Discussion: Implications and Community Recommendations

Tactical and Strategic Priorities

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.

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 knight2022fep would furnish the standardized semantic vocabulary necessary for flawless cross-study comparison.

Empirical and Theoretical Imperatives

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.

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.

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?