognitive load theory. *Explorations in the Learning Sciences, Instructional Systems and Performance Technologies*, 2011. doi: 10.1007/978-1-4419-8126-4.

Alexander Tschantz, Anil K Seth, and Christopher L Buckley. Learning action-oriented models through active inference. *PLoS Computational Biology*, 16(4):e1007805, 2020. doi: 10.1371/journal.pcbi.1007805.

Bo Wen. The missing reward: Active inference in the era of experience. *arXiv preprint arXiv:2508.05619*, 2025. doi: 10.48550/arXiv.2508.05619.

Christopher J Whyte, Ryan Smith, and Jakob Hohwy. How do inner screens enable imaginative experience? A metacognitive active inference account. *Neuroscience of Consciousness*, 2025(1):niaf009, 2025. doi: 10.1093/nc/niaf009.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

Bang Wong. Points of view: Color blindness. *Nature Methods*, 8(6):441, 2011. doi: 10.1038/nmeth.1618.