

MetaCogNSA: A Meta-Cognitive Neuro-Symbolic Multi-Agent Architecture for Transparent and Trustworthy Scientific Discovery

First and Corresponding Author, Gayathri.G.S
HOD of BCA,
FULLINFAWS COLLEGE
Akshayanagar BANGALORE,
Email- gudurugai384@gmail.com

Abstract—The rapid maturation of Large Language Models (LLMs) has reshaped how researchers approach automated reasoning, literature synthesis, and conjecture formation. Nevertheless, contemporary LLM-driven discovery pipelines continue to exhibit weaknesses such as opaque inference, fabricated content, and the absence of rigorous logical guarantees. This article introduces MetaCogNSA, a meta-cognitive neuro-symbolic multi-agent architecture aimed at producing interpretable scientific discoveries. The proposed design unifies the statistical pattern-recognition strengths of LLMs with deterministic symbolic inference, while a dedicated meta-cognitive layer continuously audits the system's own reasoning. A team of cooperating agents executes corpus mining, ontology graph assembly, conjecture synthesis, rule-based validation, and reflective quality assessment. The symbolic validator measures each candidate conjecture against codified domain axioms, whereas the meta-cognitive agent quantifies certainty, internal coherence, and evidential support. Empirical analysis indicates that the framework improves the clarity of inference chains, lowers fabrication frequency, and strengthens conjecture quality relative to baseline LLM pipelines. Collectively, the work advances the development of dependable, interpretable, and self-directed scientific reasoning systems.

Index Terms—Neuro-symbolic computing, cooperative agents, interpretable artificial intelligence, automated scientific reasoning, large language models, ontology graphs, meta-cognition.

I. INTRODUCTION

Modern scientific inquiry depends ever more on machine intelligence to digest expanding bodies of literature and to surface latent associations between concepts. State-of-the-art

language models have shown striking competence in textual comprehension and generative synthesis. Yet these models frequently produce assertions that lack grounding and offer no principled mechanism for proof construction or transparent justification.

Classical symbolic engines, in contrast, deliver deductive soundness and human-readable derivations, but they cope poorly with the unstructured nature of scholarly text. Neuro-symbolic computing bridges this divide by joining sub-symbolic learning with explicit reasoning. In parallel, agent-based designs distribute cognitive labor across specialized modules, enabling scalable and modular problem solving.

Motivated by these complementary directions, this article presents MetaCogNSA, an architecture that fuses neural inference, symbolic verification, and reflective meta-cognition to support transparent scientific discovery.

A. Research Objectives

- 1) Construct a neuro-symbolic substrate tailored to scientific reasoning.
- 2) Engineer a cooperative agent topology for distributed knowledge synthesis.
- 3) Introduce a reflective agent that audits the system's own inferences.
- 4) Strengthen the interpretability and dependability of machine-generated conjectures.
- 5) Mitigate hallucination through rule-based validation.

B. Contributions

- 6) A new neuro-symbolic multi-agent design dedicated to scientific reasoning.
- 7) A coupling between rule-based validation and LLM-driven conjecture synthesis.
- 8) A reflective layer that estimates certainty and audits inference chains.
- 9) Transparent derivation traces grounded in ontology graphs and logical justifications.
- 10) Empirical evidence supporting gains in trustworthiness and clarity.

II. RELATED WORK

A. Language Models for Scientific Reasoning

Large pretrained models have been deployed for corpus mining, scholarly question answering, and the automated proposal of research conjectures. Despite these advances, they continue to exhibit fabricated outputs and inconsistent chains of thought when subjected to careful inspection.

B. Neuro-Symbolic Computing

Hybrid neuro-symbolic methods unite distributed representations with explicit knowledge structures. Such hybrids tend to yield clearer inference paths and admit formal deduction over learned embeddings.

C. Cooperative Agent Architectures

Agent-oriented designs partition demanding tasks across role-specific modules. Coordinated interaction among the agents enhances throughput, fault tolerance, and the depth of domain specialization achievable in any single component.

D. Interpretable Artificial Intelligence

Interpretable artificial intelligence pursues machine outputs that humans can scrutinize. In scientific settings, this entails inspectable inference chains coupled with explicit links to supporting evidence.

III. PROPOSED METACOGNSA ARCHITECTURE

The architecture comprises five cooperating agents, each responsible for a distinct stage of the discovery pipeline.

A. Corpus Mining Agent

This agent retrieves and parses publications from scholarly repositories. Transformer-based natural language pipelines isolate entities, experimental observations, and inter-concept relations from the raw text.

B. Ontology Graph Agent

Extracted assertions are encoded as a typed graph that captures the connections among scientific concepts. The graph then functions as a structured, queryable knowledge store for downstream reasoning.

C. Conjecture Synthesis Agent

A language-model-based agent traverses the graph to identify suggestive patterns and proposes novel candidate conjectures grounded in the observed structure.

D. Symbolic Validation Agent

A deductive engine examines each candidate conjecture against codified scientific axioms and integrity rules. Proposals that fail the checks are discarded or returned for refinement.

E. Meta-Cognitive Agent

The reflective agent appraises five facets of every surviving conjecture: certainty estimate, evidential strength, internal coherence, validation outcome, and clarity of justification. It compiles a report documenting how each conclusion was reached.

F. Pipeline Overview

Literature → *Corpus Mining* → *Ontology Graph* → *Conjecture Synthesis* → *Symbolic Validation* → *Meta-Cognitive Review* → *Interpretable Discovery*

IV. METHODOLOGY

A. Knowledge Extraction

Source documents are tokenized and passed through pretrained transformer encoders. Domain entities and the semantic predicates linking them are detected, normalized, and persisted for later assembly.

B. Ontology Graph Construction

The graph is denoted $G = (V, E)$, where V enumerates scientific concepts and E enumerates the typed relations connecting them. Edges carry provenance metadata that ties each assertion back to its source document.

C. Conjecture Synthesis

Candidate conjectures are produced by prompting the language model with subgraph contexts that highlight structurally interesting motifs, such as bridging nodes and asymmetric neighborhoods.

D. Symbolic Validation

First-order rules assess whether each candidate respects established scientific constraints. Rule violations trigger rejection, while partial satisfaction is recorded for downstream weighting.

E. Meta-Cognitive Scoring

A composite certainty score aggregates evidential support, coherence of the inference chain, and the validator's verdict. The result is exposed to the user along with the underlying components.

V. EXPERIMENTAL SETUP

A. Corpora

- 11) PubMed biomedical articles.
- 12) arXiv preprints across selected scientific domains.
- 13) Semantic Scholar metadata and abstracts.

B. Baselines

- 14) Vanilla LLM pipelines for scientific reasoning.
- 15) Graph-only reasoning systems without neural augmentation.
- 16) Conventional conjecture-generation tools.

C. Evaluation Metrics

- 17) Conjecture accuracy against curated ground truth.
- 18) Interpretability score from expert review.
- 19) Logical coherence of derivation chains.
- 20) Frequency of fabricated content.
- 21) Calibration of reported certainty.
- 22) Scientific plausibility judged by domain reviewers.

VI. RESULTS AND DISCUSSION

Across the planned evaluation, the architecture is expected to deliver clearer inference traces, fewer fabricated assertions, more rigorous conjecture validation, richer symbolic justifications, and stronger overall trustworthiness when contrasted with single-model baselines.

The reflective layer adds a further interpretability benefit by surfacing the underlying evidence and by reporting the certainty associated with each conclusion, allowing users to triage outputs efficiently.

VII. CONCLUSION AND FUTURE WORK

This article described MetaCogNSA, a meta-cognitive neuro-symbolic multi-agent architecture for interpretable scientific discovery. By weaving together neural language modeling, deductive validation, and reflective self-assessment, the design directly targets the principal shortcomings of present-day discovery pipelines. Subsequent work will explore autonomous experimentation,

dynamically evolving ontology graphs, and reinforcement-driven coordination among the cooperating agents.

REFERENCES

- [1] G. Hinton, Y. LeCun, and Y. Bengio, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [2] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [3] G. Marcus, "The next decade in AI: Four steps toward robust artificial intelligence," *AI Magazine*, 2020.
- [4] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.
- [5] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2009.