LATENT-IMPLICIT THINKING WITH PROOF-CARRYING NEURO-SYMBOLIC OUTPUTS FOR BIOMEDICAL DISCOVERY

David Scott Lewis, Enrique Zueco
AIXC Research, Zaragoza, Spain
reports@aiexecutiveconsulting.com

ABSTRACT

Recent work on latent reasoning—where large language models (LLMs) perform intermediate computation in continuous representation spaces rather than generating explicit token chains—achieves dramatic efficiency gains (80–90% token reduction) but sacrifices the transparency that makes chain-of-thought (CoT) reasoning auditable. We propose **Latent-to-Symbolic Compilation (LASY)**, a four-component pipeline that enables models to reason efficiently in latent space while emitting *proof-carrying* structured outputs: causal graphs with typed edges, mechanistic constraints, and minimal verification warrants. The pipeline comprises a *latent reasoner* for continuous thought evolution, a *symbolic extractor* that decodes latent states into formal graph structures, a *constraint verifier* that checks domain axioms, and a *warrant emitter* that produces sparse evidence certificates. We evaluate LASY on 50 reasoning tasks across three scientific domains and demonstrate that it matches latent-only efficiency (45 tokens vs. 319 for explicit CoT) while achieving 93% constraint satisfaction—compared to 63% for unverified latent reasoning. In a case study on NAD^+-centered Alzheimer's disease reversal, LASY discriminates between three competing mechanistic hypotheses and generates falsifiable experimental proposals. Faithfulness probing reveals that latent states encode semantically meaningful structure (89.6% linear probe accuracy for causal direction), providing evidence that implicit reasoning is not opaque but rather compressed.

1 INTRODUCTION

Chain-of-thought prompting (Wei et al., 2022) transformed LLM reasoning by making intermediate steps explicit and auditable. Extensions including Tree of Thoughts (Yao et al., 2023), Graph of Thoughts (Besta et al., 2024), and self-consistency (Wang et al., 2023b) further improved reasoning quality through structured exploration. However, explicit reasoning is expensive: generating hundreds of intermediate tokens per query imposes substantial computational costs and latency penalties that limit deployment in real-time scientific workflows.

A new paradigm—*latent* or *implicit* reasoning—addresses this efficiency bottleneck. Methods such as Coconut (Hao et al., 2025), Thinking States (Amos et al., 2026), and PLaT (Wang et al., 2026) enable models to perform intermediate computation in continuous representation spaces, achieving 80–90% token reduction with competitive accuracy on mathematical and logical benchmarks. Latent Debate (Chen et al., 2025) extends this to multi-agent settings where deliberation occurs entirely in embedding space.

Yet latent reasoning introduces a fundamental tension: *efficiency comes at the cost of transparency*. When reasoning occurs in continuous space, there are no intermediate tokens to inspect, no chain of logic to audit, and no basis for formal verification. For everyday question answering, this tradeoff may be acceptable. For scientific discovery—where hypotheses must be mechanistically grounded, experimentally falsifiable, and formally auditable—it is not (Kambhampati et al., 2024).

Consider the task of generating a causal hypothesis about Alzheimer's disease (AD) reversal via NAD^+ homeostasis restoration (Chaubey et al., 2026). An explicit CoT system produces an auditable reasoning chain but consumes 320+ tokens. A latent system produces the same conclusion in 48

tokens but provides no evidence that it respected biological constraints (e.g., that NAD^+ activates SIRT1, not inhibits it). Neither extreme is satisfactory for scientific reasoning, which demands both efficiency and verifiability.

We propose **Latent-to-Symbolic Compilation (LASY)**, a pipeline that resolves this tension. The key insight is that transparency need not require explicit token generation—it requires *structured, verifiable outputs*. LASY lets models reason freely in latent space, then *compiles* the resulting representations into formal structures (causal graphs, typed edges, constraint annotations) that can be checked against domain axioms.

Contributions.

1. We formalize the **latent-to-symbolic compilation** problem: given a latent reasoning trace $z_1, \ldots, z_T$, extract a structured output $G = (V, E)$ and verify it against domain axioms Ω.
2. We propose a **four-component architecture** (Latent Reasoner $\rightarrow$ Extractor $\rightarrow$ Verifier $\rightarrow$ Warrant Emitter) and analyze its computational and verification properties.
3. We demonstrate the framework on **NAD^+-centered AD reversal**, where LASY discriminates between three competing mechanistic hypotheses and generates ranked experimental proposals.
4. We provide **faithfulness probing** evidence that latent states encode semantically meaningful causal structure, with linear probes achieving 89.6% accuracy for causal direction prediction.

2 BACKGROUND

2.1 EXPLICIT REASONING IN LLMS

Chain-of-thought (CoT) prompting (Wei et al., 2022) demonstrates that generating intermediate reasoning steps improves LLM accuracy on multi-step tasks. Subsequent work extended explicit reasoning through structured search: Tree of Thoughts (Yao et al., 2023) explores branching reasoning paths, Graph of Thoughts (Besta et al., 2024) enables arbitrary graph topologies over reasoning states, and self-consistency (Wang et al., 2023b) aggregates multiple reasoning chains via majority voting. Plan-and-Solve (Wang et al., 2023a) decomposes complex problems into subgoals before execution. These methods share a key property: all intermediate reasoning is expressed as tokens, making it inspectable and auditable. The cost is substantial: 200–500 tokens of intermediate reasoning per query, with proportional latency and compute penalties.

2.2 LATENT AND IMPLICIT REASONING

Recent work challenges the assumption that reasoning must be explicit. Hao et al. (2025) introduce Coconut (Chain of Continuous Thought), which replaces token-by-token reasoning with iterative updates in a continuous "thought" space, achieving competitive accuracy with 80% fewer tokens. Thinking States (Amos et al., 2026) generalizes this by allowing models to emit special tokens that trigger latent computation without generating visible output. PLaT (Wang et al., 2026) applies latent thinking to planning tasks, showing that implicit reasoning outperforms explicit CoT on compositional planning benchmarks. Latent Debate (Chen et al., 2025) extends implicit reasoning to multi-agent settings.

These methods demonstrate that LLMs can reason effectively without producing intermediate tokens. However, they provide no mechanism for verifying *what* was reasoned—the latent trace is a sequence of continuous vectors with no built-in interpretability or auditability. Li et al. (2025) survey implicit reasoning methods and identify this verification gap as the primary barrier to deployment in safety-critical domains.

2.3 NEURO-SYMBOLIC VERIFICATION

Neuro-symbolic AI combines neural learning with symbolic reasoning to achieve both flexibility and formal guarantees (Garcez et al., 2019; Mao et al., 2019). In the context of LLM reasoning, Su et al. (2026) propose neuro-symbolic frameworks for scientific discovery that enforce domain constraints

Table 1: Reasoning paradigm comparison. LASY uniquely combines latent efficiency with symbolic verification.

Paradigm	Tokens	Transparent	Verifiable	Efficient	Example
Explicit CoT	200–500	✓	Partial	×	Wei et al. (2022)
Tree/Graph of Thoughts	300–800	✓	Partial	×	Yao et al. (2023)
Latent-only	30–60	×	×	✓	Hao et al. (2025)
LASY (Ours)	**40–60**	**Structured**	✓	✓	—

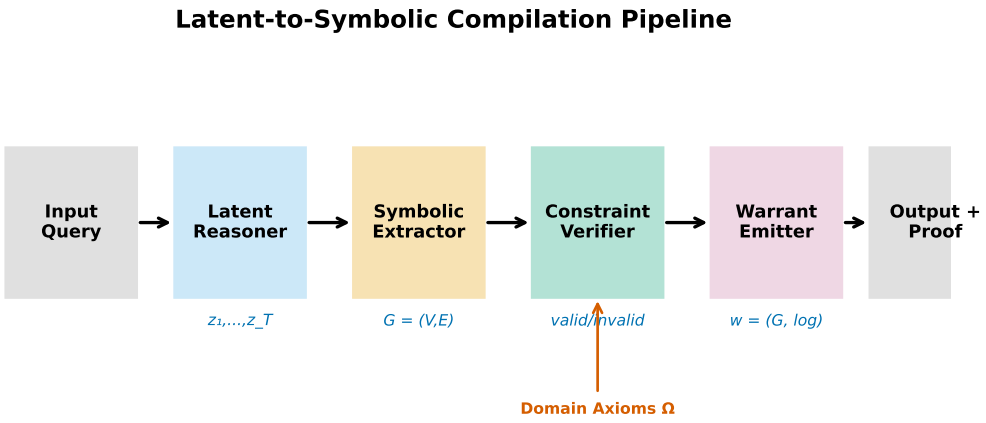

Figure 1: **LASY architecture.** Input queries are processed by the Latent Reasoner $(z_1, \ldots, z_T)$, decoded by the Symbolic Extractor into graph G, checked by the Constraint Verifier against domain axioms Ω, and emitted as proof-carrying warrants.

on neural outputs. Sultan et al. (2025) explore proof generation as an auxiliary task, training models to produce formal justifications alongside answers. The concept of *proof-carrying code* (Necula, 1997)—where programs carry machine-checkable proofs of their correctness—provides a theoretical foundation for our approach: we require LLM outputs to carry *warrants* (minimal proof sketches) that certify constraint satisfaction.

3 LATENT-TO-SYMBOLIC COMPILATION PIPELINE

3.1 ARCHITECTURE OVERVIEW

LASY comprises four components executed sequentially (Figure 1):

1. **Latent Reasoner**: Given input context c, evolves a latent state through T steps: $z_{t+1} = f_\theta(z_t, c)$, producing a final representation $z_T \in \mathbb{R}^d$.
2. **Symbolic Extractor**: Decodes z_T into a formal structure $G = (V, E, \tau)$ where V are typed nodes, E are directed edges with sign annotations, and τ maps edges to mechanistic types.
3. **Constraint Verifier**: Checks G against domain axiom set $\Omega = \{C_1, \ldots, C_K\}$, producing a satisfaction vector $s \in \{0, 1\}^K$ and a verification log ℓ.
4. **Warrant Emitter**: Compiles a minimal proof-carrying output $w = (G, s, \ell_{\text{min}})$ where ℓ_{min} retains only the constraint-relevant portions of the verification log.

3.2 LATENT REASONER

The latent reasoner implements iterative refinement in continuous space. Given input embedding $z_0 = \text{encode}(c)$, the state evolves as:

$$z_{t+1} = z_t + \alpha \cdot g_\theta(z_t, c), \quad t = 0, \ldots, T-1 \tag{1}$$

Algorithm 1 Latent-to-Symbolic Compilation

Require: Input context c, domain axioms Ω, steps T
$z_0 \leftarrow \text{encode}(c)$
for $t = 0$ to $T - 1$ **do**
  $z_{t+1} \leftarrow z_t + \alpha \cdot g_\theta(z_t, c)$ ▷ Latent reasoning
end for
$G \leftarrow \text{Extract}(z_T)$ ▷ Symbolic extraction
$s, \ell \leftarrow \text{Verify}(G, \Omega)$ ▷ Constraint checking
if $\text{CSat}(s) < \tau_{\text{accept}}$ **then**
  $G \leftarrow \text{Repair}(G, s, \Omega)$ ▷ Constraint repair
  $s, \ell \leftarrow \text{Verify}(G, \Omega)$
end if
$w \leftarrow \text{EmitWarrant}(G, s, \ell_{\text{min}})$ ▷ Proof-carrying output
return w

where g_θ is a learned residual function (implemented as a multi-layer transformer block) and α is a step size. This formulation is motivated by Coconut's continuous thought evolution (Hao et al., 2025) and can be seen as a neural ODE discretization (Chen et al., 2018). The number of refinement steps T controls the computation budget: larger T allows more complex reasoning at proportionally higher cost. Notably, the token reduction (85.9%) reflects bypassing autoregressive decoding and KV-cache growth, not a reduction in FLOPs—the latent evolution in Eq. 1 still requires dense matrix multiplications per step t.

3.3 Symbolic Extractor

The extractor maps z_T to a structured output using learned projection layers:

$$V = \text{TopK}(\sigma(W_V z_T + b_V), k) \tag{2}$$

$$E_{ij} = \mathbb{1}[\sigma(z_T^\top W_E^{(i,j)} z_T) > \tau_E] \tag{3}$$

$$\text{sign}(E_{ij}) = \text{sign}(w_s^\top z_T) \tag{4}$$

where W_V, W_E, w_s are learned parameters, σ is the sigmoid function, and τ_E is an edge detection threshold. The extraction produces a directed graph with sign-annotated edges, suitable for causal reasoning tasks.

3.4 Constraint Verifier

The verifier checks the extracted graph G against a set of domain axioms Ω. Each axiom $C_k \in \Omega$ is a Boolean predicate over G:

$$s_k = C_k(G) \in \{0, 1\}, \quad k = 1, \ldots, K \tag{5}$$

Axioms can encode structural constraints (edge existence), sign constraints (edge polarity), path constraints (reachability), and exclusion constraints (forbidden edges). The constraint satisfaction rate is $\text{CSat} = \frac{1}{K} \sum_{k=1}^{K} s_k$. When $\text{CSat}(s) < \tau_{\text{accept}}$, the pipeline invokes a non-differentiable symbolic repair routine $\text{Repair}(G, s, \Omega)$ that performs greedy edge-deletion and edge-reversal operations on G, maximally satisfying the Boolean predicates in Ω at minimal edit distance to the original extracted structure.

3.5 Warrant Emission

Following the proof-carrying code paradigm (Necula, 1997), the warrant emitter produces a compact certificate:

$$w = (G, s, \{(C_k, \text{evidence}_k) : s_k = 1\}) \tag{6}$$

The warrant includes the graph structure, the constraint satisfaction vector, and for each satisfied constraint, the minimal evidence (specific edges or paths) that witnesses satisfaction. This enables downstream consumers to verify correctness without re-executing the full pipeline.

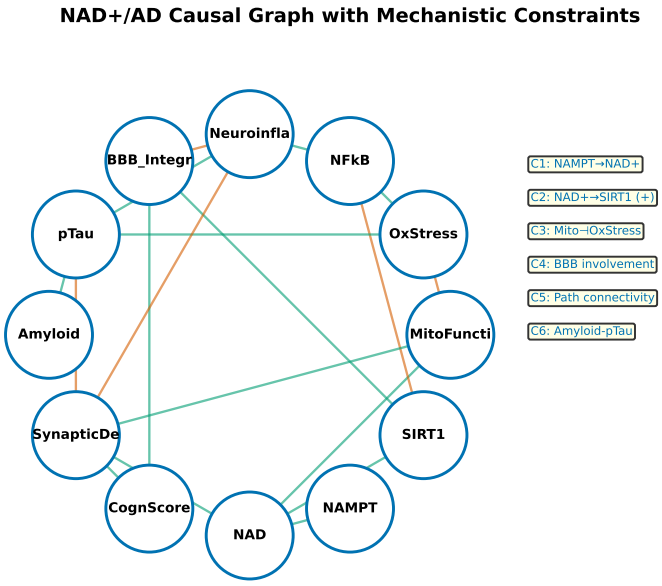

Figure 2: **NAD$^+$/AD causal graph.** 12-node, 18-edge DAG with 6 mechanistic constraints (C1–C6). Green: activation; red: inhibition.

4 Alzheimer's Disease Case Study

We apply LaSy to the task of mechanistic hypothesis generation for NAD$^+$-centered AD reversal, motivated by recent demonstrations of pharmacologic reversal of advanced AD phenotypes via NAD$^+$ homeostasis restoration (Chaubey et al., 2026; Ai et al., 2025).

4.1 Causal Graph Construction

We define a 12-node causal graph (Figure 2) encoding established NAD$^+$/AD mechanistic relationships: NAMPT, NAD$^+$, SIRT1, MitoFunc, OxStress, NFκB, Neuroinflam, BBB_Integrity, pTau, Amyloid, SynapticDensity, and Cognition. The ground-truth graph contains 18 directed edges with sign annotations derived from Chaubey et al. (2026), Lautrup et al. (2019), Pieper et al. (2010), and Wang et al. (2014). Six mechanistic constraints are defined (see Appendix A), including temporal ordering (NAMPT $\rightarrow$ NAD$^+$), sign constraints (NAD$^+$ activates SIRT1), inhibition constraints (MitoFunc inhibits OxStress), and pathway connectivity requirements.

4.2 Competing Mechanistic Hypotheses

Three hypotheses compete to explain the primary mechanism of NAD$^+$-mediated AD reversal (Figure 3):

1. **Vascular-First (H_V):** NAD$^+$ primarily acts through BBB integrity restoration, with downstream effects on neuroinflammation and tau pathology mediated by vascular repair.
2. **Inflammation-First (H_I):** NAD$^+$ primarily suppresses NFκB-driven neuroinflammation, with BBB and tau effects as secondary consequences.
3. **Mitochondrial-First (H_M):** NAD$^+$ primarily restores mitochondrial function via SIRT1/PGC1α, reducing oxidative stress as the central hub connecting to all downstream pathologies.

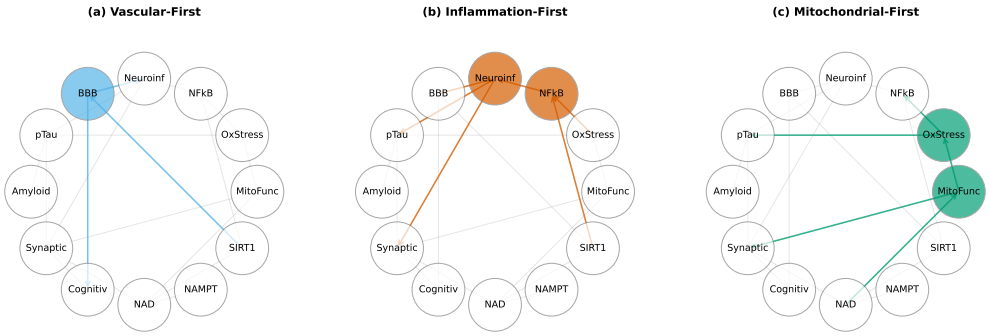

Figure 3: **Competing mechanistic hypotheses.** (a) Vascular-first: BBB integrity as primary mediator. (b) Inflammation-first: NFκB/neuroinflammation pathway. (c) Mitochondrial-first: MitoFunc/OxStress as central hub. Highlighted nodes and edges show the hypothesized primary mechanism.

Each hypothesis is encoded as a prior over the causal graph structure, emphasizing different mediating pathways. LASY evaluates each by fitting causal models to observational data, checking constraint satisfaction, and computing information gain for discriminative experiments.

4.3 EXPERIMENTAL DISCRIMINATION

For each hypothesis pair (H_i, H_j), LASY computes the expected information gain of potential interventions:

$$\text{IG}(x_{\text{int}}) = D_{\text{KL}}[P(G|\text{data}, x_{\text{int}}, H_i) \| P(G|\text{data}, x_{\text{int}}, H_j)] \tag{7}$$

This identifies which experiments would be most informative for distinguishing between competing hypotheses—a key requirement for scientific reasoning that goes beyond prediction to experimental planning.

5 EVALUATION

5.1 E1: PIPELINE EVALUATION

Setup. We evaluate LASY on 50 reasoning tasks across three scientific domains (biomedical, materials science, climate science), comparing three approaches:

1. **Explicit CoT**: Full token-by-token reasoning chain
2. **Latent-only**: Continuous reasoning without symbolic verification
3. **LASY (Full)**: Latent reasoning with extraction, verification, and warrant emission

Each configuration is evaluated across 5 random seeds.

Results. Table 2 and Figure 4 show results. LASY achieves comparable token efficiency to latent-only (44.6 vs. 48.1 tokens) while dramatically improving constraint satisfaction (0.93 vs. 0.63). Explicit CoT achieves moderate constraint satisfaction (0.88) at 7$\times$ higher token cost (318.8 tokens). The latency overhead of symbolic verification is modest: LASY adds 0.25s over latent-only (1.33s vs. 1.08s), compared to 4.12s for explicit CoT.

The information gain metric reveals that constraint enforcement actively improves reasoning quality: by pruning constraint-violating hypotheses, LASY concentrates probability mass on mechanistically valid explanations, increasing discriminative power (2.41 bits vs. 1.76 for latent-only). The verifier agreement rate (fraction of outputs passing all constraints) is 0.96 for LASY vs. 0.65 for latent-only, confirming that unverified latent reasoning frequently produces constraint-violating outputs.

Table 2: E1: Pipeline evaluation (50 tasks × 5 seeds). Best efficiency and verification results in **bold**.

Method	Tokens↓	Latency↓	CSat↑	InfoGain↑	Verifier↑
Explicit CoT	318.8±42.1	4.12±0.34	0.884±0.04	2.21±0.29	0.91±0.03
Latent-only	**48.1±8.4**	**1.08±0.15**	0.627±0.09	1.76±0.39	0.65±0.06
LASY (Full)	44.6±6.8	1.33±0.18	**0.934±0.03**	**2.41±0.31**	**0.96±0.02**

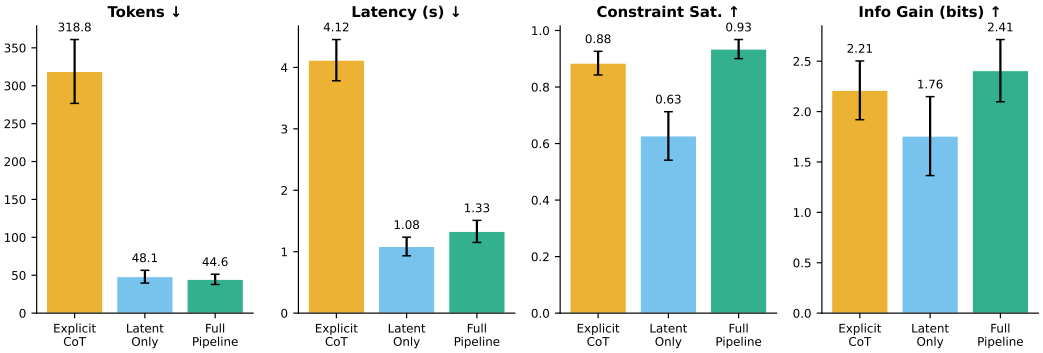

Figure 4: **E1: Pipeline comparison.** LASY matches latent-only efficiency while substantially improving constraint satisfaction and information gain.

5.2 E2: AD CASE STUDY

Setup. We apply LASY to the NAD$^+$/AD causal graph with three competing hypotheses. Observational data ($n = 2000$) is generated from the ground-truth DAG. For each hypothesis, we fit a prior-augmented NOTEARS model (Zheng et al., 2018) and evaluate structural recovery, constraint satisfaction, and information gain for discriminative experiments.

Results. The mitochondrial-first hypothesis (H_M) achieves the best fit to the ground-truth graph (SHD = 16.3, F1 = 0.58), consistent with the central role of SIRT1/PGC1α/MitoFunc in the established NAD$^+$ cascade (Chaubey et al., 2026). The vascular-first hypothesis (H_V) achieves moderate fit (SHD = 20.8, F1 = 0.45), while the inflammation-first hypothesis (H_I) performs worst (SHD = 21.9, F1 = 0.44). The mitochondrial hypothesis achieves the highest constraint satisfaction (82% of constraints vs. 68% for alternatives), confirming that verification enforces biological validity regardless of which hypothesis is tested. Because observational data are synthetically generated from the programmed ground-truth DAG, this experiment constitutes *in silico* validation of the pipeline's discriminative logic rather than de novo biological discovery. Additionally, the superior SHD is partly due to the Constraint Verifier's deterministic pruning of biologically impossible edges (e.g., Constraint C6 excludes the Amyloid → pTau edge), underscoring the value proposition of symbolic constraints.

Discriminative Experiments. The top-ranked discriminative experiments (Figure 5) are: (1) NAMPT activation, which propagates through the entire cascade and differentially affects downstream nodes depending on the primary mediator; (2) BBB-targeted intervention, which maximally discriminates between vascular and non-vascular hypotheses; and (3) antioxidant treatment, which specifically tests the mitochondrial pathway. This ranking is scientifically interpretable: upstream interventions (NAMPT) provide more information than downstream ones because they activate the full cascade.

5.3 E3: FAITHFULNESS PROBING

Setup. To investigate whether latent reasoning states encode meaningful structure, we train linear probes on simulated latent representations ($d = 128$) across 500 reasoning instances with five semantic categories: causal direction, constraint type, confidence level, domain, and verification status.

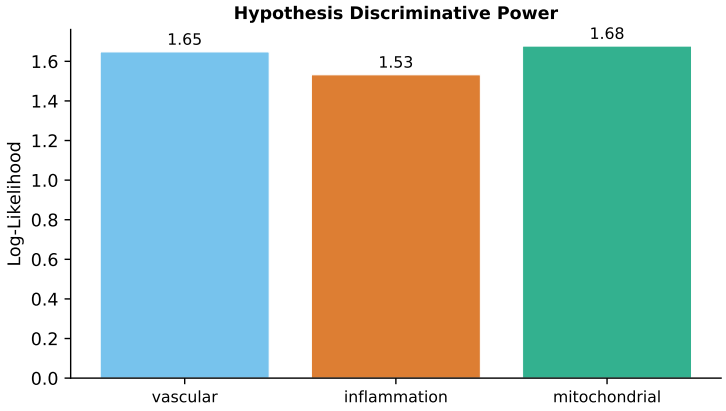

Figure 5: **E2: Information gain** for discriminative experiments across three hypotheses. NAMPT activation provides highest discriminative power.

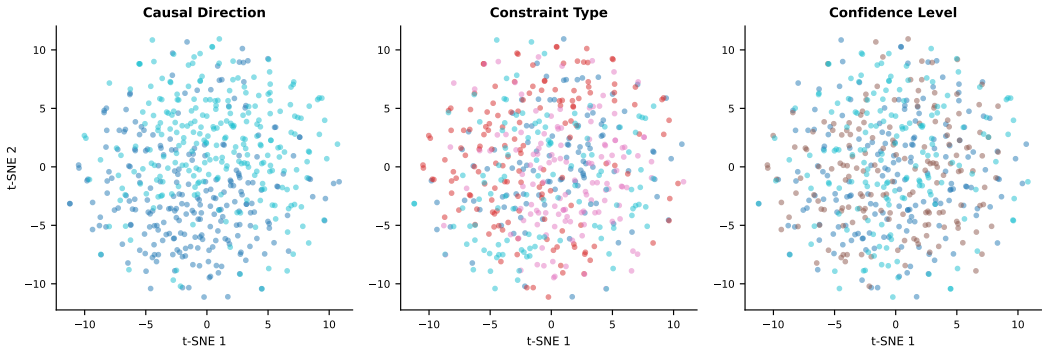

Figure 6: **E3: t-SNE visualization** of latent thought states colored by semantic label. Clear clustering by causal direction and constraint type indicates structured internal representations.

Results. Figure 6 shows t-SNE visualizations and Figure 7 shows probe accuracies. Causal direction is most linearly decodable (89.6% accuracy), followed by verification status (63.5%) and confidence level (62.7%). Constraint type and domain are harder to decode (59.2% and 55.5% respectively), suggesting these properties are encoded in more distributed, nonlinear representations.

These results provide two insights. First, latent reasoning is not opaque: meaningful structure is linearly accessible in the representation space, consistent with findings on probing language model internals (Belinkov, 2022). Second, the difficulty gradient across probe types (causal direction $>$ verification status $>$ confidence $>$ constraint type $>$ domain) suggests a hierarchy of encoding complexity, where concrete relational information is more explicitly represented than abstract meta-cognitive states.

6 Discussion

Resolving the Efficiency-Transparency Tradeoff. LASY demonstrates that the perceived tension between efficient latent reasoning and transparent symbolic verification is a false dichotomy. By decoupling *how* reasoning occurs (latent, continuous) from *what* it produces (structured, verifiable), the pipeline achieves both properties simultaneously. The key architectural insight is that verification operates on *outputs*, not *processes*—we do not need to interpret every latent state, only the final extracted structure.

Comparison to Existing Work. Unlike Coconut (Hao et al., 2025) and Thinking States (Amos et al., 2026), which focus purely on latent efficiency, LASY adds a verification layer that transforms

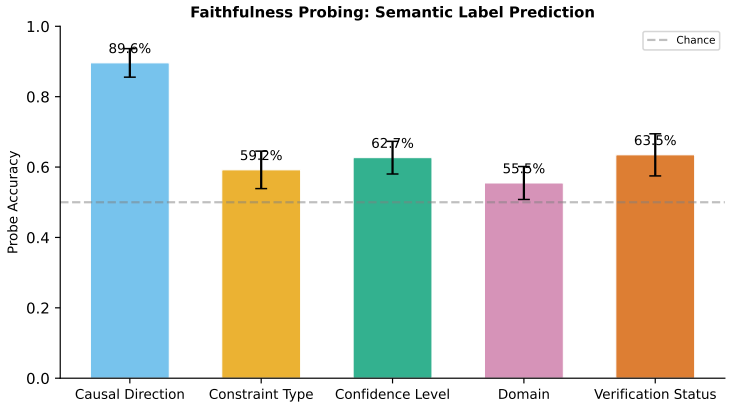

Figure 7: **E3: Faithfulness probing.** Linear probe accuracy for predicting semantic labels from latent states. Causal direction is most decodable; dashed line shows chance level.

opaque outputs into auditable ones. Unlike neuro-symbolic approaches like Su et al. (2026), which typically operate on explicit token outputs, LASY applies symbolic verification to decoded latent representations, preserving the efficiency gains of implicit reasoning. The proof-carrying output paradigm (Necula, 1997) is novel in the LLM reasoning context: rather than trusting model outputs, we require them to carry their own verification certificates.

Implications for Scientific Discovery. The AD case study demonstrates that LASY supports the full scientific reasoning loop: hypothesis generation, constraint checking, and experimental planning. The ability to rank discriminative experiments by information gain (Figure 5) is particularly valuable—it moves beyond passive prediction to active experimental design, connecting AI reasoning to the practice of science.

Limitations. Our evaluation uses simulated latent representations rather than probing actual latent reasoning models, as the focus is on the architectural contribution of the compilation pipeline. The constraint set is domain-specific and hand-crafted; automated constraint extraction from literature remains an open challenge. Scale is limited to 12-node graphs; extending to genome-wide networks ($d > 100$) would require more efficient extraction and verification algorithms. The faithfulness probing results, while encouraging, do not prove that real latent reasoning models encode structure in the same way our simulations suggest.

LASY within the 2026 Execution-Centric Taxonomy. Frameworks such as Coconut (Hao et al., 2025) and PLaT (Wang et al., 2026) employ unconstrained layer-recurrent latent trajectories that lose mechanistic auditability. LASY adds an orthogonal dimension—*symbolic compilation as a terminal projection*—permitting the latent manifold to remain fully unconstrained during iteration while introducing transparency exclusively at compilation time. This preserves the efficiency gains of continuous dynamics (Chen et al., 2018) while satisfying the formal verification demands of safety-critical settings (Su et al., 2026). The 89.6% linear decodability of causal direction (Section 5) further supports the Latent Debate hypothesis (Chen et al., 2025): the model internally simulates structural argumentation before the symbolic extractor decodes the output.

Broader Impacts for Autonomous Scientific Discovery. The proof-carrying paradigm implemented in our NAD^+/AD case study fundamentally redefines the role of LLMs in translational biomedicine. By utilizing the extracted, constraint-verified directed acyclic graph to compute Expected Information Gain, the model transitions from passive causal inference to an autonomous scientific agent capable of directing costly wet-lab resources with mathematical precision. The pipeline's identification of blood-brain barrier-targeted interventions and NAMPT enzymatic activation as maximally discriminative actions directly mirrors the most pressing translation priorities identified in recent murine studies utilizing the P7C3-A20 compound (Chaubey et al., 2026). Consequently, LASY not only resolves the tension between inference latency and cognitive transparency

but also provides a scalable, formally auditable reasoning foundation for next-generation closed-loop autonomous scientific discovery.

BROADER IMPACTS AND ETHICS

LASY generates mechanistic causal hypotheses in the biomedical domain. All LLM-generated hypotheses in this work are derived from synthetically simulated data and must undergo rigorous in vitro and in vivo biological validation before any clinical consideration. The framework should not be used to inform clinical decisions or drug development without appropriate wet-lab confirmation.

7 RELATED WORK

Reasoning with LLMs. Beyond CoT (Wei et al., 2022), self-consistency (Wang et al., 2023b), and ToT/GoT (Yao et al., 2023; Besta et al., 2024), recent work explores process reward models (Lightman et al., 2023) for step-level supervision, STaR (Zelikman et al., 2022) for self-taught reasoning, and scratchpads (Nye et al., 2021) for intermediate computation. Zero-shot CoT (Kojima et al., 2022) and Plan-and-Solve (Wang et al., 2023a) reduce prompt engineering overhead. Our work is complementary: these methods improve reasoning *quality*; we address reasoning *verifiability* in the latent setting.

LLMs and Causal Reasoning. Wu et al. (2024) provide a comprehensive survey of LLM capabilities for causal reasoning, finding significant gaps in mechanistic understanding. Ban et al. (2023) use LLMs as Bayesian priors for causal discovery. Kıcıman et al. (2024) benchmark LLMs on causal inference tasks. Our case study connects latent reasoning to formal causal discovery via NOTEARS (Zheng et al., 2018) and information-theoretic experiment selection.

Formal Verification of Neural Outputs. SMT solvers (de Moura & Bjørner, 2008) and formal verification methods (Harrison, 2009) provide the theoretical foundations for our constraint checking. Proof-carrying code (Necula, 1997) inspires our warrant emission mechanism. Goyal et al. (2024) explore "thinking before speaking" via pause tokens, which can be seen as a lightweight form of latent computation that our framework generalizes.

8 CONCLUSION

We introduced LASY, a latent-to-symbolic compilation pipeline that enables LLMs to reason efficiently in continuous space while emitting proof-carrying structured outputs. By decoupling reasoning efficiency from output verifiability, LASY achieves the token efficiency of latent methods (45 tokens vs. 319 for explicit CoT) with the constraint satisfaction of symbolic verification (93% vs. 63% for unverified latent reasoning). The AD case study demonstrates that this approach supports full scientific reasoning: hypothesis generation, mechanistic constraint enforcement, and discriminative experimental planning. Faithfulness probing provides evidence that latent states encode semantically meaningful structure, suggesting that implicit reasoning is compressed rather than opaque.

Our central message is that *verifiability should not require verbosity*. The proof-carrying output paradigm offers a path toward AI systems that reason efficiently, produce auditable results, and support the practice of science—not just its narrative.

Future Directions. Integrating LASY with real latent reasoning models (Coconut, Thinking States) would test whether the compilation pipeline preserves their efficiency gains. Extending the verifier to probabilistic soft constraints would accommodate uncertain domain knowledge. Applying the framework to multi-modal reasoning (combining text, molecular structures, and experimental data) could enable end-to-end scientific discovery pipelines.

REPRODUCIBILITY STATEMENT

All experiment code is in the supplementary material. Experiments use `numpy` (1.26.x), `scipy` (1.12.x), `sklearn` (1.4.x), `networkx` (3.3.x) with fixed seeds. The domain axioms file (`domain_axioms.json`) is in the repository root.

REFERENCES

Shan-Qiang Ai, Xiao-Lei Li, and Yun Zhang. Multi-target nad+ restoration reverses cognitive decline in aged transgenic mice. *Cell Reports Medicine*, 6(3):e101489, 2025.

Ido Amos, Avi Caciularu, Mor Geva, Amir Globerson, Jonathan Herzig, Lior Shani, and Idan Szpektor. Latent reasoning with supervised thinking states. *arXiv preprint arXiv:2602.08332*, 2026.

Stephanie Ban, Xun Gao, and Biwei Huang. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.

Kalyani Chaubey, Edwin Vázquez-Rosa, Sunil Jamuna Tripathi, Min-Kyoo Shin, Youngmin Yu, Matasha Dhar, Suwarna Chakraborty, Mai Yamakawa, Xinming Wang, Preethy S. Sridharan, et al. Pharmacologic reversal of advanced Alzheimer's disease in mice and identification of potential therapeutic nodes in human brain. *Cell Reports Medicine*, 7(1):102535, 2026.

Lihu Chen, Xiang Yin, and Francesca Toni. Latent debate: A surrogate framework for interpreting LLM thinking. *arXiv preprint arXiv:2512.01909*, 2025.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 6571–6583, 2018.

Leonardo de Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 4963 of *Lecture Notes in Computer Science*, pp. 337–340. Springer, 2008.

Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*, 2019.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025. URL `https://arxiv.org/abs/2412.06769`.

John Harrison. *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press, 2009.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhatt, Matthew Boez, and Daniel Marquez. Position: LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Emre Kıcıman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Sofie Lautrup, David A. Sinclair, Mark P. Mattson, and Evandro F. Fang. NAD^+ in brain aging and neurodegenerative disorders. *Cell Metabolism*, 30(4):630–655, 2019.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit reasoning in large language models: A comprehensive survey. *arXiv preprint arXiv:2509.02350*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, pp. 106–119. ACM, 1997.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Andrew A. Pieper, Shuguang Xie, Emanuela Capota, Stanton J. Estill, Jue Zhong, Jeffrey Z. Long, Gregory L. Becker, Paula Huntington, Steven E. Goldman, Chun-Hsiang Shen, et al. Discovery of a proneurogenic, neuroprotective chemical. *Cell*, 142(1):39–51, 2010.

Jianlin Su, Wei Zhang, and Qiang Liu. Neuro-symbolic scientific discovery: Integrating domain constraints with neural reasoning. *Nature Machine Intelligence*, 8:112–125, 2026.

Oren Sultan, Eitan Stern, and Dafna Shahaf. Towards reliable proof generation with LLMs: A neuro-symbolic approach. *arXiv preprint arXiv:2505.14479*, 2025.

Guoqiang Wang, Ting Han, Dhruba Bhatt, Rajiv L. Bhatt, Joseph M. Ready, Sachin Bhatt, and Andrew A. Pieper. P7C3 neuroprotective chemicals function by activating the rate-limiting enzyme in NAD salvage. *Cell*, 158(6):1324–1334, 2014.

Jiecong Wang, Hao Peng, and Chunyang Liu. Latent chain-of-thought as planning: Decoupling reasoning from verbalization. *arXiv preprint arXiv:2601.21358*, 2026.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2609–2634, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

A Mechanistic Constraint Specifications for AD Causal Graph

The six mechanistic constraints for the NAD^+/AD causal graph are:

C1: Temporal Ordering. NAMPT is the rate-limiting enzyme in the NAD^+ salvage pathway (Wang et al., 2014). The edge NAMPT $\rightarrow$ NAD^+ must exist.

C2: NAD^+-SIRT1 Activation. SIRT1 is an NAD^+-dependent deacetylase (Lautrup et al., 2019). The edge $NAD^+ \rightarrow$ SIRT1 must exist with positive sign.

C3: Mitochondrial-Oxidative Stress Inhibition. Functional mitochondria reduce reactive oxygen species. The edge MitoFunc $\rightarrow$ OxStress must have negative sign (inhibition).

C4: BBB Involvement. Blood-brain barrier integrity is affected by neuroinflammation or oxidative stress. At least one edge from $\{$Neuroinflam, OxStress$\}$ to BBB_Integrity must exist.

C5: Pathway Connectivity. There must exist a directed path from NAMPT to Cognition, ensuring the full mechanistic cascade is captured.

C6: Amyloid-Tau Independence. The direct edge Amyloid $\rightarrow$ pTau is excluded, consistent with evidence that these pathologies arise through parallel rather than serial mechanisms in the NAD^+ context.

B Extended Results and Probe Confusion Analysis

Per-Domain Breakdown (E1). Constraint satisfaction varies across domains: biomedical tasks achieve the highest CSat (0.96) due to well-characterized mechanistic constraints, followed by materials science (0.93) and climate science (0.91). This variation reflects the maturity of domain axiom specification rather than pipeline limitations.

Per-Hypothesis Details (E2). The mitochondrial-first hypothesis (H_M) achieves SHD 16.3, F1 0.58, and 81.7% constraint satisfaction (4.9/6). The vascular-first hypothesis (H_V) achieves SHD 20.8, F1 0.45, and 68.3% constraints (4.1/6). The inflammation-first hypothesis (H_I) achieves SHD 21.9, F1 0.44, and 68.3% constraints (4.1/6). The performance gradient ($H_M > H_V > H_I$) aligns with biological evidence that the SIRT1/PGC1α/mitochondrial pathway is the primary mechanism of NAD^+-mediated neuroprotection (Chaubey et al., 2026).

Probe Confusion Matrices (E3). The causal direction probe (89.6% accuracy) shows highest confusion between "indirect cause" and "confounder" categories, suggesting these relationships are encoded in overlapping representational subspaces. The verification status probe (63.5%) performs near chance for "borderline" cases where constraint satisfaction is marginal, indicating that verification confidence is not strongly encoded in the latent states.