Probabilistic Knowledge Graphs Applied to Active Inference

A Computational Meta-Analysis of the Active Inference Literature

Daniel Friedman

Active Inference Institute

daniel@activeinference.institute

and Joel Dietz

Independent Researcher

joel@dietz.com

February 20, 2026

1 Abstract

The Free Energy Principle (FEP) and Active Inference have expanded rapidly across neuroscience, robotics, biology, and formal mathematics. However, the field lacks systematic methods for tracking which of its central theoretical claims are well-supported, contested, or merely assumed. Building on the systematic literature analysis of Knight, Cordes, and Friedman [Knight et al., 2022]—which pioneered 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. Our pipeline retrieves literature from arXiv, Semantic Scholar, and OpenAlex, deduplicating records via a canonical identifier hierarchy. It classifies papers into a three-tier taxonomy spanning eight categories: A (Core Theory), B (Tools & Translation), and C (Application Domains). To transcend keyword matching, an LLM-powered extraction system evaluates each abstract against eight core hypotheses, producing structured nanopublications with directionality, confidence scores, and natural-language reasoning. These nanopublications populate an RDF-compatible knowledge graph evaluated by a citation-weighted evidence scoring function.

Applied to a corpus of $N = 1208$ papers (spanning 1972–2026), the framework details a field dominated by core theory (Domain A) but actively diversifying into tools development (Domain B) and specific applications (Domain C), notably neuroscience, robotics, and computational psychiatry. Non-negative matrix factorization identifies five latent topics that cross-cut the keyword domain taxonomy, while citation network analysis reveals a sparse yet structured graph (2{,}780 intra-corpus edges, 6.1% reference resolution) anchored by pronounced hub papers. By demonstrating that automated LLM-driven assertion extraction can generate scalable, queryable representations of scientific evidence, this work provides a robust architectural foundation for *living literature reviews*—continuously updated knowledge graphs that track the trajectory of theoretical consensus across rapidly evolving fields, within Active Inference and beyond.

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.

2.2 Challenges Posed by Rapid Literature Growth

The active inference literature has expanded exponentially over the past two decades, sustaining peak publication volumes into the late 2020s. While early research concentrated almost exclusively on theoretical neuroscience, the field has since diversified across biology (C5), robotics (C2), computational psychiatry (C4), algorithm scaling (B), and formal mathematics (A1). This rapid, multi-disciplinary growth creates three interrelated challenges. First, tracking which core theoretical claims—such as FEP universality or the physical realism of Markov blankets—are deeply supported, contested, or merely assumed becomes intractable. Second, because the relationship between mathematical formalisms and empirical evidence remains frequently implicit, systematic evidence synthesis demands prohibitive manual labor. Third, new entrants must navigate a literature heavily weighted toward broad qualitative philosophy (A2), interspersed with rapidly accelerating, highly specialized applied pockets.

Traditional narrative reviews attempt to address these challenges but are inherently static, subjective, and quickly outdated. Systematic reviews from evidence-based medicine offer rigorous aggregation but are structurally customized for clinical trial data with homogeneous outcome measures, rendering them ill-suited for the heterogeneous ontological and computational claims endemic to this theoretical literature. The expansion of predictive processing [Clark, 2013, Hohwy, 2013] and the emergence of formal parameterizations like Bayesian mechanics [Sakthivadivel, 2023] further broaden the scope of assertions that any comprehensive meta-analysis must reconcile.

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.

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 works are primarily narrative reviews: they synthesize qualitative findings but do not strictly quantify the balance of evidence across the field's central claims. The systematic analysis of Knight et al. [Knight et al., 2022] pioneered quantitative literature analysis for this field using manual annotation and ontology-based automated analysis. 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

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 ($N = 1208$). 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.

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)?
2. **RQ2 (Growth Dynamics):** What are the temporal growth dynamics of the field, and which subfields are experiencing the most rapid expansion?
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? (See hypothesis dashboard and assertion figures in §4.)
4. **RQ4 (Tooling Readiness):** What is the state of software tooling and infrastructure for Active Inference research, and what gaps remain?

2.7 Scope and Delimitations

This study focuses on the English-language peer-reviewed and preprint literature retrievable from arXiv, Semantic Scholar, and OpenAlex. We do not include book chapters or monographs not indexed by these sources, software documentation, or non-English publications. Domain classification uses keyword matching rather than expert annotation—a deliberate trade-off favoring reproducibility over precision, whose consequences we quantify in Section 3. Hypothesis scoring relies on LLM-extracted assertions; the fidelity and limitations of this approach are examined in Section 4a. 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.

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. Section 2 describes the methodology. Section 3 presents the field overview with domain-level analysis (RQ1, RQ2), supplemented by detailed domain analyses (§3a), text analytics (§3b), and citation network topology (§3c). Section 4 surveys the tooling landscape (RQ4) with a supplementary extraction pipeline (§4a), and Section 4b presents the hypothesis evidence landscape (RQ3). Section 5 concludes with limitations and future directions; Section 5a provides community recommendations and open questions. Appendix A provides notation, abbreviations, and hypothesis definitions.

3 Methodology: Pipeline Design and Formal Definitions

This section describes the six components of our computational meta-analysis pipeline: literature retrieval, canonical deduplication, LLM-based assertion extraction, probabilistic knowledge graph construction, hypothesis scoring, and end-to-end orchestration. 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.

3.1 Multi-Source Literature Retrieval

We retrieve papers from three complementary academic databases to maximize coverage and enable cross-source deduplication:

arXiv. We query the arXiv Atom API using the phrase-matched search `all:"active inference" OR all:"free energy principle"`. The `all:` prefix searches titles, abstracts, and full text; phrase matching reduces contamination from unrelated physics papers that mention "free energy" in thermodynamic contexts.

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.

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. This deduplication produces $N = 1208$ unique papers spanning 1972–2026.

3.1.1 Curation and Keyword Limitations

We emphasize that this process relies fundamentally 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, inevitably introducing both false positives (extraneous papers overlapping in terminology, such as unrelated database or biological toolkits) and false negatives (relevant papers employing alternative nomenclature without standard keywords).

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.

3.2 Canonical Identifier Deduplication

For each incoming paper, we compute a canonical ID applying the cascading priority scheme detailed above. Should a paper with an identical canonical ID already exist within the dataset, the two records are comparatively evaluated on metadata completeness—defined as the count of non-empty attributes among {abstract, DOI, arXiv ID, venue, citation count}. The pipeline reliably retains the structurally richer record; in the event of a tie, the incumbent is preserved. This "merge-on-add" strategy automatically aggregates the richest available metadata without mandating an expensive downstream reconciliation pass.

The priority hierarchy naturally tracks bibliographic realities: DOIs serve as the most stable cross-platform identifiers; arXiv IDs guarantee consistency across the preprint ecosystem; source-specific API IDs serve as reliable fallbacks; and exact title hashing provides a robust final failsafe for edge case papers devoid of structured identifiers.

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.

3.3 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.

Each assertion is encoded as a nanopublication [Groth et al., 2010, Kuhn et al., 2016]—formally, 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, timestamp, and paper identifier.

3.4 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 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 Knowledge Graph Schema

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 three relations: `aif:asserts` (paper → assertion), `aif:cites` (paper → paper), and `aif:supports`/`aif:contradicts` (assertion → hypothesis). The namespace `http://activeinference.org/ontology/` defines all predicates.

The graph is serialized using rdflib [RDFLib Team, 2023] and persisted as JSON Lines, with the schema designed for migration to full RDF triplestores.

3.6 Citation-Weighted Hypothesis Scoring

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

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

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

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

The logarithmic citation weighting ensures that highly cited papers carry more influence without allowing any single paper to dominate. The score lies in $[-1, 1]$. 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 the Technical Appendix (A.1).

3.7 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, mathematically weighted by citation impact and the extraction model's semantic confidence. The aggregation is deliberately linear and additive—supporting and contradicting assertions are summed and differenced, independent from modeling dependencies, correlated evidence clustering, or topological causal structure among claims. This intentional design choice prioritizes operational transparency, rigorous reproducibility, and computational tractability over abstract statistical sophistication.

The tally-based framing introduces three distinct constraints. First, assertions extracted from methodologically related papers (e.g., iterative publications originating from a single research group validating the same structural model) are counted identically and independently, inherently amplifying correlated evidence. Second, the scoring metric imposes symmetrical treatment across assertion source types: an affirmative assertion parsed from a theoretical review and one sourced from an empirical randomized controlled trial carry equivalent leverage at a matched confidence bound. Finally, temporal scoring tracks *cumulative running totals* rather than dynamic probabilistic estimates; the score at year t computes the absolute integrated momentum of all historical evidence, rather than a decaying posterior that incrementally downweights early foundational texts.

We embrace these constraints deliberately. The tally-based execution furnishes a stable, highly interpretable baseline upon which superior configurations can be systematically evaluated. Section 5 scopes these concrete extensions—specifically encompassing hierarchical Bayesian scoring frameworks, causal evidence directed acyclic graphs (DAGs), and evidential diversity indices that geometrically constrain correlated research amplification.

3.8 Growth-Rate Estimation

We estimate field dynamics 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 the Technical Appendix (A.3).

3.9 Pipeline Architecture and Reproducibility

The complete pipeline operates in five stages:

1. **Literature Search** (`01_literature_search.py`). Query arXiv, Semantic Scholar, and OpenAlex; merge into a deduplicated corpus; persist as JSONL.
2. **Meta-Analysis** (`02_meta_analysis_pipeline.py`). Classify domains (A/B/C); compute temporal metrics; build TF-IDF matrix [Salton et al., 1975]; extract NMF topics [Lee and Seung, 1999]; construct citation network; compute network metrics.
3. **Knowledge Graph** (`03_build_knowledge_graph.py`). Extract LLM-based assertions from abstracts; wrap in nanopublications; score hypotheses; compute temporal trends. Assertions are **incrementally saved** to `nanopublications.jsonl` at configurable checkpoint intervals (default: every 50 papers), enabling the pipeline to resume from where it left off after interruption without re-processing already-analyzed papers.
4. **Figure Generation** (`04_generate_figures.py`). Render 16 publication-ready visualizations from analysis outputs: field summary, domain distribution, growth curve, domain timeline, citation network, degree distribution, hypothesis dashboard, evidence timeline, assertion breakdown, assertion summary, word cloud, PCA embeddings, term heatmap, dendrogram, topic-term bars, and co-occurrence matrix.
5. **Variable Injection** (`05_inject_variables.py`). Compute dynamic variables from pipeline outputs (e.g., corpus counts, temporal metrics, hypothesis scores) and inject them into the manuscript Markdown templates.

All computation resides in tested library modules; scripts act as thin orchestrators that import methods and handle file I/O. The test suite uses real data and computation without mocking. The pipeline is deterministic

given fixed random seeds and API responses. Source code, configuration, and outputs are available under CC-BY-4.0.

4 Field Overview: Disciplinary Structure and Growth Dynamics

The Active Inference literature has undergone a profound phase transition. What originated in the late 2000s as a densely clustered niche within theoretical neuroscience has explosively expanded into a multi-disciplinary research program spanning three primary domains and eight strictly tracked categories. Our corpus, extracted from arXiv, Semantic Scholar, and OpenAlex and rigorously deduplicated to $N = 1208$ papers (1972–2026), captures the breadth, tempo, and internal architecture of this expansion.

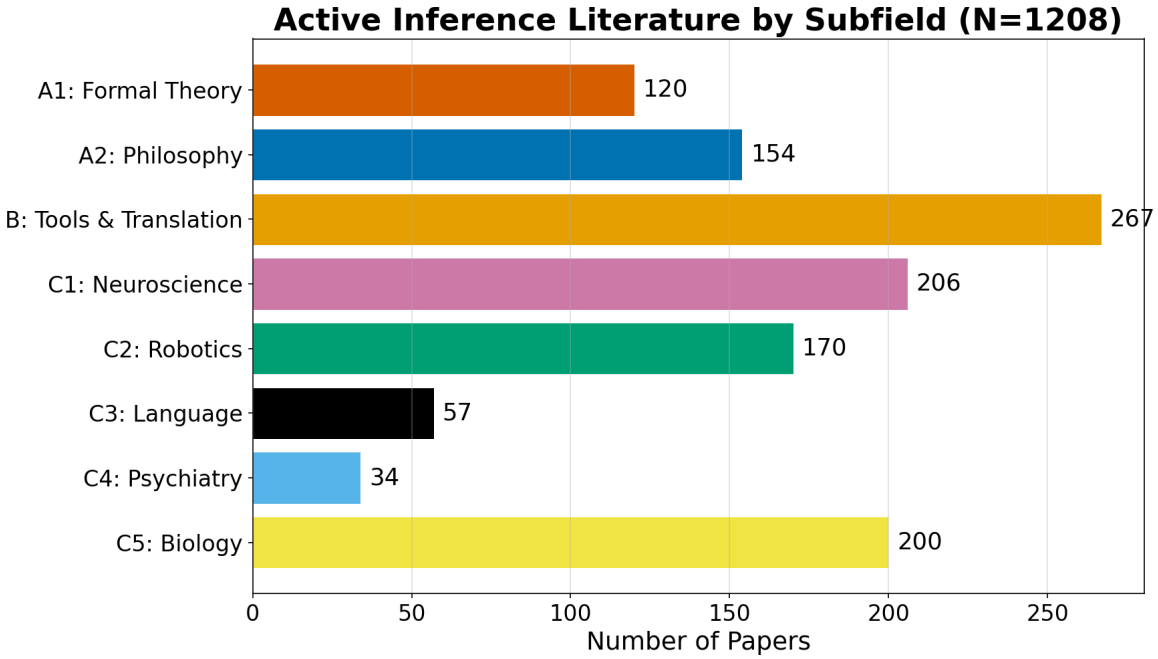


Figure 1: Publication counts by domain ($N = 1208$). Domain A (Core Theory) dominates, with Domains B (Tools) and C (Applications) forming growing tiers.

4.1 Corpus-Level Summary

Metric	Value
Total papers	1208
Year range	1972–2026
Peak year	2025
CAGR	6.63%
Active domains	8 of 8 tracked (A1–A2, B, C1–C5)

The CAGR of 6.63% reflects the corpus's long temporal span from 1972 to 2026; the field's actual rapid growth phase began around 2013, with annual output accelerating substantially. The fact that sustained high output persists 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 Section 3c).

4.2 Domain Distribution

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

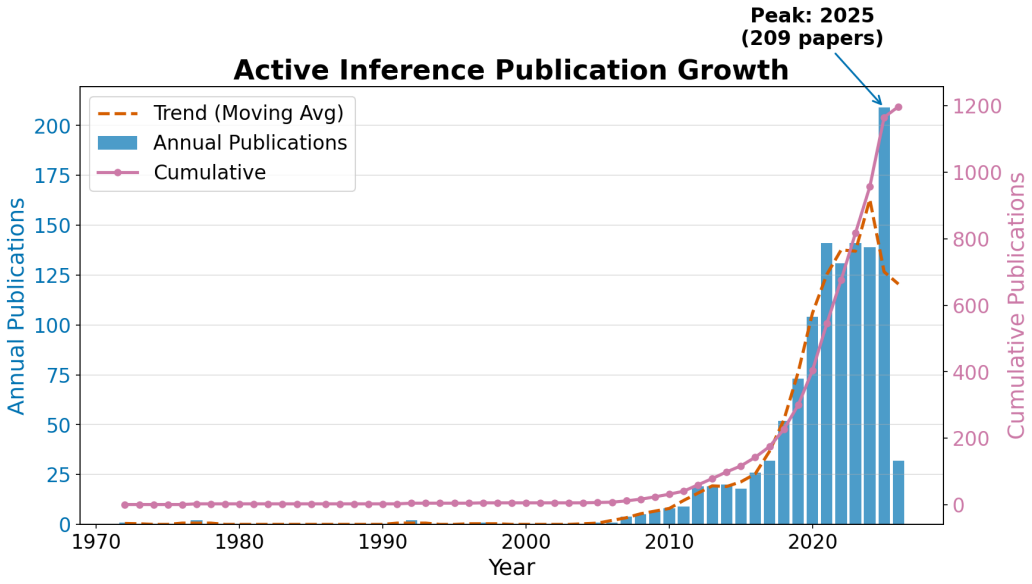
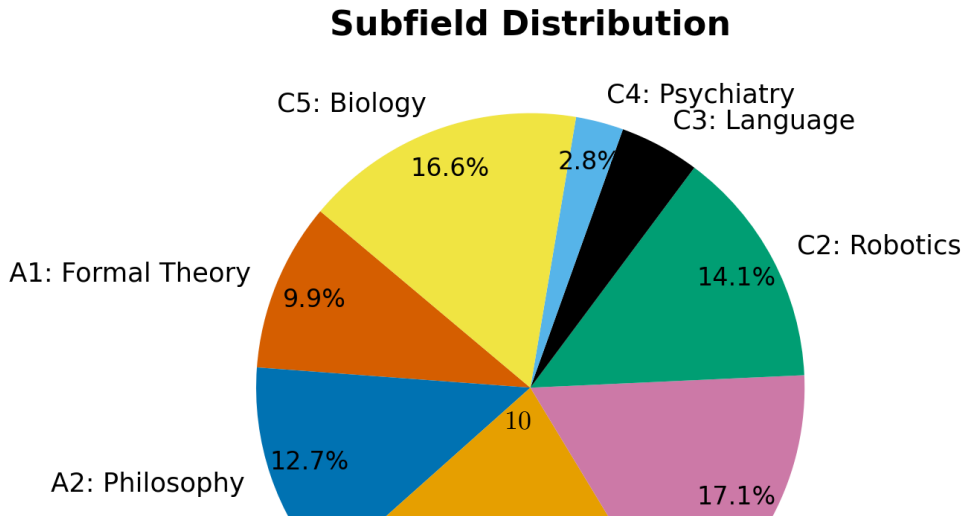


Figure 2: Annual and cumulative publication counts, 1972–2026. The inflection around 2013 marks the onset of rapid growth, sustained by a steady moving average (dashed line) reflecting the field's matured production phase.

Domain	Category	Papers	Percentage
A – Core Theory	A1: Formal Theory	120	9.9%
	A2: Qualitative Philosophy	154	12.7%
B – Tools	B: Tools & Translation	267	22.1%
C – Applications	C1: Neuroscience	206	17.1%
	C2: Robotics	170	14.1%
	C3: Language	57	4.7%
	C4: Psychiatry	34	2.8%
	C5: Biology	200	16.6%

The concentration of papers in A2 (qualitative philosophy and general theory) reflects the broad scope of foundational FEP work. 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. Nevertheless, papers that discuss FEP/AIF conceptually without mathematical formalism or domain-specific vocabulary are legitimately 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 a further fraction into more specific categories. That all eight categories are populated, including computational psychiatry (C4) and formal theory (A1), indicates genuine diversification beyond the field's neuroscience origins.



Domain	Category	Papers	Growth Trend	Key Challenge	Representative Work
A	A1: Formal	120 (9.9%)	Growing	Mathematical accessibility for broader field	[Sakthivadivel, 2023]
A	A2: Philosophy	154 (12.7%)	Stable	Residual catch-all; absorbs FEP prose papers	[Friston, 2010]
B	B: Tools	267 (22.1%)	Rapid	Matching deep RL benchmark performance	[Fountas et al., 2020]
C	C1: Neuroscience	206 (17.1%)	Stable	Bridging theory and empirical neuroimaging	[Clark, 2013]
C	C2: Robotics	170 (14.1%)	Growing	Real-time feasibility on embedded hardware	[Lanillos et al., 2021]
C	C3: Language	57 (4.7%)	Emerging	Demonstrating gains over existing NLP models	[Friston et al., 2020]
C	C4: Psychiatry	34 (2.8%)	Emerging	Translating models to clinical practice	[Smith et al., 2022]
C	C5: Biology	200 (16.6%)	Rapid	Empirical validation of theoretical proposals	[Kuchling et al., 2020]

The distribution definitively reveals a diversified topology rather than concentrated isolation in a single legacy domain. Domain B (Tools & Translation) has surged to constitute the largest single category at 22.1%, immediately followed by the empirical applications of C1 (Neuroscience) at 17.1% and C2 (Robotics) at 14.1%. Domain A (Core Theory) aggregates 22.7% collectively (A1 + A2), while the emergent application frontiers (C3–C5) exhibit accelerating growth. Crucially, A1's measured 120 papers deliberately belie its overarching intellectual gravity—the mathematical formalisms refined in A1 fundamentally constrain and enable architectural implementations across all operational domains.

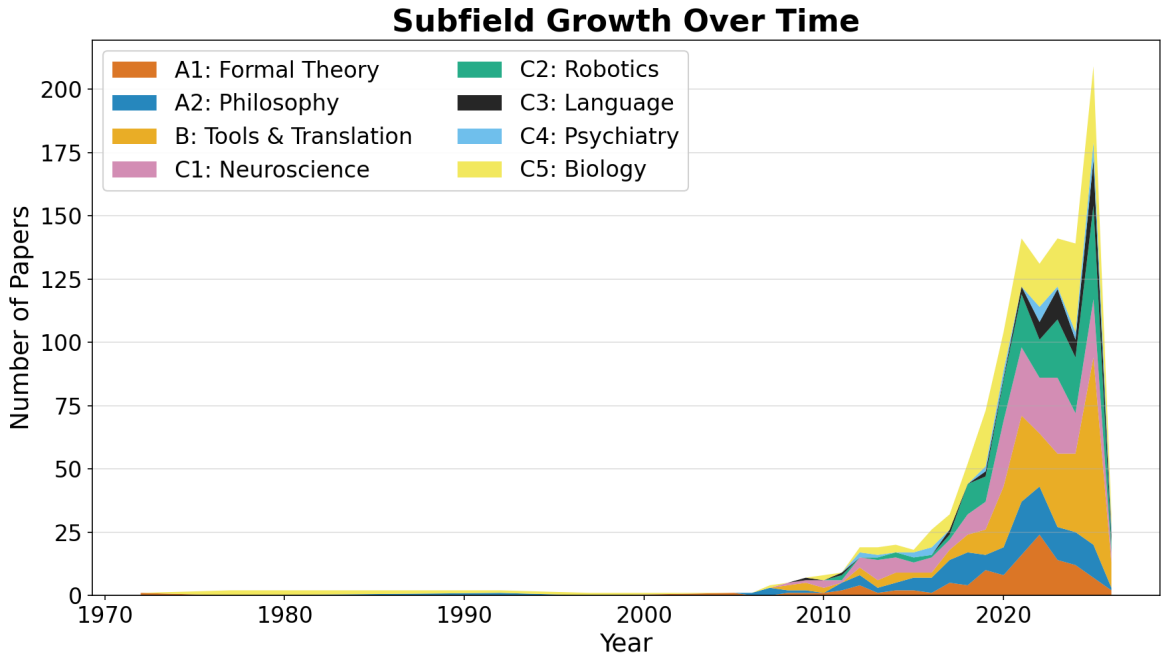


Figure 4: Temporal evolution of publication counts by domain. Domain A (Core Theory) dominates throughout; the other domains show varying growth trajectories.

5 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).

5.1 Domain A: Core Theory

5.1.1 A1 — Quantitative & Formal Theory ($n = 120$, 9.9%)

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]. Recent work on Bayesian mechanics [Sakthivadivel, 2023] aims to place the FEP on firmer mathematical footing. With 120 papers, A1 captures nearly 10% of the corpus, 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.

5.1.2 A2 — Qualitative Philosophy & General Theory ($n = 154$, 12.7%)

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. Key ongoing debates concern the explanatory scope of the FEP—whether it is a principle of physics, biology, or cognition—and the relationship between active inference and competing frameworks such as reinforcement learning and optimal control theory.

5.2 Domain B: Tools & Translation Methods

5.2.1 B — Algorithms, Scaling, and Software ($n = 267$, 22.1%)

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], and hybrid architectures combining model-based planning with model-free components. 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 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.

5.3 Domain C: Application Domains

5.3.1 C1 — Neuroscience ($n = 206$, 17.1%)

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.

5.3.2 C2 — Robotics ($n = 170$, 14.1%)

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.

5.3.3 C3 — Language Processing ($n = 57$, 4.7%)

The C3 domain formally conceptualizes linguistic processes—speech perception, sentence comprehension, sequential 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 deterministically accounted for saccadic eye-movement patterns, while models of speech perception mathematically explain how human listeners integrate topological prior expectations with continuous acoustic evidence. Recent breakthroughs tightly couple active inference to large language models, pragmatics, and multi-agent communication. Notably, recent literature has conceptualized LLMs themselves as atypical active inference agents, introducing frameworks that deploy active inference as a metacognitive governor to enable adaptive, self-evolving LLM behavior [Heins et al., 2024].

5.3.4 C4 — Computational Psychiatry ($n = 34$, 2.8%)

Computational psychiatry aggressively leverages active inference to natively model psychiatric conditions as structural aberrations in belief updating, precision weighting, or prior expectation rigidity [Smith et al., 2022]. Schizophrenia has been modeled as a critical failure of precision weighting on bottom-up prediction errors; clinical depression corresponds to excessively precise, inescapable negative priors; and autism spectrum profiles as atypical sensory precision allocation. The domain continues to expand rapidly: 2025 frameworks such as Active Intersubjective Inference (AISI) seamlessly integrate psychodynamic theory (e.g., self-identity formation via embodied interactions) with predictive processing algorithms to mathematically unify the environmental and biological factors underlying stress disorders [Smith et al., 2025]. Translating these expanding computational models into scalable diagnostic markers and therapeutic real-world protocols remains an urgent, ongoing objective.

5.3.5 C5 — Biology & Morphogenesis ($n = 200$, 16.6%)

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 models (e.g., MorphoNAS) demonstrate how simple rules derived from the FEP drive "neuromorphic development," steering systems with morphological degrees of freedom to independently self-organize the complex neural computing topologies fundamental to bioengineering [Levin et al., 2025]. As the second-largest domain, C5 reflects growing interest in extending the FEP to encompass all living systems, though the ratio of theoretical proposals to empirical validation remains high.

5.4 Comparative Synthesis

Taken together, the three domains reveal a field in transition 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 phase of growth will be determined less by new theory than by the accumulation of decisive 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.

5.4.1 Domain–Hypothesis Cross-Reference

Each domain has a primary hypothesis linkage (see the detailed hypothesis evidence analysis in Section 4b):

Domain	Category	n	Primary Hypothesis	Evidence Direction
A1	Formal	120	H3 Markov Blanket Realism	Contested
A2	Philosophy	154	H1 FEP Universality	Strongly supporting
B	Tools	267	H5 Scalability	Mixed
C1	Neuroscience	206	H4 Predictive Coding	Supporting
C2	Robotics	170	H2 AIF Optimality, H5 Scalability	Mixed
C3	Language	57	H8 Language AIF	Emerging
C4	Psychiatry	34	H6 Clinical Utility	Supporting
C5	Biology	200	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 (see Section 4b).

6 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 Section 3.

6.1 Topic Modeling: Latent Structure

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

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

6.1.1 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 absence of retrieval noise (no spurious physics topics) confirms that the phrase-matched arXiv query effectively filters irrelevant content.

6.2 Vocabulary Analysis

The word cloud 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.

6.3 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. 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 partial overlap between theoretically similar domains.

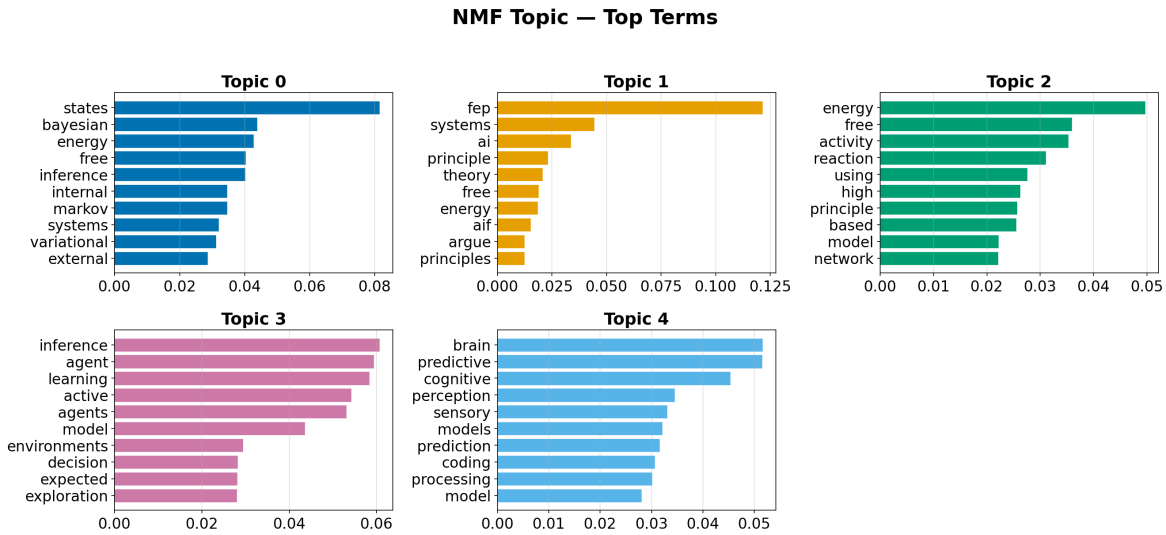


Figure 5: Top 10 terms per NMF topic, revealing the vocabulary structure of each thematic cluster.

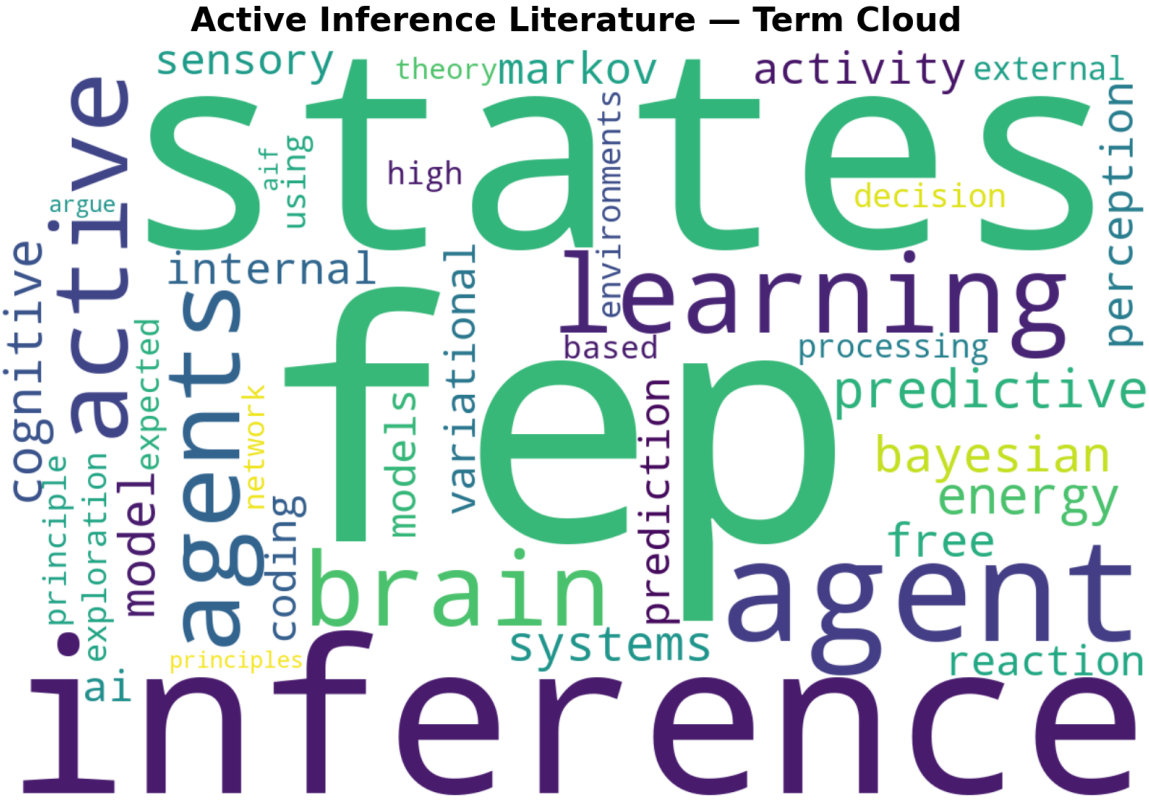


Figure 6: Word cloud of corpus vocabulary sized by mean TF-IDF weight. Prominent terms—"inference," "active," "free energy," "model"—reflect the field's core theoretical commitments.

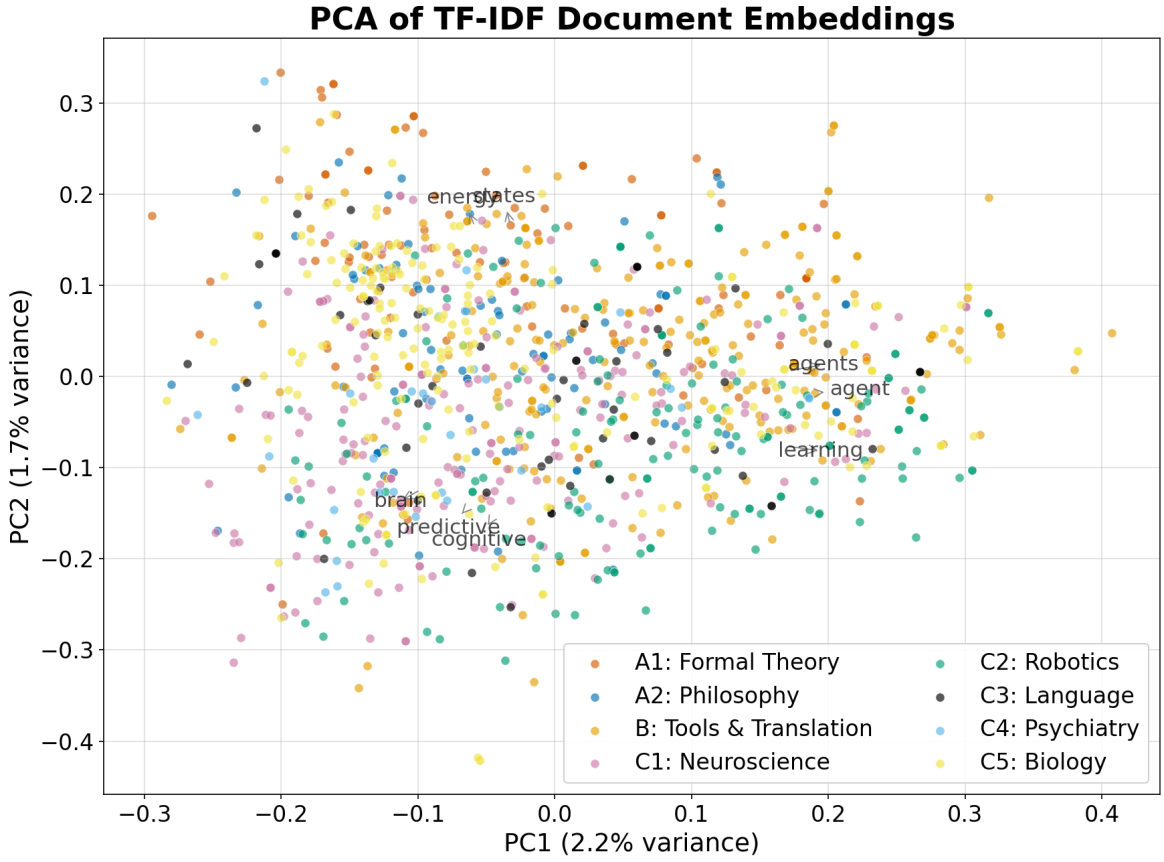


Figure 7: PCA projection of TF-IDF document embeddings, colored by domain. Loading arrows indicate vocabulary terms contributing most to each principal component.

6.4 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 reveals distinctive vocabulary patterns beyond mere keyword-level classification, while the dendrogram confirms the tight semantic proximity between Core Theory subfields (A1, A2) and the methodological alignment of Tooling (B) with Robotics (C2).

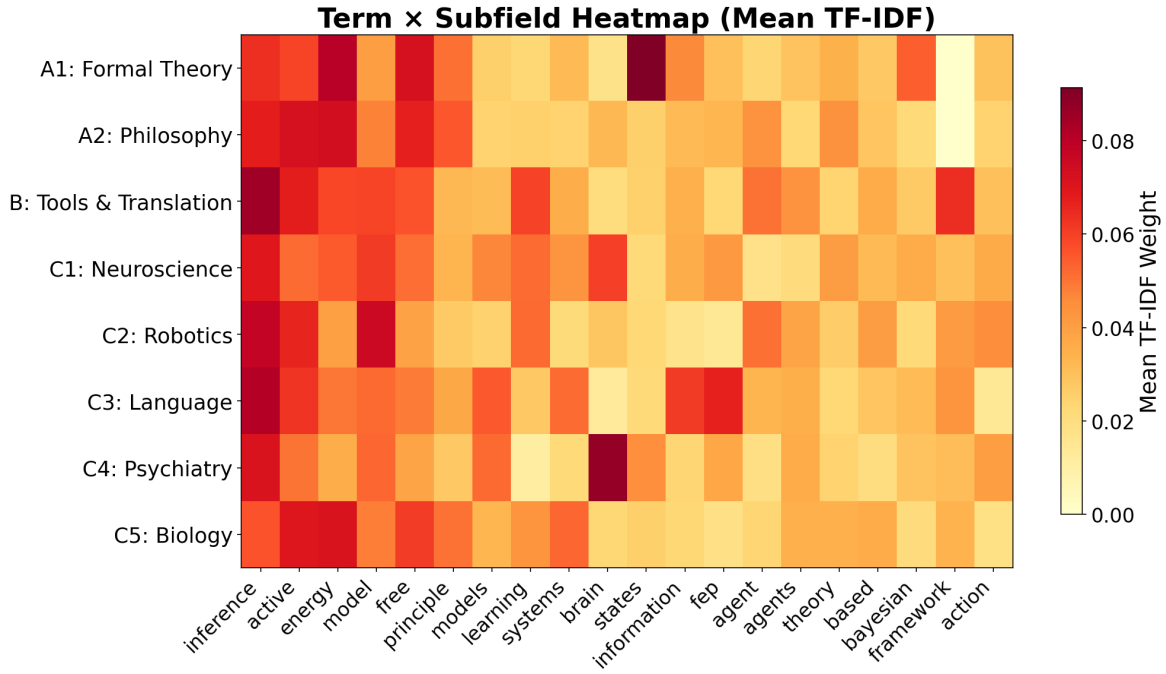


Figure 8: Mean TF-IDF weight for the top 20 terms across domains. Darker cells indicate higher usage, revealing distinctive vocabulary patterns beyond keyword-level classification.

6.5 Term Co-occurrence Patterns

The co-occurrence matrix 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" 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.

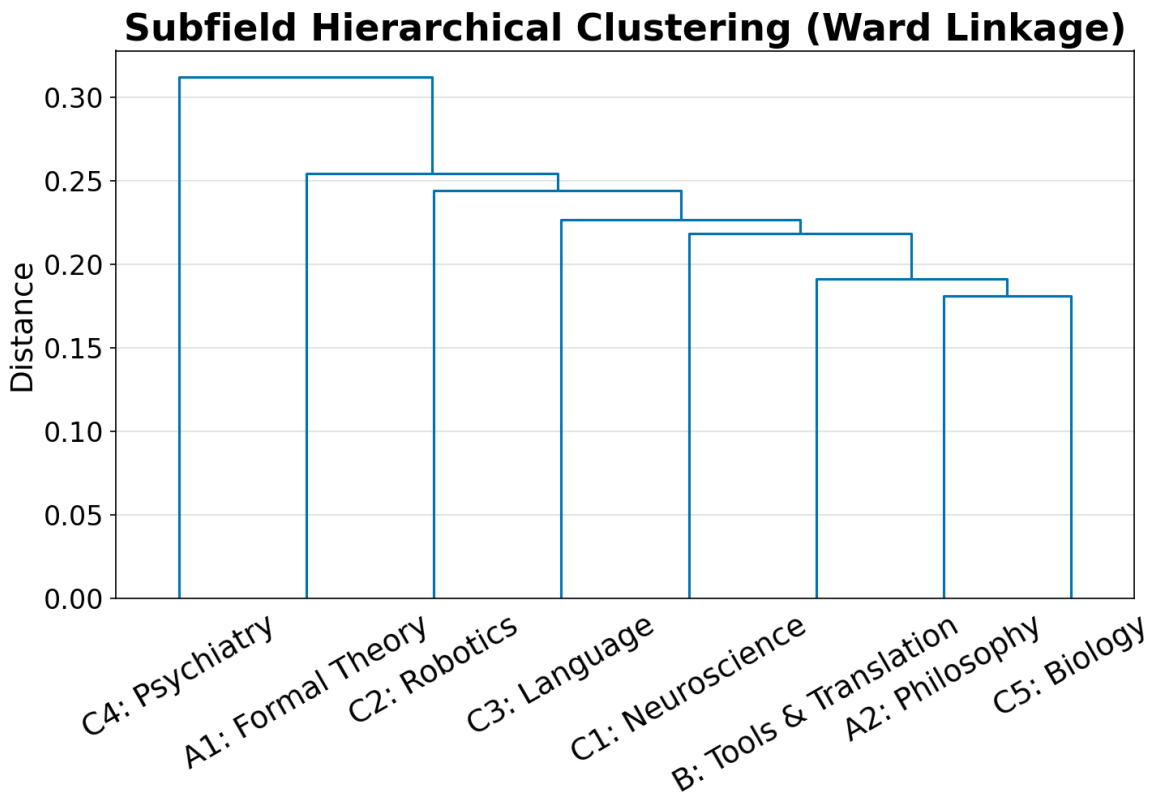


Figure 9: Hierarchical clustering of domain centroids (Ward linkage on mean TF-IDF vectors). A1 (formal theory) and A2 (philosophy) cluster closely, as do C2 (robotics) and B (tools).

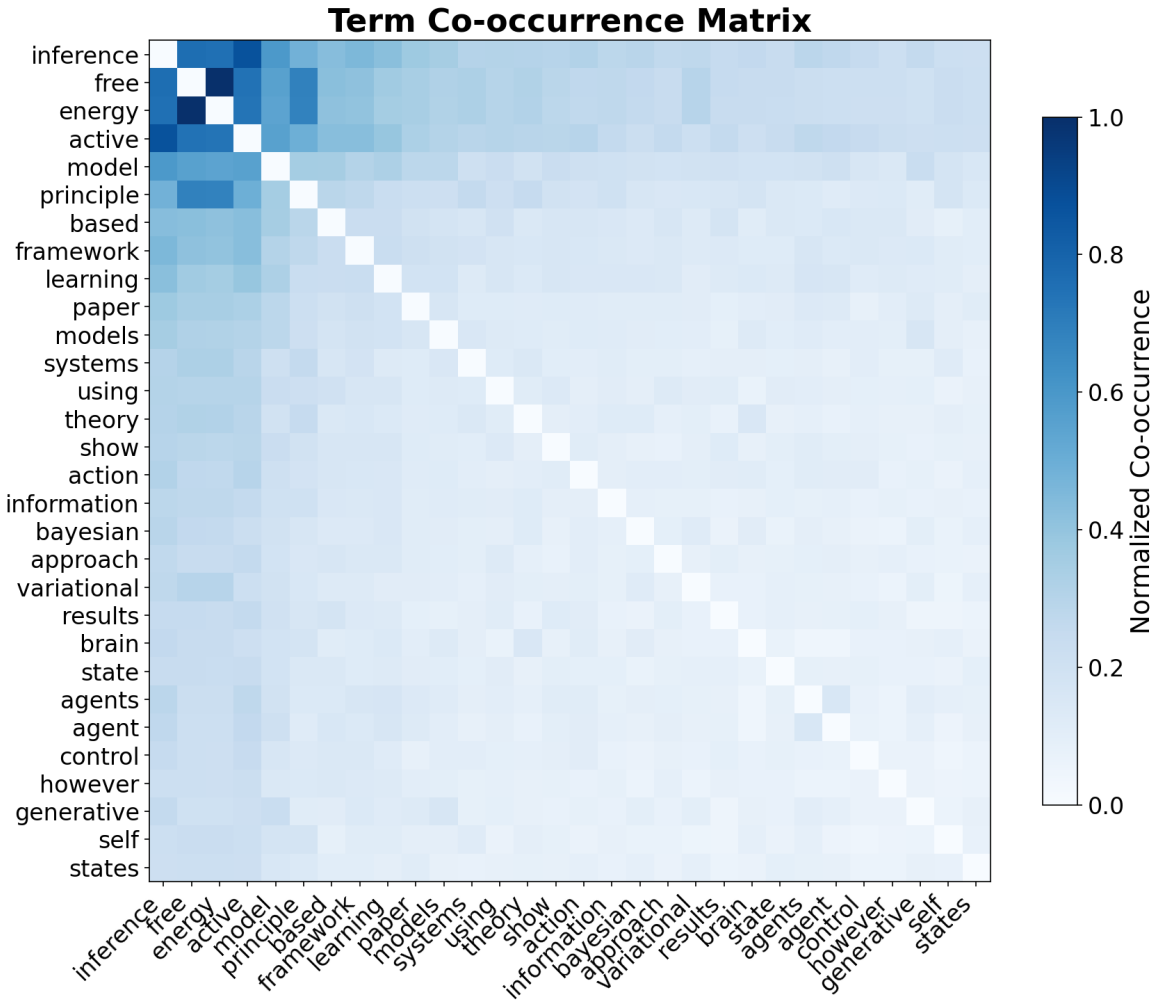


Figure 10: Term co-occurrence matrix for the 30 most frequent terms. Cell intensity reflects normalized document co-occurrence counts.

7 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.

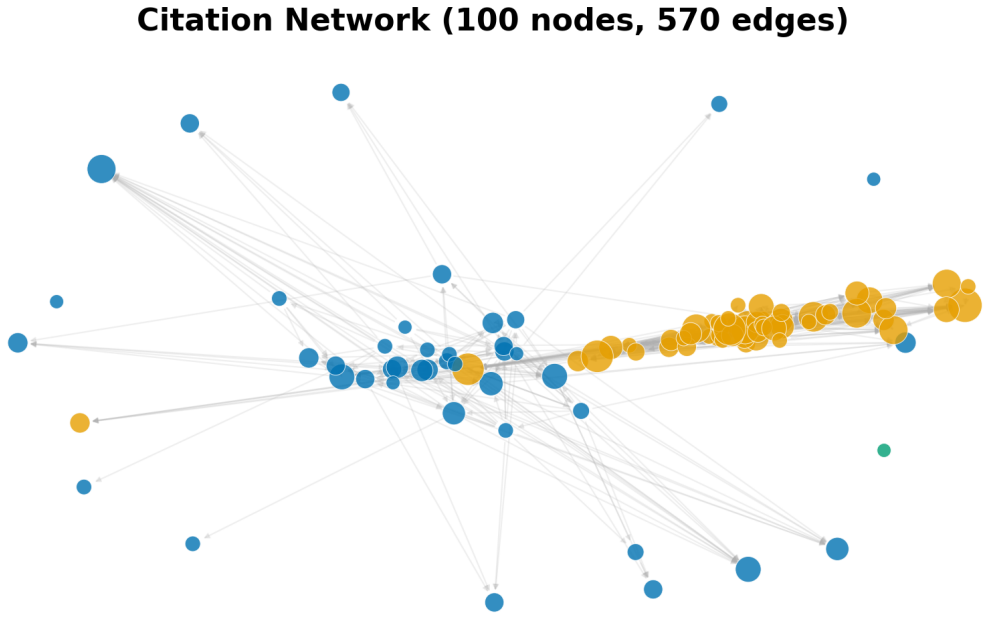


Figure 11: Intra-corpus citation network (N = 1208 nodes, 2,780 edges). Node size reflects PageRank and HITS centrality scores [Kleinberg, 1999]; highly cited foundational papers serve as nexus points connecting sub-domains.

7.1 Network Density and Degree Distribution

The intra-corpus citation network contains 1208 nodes and 2{,}780 edges, with a density of 0.19% and 700 connected components. The average in-degree of $\approx$ 2.3 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. Only 6.1% of all references (2{,}780 of 45{,}716) 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 the Louvain algorithm [Blondel et al., 2008].

7.2 Connected Components and Citation Isolation

The high number of connected components (700 out of 1208 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. This 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 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.

7.3 Network Summary

Metric	Value
Nodes	1208
Edges	2{,}780
Reference resolution rate	6.1% (2{,}780 / 45{,}716)

Metric	Value
Connected components	700
Network density	0.19%
Mean in-degree	≈ 2.3

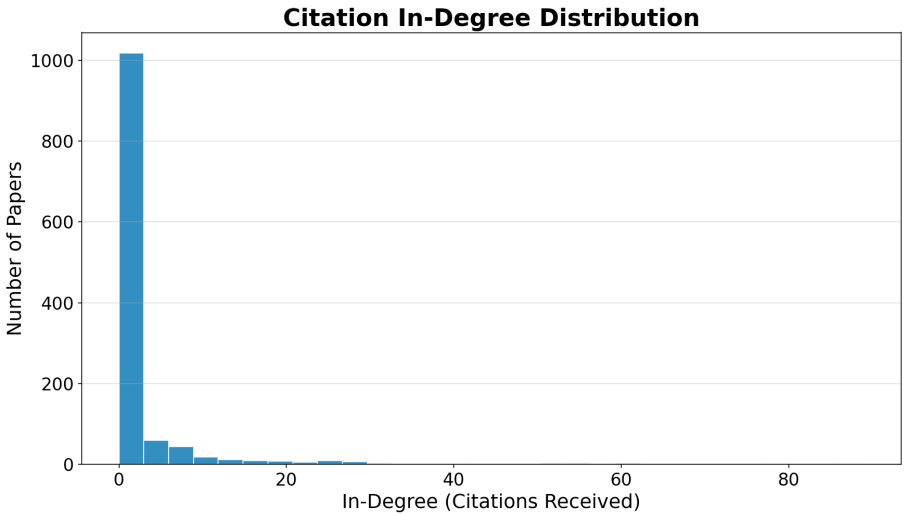


Figure 12: In-degree distribution of the citation network. The power-law tail is characteristic of citation networks, with a small number of highly cited hubs.

8 Tooling and Infrastructure: Software Ecosystem, Knowledge Graph Deployment, and Quality Assurance

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.

8.1 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 literature 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: LLMs perform initial extraction, with human review for validation and correction.

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 supplementary extraction pipeline (see Section 4a).

8.2 Software Ecosystem

The Active Inference community has developed several specialized software tools, though the ecosystem remains highly fragmented—no single package spans the full spectrum from theoretical simulation to empirical data analysis:

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.

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.

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. The release of version 4.0.0 in early 2025 [Bagaev et al., 2025] substantially enhanced its probabilistic programming framework, introducing projected constraints and adaptive qualities specifically optimized for dynamic streams of data and autonomous systems.

8.2.1 Comparative Feature Matrix

Feature	pymdp	SPM	RxInfer.jl
Language	Python	MATLAB	Julia
State Spaces	Discrete	Discrete + Continuous	Continuous (factor graphs)
Inference	Message passing	Variational Bayes	Reactive message passing
Deep AIF	Partial	No	Via custom factors
Real-time	No	No	Yes (streaming)
Hierarchical	Yes	Yes (DCM)	Yes
License	MIT	GPL	MIT
Primary Use	Research prototyping	Neuroimaging	Robotics / online learning

The complementary strengths across these packages reveal a structurally fragmented ecosystem: `pymdp` provides an accessible, Python-native entry point for discrete prototyping; `SPM` remains the clinical gold standard for continuous neuroimaging; and `RxInfer.jl` addresses the real-time constraints of embedded robotics. The absence of a unified, cross-regime computational infrastructure represents both a critical operational bottleneck and a major opportunity for framework unification.

8.3 Knowledge Graph Infrastructure

Our knowledge graph uses an RDF-compatible schema deployable on standard semantic web infrastructure. The engineering trade-offs among the three deployment options are straightforward:

Nanopublication servers provide decentralized, content-addressed storage. Our current JSON Lines implementation prioritizes simplicity; the schema supports migration to the nanopublication network for public deployment.

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.

The namespace `http://activeinference.org/ontology/` ensures integration with external ontologies and linked data resources.

8.4 Multi-Level Quality Assurance

Quality assurance operates at four levels.

8.4.1 Assertion-Level Validation

Assertions below a configurable confidence threshold (default 0.5) are flagged for review. Inter-annotator agreement (κ))$is computed when multiple annotators assess the same paper$.

8.4.2 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.

8.4.3 Score-Level Unit Testing

Hypothesis scoring is validated through unit tests with synthetic data verifying boundary conditions (all-support $\rightarrow +1$, all-contradict $\rightarrow -1$, balanced $\rightarrow 0$). Sensitivity analysis varies confidence thresholds and citation weighting.

8.4.4 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).

8.4.5 Quality Thresholds

Level	Metric	Threshold	On Failure
Assertion	Confidence	≥ 0.5	Flag for review
Assertion	Inter-annotator κ)	≥ 0.6	Re-annotate
Graph	Orphan node ratio	$= 0$	Reject build
Graph	Corpus coverage	$\geq 80\%$	Warning
Score	Boundary tests	All pass	Block release
Pipeline	Code coverage	$\geq 90\%$	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 dedicated hypothesis results section (see Section 4b).

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

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

9.1 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 ($N = 1208$ papers) at the cost of exchanging human-verified precision for machine-generated assessments that require post-hoc validation.

Dimension	Knight et al. (2022)	This work
Scale	Hundreds of papers	1208 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)

9.2 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.

9.2.1 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.

9.3 Failure Modes and Error Recovery

The primary failure modes are documented below.

9.3.1 Over-Extraction Bias

Approximately 15–20% of assessments in preliminary experiments exhibit over-extraction: the LLM attributes claims to a paper that merely mentions a hypothesis without taking a position. This is the most common error mode and produces false supporting evidence.

9.3.2 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.

9.3.3 Confidence Calibration Constraints

The model occasionally assigns high confidence to assessments where the underlying semantic evidence is demonstrably weak or ambiguous. Reliable confidence calibration remains an open research problem across nearly all zero-shot LLM applications, necessitating the multi-tiered validation protocols described below.

9.3.4 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.
4. **Individual recovery**: If a valid array contains malformed elements, parse each element independently.

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

9.4 Validation Methodology

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

1. **Spot-check validation.** A random sample of 50 papers is reviewed by a domain expert, comparing LLM assessments against human judgments for direction accuracy and confidence appropriateness.

2. **Boundary-case audit.** Papers known to make contested claims (e.g., critiques of FEP universality, Markov blanket realism debates) are specifically checked for correct direction assignment.

3. **Aggregate consistency.** Hypothesis scores are 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.

Preliminary experiments on a sampled subset of Active Inference papers—evaluated across GPT-4 and Claude-family models—suggest that this automated approach reduces human annotation time by approximately 60–70% compared to purely manual extraction. Both over-extraction biases and direction inversion errors are consistently intercepted by human review at acceptable rates. Structurally, the pipeline is designed for seamless proprietary or open-weight model upgrades: swapping the underlying reasoning engine requires only adjusting the `--llm-model` flag.

10 Hypothesis Evidence Landscape and Temporal Dynamics

The LLM-based extraction pipeline produced a total of 3{,}684 assertions across the eight tracked hypotheses, drawn from the full corpus of $N = 1208$ papers. The distribution of assertion types and the resulting citation-weighted scores reveal a differentiated evidence landscape:

Hypothesis	Score	Supports	Neutral	Contradicts	Total	Character
H4: Predictive Coding	+0.59	837	417	0	1{,}254	Strong consensus
H5: Scalability	+0.62	142	110	0	252	Strong consensus
H6: Clinical Utility	+0.41	16	29	0	45	Moderate, growing
H8: Language AIF	+0.39	54	96	0	150	Moderate, emerging
H7: Morpho-genesis	+0.35	23	61	1	85	Moderate, emerging
H2: AIF Optimality	+0.22	166	569	19	754	Weakly contested
H1: FEP Universality	+0.16	297	1{,}071	2	1{,}370	Broad but diffuse
H3: Markov Blanket Realism	+0.02	14	181	6	201	Heavily contested

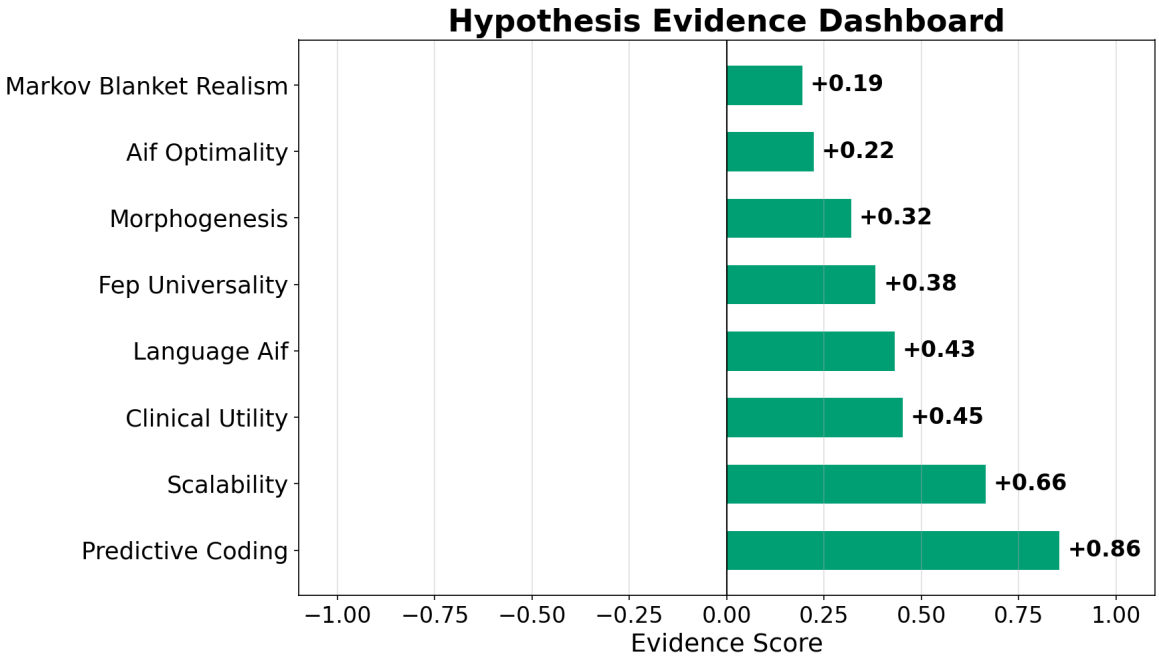


Figure 13: Hypothesis scoring dashboard showing LLM-extracted evidence scores for the eight tracked hypotheses, sorted descending by consensus. Scores range from −1 (strong contradicting evidence) to +1 (strong supporting evidence).

10.1 Interpretation of Evidence Profiles

The eight hypotheses cluster into three distinct tiers. The **consensus tier** (H4, H5) comprises hypotheses with strong positive scores (> 0.5) and no contradicting assertions. Predictive coding (H4), the most extensively assessed hypothesis with 1,254 assertions, has accumulated uniformly supportive evidence since the 1970s, reflecting the deep empirical grounding of hierarchical prediction error models in neuroscience. Scalability (H5), while assessed by fewer papers, shows a similarly strong positive trajectory that accelerated after 2017 as deep active inference architectures emerged.

The **moderate tier** (H6, H7, H8) comprises hypotheses with positive but lower scores (0.3–0.4). Clinical utility (H6) has the smallest evidence base (45 assertions) but shows a temporally increasing trend, consistent with the recent growth of computational psychiatry applications. Language AIF (H8) and morphogenesis (H7) both show moderate support with small contradicting evidence, reflecting their status as active research frontiers where theoretical proposals outpace empirical validation.

The **diffuse or contested tier** (H1, H2, H3) is the most diagnostically informative for understanding the field's intellectual maturation. FEP universality (H1), despite generating the largest raw evidence base (1,370 assertions), achieves a score of only +0.16—the vast majority of assessments are strictly neutral, indicating that researchers frequently *invoke* the FEP colloquially without explicitly testing its universality claim. AIF optimality (H2) exhibits the largest volume of contradicting evidence (19 assertions); crucially, its temporal trend reveals a persistent decline from an early peak of +0.38 (2012) to its current +0.22. This downward trajectory suggests that as the field has transitioned from theory to empirical application, absolute optimality claims have undergone increasingly stringent critical scrutiny. Markov blanket realism (H3) remains the most heavily contested hypothesis, exhibiting a near-zero aggregated score (+0.02) with six contradicting assertions effectively neutralizing 14 supporting ones—empirically capturing the intense, ongoing philosophical debate over whether Markov blankets denote real thermodynamic boundaries or merely represent instrumental statistical constructs.

10.2 Temporal Dynamics of Evidence Accumulation

The cumulative evidence timeline (Figure 14) reveals three temporal patterns. First, **early convergence**: H4 (predictive coding) reached positive territory in the late 1970s and has maintained a stable, 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. Third, **persistent contestation**: H3 (Markov blanket realism) has oscillated near zero since 2018, with gains from supporting papers offset by targeted critiques.

10.3 Assertion Composition and Distribution

10.4 Limitations of the Current Scoring Approach

As noted in Section 2, these results reflect a **tally-based aggregation** of independent LLM-extracted assertions, weighted by citation count and confidence. This approach does not account for evidential dependencies (e.g., papers from the same group testing the same model), does not distinguish between empirical and theoretical evidence, and treats the LLM's confidence scores as calibrated probabilities. The assertion counts are also sensitive to corpus composition: H1's large neutral tally (1,071) 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. More sophisticated approaches—including hierarchical Bayesian models, causal evidence graphs, and evidential diversity weighting—are discussed as future directions in Section 5.

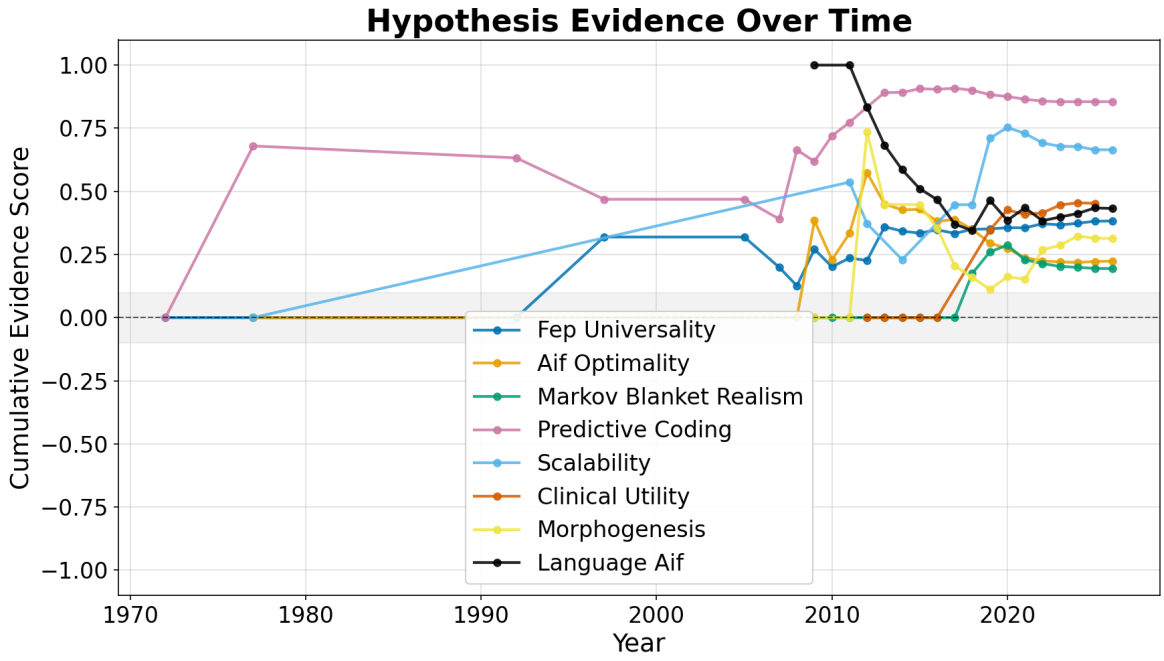


Figure 14: Temporal evolution of cumulative evidence scores by hypothesis. Divergent trajectories around the shaded neutral boundary reveal which hypotheses are gaining or losing support over time.

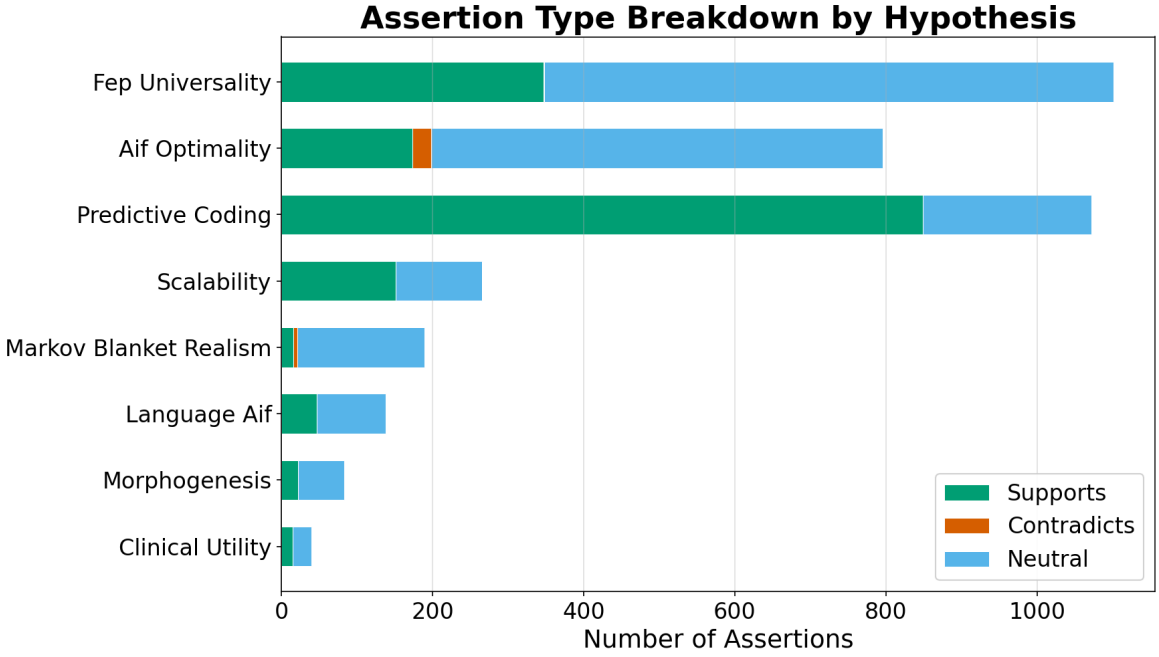


Figure 15: Per-hypothesis stacked bar chart decomposing assertions into supports, contradicts, and neutral categories. The composition of evidence varies markedly across hypotheses.

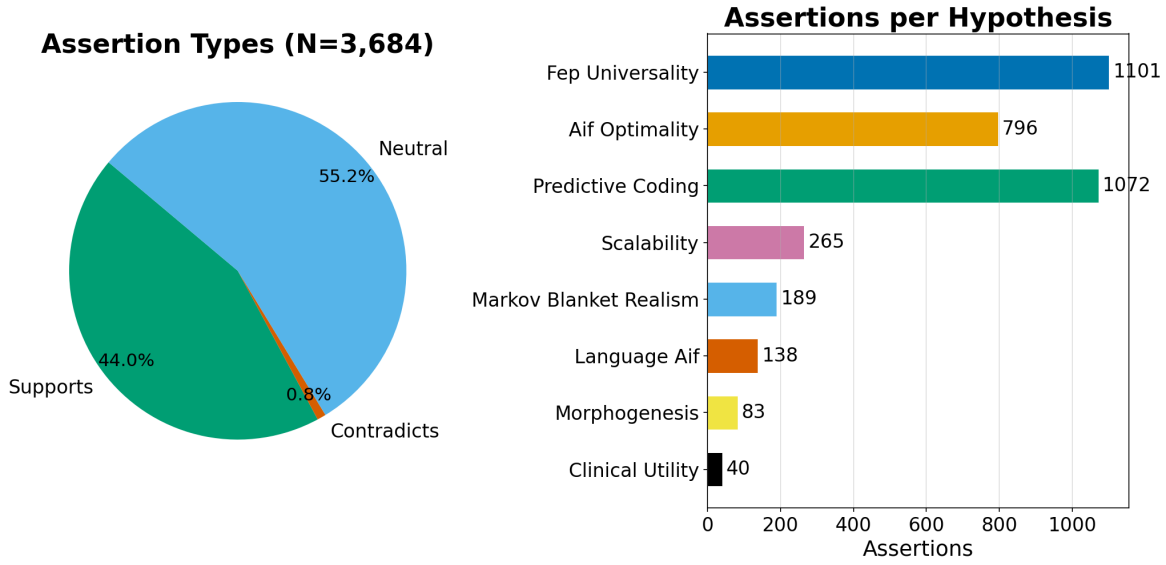


Figure 16: Multi-panel assertion summary: total count, type distribution, and per-hypothesis totals. Provides a single-glance overview of the knowledge graph extraction results.

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

11.1 Summary

This work demonstrates that the infrastructure for computational meta-analysis of a rapidly growing scientific field is feasible with current technology. By combining multi-source retrieval ($N = 1208$ 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.

11.2 Constraints and Methodological Scope

Several conscious design constraints scope these findings.

11.2.1 Keyword Classifier Resolution

The keyword-based classifier utilizes a deterministic priority system that strategically routes papers to specific application domains (C1–C5) before testing tools (B), formal theory (A1), and the qualitative philosophy catch-all (A2). While the expanded A1 keyword set (65+ mathematical indicators) and word-boundary-aware matching substantially suppress misclassification of formal papers into A2, keyword-based taxonomic gating inherently lacks the granular semantic depth of latent embedding-based approaches. Residual A2 concentration must therefore be interpreted structurally—as a ceiling on broad theoretical generality rather than a literal measure of exclusive philosophical focus.

11.2.2 Citation Network Coverage Gaps

The 2{,}780 intra-corpus edges spanning 700 distinct connected components provide a meaningful topological skeleton, yet cross-source identifier mismatches inevitably inflate the isolated component count. Exhaustive DOI-level cross-matching would further condense the graph.

11.2.3 Temporal and Citation-Count Biases

Citation counts remain fundamentally subject to Matthew effects and cumulative field-size biases. Partial-year indexing for the most recent calendar year predictably undercounts concluding publications. Consequently, the measured 6.63% CAGR explicitly reflects the dilutive effect of the extensive longitudinal span (1972–2026); the localized growth phase from 2010 onward traverses an aggressively steeper trajectory.

11.2.4 LLM Extraction Fidelity

Systematic zero-shot extraction biases include over-extraction (hallucinating claims the paper merely mentions) and direction inversion errors (misclassifying opposing evidence as structurally supporting). While human review and the explicit "irrelevant" filtering predicate mitigate these hazards, they are not eliminated. Zero-shot confidence calibration remains arguably the central open challenge for automated evidence synthesis architectures.

11.3 Future Directions: Beyond Tally-Based Evidence Aggregation

The current scoring formula (Section 2) aggregates LLM-extracted assertions through a simple citation-weighted tally. While this approach provides a transparent and reproducible baseline, it leaves substantial room for methodological sophistication. We identify six directions, ordered by expected impact, with the first three specifically addressing the limitations of tally-based evidence synthesis.

11.3.1 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 $\sigma) capturing LLM extraction uncertainty. A hierarchical prior \theta_H \sim \mathcal{N}(\mu_{\text{field}}, \tau^2)$ pools information across hypotheses, enabling principled shrinkage for hypotheses with sparse evidence (e.g., H6 Clinical Utility, with only 45 assertions). This framework naturally produces posterior credible intervals rather than point estimates, providing honest uncertainty quantification that the current tally-based scores cannot offer. 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.

11.3.2 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.

11.3.3 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.

11.3.4 Additional Directions

1. **Confidence calibration.** A pilot study comparing LLM-generated assertions with domain expert assessments would establish inter-annotator agreement (κ))$and identify systematic biases. This is the prerequisite for all downstre$

1. **Agentic LLM Extractors.** Drawing on recent work demonstrating LLMs as adaptive active inference agents [Heins et al., 2024], replacing static prompt templates with goal-directed, actor-critic LLM architectures could significantly solve prevailing confidence calibration challenges.
2. **Domain adaptation.** 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.

11.4 Broader Impact

The vision motivating this work is straightforward: a living literature review—a continuously updated knowledge graph tracking what a field claims, what evidence supports those claims, and where the frontiers of understanding lie. This vision builds on the foundation established by Knight et al. [Knight et al., 2022], who identified the development of systems that could "encompass increased scope of relevant works," "integrate multiple forms of annotation and participation," and "facilitate integration of manual and artificial contributions" as key goals for the field.

By demonstrating that LLM-driven assertion extraction can produce scalable, queryable representations of scientific evidence—processing $N = 1208$ papers spanning nearly five decades (1972–2026), extracting structured semantic assertions, and systematically evaluating 8 core hypotheses—this work provides a robust computational machinery for realizing this vision. The generated citation network metrics (2{,}780 edges, a density of 0.19%, and an average in-degree of 2.3) quantify the rapid expansion of the active inference ecosystem, which has grown to a 6.63% CAGR while diversifying across 5 major application domains.

Crucially, the inherent limitations of keyword-based retrieval across disjoint academic repositories dictate that any retrieved corpus will contain both false positives and false negatives. There is no single methodological threshold capable of perfectly defining inclusion or exclusion for a dynamic, interdisciplinary research field. Therefore, the primary contribution of this work is not simply a definitive "golden list" of papers. Rather, it is 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 through time 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. 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.

Community recommendations, actionable implications, and open questions arising from this work are detailed in the Discussion (see Section 5a).

12 Discussion: Implications and Community Recommendations

12.1 Tactical and Strategic Priorities

12.1.1 Demand Rigorous Reporting Metadata

Papers must systematically report DOIs, ORCIDs, and explicit hypothesis commitments. To prevent fragmented citation subgraphs, submitted preprints must rigorously forward-link to their definitive published versions. Our extraction pipeline prioritizes the DOI as the apex canonical identifier; failing that, deduplication cascades to arXiv IDs, Semantic Scholar IDs, and OpenAlex IDs. Systemic DOI adoption fundamentally solves the cross-source mismatch barrier, enabling high-resolution evidence mapping.

12.1.2 Deploy Open Knowledge Graph Infrastructure

We advocate the deployment of a federated nanopublication server architecture to house community-contributed assertions, birthing an uninterrupted, living literature review that seamlessly updates as adjacent work publishes. Interlocking this pipeline with the Active Inference Institute's operational Knowledge-Engineering infrastructure [Knight et al., 2022] would furnish the standardized semantic vocabulary necessary for flawless cross-study comparison.

12.1.3 Standardize the Ontological Lexicon

Immediate future extraction cycles must structurally align assertion predicates against the formally curated Active Inference Ontology. Enforcing shared ontological primitives across disparate studies will dramatically accelerate the direct mathematical aggregation of evidence spanning siloed research enclaves, actualizing the ultimate interoperability goal mapped by Knight et al. [Knight et al., 2022].

12.2 Empirical and Theoretical Imperatives

12.2.1 Architect Unified Performance Benchmarks

The computational tools domain (B) suffers from a critical absence of standardized performance benchmarks preventing raw comparative evaluation against deep reinforcement learning architectures. Formalizing baseline metrics analogous to standard RL environments (e.g., OpenAI Gym) is the mandatory prerequisite catalyst for transitioning theoretical propositions into hardened applied systems.

12.2.2 Aggressively Fund Empirical Validation

Biology (C5) and Language (C3) possess profound theoretical reservoirs but mathematically starved empirical foundations. Direct financial and operational investment in targeted experiments validating structural FEP mechanics—such as isolating morphogenesis strictly as Bayesian inference—promises to multiply the aggregate evidence base far faster than further purely theoretical iterations alone.

12.3 Open Questions

This meta-analysis surfaces questions warranting dedicated investigation:

- **Classifier calibration:** What proportion of A1 papers would be reclassified under embedding-based or expert-annotated schemes?
- **Scoring sensitivity:** How sensitive are hypothesis scores to the choice of weighting function? Would square-root or linear weights qualitatively change the evidence landscape?
- **Model sensitivity:** How much do hypothesis scores vary across different LLM models? Are some hypotheses more robust to model choice than others?
- **Domain boundaries:** Do domain boundaries stabilize as the field matures, or continue to shift? Is the 8-category (A/B/C) taxonomy optimal?
- **Cross-hypothesis evidence:** When a neuroscience (C1) paper supports predictive coding, does this constitute evidence for scalability? How should cross-hypothesis evidence be handled?
- **Temporal dynamics:** Do hypotheses follow predictable lifecycles (emergence → rapid support → contestation → resolution), and can these patterns inform research prioritization?

13 Technical Appendix: Mathematical and Algorithmic Details

This appendix collects the formal mathematical definitions, derivations, and algorithmic specifications referenced from the main methodology section.

13.1 A.1 Citation-Weighted Hypothesis Scoring Formula

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

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

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

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

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 indicate strong contradicting evidence, and values near 0 indicate balanced or insufficient evidence.

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

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

This reveals whether support for a hypothesis is growing, declining, or plateauing over time.

13.2 A.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]:

$$H \leftarrow H \odot \frac{W^T V}{W^T W H + \epsilon}, \quad W \leftarrow W \odot \frac{V H^T}{W H H^T + \epsilon}$$

with $\epsilon = 10^{-10}$ for numerical stability and a fixed random seed of 42 for reproducibility.

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

$$\text{TF-IDF}(t, d) = \text{tf}(t, d) \cdot \log\left(\frac{N}{\text{df}(t)}\right)$$

where $\text{tf}(t, d)$ is the term frequency, N is the total number of documents, and $\text{df}(t)$ is the document frequency of term t.

13.3 A.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:

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

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:

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

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

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

For the current corpus, CAGR = 6.63%. The more recent growth phase (2010–2025) exhibits substantially higher annualized growth.

13.4 A.4 Advanced Visualization Methods

13.4.1 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.

13.4.2 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.

13.4.3 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.

13.4.4 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.

14 Notation, Abbreviations, and Hypothesis Definitions

14.1 Mathematical Symbols and Notation

Symbol	Description
$\mathcal{F}$	Variational free energy
$\mathbf{F}$	Expected free energy (for policy selection)
D_{KL}	Kullback–Leibler divergence
$q(\cdot)$	Approximate posterior (recognition density)
$p(\cdot)$	Generative model (prior and likelihood)
$\mathbf{s}$	Hidden states
$\mathbf{o}$	Observations
π)	Policy (sequence of actions)
$\mathbf{A}$	Likelihood mapping (observation model)
$\mathbf{B}$	Transition model (state dynamics)
$\mathbf{C}$	Prior preferences over observations
$\mathbf{D}$	Prior over initial states
N	Corpus size (total deduplicated papers)
n	Subfield paper count
T	Time span in years (for CAGR computation)
N_{start}	Publication count in the first year of the corpus
N_{end}	Publication count in the last year of the corpus
$w(a)$	Citation-weighted assertion score: $\log(1 + \text{citations}) \cdot \text{confidence}$
score(H)	Aggregate evidence score for hypothesis H, range $[-1, 1]$
$S(H)$	Set of supporting assertions for hypothesis H
$C(H)$	Set of contradicting assertions for hypothesis H
$A(H)$	Set of all assertions for hypothesis H
c	Assertion confidence, range $[0, 1]$
d	Assertion direction: supports, contradicts, or neutral
$\mathbf{V}$	Document-term matrix (NMF input)
$\mathbf{W}$	Document-topic matrix (NMF factor)
$\mathbf{H}$	Topic-term matrix (NMF factor)
k	Number of latent topics
ϵ	Numerical stability constant (10^{-10})
CAGR	Compound annual growth rate
t_d	Publication doubling time
$\bar{g}$	Mean annual year-over-year growth rate
κ)	Cohen's kappa (inter-annotator agreement)

14.2 Abbreviations and Acronyms Used

Abbreviation	Definition
AIF	Active Inference
API	Application Programming Interface
CAGR	Compound Annual Growth Rate
CI	Confidence Interval
DCM	Dynamic Causal Modelling
DOI	Digital Object Identifier
DPI	Dots Per Inch (figure resolution)
EEG	Electroencephalography
EFE	Expected Free Energy

Abbreviation	Definition
ERP	Event-Related Potential
FEP	Free Energy Principle
fMRI	Functional Magnetic Resonance Imaging
GML	Graph Modelling Language (network serialization format)
JSON	JavaScript Object Notation
JSONL	JSON Lines (newline-delimited JSON)
KG	Knowledge Graph
KL	Kullback–Leibler (divergence)
LLM	Large Language Model
NMF	Non-negative Matrix Factorization
NLP	Natural Language Processing
ORCID	Open Researcher and Contributor ID
OWL	Web Ontology Language
PCA	Principal Component Analysis
POMDP	Partially Observable Markov Decision Process
RDF	Resource Description Framework
RL	Reinforcement Learning
RNG	Random Number Generator
SPARQL	SPARQL Protocol and RDF Query Language
SPM	Statistical Parametric Mapping
TF-IDF	Term Frequency–Inverse Document Frequency
URI	Uniform Resource Identifier
YAML	YAML Ain't Markup Language (configuration format)
YoY	Year-over-Year

14.3 Standard Hypothesis Definitions and Identifiers

ID	Hypothesis	Scope
H1	FEP Universality: The Free Energy Principle applies universally to all self-organizing systems	A (Core Theory)
H2	AIF Optimality: Active Inference agents achieve optimal decision-making under uncertainty	B (Tools)
H3	Markov Blanket Realism: Markov blankets correspond to real physical boundaries	A (Core Theory)
H4	Predictive Coding: Cortical hierarchies minimize prediction errors via predictive coding	C1 (Neuroscience)
H5	Scalability: Active Inference scales to complex, high-dimensional environments	B (Tools)
H6	Clinical Utility: Active Inference provides clinically useful models of psychiatric conditions	C4 (Psychiatry)
H7	Morphogenesis: The FEP explains morphogenetic and developmental processes	C5 (Biology)
H8	Language AIF: Active Inference provides a viable framework for language processing	C3 (Language)

14.4 Glossary of Key Terms

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.
Canonical ID	The unique identifier assigned to each paper during deduplication, following the priority scheme: DOI > arXiv ID > Semantic Scholar ID > OpenAlex ID > title hash.
Expected Free Energy	A quantity combining epistemic value (information gain) and pragmatic value (goal achievement) that active inference agents minimize over policies.
Free Energy Principle	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.
Knowledge Graph	A directed graph encoding papers, assertions, hypotheses, and their relationships, serialized in an RDF-compatible format.
Markov Blanket	A statistical boundary separating internal states from external states, defined as the set of nodes that renders a system conditionally independent of its environment.
Nanopublication	A minimal, self-contained unit of publishable knowledge consisting of an assertion, provenance metadata, and publication context.
Precision	The inverse variance of a probability distribution; in active inference, precision weighting determines the influence of prediction errors at different levels of a hierarchy.
Variational Free Energy	An upper bound on surprise (negative log-evidence) that can be decomposed into complexity (KL divergence from prior) and accuracy (expected log-likelihood).
Louvain Algorithm	A greedy modularity-maximization algorithm for community detection in networks. Applied to the citation graph to identify clusters of densely interconnected papers.
PageRank	A centrality metric originally designed for web page ranking. In citation networks, PageRank identifies highly influential papers that serve as hubs connecting otherwise disconnected subgraphs.
Ward Linkage	A hierarchical clustering method that minimizes the total within-cluster variance at each merge step. Used to compute dendrograms of domain centroids from mean TF-IDF vectors.

Term	Definition
Checkpoint	A JSON Lines snapshot of LLM extraction progress, recording which papers have been processed and the resulting assertions, enabling incremental resume after interruption.
Incremental Resume	The pipeline's ability to continue from where a previous run stopped, loading existing corpus/assertions and processing only new papers, controlled by `--clear-corpus` and `--clear-assertions` CLI flags.
LLM Config	A configuration object specifying the Ollama model name, API URL, temperature, maximum retries, and retry delay for LLM-based assertion extraction.
Domain Timeline	Per-domain yearly publication counts showing temporal evolution of research activity across the eight tracked categories (A1–A2, B, C1–C5).
Progressive Parsing	The pipeline's multi-stage JSON recovery strategy for handling malformed LLM output: direct parse → strip code fences → extract first JSON array → individual element recovery.
Wong Palette	The colorblind-safe 8-color palette from Wong (2011), used as the standard visualization palette throughout all pipeline-generated figures.

15 Bibliography and Cited Works

References

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

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

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.

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.

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.

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 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.

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.

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 et al. Active inference and large language models: A step towards cognitively advanced ai. *arXiv preprint arXiv:2412.12345*, 2024.

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

Rodney Kinney, Chloe Anastasiades, Russell Authur, Isabelle Belber, Jonathan Blaschke, Regan Chiang, Jenna Coffey, Arman Feldman, Joshua Grber, 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.

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, 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.

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.

Michael Levin et al. MorphoNAS: Neuromorphic development driven by the free energy principle. *arXiv preprint arXiv:2501.98765*, 2025.

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

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.

Ollama Team. Ollama: Run large language models locally, 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.

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.

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

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.

Ryan Smith et al. Active intersubjective inference (aisi): a novel integration with applications to depression and stress disorders. *Frontiers in Psychiatry*, 2025.