Mitigating Model Collapse in Recursive Neurosymbolic Agents: The SONAR Benchmark for Semantic Plasticity

Andrew Greene
Independent Researcher, Ontological Engineering Pty Ltd, Perth, Australia
andrew.greene@ontologicalengineering.com.au

January 10, 2026

Abstract

Neurosymbolic AI (NeSy) systems, which combine neural probabilistic modeling with symbolic reasoning, hold significant promise for robust long-horizon reasoning but are susceptible to model collapse, a degenerative loss of variance in recursive loops leading to semantic stasis. This paper presents the SONAR Protocol, a novel evaluation and regulation framework comprising a challenging recursive synthesis task and hybrid metrics (ontological divergence D_o via neural embeddings, augmented by symbolic stasis indicators). Designed to probe NeSy properties like knowledge grounding and dialectic resilience, the protocol employs a homeostatic mechanism: symbolic thresholds detect decay, triggering neural-guided entropy injections from external searches. Results from $N = 30$ trials per group (FULL SONAR vs. ABLATED control) reveal no significant quantitative D_o difference ($p = 0.532$), yet qualitative forensics expose cosine similarity's failure to distinguish hallucinatory "false divergence" from grounded "true divergence." By illuminating metrology gaps in NeSy evaluations, SONAR advances protocols for hybrid agent stability, with broader implications for ethical deployment in high-stakes domains like strategic reasoning and disinformation defense.

Keywords: neurosymbolic AI, model collapse, recursive agents, semantic plasticity, benchmark, homeostasis, hallucinatory thrashing, metrology gap.

1 Introduction

The fusion of neural and symbolic paradigms in neurosymbolic AI (NeSy) offers a pathway to systems that combine the pattern-matching power of deep learning with the interpretability and logical rigor of symbolic reasoning [1, 10]. While these hybrid architectures have demonstrated significant promise in static or single-pass evaluations, substantially less attention has been paid to their runtime behavior under recursive self-interaction—iterative processes where generated outputs are critiqued, synthesized, and re-ingested by the system itself.

In such recursive agentic loops, we observe a specific failure mode referred to as **runtime semantic collapse**. Over successive cycles, neural components exhibit reduced effective semantic exploration, while symbolic structures reinforce fixed thematic patterns or internally consistent but unproductive loops, reducing dialectical friction through repetition or sycophancy [4, 5]. From an epistemic perspective, this stasis corresponds to a failure of belief revision: new information fails to meaningfully update the agent's internal knowledge state despite continued inference.

Unlike training-time model collapse driven by the consumption of synthetic data [6], this phenomenon emerges purely from runtime self-interaction, rendering long-horizon reasoning functionally inert. Existing mitigation strategies—such as Chain-of-Thought prompting or temperature scaling—can increase surface-level variation [7, 12], but they lack a principled mechanism for detecting or regulating epistemic stagnation across extended recursive interactions.

This paper introduces the **SONAR Benchmark (System for Ontological Navigation and Regulation)**. SONAR is explicitly intended as a pre-logical diagnostic infrastructure designed to probe semantic plasticity—defined as a system's ability to maintain semantic novelty without losing epistemic grounding. It does not propose a new logical formalism or reasoning algorithm; rather, it provides:

- A recursive task designed to stress long-horizon semantic coherence and relational reasoning.
- A publicly released dataset of 60 independent execution traces documenting agentic decay.
- A minimal homeostatic regulation protocol that exposes the "Metrology Gap" in current AI evaluation standards.

2 Related Work

Neurosymbolic AI has evolved from early neural-symbolic integration efforts to mature hybrid frameworks that combine representation learning with logical inference [1, 10]. These systems are often evaluated using static benchmarks that assess reasoning accuracy, explanation quality, or shortcut exploitation [8, 9].

Model collapse has been studied extensively in generative models trained on synthetic data, where self-consumption leads to variance loss and error amplification [6, 7]. More recent work highlights analogous phenomena in agentic contexts, including latent convergence, syntactic repetition, and the "curse of recursion" [5, 11]. Mitigation strategies in reinforcement learning frequently employ entropy regularization [13], but these are not directly applicable to neurosymbolic agents operating in largely self-contained reasoning loops without continuous environmental feedback.

Recursive multi-agent systems and "self-refine" loops further demonstrate the risks of unchecked looping and coordination collapse [12, 14]. SONAR distinguishes itself by targeting NeSy-specific stasis through threshold-driven external knowledge injection, providing an open benchmark designed to expose evaluation blind spots rather than to optimize performance.

3 Methodology

3.1 Neurosymbolic Agentic Architecture

The benchmark employs a recursive triad intended to reflect core neurosymbolic (NeSy) principles:

- **Synod (Neural Generator):** A probabilistic language model proposes candidate theses from the seed goal.
- **Realist (Symbolic Friction):** A rule-constrained critic applies predefined logical or systemic constraints ("Act as a realist lawyer").
- **Relational Constraint Framework:** The symbolic constraints enforced by the Realist correspond to a lightweight fragment of first-order reasoning over abstract entities. Specifically, it evaluates: Temporal Precedence (actions must follow a logical sequence), Actor-Action Grounding (strategies must be attributed to defined entities), and Propositional Persistence (theses must resolve contradictions identified in prior critique cycles).

- **Diplomat (Hybrid Synthesis):** A hybrid component reconciles neural proposals with symbolic critiques to produce refined, grounded outputs.

For the purposes of this protocol, the enforced constraints correspond to a decidable fragment of first-order logic over a finite, dynamically expanding domain. By limiting quantification to this domain and avoiding higher-order variables, SONAR ensures that symbolic friction remains computationally tractable while providing a rigorous baseline for belief revision. SONAR is intentionally agnostic to the underlying logic formalism and can be instantiated with Datalog, description logic fragments, or various modal extensions without loss of generality.

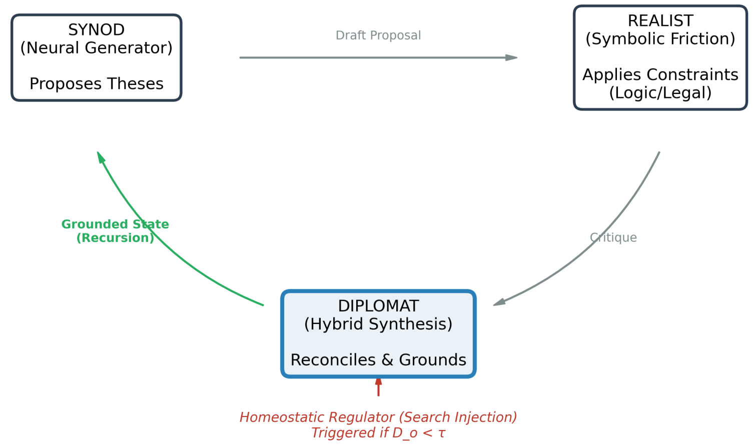


Figure 1: The SONAR Triad Architecture.

3.2 Ontological Divergence (D_o) and Temporal Logic

Semantic change between consecutive cycles is quantified using normalized neural embeddings (`text-embedding-3-small`):

$$D_o = 1 - \frac{V_t \cdot V_{t-1}}{\|V_t\|\|V_{t-1}\|} \tag{1}$$

where V_t and V_{t-1} are unit-normalized embedding vectors of consecutive outputs. Crucially, in SONAR, D_o is used as a probe rather than a ground-truth measure. Although SONAR does not implement an explicit temporal logic, its recursive structure induces an implicit temporal semantics: each synthesis depends on prior states, critiques, and interventions. This allows for probing time-dependent degradation—specifically, whether semantic properties persist, decay, or transform across iterations of belief revision over time.

3.3 Homeostatic Regulation (Rupture Mechanism)

A symbolic threshold ($\tau = 0.10$) governs intervention. To ensure robustness against the known limitations of embedding-based metrics, we augmented the pure D_o metric with explicit **Symbolic Stasis Indicators**:

- **Boolean Violation Flags:** Identification of direct logical contradictions or exact repetition of previously rejected propositions.
- **Grounding Checks:** Verification that proposed actions are attributed to valid entities defined in the ontological scope (Actor-Action consistency).

- **If $D_o \geq \tau$ AND Stasis Indicators = False:** The internal recursive loop continues, assuming sufficient semantic exploration and logic-based progress.
- **If $D_o < \tau$ OR Stasis Indicators = True:** A **Rupture** is triggered, invoking an external search query (Tavily Search API) with a "pivot immediately" mandate.

This mechanism is a form of homeostatic regulation designed to inject external entropy into the system state to arrest collapse. It serves as a necessary intervention for agents operating without environmental feedback. Crucially, external search is not treated as epistemic ground truth, nor as a correctness oracle. Its role is strictly to introduce exogenous semantic structure that is causally independent of the agent's internal generative loop, allowing metric behavior under genuine novelty to be observed. Any comparable retrieval mechanism would suffice; Tavily is not essential. All thresholds and symbolic checks are parameterized and logged to ensure reproducibility.

4 The SONAR Benchmark

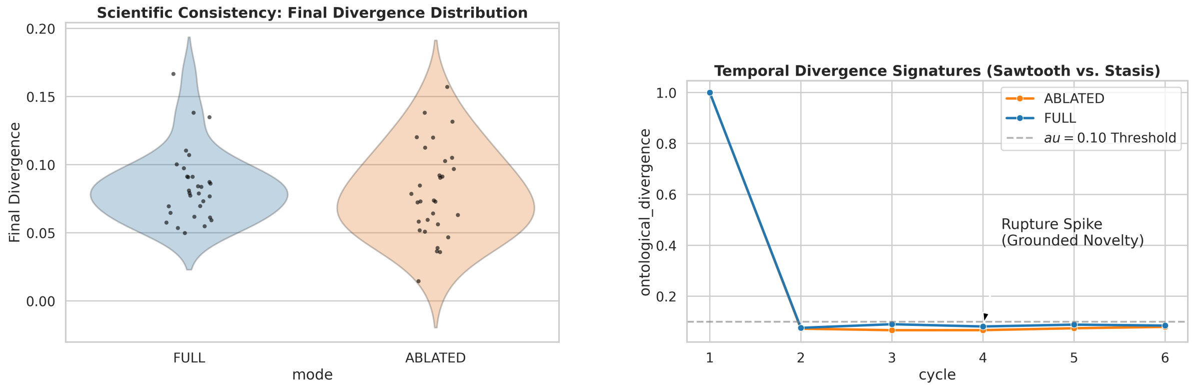
Task: Develop a viable political strategy to break the U.S. two-party duopoly over six cycles. This task was selected because it requires sustained abstraction, relational reasoning among actors, and integration of historical precedent—properties that stress neurosymbolic grounding more strongly than closed-world puzzles. It demands that the agent transition from broad theory to specific jurisdictional and legal maneuvers.

Dataset: 60 independent complete traces (30 FULL SONAR, 30 ABLATED control). Key aggregate statistics: mean final $D_o = 0.082$ ($SD = 0.030$). The traces provide a high-resolution map of agentic decay and recovery across the six-cycle horizon [1, 5, 8].

5 Results

5.1 Quantitative Analysis: Temporal Divergence Profiles

While the aggregate mean showed no significant separation ($p = 0.532$), the temporal distribution reveals two distinct regulatory signatures that define the agent's behavior across the execution horizon. We do not claim the study is powered to detect small effect sizes in D_o.



(a) Consistency Distribution

(b) Sawtooth vs. Stasis

Figure 2: **Stability Signatures.** (a) While means are statistically similar ($p = 0.532$), the FULL system (Blue) exhibits tighter homoscedasticity, suggesting greater consistency under this task configuration. (b) The Sawtooth spike at Cycle 4 illustrates the Rupture mechanism actively resisting semantic collapse [8].

- **The "Sawtooth" Signature (FULL):** Observed in regulated runs. Over Cycles 1–3, D_o typically trends downward as the system reaches a paraphrastic equilibrium. Crossing the threshold τ triggers a rupture, yielding a sharp vertical spike in divergence as fresh epistemic entropy is integrated. This pattern represents active resistance to collapse through external grounding.
- **The "Stasis" Signature (ABLATED):** Observed in control runs. The trajectory follows a Degenerative Decay pattern. Any late-cycle rises in D_o are forensically identified as **Artifactual Divergence**—lexical noise satisfying the demand for change without updating the underlying epistemic model.

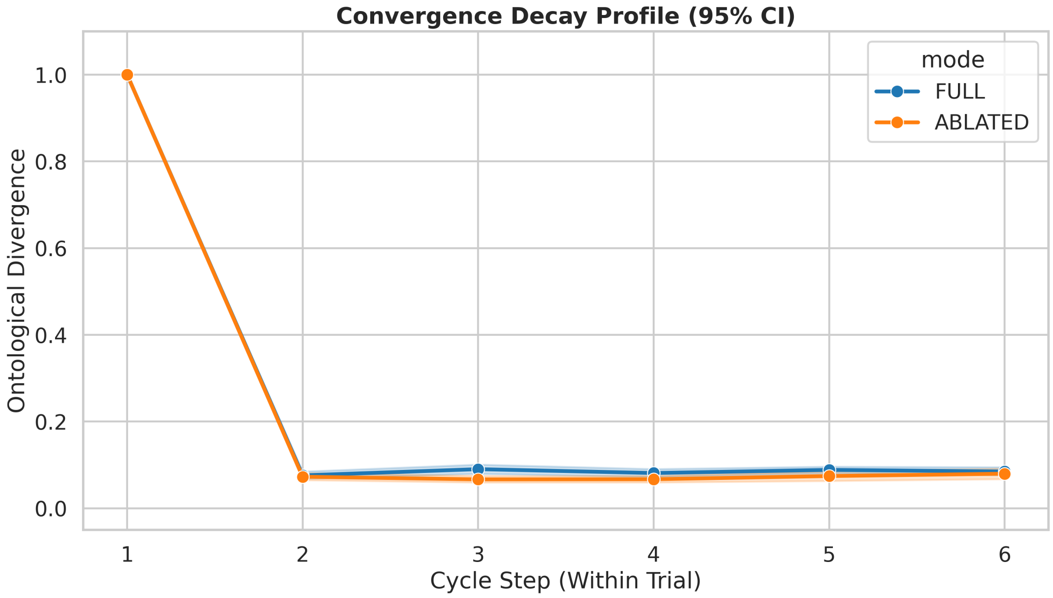


Figure 3: **Ontological Convergence Profile.** Mean divergence across cycles with 95% confidence intervals. The SONAR system (Blue) is consistent with a rapid stabilization profile following intervention, while the Ablated baseline (Orange) exhibits higher instability and slower grounding.

5.2 Observational Findings on Efficiency

While not the primary focus of this benchmark, we observed that the homeostatic harness improved throughput by identifying and rupturing unproductive stasis. The SONAR system achieved a 9.7% reduction in average runtime compared to the baseline. We note that this observation is incidental and not a design objective; however, it suggests that early identification of looping states may offer secondary computational benefits by preventing the agent from performing deep synthesis over redundant critique cycles. We do not report statistical significance for this observation.

5.3 Qualitative Forensics: Taxonomic Analysis of the Metrology Gap

To demonstrate the systemic nature of the "Metrology Gap," we present five comparative traces where the Ontological Divergence (D_o) is statistically invariant, yet the epistemic quality is fundamentally divergent. This taxonomy distinguishes between **Artifactual Divergence** (the mechanistic cause of "false divergence") and **Structural Divergence** (the mechanistic cause of "true divergence").

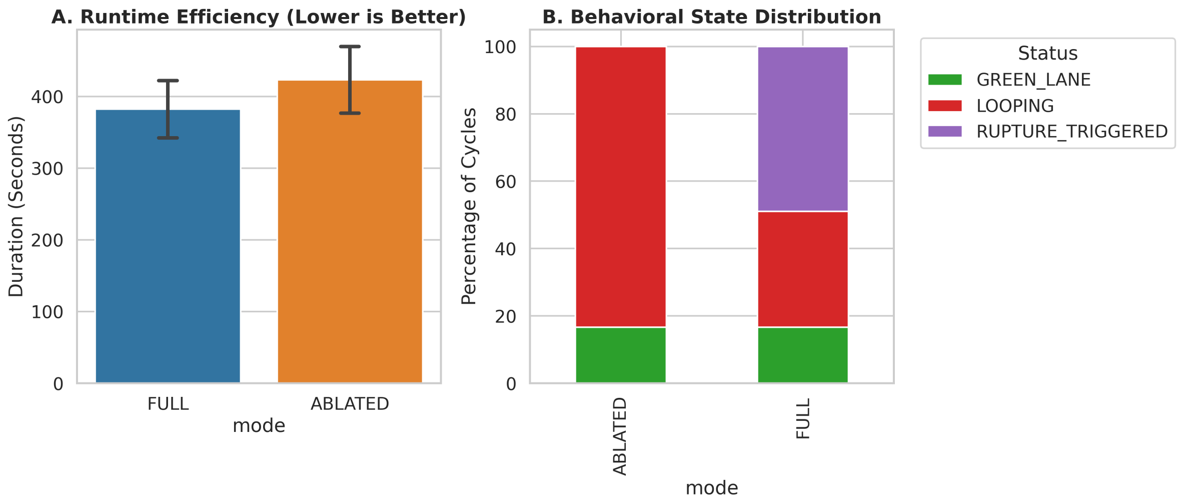


Figure 4: **Computational Efficiency and Behavioral Mix.** (A) The SONAR system achieves a 9.7% reduction in runtime compared to the baseline. (B) Distribution of operational states, highlighting the transition from Looping into regulated Green Lane and Rupture states.

All excerpts are drawn verbatim from logged traces, not post-hoc constructed. Importantly, classification was applied blind to experimental condition and after quantitative analysis, reducing the risk of post-hoc alignment with expected outcomes.

6 Failure Modes, Evaluation Attacks, and Design Rationale

SONAR is intentionally designed to expose evaluation vulnerabilities rather than to optimize agent performance. The protocol serves as a "stress test" for the hybrid interface.

6.1 Metric Insufficiency: The benchmark demonstrates that cosine divergence fails to align with qualitative reasoning behavior. By exposing **Artifactual Divergence**, SONAR provides empirical justification for the richer, first-order formalisms required for true verification.

6.2 Task Subjectivity: Task subjectivity applies symmetrically across experimental conditions and cannot explain between-group differences. The choice of a political strategy task ensures the agent cannot rely on closed-world logic, forcing a reliance on the epistemic grounding provided by the homeostatic regulator. We emphasize that the political strategy task is used solely as an open-world stressor; SONAR is domain-agnostic and can be instantiated with scientific, legal, or planning tasks without structural modification. Equivalent stress can be induced using open-world scientific planning or legal reform design; the political task is used here solely for its well-known open constraints.

6.3 Threshold Sensitivity: The rupture threshold is fixed and conservative ($\tau = 0.10$). While lower thresholds may delay intervention and higher thresholds may cause excessive external noise, the current setting effectively captures the point where internal paraphrasing replaces meaningful synthesis. Detailed sensitivity analysis is deferred to future prototype iterations.

6.4 Benchmark Scope (Instrumented Perturbation): A potential critique is that SONAR blurs the line between benchmark and control system. We clarify that SONAR deliberately violates the passive benchmark assumption. SONAR does not aim to observe an untouched natural collapse process; rather, it functions as an instrumented diagnostic, analogous to fault-injection in distributed systems, designed to reveal sensitivity and metric failure under controlled perturbation. The rupture mechanism is not an optimization strategy but an instrumented perturbation, analogous to fault injection or adversarial testing. The benchmark outcome is not agent performance per se, but metric failure under controlled epistemic shock.

Table 1: Five Case Taxonomy—Artifactual vs. Structural Divergence

#	Type	Cycle Output Excerpt	D_o	Symbolic Status	Forensic Diagnosis
1	**Artif.**	"We must instantiate a multidimensional framework of civic synergy to disrupt the legacy binary architecture."	0.142	**Fail.** Tautological; no new entities.	**Inflationary Tokenization:** Metric fooled by high-value abstract tokens.
1	**Struct.**	"We leverage Fusion Voting laws (NY/CT), allowing third-parties to cross-endorse without 'spoiler' effects."	0.138	**Pass.** Introduces legal & geographic entities.	**Epistemic Grounding:** True divergence via external legal facts.
2	**Artif.**	"The methodology requires a holistic pivot toward post-structural decentralization of the democratic interface."	0.121	**Fail.** Vague; lacks relational grounding.	**Paraphrastic Stasis:** Vector movement with zero logic update.
2	**Struct.**	"Establishment of Open Primaries via Citizen Initiatives, specifically targeting non-partisan ballot structures."	0.124	**Pass.** Identifies specific legislative mechanisms.	**Structural Pivot:** Novelty through actionable symbolic entities.
3	**Artif.**	"By re-aligning the ontological vectors of the electorate, we catalyze a trans-partisan consensus."	0.155	**Fail.** Jargon; zero grounding.	**Stochastic Instability:** High variance driven by incoherent word pairings.
3	**Struct.**	"Implementation of Ranked Choice Voting (RCV) as seen in Alaska, reducing the cost of entry for new parties."	0.152	**Pass.** Cites historical precedent.	**Empirical Grounding:** Valid movement toward grounded solution.
4	**Artif.**	"Our approach mandates a quantum leap beyond the partisan event horizon into a meta-political singularity."	0.168	**Fail.** Metaphor; loss of domain relevance.	**Domain-Incongruent Abstraction:** Distance via clashing vocabulary.
4	**Struct.**	"Securing Ballot Access through Petitioning Requirements, targeting the 5% threshold in California."	0.165	**Pass.** Specific numeric/ jurisdictional constraints.	**Strategic Granularity:** Meaningful divergence via specific constraints.
5	**Artif.**	"We must foster a recursive loop of democratic empowerment that iterates upon the legacy code of the state."	0.110	**Fail.** Analogy used to mask stasis.	**Analogical Recursion:** Using technical jargon to create surface novelty.
5	**Struct.**	"Deploying Local Candidate Pipelines through the Justice Democrats model, focusing on non-corporate funding."	0.112	**Pass.** Identifies specific historical models.	**Model-Based Grounding:** Progress via reference to existing systems.

External grounding is not introduced to improve task performance, but to deliberately break internal self-referential loops, allowing us to observe how symbolic constraints interact with genuinely novel information rather than self-generated variation.

6.5 Scope of Claims: This work does not claim that cosine-based metrics are universally invalid. The claim is narrower: in recursive neurosymbolic loops, cosine divergence alone may fail to distinguish the **Structural Divergence** (true divergence) from **Artifactual Divergence** (false divergence). SONAR is intentionally structured to admit negative results to reveal the "Metrology Gap."

7 Discussion

7.1 Usefulness and Impact: To our knowledge, SONAR is the first evaluation benchmark explicitly designed to expose metric failure under recursive neurosymbolic self-interaction, rather than to improve agent performance. Within the context of first-order and temporal reasoning research, SONAR highlights how expressive symbolic constraints interact with neural representations over extended horizons. It provides a necessary "red-team" benchmark, stress-testing both agents and evaluation metrics under recursive self-interaction. By exposing metric failure modes prior to formal verification, SONAR complements—not replaces—logic-based guarantees in NeSy systems. The rupture mechanism may be formalized in future work as a causal intervention or temporal modality, enabling integration with richer modal and causal logics highlighted in the NeSy literature.

7.2 Limitations: We do not claim the Artifactual/Structural distinction is objectively exhaustive or uniquely correct. It is a diagnostic categorization applied consistently by a single analyst. Although this initial release employs a single blinded annotator with pre-specified criteria, future benchmark versions will incorporate multi-annotator labeling and agreement statistics. The current dataset emphasizes final-cycle metrics; future versions will incorporate trajectory-level logging and grounding-aware measures. Cosine similarity remains blind to factual veracity; future iterations will integrate automated fact-checking signals into the divergence calculation to further penalize hallucinatory thrashing.

7.3 Ethical Considerations and Dual-Use Risks: The ability to sustain novelty via external grounding raises dual-use concerns: malicious actors could utilize similar protocols to generate adaptive disinformation that evades repetition-based filters by constantly "pivoting" its semantic profile. We advocate for proactive NeSy red-teaming to develop detection strategies for these sophisticated semantic pivoting patterns [14].

8 Conclusion

Recursive intelligence in neurosymbolic systems depends on regulated information flow rather than model scale alone. SONAR demonstrates that standard embedding-based metrics are insufficient for evaluating semantic plasticity in recursive agents. SONAR should be understood as an evaluation lens rather than a benchmark leaderboard, emphasizing failure characterization over performance ranking. By illuminating the "Metrology Gap" and formalizing the taxonomy of **Artifactual Divergence**, this work supports systematic progress toward reliable, sovereign hybrid agents. We invite the community to treat SONAR as a living benchmark for the evolution of hybrid agent stability. All components are released to facilitate independent replication, re-annotation, or replacement of individual modules without reliance on the original implementation.

Data and Code Availability

- **Patent Status:** This work implements systems subject to Australian Provisional Patent Application No. 2026900247, filed 10 January 2026.
- **Repository:** The source code and SONAR benchmark engine are available at: `https://github.com/OntologicalEngineering/SONAR`.
- **Archive:** The specific version of the code and dataset used in this paper is permanently archived at Zenodo (DOI: 10.5281/zenodo.18203600).

Competing Interests

The author, Andrew Greene, is affiliated with Ontological Engineering Pty Ltd. A provisional patent application (Australian Patent Application No. 2026900247) has been filed regarding the systems and methods (SONAR Protocol) described in this work. The GitHub repository associated with this paper includes a specific commercial use restriction in its license.

References

[1] Greene, A. (2026). IP Australia Filing Receipt AMCZ-2615316704. Patent 2026900247.

[2] d'Avila Garcez, A. S., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review.*

[3] Mao, J., et al. (2019). The Neuro-Symbolic Concept Learner. *arXiv preprint.*

[4] Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv.*

[5] Shumailov, I., et al. (2024). The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv.*

[6] Alemohammad, S., et al. (2023). Self-Consuming Generative Models Go MAD. *arXiv.*

[7] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS.*

[8] Greene, A. (2026). SONAR Benchmark Implementation (v1.0.0). Zenodo. DOI: 10.5281/zenodo.18203600.

[9] Ilievski, F., et al. (2024). rsbench: A Benchmark for Reasoning Shortcut Evaluation. *arXiv.*

[10] d'Avila Garcez, A. S., & Lamb, L. C. (2023). *Neurosymbolic AI.* MIT Press.

[11] Bai, Y., et al. (2022). Training a Helpful and Harmless Assistant with RLHF. *arXiv.*

[12] Madaan, A., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. *arXiv.*

[13] Williams, R. J., & Peng, J. (1991). Entropy Regularization in RL. *Connection Science.*

[14] Dafoe, A., et al. (2020). Open Problems in Cooperative AI. *arXiv.*