Conclusion: Evidence Landscape, Methodological Limitations, and Research Agenda

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.

Constraints and Methodological Scope

Several conscious design constraints scope these findings.

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.

Citation Network Coverage Gaps

The 2{,}780 intra-corpus edges spanning 700 distinct connected components provide a meaningful topological skeleton, yet

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.

Hierarchical Bayesian Hypothesis Scoring

The most direct extension replaces the additive tally with a **hierarchical Bayesian model** that treats each hypothesis score as a latent variable inferred from noisy assertion observations. Under this formulation, each assertion a_i contributes a likelihood term $P(a_i|\theta_H, \sigma)$ parameterized by the hypothesis-level evidence strength θ_H and an observation noise term σ capturing LLM extraction uncertainty. A hierarchical prior $\theta_H \sim \mathcal{N}(\mu_{\text{field}}, \tau^2)$ pools information across hypotheses, enabling principled shrinkage for hypotheses with sparse evidence (e.g., H6 Clinical Utility with

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. knight2022fep, 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