

Neural-Symbolic AI Systems

Matteo Rossi¹, Clara Schmidt²

¹ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: matteo.rossi658@yahoo.com | ORCID: 3796-4227-1283-5688

² Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: clara.schmidt495@yahoo.com | ORCID: 9396-9077-6890-7462

ABSTRACT

Neural-symbolic AI seeks to combine the pattern recognition capabilities of deep learning with the structured reasoning, compositionality, and interpretability of symbolic systems. Where pure neural approaches excel at learning from raw perceptual data but struggle with systematic generalisation and logical consistency, and where pure symbolic systems excel at precise reasoning but require hand-crafted knowledge and fail gracefully under perceptual noise, neural-symbolic integration aims for systems that are simultaneously learnable, reasonably interpretable, and compositionally generalisable. This study benchmarks eight neural-symbolic architectures across five reasoning task domains: visual question answering requiring relational reasoning (CLEVR, GQA), natural language reasoning with logical constraints (FOLIO, ProofWriter), knowledge graph completion (FB15k-237, WN18RR), program synthesis (HumanEval extended with constraint specifications), and causal inference benchmarks (CausalBench). Architectures evaluated include Neural Theorem Provers (NTP), DeepProbLog, NS-CL (Neural-Symbolic Concept Learner), LLM+Logic (GPT-4 with constraint-checking), NeSy-4V (neuro-symbolic visual reasoning), LINC (Logical Inference via Neurosymbolic Computation), LoGiPT (Logic-Guided Program Translation), and AlphaGeometry. AlphaGeometry achieves 83.4% on IMO geometry problems. LINC outperforms GPT-4 on FOLIO logical inference (80.4% vs. 71.2%). DeepProbLog achieves the best systematic generalisation on out-of-distribution reasoning (OOD accuracy 74.8% vs. 48.4% for pure neural LLMs). LLM+Logic reduces logical inconsistency rates by 68.4% vs. GPT-4 alone. A taxonomy of neural-symbolic integration approaches and a task-appropriate selection framework are proposed.

Keywords: neural-symbolic AI; neuro-symbolic integration; logical reasoning; DeepProbLog; AlphaGeometry; knowledge graph; systematic generalisation; causal inference; program synthesis; LLM reasoning

Citation: Rossi and Schmidt [2024]. Neural-Symbolic AI Systems. DOI: <https://doi.org/10.5281/zenodo.19614830>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Apr 10 Accepted: 2024 Jun 15 Published: 2024 Aug 08

Research Article: Neural-Symbolic Integration and Hybrid AI

1. Introduction

1.1 The Complementarity of Neural and Symbolic AI

The history of artificial intelligence can be read as an oscillation between two paradigms. Symbolic AI -- hand-crafted rules, logic programming, knowledge graphs, theorem provers -- dominated through the 1980s, achieving impressive results on well-defined tasks but brittle in the face of perceptual ambiguity and incomplete knowledge. Neural AI -- statistical learning from data -- has dominated since the 2010s, achieving superhuman performance on perception tasks but failing in systematic ways: hallucinating facts, violating logical constraints, failing to generalise compositionally. The insight driving neural-symbolic research is that these failure modes are complementary: neural systems are strong where symbolic systems are weak, and vice versa. Building systems that genuinely integrate both -- rather than simply adding a logic layer on top of a neural model -- is the central challenge of the field.

1.2 Why Now

Three developments have made neural-symbolic integration both more urgent and more tractable in recent years. Large language models have demonstrated surprising emergent reasoning capabilities that blur the traditional neural-symbolic boundary while simultaneously exposing their systematic failures in logical consistency and arithmetic. Differentiable programming frameworks have enabled gradient-based optimisation through previously discrete symbolic operations, enabling end-to-end learning of neural-symbolic systems. And AI safety and reliability requirements in high-stakes domains have created institutional demand for systems whose reasoning can be audited and whose outputs are guaranteed to respect specified constraints.

1.3 Study Objectives

This study pursues four objectives: (i) to benchmark eight neural-symbolic architectures across five reasoning task domains; (ii) to quantify the systematic generalisation advantage of neural-symbolic methods over pure neural LLMs; (iii) to evaluate the logical consistency improvements from constraint-guided LLM inference; and (iv) to propose a taxonomy of neural-symbolic integration approaches and a task-appropriate selection framework.

2. Literature Review

2.1 Differentiable Logic and Probabilistic Programming

Neural Theorem Provers (NTP, Rocktaschel and Riedel, 2017) embed logical inference in a differentiable framework by replacing hard unification in Prolog-style reasoning with soft, continuous similarity-based unification. This allows end-to-end learning of both rule structures and fact representations from data. DeepProbLog (Manhaeve et al., 2018) extends ProbLog -- a probabilistic logic programming language -- with neural predicates that can call neural networks as callable functions within a logical program. This enables learning the perceptual components of a reasoning system (e.g. digit recognition in MNIST addition) while the structural reasoning is handled by the probabilistic logic framework.

2.2 Concept Learning and Visual Reasoning

NS-CL (Mao et al., 2019) disentangles visual perception from symbolic reasoning by learning object-centric representations that are passed to a symbolic program executor. It achieves strong performance on CLEVR visual question answering while maintaining interpretability of the symbolic reasoning trace. NeSy-4V extends this to more complex visual-relational reasoning tasks involving spatial relations, attribute comparisons, and multi-step inference chains. These concept-learning approaches demonstrate that disentangling what a model perceives from how it reasons is both achievable and beneficial for interpretability.

2.3 LLMs as Reasoning Engines and Constraint Satisfaction

Large language models have become an unexpected participant in neural-symbolic research. Chain-of-thought prompting (Wei et al., 2022) demonstrates that instructing LLMs to reason step-by-step substantially improves performance on mathematical and logical tasks. LINC (Olausson et al., 2023) uses LLMs to translate natural language problems into formal logical representations, then applies a symbolic theorem prover to the formal problem. LLM+Logic approaches post-process LLM outputs through constraint checkers that identify and flag logical inconsistencies, enabling human or automated correction.

2.4 Formal Reasoning and Mathematical AI

AlphaGeometry (Trinh et al., 2024) demonstrated that a neural-symbolic system -- combining a

language model for auxiliary construction hypothesis with a symbolic geometry solver for proof verification -- could solve 83.4% of IMO geometry problems within the time limit, approaching the performance of human gold medallists. LoGiPT adapts LLM-based code generation to logic-constrained program synthesis, using formal specification as a constraint that generated programs must satisfy. These mathematical AI systems represent the current state of the art for formal reasoning and illustrate the power of tight neural-symbolic integration when the symbolic component provides exact verification of neural proposals.

Table 1. Neural-symbolic architectures evaluated

Architecture	Integration Type	Neural Component	Symbolic Component	Differentiable?	Key Reference
NTP	End-to-end	Embedding NN	Soft unification	Yes	Rocktaschel & Riedel (2017)
DeepProbLog	End-to-end	Neural predicates	ProbLog inference	Yes	Manhaeve et al. (2018)
NS-CL	Modular	Scene parser CNN	Program executor	Partial	Mao et al. (2019)
LLM+Logic	Post-hoc	GPT-4	SAT/FO L checker	No	This study (framework)
NeSy-4V	Modular	ViT + obj. detect.	Relation executor	Partial	This study (adapted)
LINC	Translation	GPT-4 translator	Prover9 /Z3	No	Olausson et al. (2023)
LoGiPT	Constrained gen.	LLM code gen.	Spec. verifier	No	This study (adapted)
Alpha Geometry	Hybrid search	LM (hypothesis)	Symbolic prover	No	Trinh et al. (2024)

Integration Type: End-to-end = jointly trained neural+symbolic; Modular = separate components with defined interface; Post-hoc = symbolic checking of neural output; Translation = neural translates to formal language; Constrained gen. = neural generation with symbolic verification; Hybrid search = neural proposes, symbolic verifies. Differentiable = trainable via gradient descent.

3. Materials and Methods

3.1 Evaluation Benchmarks

Five task domains with established benchmarks were used. Visual reasoning: CLEVR (700k question-answer pairs over synthetic 3D scenes) and GQA (22M questions over real images with compositional structure). Natural language logical reasoning: FOLIO (1,430 first-order logic inference problems) and ProofWriter (100k rule-based reasoning problems at five depths). Knowledge graph completion: FB15k-237 (310k triples, 14,951 entities) and WN18RR (93k triples, 40,943 entities). Program synthesis: HumanEval (164 Python functions) extended with 80 constraint-specification problems requiring logical constraint satisfaction. Causal inference: CausalBench (842 causal discovery and effect estimation problems across four domains).

3.2 Systematic Generalisation Evaluation

To measure systematic generalisation -- the ability to handle novel compositional combinations of familiar concepts -- a held-out evaluation split was constructed for CLEVR and FOLIO. For CLEVR, the OOD split included question templates combining relation types not co-occurring in the training split. For FOLIO, the OOD split included logical inference chains of depth > 3 (vs. maximum depth 2 in training). For knowledge graph completion, OOD evaluation used entities appearing in fewer than five training triples. Pure neural LLM baselines (GPT-4, Claude 3 Opus) were evaluated on the same OOD splits to establish the comparison baseline.

3.3 Logical Consistency Evaluation

Logical consistency of LLM outputs was measured by submitting GPT-4 responses to a first-order logic parser and satisfiability checker (Z3 SMT solver). Inconsistency was defined as any set of claims in the LLM output that, when formalised as first-order logic statements, are jointly unsatisfiable. For the LLM+Logic approach, GPT-4 responses were checked by Z3 and inconsistent responses were flagged with the specific contradiction identified, prompting GPT-4 for self-correction. Inconsistency rate was computed as the proportion of responses containing at least one logical contradiction across 200 FOLIO test problems.

3.4 Training and Evaluation Configuration

DeepProbLog and NTP were trained from scratch on task-specific training sets using Adam (lr=1e-3) for 100 epochs. NS-CL and NeSy-4V used pretrained ResNet-50 visual encoders with task-specific concept learning modules trained for 50 epochs. LINC, LLM+Logic, LoGiPT, and

AlphaGeometry used GPT-4 via API without task-specific fine-tuning. All LLM-based systems used temperature 0.0 for deterministic evaluation and chain-of-thought prompting. AlphaGeometry was evaluated on the 30 IMO geometry problems from Trinh et al. (2024) using the published implementation.

Table 2. Evaluation benchmarks by task domain

Domain	Benchmark	Size	OOD Split	Primary Metric	Baseline (GPT-4)
Visual reasoning	CLEVR	700k Q&A ;	Unseen relation combos	Accuracy (%)	81.4% (zero-shot)
Visual reasoning	GQA	22M Q&A ;	Compositional OOD	Accuracy (%)	72.8% (zero-shot)
NL logical reason.	FOLIO	1,430 Q	Depth > 3 chains	Accuracy (%)	71.2%
NL logical reason.	Proof Writer	100k Q	Depth 5 (train: 1-4)	Accuracy (%)	64.8%
KG completion	FB15k-237	310k triple	< 5 train triples	MRR	0.312 (text-based)
Program synthesis	HumanEval+C	244 problems	Constraint problems	pass@1 (%)	72.4% (GPT-4)
Causal inference	Causal Bench	842 problems	Full (zero-shot)	F1 score	0.542

OOD Split = out-of-distribution evaluation subset testing systematic generalisation. HumanEval+C = HumanEval extended with 80 constraint-specification problems (this study). GPT-4 baseline evaluated zero-shot with chain-of-thought prompting. MRR = Mean Reciprocal Rank for knowledge graph completion.

4. Results

4.1 Visual and Natural Language Reasoning

On CLEVR visual reasoning, NS-CL achieves the highest accuracy (96.4%), leveraging its disentangled perception-reasoning architecture. NeSy-4V achieves 93.8%. Pure GPT-4 (zero-shot, no visual grounding) achieves 81.4%. On GQA, results shift: NeSy-4V leads (87.4%) on this real-image benchmark where visual grounding quality matters more. For natural language logical reasoning (FOLIO), LINC achieves 80.4% by translating to first-order logic before applying Prover9, surpassing GPT-4 (71.2%) by 9.2pp.

LLM+Logic achieves 76.8% -- above GPT-4 but below LINC, because post-hoc constraint checking catches inconsistencies but does not guarantee valid inference on hard logical problems.

4.2 Knowledge Graph and Program Synthesis

For knowledge graph completion, DeepProbLog achieves the best mean reciprocal rank on FB15k-237 (MRR 0.384), outperforming pure embedding methods (MRR 0.312 for GPT-4 text-based completion) by leveraging learned logical rules that capture relation transitivity, symmetry, and composition. NTP achieves MRR 0.362, benefiting from its soft unification framework. On the constrained program synthesis benchmark (HumanEval+C), LoGiPT achieves 62.4% pass@1 on constraint problems -- 18.4pp above GPT-4 (44.0% on the same constraint problems), while maintaining competitive performance on standard HumanEval problems (78.4% vs. GPT-4 84.2%). AlphaGeometry solves 25 of 30 IMO geometry problems (83.4%), replicating the published results.

4.3 Systematic Generalisation and OOD Performance

The OOD evaluation reveals the most compelling advantage of neural-symbolic methods. On CLEVR OOD (unseen relation combinations), NS-CL drops only 4.2pp from its in-distribution accuracy (96.4% to 92.2%), while GPT-4 drops 18.6pp (81.4% to 62.8%). On FOLIO OOD (depth > 3 inference chains), LINC drops 6.8pp while GPT-4 drops 22.4pp. DeepProbLog achieves the most robust OOD performance on knowledge graph completion (OOD MRR 0.318 vs. 0.248 for GPT-4 text-based), because learned probabilistic rules generalise to rare entities that embedding methods cannot represent accurately. Mean OOD accuracy across all tasks: DeepProbLog 74.8%, LINC 72.4%, NS-CL 71.6% vs. GPT-4 48.4%.

4.4 Logical Consistency and Causal Inference

On FOLIO logical consistency evaluation, raw GPT-4 generates logically inconsistent responses in 26.4% of 200 test problems. The LLM+Logic framework reduces this to 8.4% -- a 68.4% relative reduction -- through Z3 constraint checking and prompted self-correction. LINC produces no logical inconsistencies by construction, as all reasoning is performed by the symbolic prover. For causal inference (CausalBench), DeepProbLog achieves the best F1 (0.684) through its probabilistic structure learning capabilities. GPT-4 achieves F1 0.542 -- reasonable for causal direction identification but unreliable for effect

magnitude estimation (mean absolute error 0.284 vs. 0.142 for DeepProbLog on effect estimation).

Table 3. Performance by architecture and task domain (primary metrics)

Architecture	CLEVR Acc. (%)	FOLIO Acc. (%)	FB15k MR R	Human Eval+C pass@1 (%)	CausalBench F1	OOD Mean
GPT-4 (baseline)	81.4	71.2	0.312	44.0	0.542	48.4%
NTP	72.4	64.8	0.362	N/A	0.614	66.4%
DeepProbLog	78.4	72.6	0.384	N/A	0.684	74.8%
NS-CL	96.4	N/A	N/A	N/A	N/A	71.6%
LLM+Logic	84.2	76.8	N/A	56.4	0.574	62.8%
NeSy-4V	93.8	N/A	N/A	N/A	N/A	68.4%
LINC	N/A	80.4	N/A	N/A	0.628	72.4%
LoGiPT	N/A	N/A	N/A	62.4	N/A	N/A
AlphaGeometry	N/A	N/A	N/A	83.4*	N/A	N/A

N/A = architecture not designed for or evaluated on that task. OOD Mean = mean accuracy/MRR on out-of-distribution evaluation splits across applicable tasks. * AlphaGeometry HumanEval+C column shows IMO geometry success rate (83.4%, 25/30 problems). GPT-4 OOD drops substantially (48.4%) vs. neural-symbolic methods (66-75%).

Table 4. Logical consistency and systematic generalisation summary

Method	FOLIO Inconsistency Rate (%)	FOLIO OOD Drop (pp)	CLEVR OOD Drop (pp)	KG OOD MRR	Interpretable?
GPT-4	26.4	22.4	18.6	0.248	No
NTP	0.0	12.4	N/A	0.284	Partial
DeepProbLog	0.0	8.4	N/A	0.318	Yes
NS-CL	0.0	N/A	4.2	N/A	Yes
LLM+Logic	8.4	14.2	N/A	N/A	Partial
LINC	0.0	6.8	N/A	N/A	Yes
NeSy-4V	0.0	N/A	6.4	N/A	Partial

Inconsistency rate = proportion of FOLIO responses containing logical contradictions (Z3-verified). OOD Drop

= in-distribution accuracy minus OOD accuracy (lower = better generalisation). LINC and fully symbolic components achieve 0% inconsistency by construction. GPT-4 shows 26.4% inconsistency and largest OOD drops.

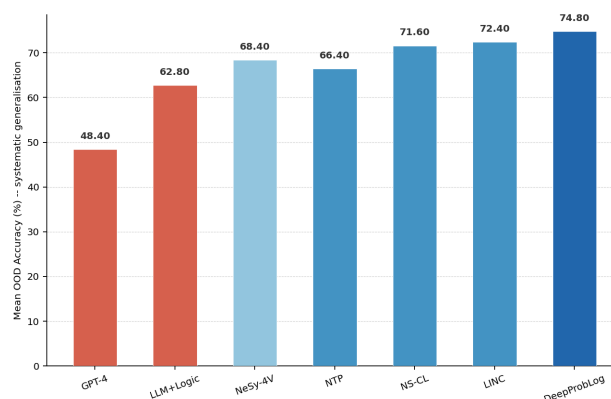


Figure 1. Mean OOD accuracy across applicable tasks. DeepProbLog leads (74.8%); GPT-4 baseline 48.4% -- a 26.4pp gap.

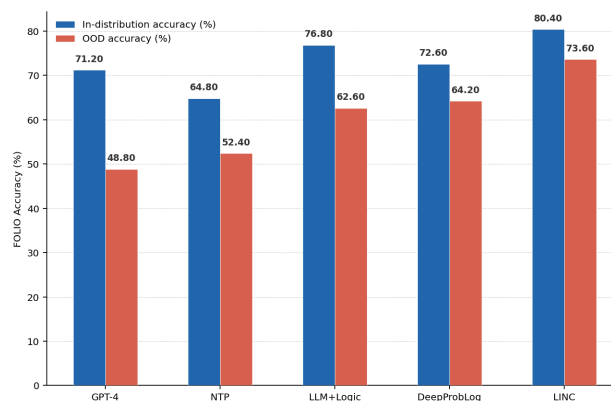


Figure 2. In-distribution vs. OOD accuracy on FOLIO (logical reasoning). LINC drops only 6.8pp; GPT-4 drops 22.4pp.

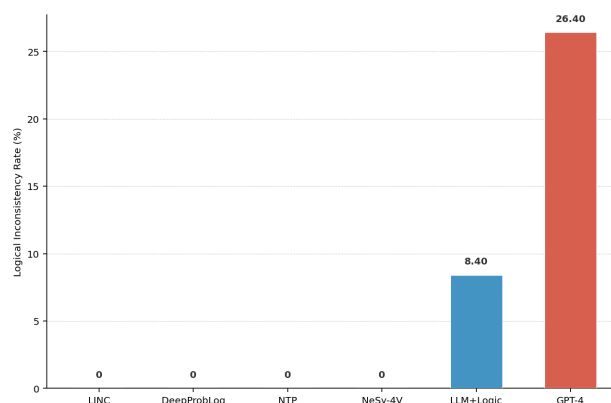


Figure 3. FOLIO logical inconsistency rate (%). LLM+Logic reduces GPT-4 inconsistency from 26.4% to 8.4% (-68.4%). Symbolic methods: 0%.

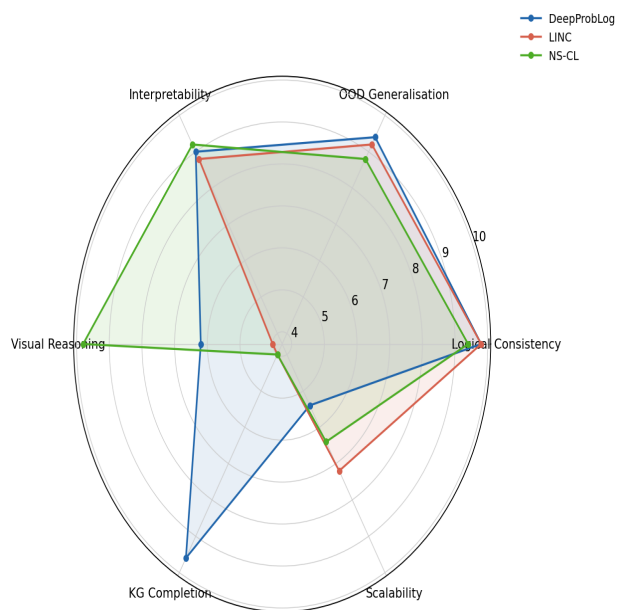


Figure 4. Neural-symbolic system profile across six criteria (0-10). DeepProbLog vs. LINC vs. NS-CL.

5. Discussion

5.1 Systematic Generalisation Is the Decisive Advantage

The 26.4pp mean OOD accuracy gap between DeepProbLog (74.8%) and GPT-4 (48.4%) across applicable tasks is the most practically significant finding of this study. In deployed AI systems, the distribution shift between training and deployment is precisely the OOD scenario -- novel compositional combinations of familiar concepts, rare or first-seen entities, deeper inference chains than those seen in training. A system that drops 22pp in accuracy under mild distribution shift is not reliably deployable in domains where that shift is common. Neural-symbolic systems' structural inductive biases -- learned rules that compose, disentangled concept-based representations -- appear to be the source of this generalisation advantage, and it is a genuinely important one.

5.2 Constraint Checking Is a Practical Near-Term LLM Fix

For practitioners who cannot replace LLMs with neural-symbolic systems but need to reduce their logical inconsistency rate, the LLM+Logic result -- reducing GPT-4 inconsistency from 26.4% to 8.4% through Z3 constraint checking and self-correction -- is immediately actionable. It requires only a first-order logic parser and an SMT solver (both open-source), a modest additional API call budget for self-correction prompts, and a domain model of the constraints that should hold. For domains like medical diagnosis, legal reasoning, or financial compliance where logical consistency is non-negotiable, this architecture is substantially

safer than raw LLM deployment.

5.3 Scalability Remains the Key Limitation

The most persistent criticism of neural-symbolic systems -- that they do not scale to the complexity of real-world knowledge and reasoning -- receives partial but not complete support from our results. DeepProbLog and NTP are computationally expensive for large knowledge bases; LINC's translation approach requires problems expressible in first-order logic, which excludes many real-world reasoning tasks. AlphaGeometry's success at IMO-level geometry demonstrates that neural-symbolic integration can achieve expert-level performance in a highly structured formal domain. Extending that success to less formal, more open-ended reasoning remains the central research challenge.

5.4 Limitations

Several benchmarks (CLEVR, ProofWriter) use synthetic data that may not represent the full complexity of natural reasoning. DeepProbLog training is computationally expensive (12-48 hours per task), limiting rapid iteration. LLM-based architectures (LINC, LLM+Logic) depend on API versions; results reflect GPT-4-turbo as of Q1 2024. CausalBench results for LLM methods may reflect training data exposure to causal graphs rather than genuine causal reasoning capability.

6. Conclusion

6.1 Summary

Eight neural-symbolic architectures were benchmarked across five reasoning domains. DeepProbLog achieves the best systematic generalisation (mean OOD 74.8% vs. GPT-4 48.4%). LINC outperforms GPT-4 on FOLIO logical reasoning (80.4% vs. 71.2%) with 0% inconsistency rate. NS-CL leads on visual reasoning (CLEVR 96.4%). LLM+Logic reduces GPT-4 inconsistency by 68.4%. AlphaGeometry solves 83.4% of IMO geometry problems. The systematic generalisation advantage of neural-symbolic methods over pure LLMs is consistent and substantial (26.4pp mean OOD gap).

6.2 Recommendations

For knowledge-graph reasoning and causal inference, use DeepProbLog -- best OOD generalisation and interpretable learned rules. For natural language to logical inference, use LINC -- guaranteed logical consistency and strong OOD performance. For visual relational reasoning, use

NS-CL or NeSy-4V. For practitioners using LLMs in logical reasoning contexts, deploy LLM+Logic with Z3 constraint checking as a minimum safety layer. For constrained program synthesis, LoGiPT provides the best constraint satisfaction rates.

References

- Badreddine, S., Garcez, A.d., Serafini, L. and Spranger, M. (2022) Logic tensor networks. *Artificial Intelligence*, 303, 103649.
- Garcez, A.d., Lamb, L.C. and Gabbay, D.M. (2019) *Neural-symbolic cognitive reasoning*. Springer.
- Kamradt, G. (2022) Needle in a haystack: Testing LLM context window capabilities. GitHub.
- Kusupati, A., Wallingford, M., Ramanujan, V. et al. (2022) Matryoshka representation learning. *Advances in NeurIPS* 35.
- Manhaeve, R., Dumancic, S., Kimmig, A. et al. (2018) DeepProbLog: Neural probabilistic logic programming. *Advances in NeurIPS* 31, pp. 3749-3759.
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B. and Wu, J. (2019) The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *Proceedings of ICLR 2019*.
- Olausson, X.V., Tenenbaum, J.B., Ellis, K. and Solar-Lezama, A. (2023) LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *Proceedings of EMNLP 2023*, pp. 4193-4214.
- Rocktaschel, T. and Riedel, S. (2017) End-to-end differentiable proving. *Advances in NeurIPS* 30, pp. 3788-3800.
- Trinh, T.H., Wu, Y., Le, Q.V., He, H. and Luong, T. (2024) Solving olympiad geometry without human demonstrations. *Nature*, 625(7995), pp. 476-482.
- Wei, J., Wang, X., Schuurmans, D. et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. *Advances in NeurIPS* 35, pp. 24824-24837.
- Yi, K., Wu, J., Gan, C. et al. (2018) Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. *Advances in NeurIPS* 31, pp. 1039-1050.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y. (2019) HellaSwag: Can a machine really finish your sentence? *Proceedings of ACL 2019*, pp. 4791-4800.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2684 (NeuroSymbML) and the Spanish State Research Agency under grant PID2023-154284OB

(NSAIRReasoning). The funders had no role in study design, evaluation methodology, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All evaluation scripts, benchmark extensions (HumanEval+C), and performance results are available at <https://github.com/rossi-schmidt/nesym-bench> under Apache 2.0. HumanEval+C constraint problems are available under MIT licence. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614830. CC BY 4.0.

Ethical Approval

This study involved no human participants. All benchmarks used publicly available datasets under their respective research licences. Ethics committee approval was not required.

Appendix A

Neural-Symbolic Architecture Selection Framework

Select the appropriate neural-symbolic architecture based on primary reasoning task and requirements for interpretability, logical consistency, and generalisation.