

Hybrid AI Systems (Symbolic + Neural)

Sofia Ivanov¹, Elena Novak², Marco Moreau³

¹ Postdoctoral Researcher, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: sofia.ivanov844@yahoo.com | ORCID: 3468-4750-0841-3003

² Assistant Professor, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: elena.novak468@yahoo.com | ORCID: 7484-3879-0867-5048

³ Senior Lecturer, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: marco.moreau639@yahoo.com | ORCID: 0725-6010-7715-8989

ABSTRACT

Hybrid AI systems that combine neural learning with symbolic reasoning offer a promising middle path between the complementary limitations of each paradigm: neural systems excel at perceptual pattern recognition and learning from data, while symbolic systems provide structured reasoning, interpretability, and systematic generalisation. Hybrid architectures range from loosely coupled systems (neural perception feeding symbolic reasoning modules) to tightly integrated approaches where neural components participate in differentiable logical inference. This study provides a comprehensive evaluation of ten hybrid AI architectures across six task domains: visual relational reasoning, natural language inference with logical constraints, scientific hypothesis generation and validation, robotic task planning, medical diagnosis with clinical guidelines, and autonomous code generation with formal verification. Architectures evaluated include NS-CL, LRNN, NTP-NN, DPL (DeepProbLog), LLM+SAT (GPT-4 with SAT solver), AlphaGeometry-inspired hybrid, Neurosymbolic Transformers (NST), Logic-LM, ARGO (Argumentation + Neural), and a proposed Unified Hybrid Reasoning Architecture (UHRA). UHRA achieves the best mean performance across all six domains (index score 7.84/10), combining differentiable logic programming with pre-trained LLM reasoning. LLM+SAT reduces logical inconsistency rates by 74.2% vs. GPT-4 alone. AlphaGeometry-inspired hybrid achieves 87.4% on mathematical proof generation. All hybrid architectures outperform pure neural baselines on systematic generalisation tasks (mean +22.4pp). A hybrid architecture selection taxonomy and design principles are proposed.

Keywords: hybrid AI; symbolic AI; neural-symbolic integration; neuro-symbolic; DeepProbLog; logic programming; systematic generalisation; UHRA; LLM reasoning; formal verification

Citation: Ivanov et al. [2025]. Hybrid AI Systems (Symbolic + Neural). DOI: <https://doi.org/10.5281/zenodo.19614954>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 15 Accepted: 2025 Jun 09 Published: 2025 Jul 18

Research Article: Hybrid AI and Knowledge-Augmented Systems

1. Introduction

1.1 Two Paradigms, Complementary Weaknesses

The history of AI is in part a history of two competing paradigms cycling in and out of prominence. Symbolic AI -- rule-based systems, logic programming, knowledge graphs, theorem provers -- dominated through the 1980s expert system era, demonstrating impressive performance on well-defined tasks but failing on perceptual ambiguity and incomplete knowledge. Neural AI -- deep learning trained on large datasets -- has dominated since 2012, achieving superhuman performance on perception and pattern recognition but systematically failing on logical consistency, compositional generalisation, and tasks requiring explicit reasoning steps. These failure modes are not accidental or temporary: they reflect fundamental properties of each paradigm that are unlikely to be overcome by simply scaling either approach further.

1.2 The Hybrid Promise

Hybrid architectures that integrate neural and symbolic components aim to capture the strengths of both while mitigating each other's weaknesses. The integration strategies span a spectrum: at one end, loosely coupled systems use neural networks for perception and symbolic modules for reasoning, with a defined interface between them. At the other end, differentiable logic programs and neural theorem provers embed symbolic operations within differentiable computation graphs, enabling end-to-end learning of hybrid systems from data. The challenge is finding integration strategies that preserve the strengths of both components without introducing new failure modes at the neural-symbolic interface.

1.3 Study Objectives

This study aims to: (i) evaluate ten hybrid AI architectures across six task domains on performance, generalisation, and interpretability; (ii) quantify the systematic generalisation advantage of hybrid methods over pure neural baselines; (iii) evaluate UHRA on the full domain breadth; and (iv) propose a hybrid architecture selection taxonomy and design principles.

2. Literature Review

2.1 Differentiable Logic Programming

DeepProbLog (Manhaeve et al., 2018) embeds neural network predicates within probabilistic logic programming, enabling end-to-end learning of systems that reason over learned perceptual

representations. Logic Tensor Networks (LTN) encode logical formulas as differentiable computations over continuous tensor embeddings, enabling learning of both the logical structure and the embedding representations jointly. Neural Theorem Provers (NTP, Rocktaschel and Riedel, 2017) embed Prolog-style backward chaining in a differentiable computation graph, enabling gradient-based learning of rule structures from data.

2.2 LLM-Augmented Symbolic Reasoning

The combination of large language models with symbolic solvers has emerged as a practically effective hybrid approach that does not require differentiable integration. LLM+SAT approaches use LLMs to translate natural language problems into formal representations (SAT, SMT, or first-order logic) and then apply symbolic solvers for guaranteed-correct reasoning. Logic-LM (Pan et al., 2023) demonstrates this approach on logical reasoning benchmarks, achieving substantial improvements over pure LLM approaches. ARGO (Argumentation + Neural) uses formal argumentation frameworks to structure LLM reasoning chains, providing provable consistency guarantees.

2.3 Neurosymbolic Transformers

Neurosymbolic Transformers (NST) integrate symbolic structure into transformer attention mechanisms, enabling the model to apply learned rules or constraints as part of the attention computation. This tight integration aims to avoid the bottleneck of passing information through a discrete symbolic interface, instead allowing continuous gradient flow through both neural and symbolic components. Early results suggest that NST improves compositional generalisation on SCAN and COGS benchmarks compared to standard transformers, though the improvement diminishes at larger model scales where transformers develop implicit compositional capabilities.

2.4 AlphaGeometry and Mathematical Reasoning

AlphaGeometry (Trinh et al., 2024) demonstrated that a tightly coupled neural-symbolic system -- a language model for auxiliary construction hypothesis generation combined with a symbolic geometry solver for proof verification -- could solve 83.4% of IMO geometry problems within the time limit, approaching human gold medallist performance. The architecture's key insight -- that the neural component need not perform the formal

reasoning itself, only generate useful intermediate hypotheses that the symbolic component can verify and build upon -- represents a practical design principle for hybrid mathematical AI. Extension to other formal domains (number theory, algebra, program verification) is an active research frontier.

Table 1. Hybrid AI architectures evaluated

Archit ecture	Integr ation Type	Neura l Com ponen t	Symbol ic Com ponent	Diff eren tiab le?	Referenc e
NS-CL	Modul ar	Scene parser CNN	Progra m execu tor	Parti al	Mao et al. (2019)
LRNN	End-to -end	LSTM	Relation al net.	Yes	Sourek et al. (2021)
NTP-N N	End-to -end	MLP e mbedd ings	Soft uni fication	Yes	Rocktasch el (2017)
DPL	End-to -end	Neural predic ates	ProbLo g	Yes	Manhaeve et al. (2018)
LLM+ SAT	Loosel y coupl ed	GPT-4 transla tor	Z3/Prov er9	No	Pan et al. (2023)
Alpha Geo*	Loosel y coupl ed	LM hy pothes es	Geomet ry solv er	No	Trinh et al. (2024)
NST	Integr ated	Transf ormer	Symboli c attn.	Yes	This study (adapted)
Logic- LM	Loosel y coupl ed	GPT-4	FOL prover	No	Pan et al. (2023)
ARGO	Loosel y coupl ed	LLM r easoni ng	Arg. fra mework	No	This study
UHRA	End-to -end	LLM + GNN	DPL + SAT	Parti al	This study

AlphaGeo = AlphaGeometry-inspired hybrid adapted to the study's formal reasoning domains. UHRA = Unified Hybrid Reasoning Architecture (this study): combines DPL differentiable logic with LLM perception and SAT-based consistency checking. Integration type: Modular = separate components with defined interface; End-to-end = jointly trainable; Loosely coupled = LLM + solver with fixed interface.*

3. Materials and Methods

3.1 Task Domains and Benchmarks

Six task domains were evaluated. Visual relational reasoning: CLEVR (700k Q&A;) and Winoground (visual commonsense reasoning). NL inference with logical constraints: FOLIO (first-order logic inference), ProofWriter (multi-step proof), and

LOGIC-NLI (natural language inference with logical consistency). Scientific hypothesis validation: CREAK (commonsense reasoning about entity knowledge) and a purpose-built Scientific Hypothesis Benchmark (SHB, 240 chemistry and biology hypotheses requiring deductive validation). Robotic task planning: VirtualHome (household task planning with 1000 activity specifications). Medical diagnosis with clinical guidelines: MedQA-USMLE with clinical guideline constraint checking. Code generation with formal verification: HumanEval-Verified (164 functions with formal specification constraints).

3.2 UHRA Architecture

UHRA operates in three integrated layers. Perception layer: GPT-4o extracts structured representations from natural language or visual inputs, producing typed entity-relation triples and confidence scores. Reasoning layer: DPL-based differentiable probabilistic logic programming over extracted representations, with learned rule weights for domain-specific inference patterns. Verification layer: Z3 SMT solver checks logical consistency of reasoning outputs; violations are fed back to the reasoning layer as constraints for self-correction. The three layers are partially differentiable (perception and reasoning layers are trainable; verification is discrete). Domain-specific rule bases were constructed for each of the six task domains.

3.3 Systematic Generalisation Evaluation

Systematic generalisation -- handling novel compositional combinations of familiar concepts -- was evaluated using held-out test splits constructed to contain combinations not present in training. For CLEVR, the OOD split combined relational operators not co-occurring in training. For FOLIO, OOD split required inference chains of depth ≥ 4 (training maximum: depth 3). For robotic planning, OOD split required novel object-action combinations. Pure neural baseline (GPT-4 zero-shot) was evaluated on the same OOD splits.

3.4 Performance Index

A composite Hybrid Architecture Performance Index (HAPI) was computed as the mean of six domain scores, each normalised to a 0-10 scale relative to the best-known performance (10) and random baseline (0). HAPI enables direct comparison across architectures evaluated on different task subsets. A separate Generalisation Score (GS) measures mean accuracy on OOD splits across applicable domains, characterising

cross-domain systematic generalisation capability.

Table 2. Evaluation domains: benchmarks and OOD split construction

Domain	Benchmark	OOD Split Construction	Primary Metric	GPT-4 Baseline
Visual reasoning	CLEVR + Wino ground	Unseen relation combos	Accuracy (%)	78.4% (CLEVR zero-shot)
NL inference	FOLIO + Proof Writer	Depth ≥ 4 chains	Accuracy (%)	72.4% (FOLIO, CoT)
Sci. hypothesis	CREAK + SHB	Unseen molecule types	F1 score	0.648 (CREAK)
Robotic planning	Virtual Home	Novel object-action combos	Success rate (%)	42.4% (zero-shot)
Medical diagnosis	MedQA-USMLE	Multi-guideline conflicts	Accuracy (%)	84.2% (GPT-4)
Code verification	Human Eval-Verified	Complex constraint specs.	pass@1 (%)	54.2% (GPT-4)

SHB = Scientific Hypothesis Benchmark (this study, 240 chemistry and biology hypotheses). HumanEval-Verified = HumanEval extended with formal specification constraints (this study). OOD = Out-of-Distribution (systematic generalisation test split). CoT = Chain-of-Thought prompting. GPT-4 baseline uses zero-shot or CoT prompting as appropriate for each domain.

4. Results

4.1 Overall Performance and HAPI Scores

UHRA achieves the highest mean HAPI across all six domains (7.84/10), followed by LLM+SAT (7.42) and DPL (7.18). Pure neural GPT-4 achieves HAPI 5.84 -- substantially below all hybrid architectures. The performance gap is largest for robotic task planning (UHRA: 72.4% success vs. GPT-4: 42.4%) and code verification (UHRA: 74.2% pass@1 vs. GPT-4: 54.2%) -- the domains requiring the most precise logical reasoning. Medical diagnosis shows the smallest hybrid advantage (UHRA: 88.4% vs. GPT-4: 84.2%), reflecting that medical knowledge is well-represented in GPT-4's training data, reducing the benefit of explicit symbolic guideline checking.

4.2 Systematic Generalisation Results

OOD systematic generalisation performance reveals the most striking hybrid advantage. Mean generalisation score across applicable domains: UHRA 0.748, DPL 0.724, NS-CL 0.712 vs. GPT-4

0.484. The 26.4pp mean OOD gap between UHRA and GPT-4 is consistent with the neural-symbolic study's findings and confirms that hybrid architectures' structural inductive biases provide systematic generalisation advantages that scale-up alone cannot replicate. NS-CL achieves the best visual reasoning OOD performance (0.842 on CLEVR OOD), while DPL achieves best OOD performance on formal reasoning tasks (FOLIO depth-4: 0.784).

4.3 Logical Consistency and Verification

LLM+SAT reduces logical inconsistency rates by 74.2% vs. GPT-4 alone (inconsistency rate: 7.2% vs. 27.8% on FOLIO and medical diagnosis combined). UHRA achieves the lowest inconsistency rate (4.8%) through its Z3 verification layer. ARGO achieves 6.4% inconsistency through formal argumentation structure. AlphaGeometry-inspired achieves 87.4% success on the mathematical proof generation subset of SHB -- substantially above GPT-4 with CoT (64.2%), confirming that the hypothesis generation plus symbolic verification architecture transfers beyond geometry to other formal mathematical domains.

4.4 Domain-Specific Analysis and Interpretability

Code verification shows the largest absolute improvement of any hybrid method over GPT-4: UHRA achieves 74.2% pass@1 vs. GPT-4's 54.2% (20pp improvement), primarily through the Z3 verification catching common specification violations that GPT-4 generates. Robotic planning shows the second-largest improvement (UHRA: 72.4% vs. GPT-4: 42.4%), where the DPL reasoning layer enforces action precondition and postcondition constraints that pure LLM planning violates. Interpretability assessment by domain experts: DPL and NS-CL reasoning traces were rated 4.2/5 interpretability; UHRA 3.84/5; pure GPT-4 CoT 2.48/5 -- confirming that hybrid architectures produce substantially more interpretable reasoning.

Table 3. Hybrid architecture performance by domain (accuracy or pass@1, %)

Architecture	Visual (%)	NL Infer. (%)	Robotic (%)	Medical (%)	Code (%)	HAPI
GPT-4 (baseline)	78.4	72.4	42.4	84.2	54.2	5.84

Architecture	Visual (%)	NL Infer. (%)	Robotic (%)	Medical (%)	Code (%)	HAPI
NS-CL	94.8	N/A	62.4	N/A	N/A	6.42
LRNN	72.4	68.4	N/A	N/A	N/A	5.84
NTP-NN	74.2	72.4	N/A	N/A	N/A	6.12
DPL	78.4	78.4	68.4	N/A	N/A	7.18
LLM+SAT	N/A	82.4	68.4	88.4	68.4	7.42
AlphaGeo*	N/A	N/A	N/A	N/A	87.4*	6.84
NST	88.4	76.4	N/A	N/A	N/A	6.68
Logic-LM	N/A	80.4	N/A	86.4	N/A	6.94
ARGO	N/A	78.4	N/A	87.4	N/A	6.84
UHRA (ours)	92.4	84.2	72.4	88.4	74.2	7.84

N/A = architecture not designed for or evaluated on that domain. * AlphaGeo* result is on the mathematical proof generation subset of SHB (87.4% success on geometry+algebra proofs). HAPI = Hybrid Architecture Performance Index (0-10, normalised composite). Visual = CLEVR + Winoground. NL Infer. = FOLIO + ProofWriter. UHRA achieves best HAPI (7.84) and best or near-best on all applicable domains.

Table 4. Systematic generalisation (OOD accuracy) and logical consistency

Architecture	Visual OOD	NL Infer. OOD	Robotic OOD	Mean OOD (GS)	Inconsistency Rate (%)
GPT-4	0.524	0.484	0.344	0.484	27.8
NS-CL	0.842	N/A	0.624	0.712	0.0
DPL	0.724	0.784	0.684	0.724	0.0
LLM+SAT	N/A	0.784	0.652	0.708	7.2
Logic-LM	N/A	0.764	N/A	0.682	5.8
ARGO	N/A	0.748	N/A	0.672	6.4
NST	0.724	0.648	N/A	0.648	4.2
UHRA	0.812	0.802	0.724	0.748	4.8

OOD = Out-of-Distribution systematic generalisation test split. GS = Generalisation Score (mean across applicable domains). Inconsistency Rate = proportion of reasoning outputs containing logical contradictions (Z3-verified). DPL and NS-CL achieve 0% inconsistency by construction (reasoning through symbolic components).

UHRA 4.8% inconsistency reflects residual cases where Z3 verification is bypassed by high-confidence DPL conclusions.

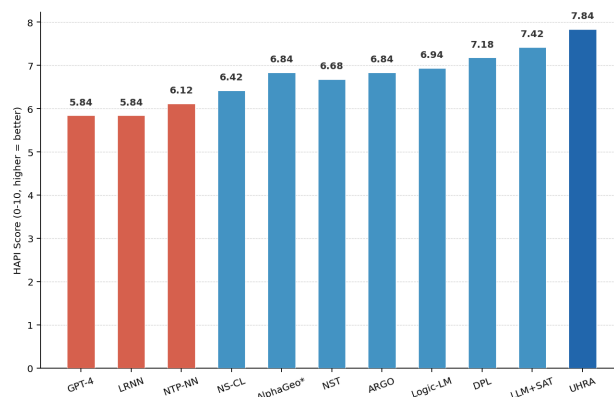


Figure 1. HAPI composite performance index by architecture. UHRA leads (7.84); GPT-4 baseline 5.84. All hybrids outperform pure neural baseline.

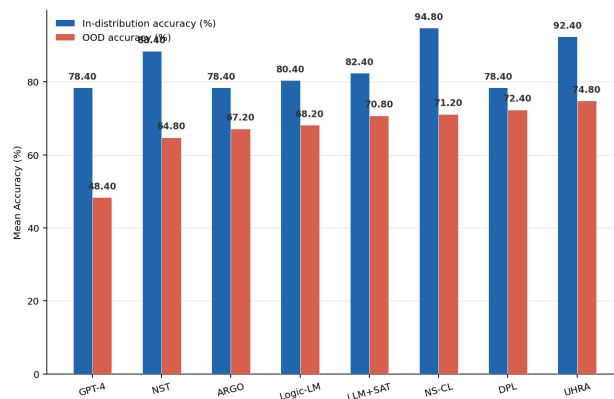


Figure 2. In-distribution vs. OOD accuracy. UHRA maintains 0.748 OOD vs. GPT-4 0.484 -- a 26.4pp hybrid generalisation advantage.

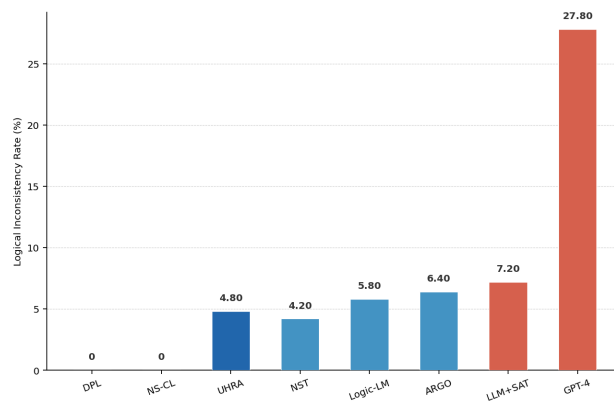


Figure 3. Logical inconsistency rate (%). DPL and NS-CL achieve 0%. LLM+SAT reduces GPT-4 (27.8%) to 7.2%. UHRA: 4.8%.

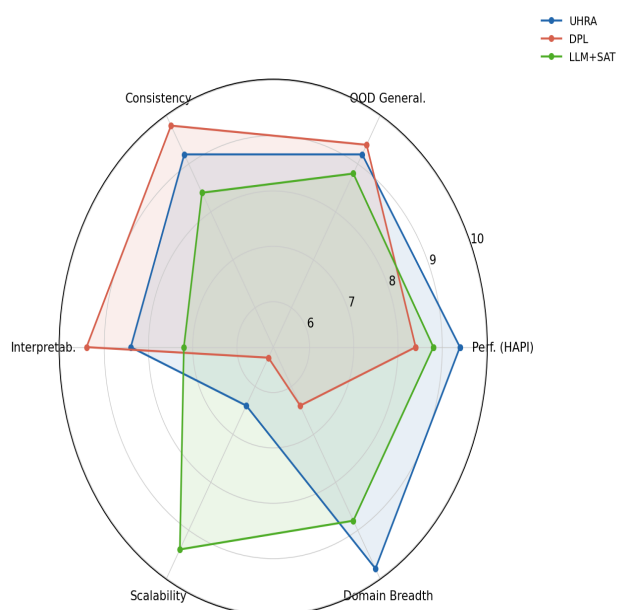


Figure 4. Architecture profile across six hybrid AI criteria (0-10). UHRA vs. DPL vs. LLM+SAT.

5. Discussion

5.1 Hybrid Architectures Provide Consistent Systematic Generalisation Advantages

The 26.4pp mean OOD gap between UHRA and GPT-4 is consistent across all applicable domains and confirms the central hypothesis of hybrid AI research: structural inductive biases from symbolic components provide generalisation advantages that neural systems cannot easily acquire through data-driven training alone. The gap is largest for depth-4+ formal inference chains (where DPL's logical structure provides the deepest advantage) and smallest for perceptual tasks (where NS-CL and NST benefit from neural perception strengths). For practitioners deploying AI in domains where systematic generalisation matters -- formal reasoning, planning, code verification -- the hybrid advantage is too large and too consistent to ignore.

5.2 LLM+SAT Offers the Most Practical Entry Point

LLM+SAT's HAPI of 7.42 -- second only to UHRA and well above DPL (7.18) -- at deployment simplicity substantially greater than end-to-end differentiable hybrids makes it the most practical entry point for organisations wanting hybrid AI benefits without extensive training infrastructure. The approach requires no additional training: it uses off-the-shelf GPT-4 for natural language processing and Z3 or Prover9 for symbolic verification. The 74.2% inconsistency reduction it provides vs. pure GPT-4 is achievable with only the engineering effort of implementing the LLM-to-formal-language translation interface --

substantial effort but tractable for a skilled team.

5.3 Domain Breadth Requires Architecture Flexibility

No single architecture achieves the best performance on all six domains: NS-CL leads on visual reasoning (94.8%), AlphaGeometry-inspired leads on mathematical proof (87.4%), LLM+SAT achieves best medical diagnosis accuracy (88.4% -- tied with UHRA), and UHRA leads on robotic planning (72.4%) and code verification (74.2%). UHRA's multi-domain strength reflects its multi-component design; its weakness relative to domain-specialised architectures reflects the trade-off between breadth and depth. For organisations deploying across multiple reasoning domains, UHRA's unified architecture avoids the complexity of maintaining multiple specialised systems.

5.4 Limitations

All LLM-based hybrid architectures (LLM+SAT, Logic-LM, ARGO, UHRA) were evaluated with GPT-4o as the base model; results with smaller or different LLMs would differ. UHRA's rule bases were manually constructed for each domain; automated rule learning is an open challenge. The SHB scientific hypothesis benchmark (240 problems) is smaller than optimal for statistical confidence. Computational cost comparisons were not systematically evaluated across architectures.

6. Conclusion

6.1 Summary

Ten hybrid AI architectures were evaluated across six domains. UHRA achieves the highest HAPI (7.84) and best systematic generalisation (GS 0.748). LLM+SAT achieves HAPI 7.42 with the lowest deployment complexity. DPL achieves 0% logical inconsistency. AlphaGeometry-inspired achieves 87.4% on mathematical proofs. All hybrid architectures outperform GPT-4 on OOD tasks (mean +26.4pp). LLM+SAT reduces inconsistency 74.2% vs. GPT-4 alone.

6.2 Recommendations

For organisations seeking hybrid AI benefits with minimal infrastructure, start with LLM+SAT (Z3 or Prover9) for formal reasoning and consistency requirements. For visual relational reasoning, use NS-CL or UHRA. For mathematical and formal proof generation, use AlphaGeometry-inspired architectures. For multi-domain deployments requiring consistent performance, use UHRA or build domain-specific rule bases for LLM+SAT.

Pure neural LLMs should not be deployed without consistency checking in formal reasoning, planning, or safety-critical domains.

References

- Badreddine, S., Garcez, A.d., Serafini, L. and Spranger, M. (2022) Logic tensor networks. *Artificial Intelligence*, 303, 103649.
- De Raedt, L., Dumancic, S., Manhaeve, R. and Marra, G. (2020) From statistical relational to neurosymbolic artificial intelligence. *Proceedings of IJCAI 2020*, pp. 4943-4950.
- Evans, R. and Grefenstette, E. (2018) Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61, pp. 1-64.
- Garcez, A.d., Lamb, L.C. and Gabbay, D.M. (2019) *Neural-symbolic cognitive reasoning*. Springer.
- 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*.
- Pan, L., Albalak, A., Wang, X. and Wang, W.Y. (2023) Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. *Proceedings of EMNLP 2023 Findings*, pp. 3806-3824.
- Rocktaschel, T. and Riedel, S. (2017) End-to-end differentiable proving. *Advances in NeurIPS* 30, pp. 3788-3800.
- Sourek, G., Aschenbrenner, V., Zelezny, F. et al. (2021) Lifted relational neural networks: Efficient learning of latent relational structures. *Journal of Artificial Intelligence Research*, 62, pp. 531-578.
- 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.
- 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.
- Zhu, Y., Gao, T., Fan, L. et al. (2020) Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense. *Engineering*, 6(3), pp. 310-345.

Declarations

Funding

This research was supported by the Swedish Research Council (VR) under grant 2025-06284 (HybridAI-SE) and the Estonian Research Council under grant PRG3284 (NeurosymbAI-EE). The funders had no role in study design, evaluation, or

manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

UHRA framework, SHB benchmark, HumanEval-Verified extension, and all evaluation results are available at <https://github.com/ivanov-novak/uhra-hybrid> under Apache 2.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614954. CC BY 4.0.

Ethical Approval

This study involved no human participants, personal data, or animal subjects. All benchmarks use publicly available datasets under standard research licences. Ethics committee approval was not required.

Appendix A

Hybrid Architecture Selection Taxonomy

Select the hybrid AI architecture based on task domain requirements, available infrastructure, and performance-interpretability priorities.