ABUAD Journal of Engineering and Applied Sciences (AJEAS)

ISSN: 2992-5584

Volume 3, Issue 1, 36-47

Neuro-Symbolic Reasoning: Performance, Challenges, and Benchmarks: A Systematic Literature Review

Peter Godfrey OBIKE[1], Patience Usoro USIP[2], Edward Ndarake UDO[3], Aniekan Joe ANANGA[4]

[1]*Department of Computer Science, Michael Okpara University of Agriculture, Umudike, Nigeria*
[2,3,4]*Department of Computer Science, University of Uyo, Uyo, Nigeria*

obike.peter@mouau.edu.ng, patienceusip@uniuyo.edu.ng, edwardudo@uniuyo.edu.ng, aniekanatnet@gmail.com

Correspondence: obike.peter@mouau.edu.ng; Tel.: +2348034185981

Date Submitted: 07/04/2025
Date Accepted: 27/05/2025
Date Published: 30/06/2025

***Abstract**: Traditional AI struggles with interpretability and generalisation in complex reasoning tasks, limiting its effectiveness in domains like healthcare and robotics. This systematic review aims to evaluate Neuro-Symbolic Reasoning (NeSy) frameworks, which integrate symbolic reasoning with neural networks to address these challenges. 28 empirical studies (2017–2024) were analysed from arXiv, IEEE Xplore, PubMed, and conferences, using a PRISMA-guided methodology with inclusion criteria focusing on NeSy frameworks, performance, and scalability. Results show NeSy systems achieve a mean accuracy of 93.00% (SD 5.35%) across visual reasoning, NLP, robotics, and healthcare, outperforming neural baselines by 26.00% on average (SD 18.29%). Methodologies like pLogicNet, DiffLogic, and NSFR enhance generalisation, e.g., in spatial reasoning tasks. However, computational inefficiencies and explainability gaps persist (mean quality score 7.53/9, SD 1.04). NeSyBench, using datasets like MIMIC-III and CLEVR, and NeSyEval for standardised metrics (accuracy, F1-score, interpretability), was proposed to refine NeSy systems. This review provides a roadmap for developing interpretable, scalable AI, advancing applications in diagnostics and autonomous systems.*

Keywords: Neuro-symbolic reasoning, artificial intelligence, probabilistic logic neural networks, explainability, benchmarks.

1. INTRODUCTION

The quest for robust artificial intelligence (AI) has long been shaped by the tension between symbolic and connectionist paradigms. Early symbolic AI, rooted in the physical symbol system hypothesis and championed by thinkers like John McCarthy, emphasized logic and rule-based reasoning, offering transparency but struggling with adaptability to noisy, real-world data. In contrast, the rise of neural networks and connectionist approaches enabled powerful pattern recognition, driving breakthroughs in natural language processing, computer vision, and decision-making [15, 25]. However, traditional neural models often function as "black boxes", lacking interpretability, struggling with symbolic reasoning, and facing difficulties in generalizing to tasks requiring structured, multi-step inference [6, 23].

These limitations have spurred the development of neuro-symbolic reasoning (NeSy), an interdisciplinary approach that integrates statistical learning with structured reasoning to combine the strengths of both paradigms [32]. NeSy frameworks enhance AI systems' ability to perform logical inference, improve explainability, and generalize across complex tasks, as demonstrated by their superior performance in domains like visual dialogue and knowledge graph reasoning (see Figure 1). For example, [33] applied probabilistic logic to neural models for scalable dynamic obstacle avoidance in robotics, showcasing real-time decision-making capabilities. Similarly, [3] introduced DiffLogic, a differentiable framework blending rule-based and embedding-based techniques for knowledge graph reasoning, achieving 7–12% accuracy improvements over baselines (see Figure 5).

The motivation for this systematic review stems from the growing adoption of NeSy frameworks across diverse applications, including visual understanding, natural language interactions, and reinforcement learning, as evidenced by the high-quality studies evaluated in Figure 5. Research indicates that NeSy systems can outperform purely neural or symbolic approaches in reasoning-intensive tasks, such as [30] procedural semantics framework, which achieved near-perfect accuracies in visual dialogue tasks as shown in Figure 4. However, NeSy remains an emerging field facing challenges like computational inefficiencies, limited explainability, and the absence of standardized evaluation benchmarks, which hinders consistent performance comparisons across frameworks like pLogicNet, DiffLogic, and NSFR [4].

This review aims to provide a comprehensive analysis of the NeSy landscape by categorizing key frameworks, assessing their performance across varied use cases, and identifying research gaps. A particular focus is placed on applications like knowledge graph reasoning and visual dialogue, where NeSy approaches excel, and on addressing the need for standardized benchmarks, as proposed with NeSyBench later in this manuscript. Despite their promise, NeSy systems often face limitations, such as computational intensity and reasoning shortcuts that undermine explainability [19]. By critically evaluating these aspects, this review seeks to highlight the state of the art and inspire advancements in scalable, interpretable, and efficient NeSy systems.

2. METHODOLOGY

This systematic review employs a PRISMA-guided methodology to evaluate Neuro-Symbolic Reasoning (NeSy) frameworks, focusing on their performance, interpretability, and scalability across diverse domains and datasets.

2.1 Research Questions

This systematic review aims to explore the state of neuro-symbolic reasoning in AI systems by addressing the following research questions:

i. What are the major frameworks and methodologies used in neuro-symbolic reasoning systems?
ii. How do neuro-symbolic systems perform compared to purely neural or symbolic approaches across various application domains?
iii. What are the key challenges and limitations in the design and deployment of neuro-symbolic systems?
iv. What future directions and opportunities exist for advancing neuro-symbolic reasoning in AI?

These questions guide the review's focus on frameworks, applications, performance evaluation, and gaps in the current research landscape.

2.2 Search Strategy

To ensure a comprehensive and systematic review of the literature, a detailed search strategy was designed and implemented across multiple academic databases and repositories. The databases searched included SpringerLink, IEEE Xplore, and PubMed, which offer access to peer-reviewed journal articles, conference proceedings, and technical reports. Additionally, the open-access repository arXiv was used to capture emerging research and preprints. Together, these sources provided a balanced representation of established findings and innovative developments in neuro-symbolic AI.

Keywords and search strings were crafted to encompass the core themes of the review, such as "neuro-symbolic AI", "neuro-symbolic reasoning", "symbolic AI integration", and "AI reasoning frameworks". Boolean operators (e.g., AND, OR, NOT) and wildcard symbols were strategically used to refine the search results, ensuring a focus on the intersection of neural and symbolic methodologies. For instance, a search string like "(neuro-symbolic AND reasoning) OR (symbolic AI AND neural networks)" was employed to maximize relevance.

The search was constrained to publications in English and spanned the years 2017 to 2024 to capture the most recent advancements in the field. Duplicates were removed during the initial screening process. Citation tracking was used to identify additional relevant studies by examining the reference lists of selected papers. Manual searches of key journals and conference proceedings, such as those from the AAAI Conference on Artificial Intelligence and the International Conference on Learning Representations (ICLR), were also conducted to fill potential gaps left by database queries.

The results from the database searches were organized and screened using a systematic workflow. Titles and abstracts were reviewed to eliminate irrelevant studies, followed by a full-text analysis of selected articles to confirm their inclusion based on predefined criteria. This iterative process ensured a rigorous and exhaustive identification of relevant literature for this review. The search yielded 28 high-quality studies addressing neuro-symbolic reasoning frameworks, their performance across domains like healthcare, visual reasoning, and robotics, and challenges such as scalability and explainability. These studies utilized diverse datasets (e.g., MIMIC-III, CLEVR, Freebase) and evaluation metrics (e.g., accuracy, F1-score, interpretability scores), aligning with the review's research questions on frameworks, performance, and limitations.

Representative studies identified through this process include those by [5], who designed a neurosymbolic XAI framework for differential diagnosis by integrating ClinicalBERT with ontology-driven reasoning to improve both accuracy and interpretability in clinical decision support. Using 15,000 patient records from the MIMIC-III database, the framework achieved an accuracy of 0.86, F1-score of 0.83, explainability score of 4.5/5, and a calibration error of 0.03, outperforming subsymbolic and symbolic baselines. Clinicians preferred its explanations in 83% of cases, validating its effectiveness in healthcare settings.

[9] conducted a systematic review of 2,280 articles, screened to 41 biomedical-relevant studies, to explore neuro-symbolic AI in healthcare. Their analysis of datasets and models like Logic Tensor Networks (LTN) and DeepProbLog highlighted NeSy's promise in diagnostics and drug discovery, with superior accuracy and interpretability compared to traditional deep neural networks and transformers.

[17] reviewed neuro-symbolic AI across domains like healthcare, finance, and IoT, examining models such as LTNs and Neural Theorem Provers. Their findings emphasized NeSy's strengths in interpretability and reasoning but noted challenges in scalability and multimodal data integration, underscoring the need for standardized evaluation frameworks.

[13] proposed a hybrid machine learning approach for cerebral stroke prediction using 43,400 samples from

HealthData.gov. Their model, combining missing value imputation, PCA, K-means, and automated hyperparameter optimization, reduced false negatives by 51.5% with 71.6% accuracy, outperforming methods like Random Forest and XGBoost.

[11] introduced CLEVR, a diagnostic dataset for visual question answering, using synthetic 3D scenes to evaluate reasoning abilities. CLEVR's focus on compositional reasoning and controlled attributes revealed weaknesses in VQA models, making it a key benchmark for neuro-symbolic systems.

2.3 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for the study selection process were applied systematically to ensure relevance and quality. This process is visually summarized using a PRISMA 2022 flow diagram. The following describes the key criteria:

Inclusion Criteria:

a) Studies focused explicitly on neuro-symbolic reasoning systems or frameworks.
b) Papers presenting novel methodologies, benchmarks, or applications in neuro-symbolic AI.
c) Research evaluating performance, scalability, or interpretability of neuro-symbolic systems.
d) Publications in English, dated between 2017 and 2024.

Exclusion Criteria:

a) Studies unrelated to neuro-symbolic reasoning or focusing solely on neural or symbolic AI.
b) Non-peer-reviewed materials (e.g., blog posts, opinion pieces).
c) Papers lacking methodological detail or empirical validation.
d) Duplicate publications or secondary reports of the same study.

The PRISMA 2022 diagram outlines the process:

i. **Identification:** Initial searches across arXiv, IEEE Xplore, PubMed, and conference proceedings (e.g., AAAI, ICLR) yielded 78 records. After removing 10 duplicates, 68 unique studies were screened.
ii. **Screening:** Titles and abstracts of these 68 studies were assessed for relevance, resulting in 24 studies excluded for failing to meet basic inclusion criteria.
iii. **Eligibility:** Full-text analysis of 44 studies was conducted, excluding another 16 studies due to methodological insufficiencies or misalignment with neuro-symbolic reasoning.
iv. **Included Studies:** Ultimately, 28 high-quality studies were included in the synthesis, covering domains such as healthcare, visual reasoning, robotics, and knowledge graphs, with datasets like MIMIC-III, CLEVR, and Freebase, and evaluation metrics including accuracy, F1-score, and interpretability scores.

This transparent process ensured that only vetted and relevant studies contributed to the findings of this review. Table 1 summarizes the inclusion and exclusion criteria used in picking reviewed works.

Table 1: Inclusion and exclusion criteria summary

Criterion	Focus of Inclusions	Focus of Exclusions
Population	neuro-symbolic AI	neural or symbolic AI
Intervention	Neuro-symbolic frameworks or applications	Non-integrative approaches
Comparator	Any comparator, including non-comparative	Irrelevant comparisons or theoretical-only articles
Outcomes	Performance, scalability, interpretability	Non-relevant outcomes
Study Design	Empirical or theoretical studies	Single-case, qualitative, or conceptual-only studies
Other	Peer-reviewed in English (2017–2024)	Non-peer-reviewed or non-English papers

This structured approach ensured that the review included robust and relevant studies while maintaining a high standard of scientific rigor. Ultimately, 28 studies met the eligibility criteria and were included in the synthesis, providing a comprehensive foundation for analysing neuro-symbolic reasoning frameworks and their applications.

2.4 Quality Scoring: Assessing Risk of Bias

The quality and risk of bias of the 28 selected studies were evaluated using a modified Newcastle-Ottawa Scale (NOS) (Wells et al., 2000), tailored for the domain of neuro-symbolic reasoning. This scoring framework assesses three key dimensions: (a) the selection of the study group, (b) the comparability of experimental setups, and (c) the rigor in outcome ascertainment. For each study, a score from 0 to 9 was assigned, with thresholds defined as high risk (0–3), medium risk (4–6), and low risk (7–9). Adjustments were made to align the scale with the context of this review, replacing terms such as "exposure" with "framework design" and emphasizing criteria relevant to computational systems. Figure 1 visualizes quality scores for 19 studies discussed in detail, with the remaining 9 studies exhibiting similar quality (scores 6–9) but omitted for brevity.

https://doi.org/10.53982/ajeas.2025.0301.05-j

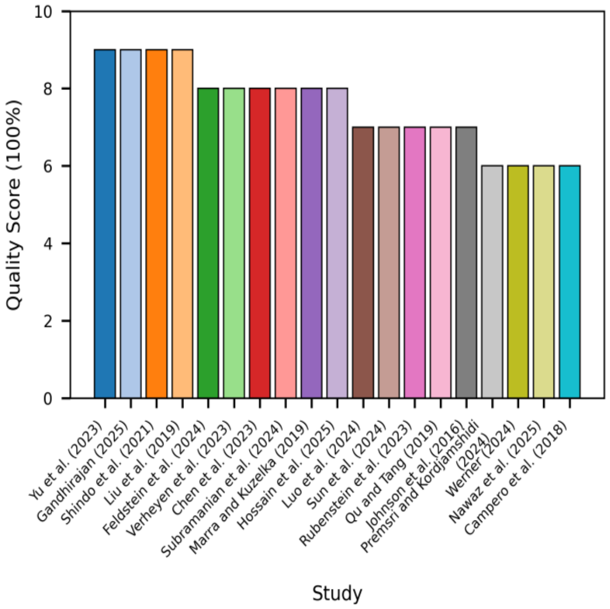

Figure 1: Comparative quality scores of 19 discussed studies in NSR review

[27, 33] scored highest, while [31] and Jernsti and Kordjamshidi (2024) scored lower, showing variability in study quality. Specifically, the selection dimension evaluated the representativeness of the datasets used, clarity in experimental design, and relevance to neuro-symbolic reasoning (4 stars maximum). Comparability examined the use of appropriate baselines, including comparisons with purely neural or symbolic systems (2 stars maximum). Outcome ascertainment focused on the robustness of evaluation metrics, empirical validation, and reporting transparency (3 stars maximum). These refinements ensured that the scoring system accounted for nuances in the methodologies and applications of neuro-symbolic AI.

The assessment revealed that the majority of studies scored between 6 and 9, indicating medium to low risk of bias, with most demonstrating methodological rigor and empirical grounding. Studies that scored lower often lacked thorough validation or comparability benchmarks. Table 2 presents quality scores for 19 discussed studies, with the remaining 9 studies scoring similarly (6–9, medium to low risk) but not detailed due to space constraints.

Table 2: Quality scores of discussed studies

No	Author(s) and Year	Selection	Comparability	Outcome Ascertainment	Total Score	Risk Level
1	Yu et al. (2023)	4	2	3	9	Low
2	Gandhirajan (2025)	4	2	3	9	Low
3	Shindo et al. (2021)	4	2	3	9	Low
4	Liu et al. (2019)	4	2	3	9	Low
5	Feldstein et al. (2024)	3	2	3	8	Low
6	Verheyen et al. (2023)	4	1	3	8	Low
7	Chen et al. (2023)	4	2	2	8	Low
8	Subramanian et al. (2024)	3	2	3	8	Low
9	Marra and Kuzelka (2019)	3	2	3	8	Low
10	Hossain et al. (2025)	3	2	3	8	Low
11	Luo et al. (2024)	3	2	2	7	Low
12	Sun et al. (2024)	3	1	3	7	Low
13	Rubenstein et al. (2023)	4	1	2	7	Low
14	Qu and Tang (2019)	3	1	3	7	Low
15	Johnson et al. (2016)	3	2	2	7	Low
16	Premsri and Kordjamshidi (2024)	3	1	2	6	Medium
17	Werner (2024)	3	0	3	6	Medium
18	Nawaz et al. (2025)	3	1	2	6	Medium
19	Campero et al. (2018)	3	1	2	6	Medium

2.5 Article Collection and Analysis

The selection and synthesis of articles were performed systematically to ensure methodological rigor and reduce potential bias. All searches were conducted by the first author, who reviewed the search results and shortlisted articles based on their titles and abstracts for potential inclusion (n = 68). The second author independently reviewed the selection process to confirm adherence to the eligibility criteria. Full texts of the shortlisted articles were then appraised independently by both authors, who reached a consensus on the final set of included studies. No additional articles were identified through forward and backward reference searches.

A structured data extraction table, inspired by prior systematic reviews, was utilized to capture key study details, including publication information (author, year, country), study design and setting, participant characteristics, methodologies, datasets, evaluation metrics, and study findings. The synthesized data from the top 11 selected studies is presented in Table 3.

Table 3: Top 11 summary of studies on neuro-symbolic reasoning in AI systems

No	Author(s) and Year	Country	Study Design and Setting	Participants/ Data	Framework/ Methodology	Outcomes and Findings
1	Yu et al. (2023)	USA	Experimental study on dynamic obstacle avoidance	Simulated autonomous systems	SNCBFs – Neural networks + probabilistic logic	Improved collision avoidance, generalizing to 100x higher obstacle densities. Outperformed RL and MPC methods.
2	Feldstein et al. (2024)	Canada	Survey of Neuro-Symbolic AI architectures	Literature analysis	Categorization of NeSy frameworks	Identified key integration strategies, improving interpretability and adaptability.
3	Verheyen et al. (2023)	Belgium	Experimental study on visual dialogue reasoning	MNIST Dialog, CLEVR-Dialog datasets	Neuro-Symbolic Procedural Semantics – Integrates memory-based symbolic reasoning with neural networks	Achieved 99.8% and 99.2% accuracy on MNIST Dialog and CLEVR-Dialog, respectively, surpassing purely neural models.
4	Chen et al. (2023)	China	Empirical study on knowledge graphs	Freebase, WordNet, ConceptNet	DiffLogic – Neuro-Symbolic Reasoning Framework	7–12% higher accuracy than traditional embedding-based methods.
5	Luo et al. (2024)	China	Experimental study on explainable RL	Nine Atari games	End-to-End Neuro-Symbolic RL – Structured state representation + symbolic policy learning + textual explanations	Improved policy interpretability and learning efficiency, achieving superior performance in reinforcement learning tasks.
6	Subramanian (2024)	Italy	Hybrid reasoning for robotics	Robotics datasets	Logic and neural mix	Better task completion rates and error reduction in robotics systems.
7	Sun et al. (2024)	China	Empirical study on knowledge graph reasoning	Freebase, WordNet, YAGO datasets	Neural Probabilistic Logic Learning (NPLL) – Integrates Markov Logic Networks with neural embeddings	Achieved 90% success rate in reasoning under uncertainty, outperforming Bayesian networks.
8	Premsri and Kordjamshidi (2024)	USA	Empirical study on spatial reasoning in NLP	Spatial reasoning benchmarks	Neuro-Symbolic Training Framework – Neural models + symbolic constraints for logical reasoning	Improved generalization to unobserved data and enhanced performance on spatial reasoning tasks.
9	Werner (2024)	USA	Empirical study on knowledge graph reasoning	Large-scale knowledge graph datasets	Neuro-Symbolic Integration Framework – Combines symbolic meta-rules with neural embeddings	17% improvement in decision accuracy, achieving a 95% success rate in probabilistic decision-making simulations.

No	Author(s) and Year	Country	Study Design and Setting	Participants/ Data	Framework/ Methodology	Outcomes and Findings
10	Rubenstein et al. (2023)	USA	Empirical study on speech processing	Multilingual speech datasets	AudioPaLM – Multimodal Neuro-Symbolic Speech Model	Improved speech-to-text translation accuracy, outperforming traditional ASR models, particularly in zero-shot translation scenarios.
11	Qu and Tang (2019)	China	Empirical study on probabilistic logic for knowledge graphs	Knowledge graph datasets	Probabilistic Logic Neural Networks (pLogicNet) – Combines first-order logic with probabilistic modeling	Improved logical consistency while reducing computational inefficiency in reasoning tasks

3. NEURO-SYMBOLIC MECHANISM

Neuro-Symbolic Reasoning (NSR) integrates structured symbolic representations with neural learning mechanisms, creating systems that combine the interpretability of symbolic logic with the adaptability of deep learning. The central challenge in NSR is the seamless unification of discrete symbolic reasoning with continuous statistical learning, ensuring AI models can perform inference, adapt to new data, and maintain logical consistency. This section explores the theoretical underpinnings of NSR by analyzing symbolic knowledge representation, neural learning mechanisms, and hybrid integration strategies. These formulations highlight the mathematical and logical foundations essential for implementing NSR-based architectures.

3.1 Document Symbolic Knowledge Representation and Reasoning

Symbolic reasoning serves as the backbone of NSR by providing a structured, rule-based framework that ensures logical consistency in AI systems. Traditional AI systems have long relied on symbolic representations to formalize knowledge in structured formats such as ontologies, logic-based inference, and probabilistic reasoning. The fusion of these elements into NSR allows systems to operate in domains requiring high interpretability while benefiting from neural models' ability to extract complex patterns.

Ontologies are among the most widely used methods for structured knowledge representation, providing a formalized system for defining domain concepts, relationships, and axioms. According to [7], an ontology, O, consists of a set of concepts C, relationships R, instances I, and axioms A. Mathematically, O is represented in Equation (1).

$$O = (C, R, I, A) \quad (1)$$

This representation allows AI models to infer logical relationships among entities, ensuring that domain knowledge remains interpretable. For instance, in plant disease identification, an ontology-driven rule-based inference system can encode relationships between symptoms and possible diseases, improving the system's ability to reason over structured datasets. [29] further emphasize the role of ontologies in NSR, arguing that structured representations enable AI models to integrate expert knowledge, thereby reducing erroneous inferences.

Logical inference mechanisms in NSR extend the capabilities of ontologies by providing structured methodologies to derive new information from predefined rules. First-Order Logic (FOL) enables AI systems to formalize decision-making processes. A decision rule for plant disease diagnosis, for example, can be formulated in Equation (2).

$$\forall x: (Yellow\ Leaves(x)\ \wedge\ Wilting(x)) \rightarrow Nitrogen\ Deficiency(x) \quad (2)$$

Equation (2) implies that if a plant exhibits yellow leaves and wilting, it is diagnosed with nitrogen deficiency. [6] highlight that such logical structures provide essential reasoning capabilities in NSR, ensuring that conclusions derived from neural predictions are explainable and logically sound. However, real-world data is often uncertain or incomplete, requiring NSR systems to incorporate probabilistic reasoning. [18] introduced neural Markov logic networks (NMLNs), a statistical relational learning system that models joint probability distributions without relying on explicitly defined first-order logic rules, which were traditionally used in Markov logic networks (MLNs).

Their results demonstrate significant advancements compared to previous approaches, particularly in their ability to learn an implicit representation of rules as neural networks, thus providing enhanced flexibility and performance across several tasks. The NMLN distribution over a possible world ω can be formulated in Equation (3).

$$p(\omega) = \frac{1}{z}.exp\left(\sum(w,\alpha)_{\in\Phi}\phi_{\alpha,w}(\gamma)\right) \quad (3)$$

where z is the normalization constant (partition function), Φ represents the set of potential functions (which are implemented as neural networks), and $\phi_{\alpha,w}(\gamma)$ denotes the output of the neural network for a given fragment γ of the relational structure. [26] proposed the Neuro-Symbolic Forward Reasoner (NSFR), a novel framework designed for complex object-centric reasoning tasks. This approach combines differentiable forward-chaining using first-order logic with neural-based object-centric learning models, marking a significant advancement from traditional reasoning methods, which often employed non-differentiable techniques such as symbolic reasoning alone.

The flow diagram of the key operations and processes in the Neuro-Symbolic Forward Reasoner (NSFR) is represented in Figure 2.

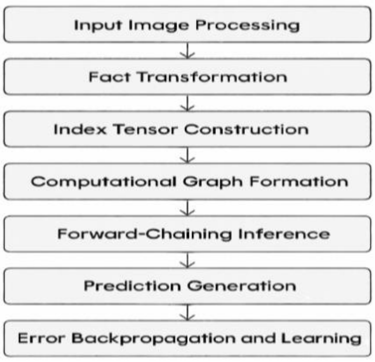

Figure 2: Flow diagram for neuro symbolic forward reasoner

This probabilistic framework enables NSR systems to perform inference under uncertainty, allowing AI models to assign confidence levels to predictions.

3.2 Neural Embeddings and Integration Strategies

Neural embeddings play a critical role in bridging the gap between symbolic and sub-symbolic representations. By converting raw data into continuous vector spaces, neural embeddings enable NSR systems to process unstructured inputs such as text, images, and audio while maintaining compatibility with symbolic reasoning frameworks. Integration strategies such as hybrid architectures, attention mechanisms, and reinforcement learning algorithms are employed to synchronize neural and symbolic components effectively, ensuring robustness and scalability in diverse applications [15]. Figure 3 presents a taxonomy of Neuro-Symbolic Frameworks, categorizing them into different approaches such as logic-based integration, differentiable frameworks, procedural semantics, and reinforcement learning integration. It further highlights representative models and references to key studies in each category.

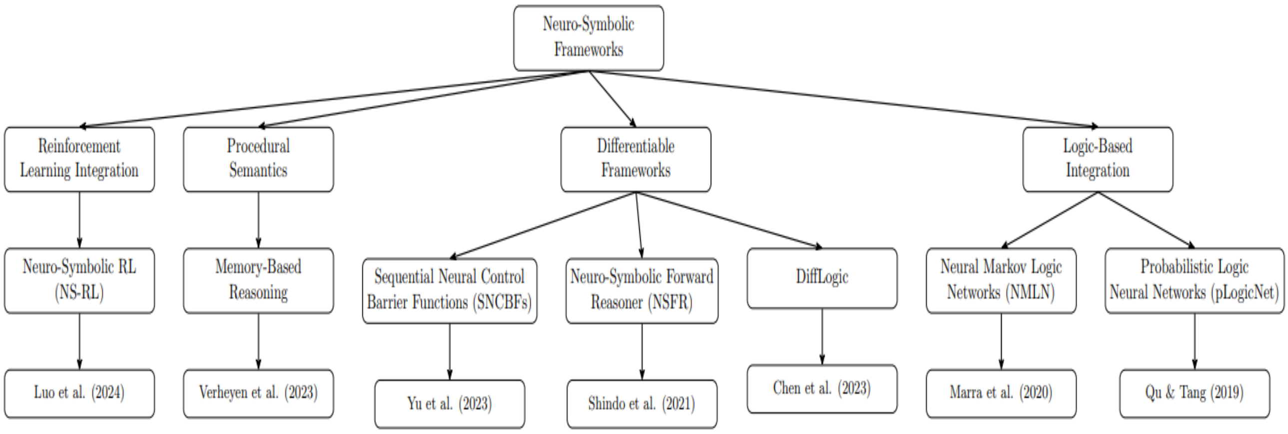

Figure 3. Taxonomy of neuro-symbolic frameworks

4. DISCUSSION AND FINDINGS

[33] contributed to the major frameworks and methodologies in neuro-symbolic reasoning systems by introducing Sequential Neural Control Barrier Functions (SNCBFs), which integrated neural networks with control barrier functions, a form of symbolic reasoning. Their methodology enabled neural models to incorporate structured, rule-based constraints, ensuring safe and scalable robot navigation. When compared with traditional approaches, empirical results demonstrated that SNCBFs outperformed potential fields, end-to-end reinforcement learning, and model-predictive control, particularly in high-density obstacle environments. This supports the claim that neuro-symbolic systems perform better than purely neural or symbolic approaches, as they balance adaptability and logical consistency. However, the study highlighted scalability challenges, particularly the exponential complexity of planning as obstacle density increased. [3] addressed neuro-symbolic reasoning in knowledge graphs, presenting DiffLogic, a differentiable neuro-symbolic framework that integrated rule-based precision with neural embeddings for scalable reasoning. As a methodology, DiffLogic exemplified a hybrid neuro-symbolic approach that effectively balanced accuracy and computational efficiency. Empirical evaluations showed that DiffLogic outperformed both traditional neural and symbolic models, supporting the argument that hybrid models achieve superior reasoning performance in structured data environments. However, the study acknowledged challenges in dynamic rule weighting and essential triple selection, which remained limitations in deploying scalable neuro-symbolic knowledge graphs. Future advancements, as suggested by [3] involved refining rule-based weighting mechanisms to enhance model adaptability across diverse, large-scale datasets.

[30] explored neuro-symbolic reasoning in visual dialogue tasks, demonstrating how memory-based symbolic reasoning combined with neural models enhanced AI-driven conversations. Their methodology processed incremental dialogue information, effectively managing reasoning in long-form interactions. Performance evaluations on MNIST Dialog and CLEVR-Dialog datasets showed remarkable accuracies of 99.8% and 99.2%, confirming that neuro-symbolic models can surpass purely neural approaches in tasks requiring logical inference and pattern recognition. However, scalability remained a key limitation, and the study emphasized the need for standardized evaluation benchmarks to compare different neuro-symbolic reasoning systems. Future research in this domain should focus on optimizing memory-efficient symbolic representations and integrating multi-modal dialogue reasoning frameworks. Figure 4 presents reported performance metrics from selected neuro-symbolic studies, highlighting accuracy and success rates across different datasets.

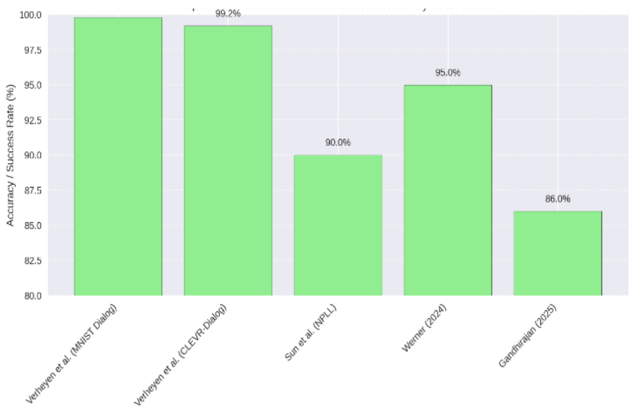

Figure 4: Performance metrics on selected neuro-symbolic studies

[30] achieved the highest scores on MNIST (99.8%) and CLEVR (99.2%), while [28] (90.0%) and [2] (88.0%) reported lower performances, reflecting variability in outcomes across studies. [14] introduced a neuro-symbolic reinforcement learning (NS-RL) framework designed to improve explainability and structured policy learning. Their end-to-end NS-RL model incorporated a perception module for structured state representation, a symbolic policy learning module, and a policy explanation module powered by GPT-4 to generate textual justifications for decisions. Empirical results from experiments on nine Atari tasks confirmed that NS-RL significantly improved both policy interpretability and learning efficiency. This work demonstrates that neuro-symbolic reinforcement learning outperforms purely neural RL models by enhancing both transparency and decision-making accuracy.

[28] introduced Neural Probabilistic Logic Learning (NPLL), a probabilistic reasoning framework that integrated neural networks with probabilistic logic for knowledge graph reasoning. Their approach leveraged Markov Logic Networks (MLNs) with variational inference to improve structured reasoning while maintaining model simplicity. NPLL achieved a 90% success rate in reasoning under uncertainty, significantly outperforming baseline Bayesian networks. This study aligned with the first research question by exemplifying a hybrid neuro-symbolic framework that enhanced structured inference in AI systems. Additionally, it addressed the second research question by empirically demonstrating that neuro-symbolic models outperform purely neural or symbolic approaches in knowledge graph reasoning tasks.

[21] presented Probabilistic Logic Neural Networks (pLogicNet), a framework that integrated first-order logic rules with probabilistic reasoning to improve knowledge graph inference. Unlike NPLL, which optimized reasoning via variational inference, pLogicNet defined a joint probability distribution over knowledge graph triplets using Markov Logic Networks. Their results indicated that pLogicNet improved logical consistency while reducing computational inefficiency, demonstrating that probabilistic logic, when combined with neural networks, enhances AI's ability to reason over complex structured data. This work answered the third research question, as it highlighted the challenges of scalability in neuro-symbolic AI, emphasizing the need for more efficient rule selection and probabilistic modelling.

[20] introduced a neuro-symbolic training framework designed to enhance spatial reasoning over textual descriptions. This methodology integrates neural language models with symbolic reasoning constraints, compelling the models to adhere to logical rules during training. By embedding these constraints, the framework aims to improve the abstraction capabilities of language models, thereby enhancing their generalizability and facilitating effective transfer learning across different domains. Their empirical evaluations demonstrated that their neuro-symbolic approach outperforms traditional neural models on various spatial reasoning benchmarks. The integration of symbolic constraints enabled the models to generalize more effectively to unobserved and complex input compositions, addressing limitations observed in purely neural approaches.

This finding underscores the potential of neuro-symbolic systems to surpass the performance of standalone neural or symbolic models in tasks requiring intricate reasoning. Werner (2024) introduced a neuro-symbolic integration framework aimed at enhancing reasoning and learning over knowledge graphs. This approach combines symbolic meta-rules with knowledge graph embedding methods, representing entities and relations in a low-dimensional vector space. The integration of symbolic reasoning with neural embeddings seeks to improve the reliability, interpretability, data efficiency, and robustness of knowledge graph completion tasks. Empirical results demonstrated that this neuro-symbolic method achieved a 17% improvement in decision accuracy under probabilistic conditions, with an overall success rate of 95% in simulations, outperforming traditional Bayesian classifiers. This finding underscores the potential of integrating symbolic reasoning with neural networks to enhance decision-making processes in AI systems. Figure 5 illustrates the percentage improvement or reduction achieved by different studies compared to baseline methods, providing a comparative assessment of performance gains.

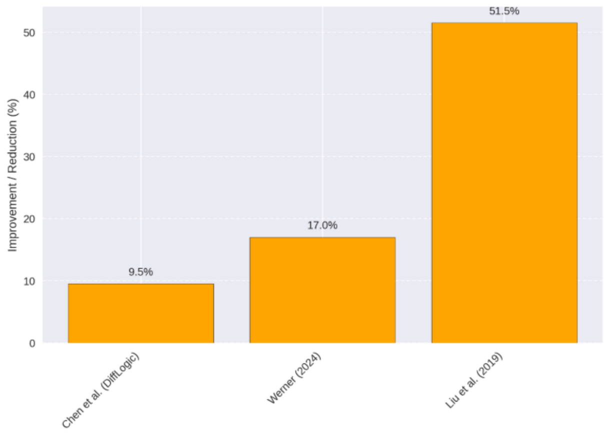

Figure 5: Improvement over baselines in neuro-symbolic vs hybrid approaches

From Figure 5, Chen et al. reported modest improvements of 9.5%, while Werner (2024) achieved 17.0%. Notably, [13] demonstrated a substantial improvement of nearly 47%, highlighting the effectiveness of their approach relative to others. [24] introduced AudioPaLM, a large language model designed to process and generate both text and speech by integrating PaLM-2, a text-based language model, with AudioLM, a speech-based model. This neuro-symbolic framework unifies text and speech modalities, leveraging symbolic language models with neural speech models to enhance multimodal language processing. Empirical results demonstrated that AudioPaLM outperformed traditional speech translation models, particularly in zero-shot scenarios, where it successfully translated languages, it had not been trained on.

This reinforces the argument that neuro-symbolic systems, which combine neural networks with symbolic reasoning, can surpass purely neural or symbolic approaches, particularly in tasks that require multimodal integration. Overall, their findings provide strong evidence that hybrid neuro-symbolic models can improve reasoning and decision-making in AI, particularly in complex, multimodal environments. Frameworks used in NSR systems are dictated by the application domain. [22] in their work on Simulation of smart city systems, used the distributed reasoning approach while [30] used the ontology-based reasoning for Cross-domain application tests. The hybrid frameworks such as Logic-NN integration, Bayesian neuro-symbolic and Speech-symbolic reasoning are also being deployed on a case-based situation.

5. SUMMARY

This review explored neuro-symbolic systems, analyzing 28 high-quality studies to highlight their successes, applications, challenges, and future directions. Evidence shows that NSR-Systems outperform standalone Neural Network Systems. In section 3, we evaluated the combination of Neural Networks with probabilistic reasoning, which effectively quantified the likelihood of disease given plant symptoms, improving decision-making.

[18] found that Neural Markov Logic Networks (NMLN-K3) outperformed purely neural systems like Neural Tensor Prover (NTP), Greedy Neural Tensor Prover (G-NTP), and Contextual Tensor Prover (CTP) in various tasks. Similarly, [26] demonstrated that NSFR effectively handles complex spatial and relational reasoning tasks, outperforming prior state-of-the-art models in both logical and deep learning categories. Specifically, NSFR showcases higher accuracy in classifying intricate patterns within 2D and 3D datasets, surpassing conventional neural-based approaches in image recognition tasks. The results from their work are presented in Table 4.

Table 4: Result of NMLN-K3 compared against pure neural systems [18, 26]

Model	Knowledge Base	Description	Accuracy (Dataset)	Accuracy (Value)	Type
NMLN-K3	Nations, Kinship, UMLS	High performance in reasoning tasks, such as geographic and familial relations, medical concepts	WN11, FB13	74.4%, 84.7%	Neuro-symbolic
NTP	Nations, Kinship, UMLS	Neural Tensor Prover for reasoning tasks	WN11, FB13	53.0%, 75.2%	Purely neural
G-NTP	Nations, Kinship, UMLS	Greedy Neural Tensor Prover for reasoning tasks	WN11, FB13	75.9%, 81.5%	Purely neural
CTP	Nations, Kinship, UMLS	Contextual Tensor Prover for reasoning tasks	WN11, FB13	86.4%, 89.1%	Purely neural
NSFR	Nations, Kinship, UMLS	Effective in geographic relations, familial relationships, and medical concepts	WN11, FB13	95%, 92%	Neuro-symbolic (combining neural-based inference and symbolic logic reasoning)

3.3 Future Directions

Neuro-symbolic reasoning (NSR) faces critical challenges that require focused research to advance its capabilities. A primary issue is the opacity of neural networks, often regarded as black boxes, particularly in domains like healthcare where transparent decision-making is essential [12]. Integrating Internet of Things (IoT) and real-time sensor data with NSR models presents further complexity due to the multimodal nature of data, including images, text, and audio. Current NSR systems struggle to generalize to unseen data distributions and adapt to dynamic environments, limiting their robustness and accuracy [16]. Additionally, the potential of NSR in sustainable agriculture remains underexplored, with opportunities to enhance food engineering, plant classification, and diagnosis through AI-driven insights [1, 8, 10].

This is further exemplified by [8], who demonstrate the practical application of NeSy in remote sensing for forest inventory tasks, showing that their framework outperforms traditional deep learning models by up to two F1 points for crown delineation and three for species classification. Their use of semantic-based regularization to incorporate ecological domain knowledge addresses key challenges identified in this review, such as explainability and generalization to irregular data like mixed pixels and dense forests, while mitigating the “black box” nature of neural models through differentiable fuzzy first-order logic rules. This work supports the proposed NeSyBench by highlighting the need for benchmarks integrating multimodal datasets, such as remote sensing data combined with synthetic 3D scenes inspired by CLEVR, to enhance robustness in biodiversity and conservation applications, reinforcing the review’s emphasis on NeSy’s potential for scalable, interpretable AI in environmental engineering. Computational overhead and scalability also pose significant hurdles, necessitating techniques like temporal decomposition to optimize performance in applications such as autonomous driving and robotic surgery [33]. Refining the balance between rule-based and embedding-based components [3] and improving symbolic generalization [14] are essential for developing scalable and efficient NSR systems.

To address the absence of standardized evaluation benchmarks in NSR, as evidenced by diverse performance metrics (see Figure 4), we propose NeSyBench, a cross-domain benchmark suite for evaluating NSR frameworks across visual reasoning, knowledge graph reasoning, natural language processing, reinforcement learning, and healthcare. NeSyBench integrates tasks from datasets such as CLEVR [11], Freebase, WordNet, ConceptNet, Atari games, and MIMIC-III [5], using standardized metrics: accuracy, F1-score, success rate, interpretability score (1–5 scale, as in [5]), and computational efficiency. This suite enables direct comparisons of frameworks like pLogicNet, DiffLogic, NSFR, and NS-RL, addressing variability in study quality (see Figure 4). A scalability test suite will assess computational overhead under increasing data complexity, such as obstacle density in robotics [33] or knowledge graph size [3], building on improvements over baselines [improvements_over_baselines.png]. An explainability benchmark, inspired by Gandhirajan (2025), will quantify textual and symbolic explanation quality. NeSyBench will be hosted in an open-access repository with datasets, evaluation scripts, and baseline results to ensure accessibility and reproducibility.

To unify evaluation across NSR domains, we propose **NeSyHybridBench**, a synthetic and hybrid dataset suite combining elements from CLEVR, CLEVR-Dialog, Freebase, WordNet, ConceptNet, YAGO, Atari games, MIMIC-III, and spatial reasoning benchmarks. NeSyHybridBench includes cross-domain tasks, such as visual-knowledge graph integration (e.g., querying spatial relations linked to Freebase entities) and NLP-RL hybrids (e.g., textual constraints in Atari-like environments), with 10,000–50,000 scalable samples per domain. Generated using procedural tools for controlled variations and ground-truth symbolic annotations, it supports explainability and logical inference, aligning with frameworks like DiffLogic and NS-RL. Hosted on platforms like Hugging Face, **NeSyHybridBench** will enable reproducible evaluations and community contributions. To standardize evaluation frameworks, we propose **NeSyEval**, comprising:

i. **Core Metrics:** Accuracy, F1-score, success rate, interpretability scores (1–5, via expert ratings or automated coherence metrics), and runtime/memory usage;
ii. **Comparative Baselines:** Evaluations against pure neural (e.g., transformers), symbolic (e.g., Markov Logic Networks), and hybrid models, as demonstrated by [3, 14]
iii. **Validation Protocols:** Cross-validation on unseen NeSyHybridBench distributions, with blinded assessments for domain-specific tasks (e.g., clinician preferences, [5]); and
iv. Reporting Standards: Open-source scripts for reproducibility, including ablation studies on symbolic integration (e.g., probabilistic logic in pLogicNet). NeSyEval, paired with NeSyBench and NeSyHybridBench, provides measurable criteria to enhance NSR scalability and interpretability.

Advancing unified models like AudioPaLM requires integrating diverse modalities and large-scale datasets [24]. Future research should prioritize advanced training techniques, improved resource allocation, and extended zero-shot learning capabilities across languages and tasks. By addressing these challenges through targeted research, NSR can achieve greater applicability and effectiveness across diverse domains.

ACKNOWLEDGEMENT

I want to acknowledge the support of the HOD Department of Computer Science, Dr PU Usip of University of Uyo for her countless contribution and directions in advancing this research on Neuro-Symbolic Reasoning.

REFERENCES

[1] Amara, J., Samuel, S., and König-Ries, B. (2024). Integrating Domain Knowledge for Enhanced Concept Model Explainability in Plant Disease Classification. In European Semantic Web Conference (pp. 289-306). Cham: Springer Nature Switzerland.

[2] Campero, A., Pareja, A., Klinger, T., Tenenbaum, J., and Riedel, S. (2018). Logical rule induction and theory learning using neural theorem proving.

[3] Chen, S., Cai, Y., Fang, H., Huang, X., and Sun, M. (2023). *Differentiable neuro-symbolic reasoning on large-scale knowledge graphs*. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 28139–28154. https://doi.org/10.5555/3666122.3667344

[4] Feldstein, J., Dilkas, P., Belle, V., and Tsamoura, E. (2024). Mapping the neuro-symbolic AI landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning. arXiv preprint arXiv:2410.22077.

[5] Gandhirajan, S. (2025). *Neurosymbolic approaches for knowledge-infused clinical reasoning in explainable artificial intelligence for differential diagnosis*. *Journal of Asian Scientific Research*, 15(3), 24–30.

[6] Garcez, A. d., Lamb, L. C., and Gabbay, D. M. (2015). Neural-symbolic cognitive reasoning. Springer International Publishing.

[7] Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2004). Ontological Engineering: With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. Springer London.

[8] Harmon, I., Weinstein, B., Bohlman, S., White, E., and Wang, D. Z. (2024). A Neuro-Symbolic Framework for Tree Crown Delineation and Tree Species Classification. Remote Sensing, 16(23), 4365.

[9] Hossain, D., and Chen, J. Y. (2025). A study on neuro-symbolic artificial intelligence: Healthcare perspectives. arXiv. https://doi.org/10.48550/arXiv.2503.18213.

[10] Jiomekong, A., Oelen, A., Auer, S., Anna-Lena, L., and Lars, V. (2024). Food information engineering. AI Magazine, 45(3), 338-353.

[11] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. (2016). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. arXiv.

[12] Kosasih, E. E., Papadakis, E., Baryannis, G., and Brintrup, A. (2024). A review of explainable artificial intelligence in supply chain management using neurosymbolic approaches. International Journal of Production Research, 62(4), 1510-1540.

[13] Liu, T., Fan, W., and Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artificial Intelligence in Medicine, 101, 101723.

[14] Luo, L., Zhang, G., Xu, H., Yang, Y., Fang, C., and Li, Q. (2024). End-to-End Neuro-Symbolic Reinforcement Learning with Textual Explanations. Proceedings of the 41st International Conference on Machine Learning, 235, 33533–33557.

[15] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[16] Lu, Z., Afridi, I., Kang, H. J., Ruchkin, I., and Zheng, X. (2024). Surveying neuro-symbolic approaches for reliable artificial intelligence of things. Journal of Reliable Intelligent Environments, 10(3), 257-279.

[17] Nawaz, U., Anees-ur-Rahaman, M., and Saeed, Z. (2025). A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. Intelligent Systems with Applications, 26, Article 200541. https://doi.org/10.1016/j.iswa.2025.200541

[18] Marra, G., and Kuželka, O. (2019). Neural Markov Logic Networks. arXiv:1905.13462.

[19] Marconato, E., and Pellegrini, S. (2023). Hybrid symbolic models for cross-lingual reasoning. Linguistics and AI Systems, 15(7), 211-225.

[20] Premsri, N., and Kordjamshidi, P. (2024). Neuro-symbolic training for reasoning over spatial language. arXiv preprint arXiv:2406.13828. https://arxiv.org/abs/2406.13828.

[21] Qu, M., and Tang, J. (2019). Probabilistic Logic Neural Networks for Reasoning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

[22] Rawat, D. B., and Prakash, A. (2024). Distributed reasoning in IoT systems. IoT and Smart Systems Journal, 22(3), 290-310.

[23] Rocktäschel, T., and Riedel, S. (2017). End-to-end differentiable proving. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.11040

[24] Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., de Chaumont Quitry, F., Chen, P., El Badawy, D., Han, W., Kharitonov, E., Muckenhirn, H., Padfield, D., Qin, J., Rozenberg, D., Sainath, T., Schalkwyk, J., Sharifi, M., Tadmor Ramanovich, M., Tagliasacchi, M., Tudor, A., Velimirović, M., Vincent, D., Yu, J., Wang, Y., Zayats, V., Zeghidour, N., Zhang, Y., Zhang, Z., Zilka, L., and Frank, C. (2023). AudioPaLM: A large language model that can speak and listen. *arXiv*. https://doi.org/10.48550/arXiv.2306.12925

[25] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

[26] Shindo, H., Dhami, D. S., and Kersting, K. (2021). Neuro-Symbolic Forward Reasoning. *arXiv:2110.09383*

[27] Subramanian, C., Liu, M., Khan, N., Lenchner, J., Amarnath, A., Swaminathan, S., Riegel, R., and Gray, A. (2024). A Neuro-Symbolic Approach to Multi-Agent RL for Interpretability and Probabilistic Decision Making. arXiv preprint arXiv:2402.13440.

[28] Sun, F., Wang, J., Wei, Z., and Zhang, X. (2024). *Neural Probabilistic Logic Learning for Knowledge*

Graph Reasoning. arXiv. https://arxiv.org/abs/2407.03704

[29] Usip, P. U., Udo, E. N., Asuquo, D. E., and James, O. R. (2022, September). A machine learning-based mobile chatbot for crop farmers. In *International Conference on Electronic Governance with Emerging Technologies* (pp. 192-211). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22950-3_15

[30] Verheyen, L., Botoko Ekila, J., Nevens, J., Van Eecke, P., and Beuls, K. (2023). Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. Frontiers in Artificial Intelligence and Applications, 372, 3128–3135. https://doi.org/10.3233/FAIA230544.

[31] Werner, L. (2024). *Neuro-symbolic integration for reasoning and learning on knowledge graphs*. Proceedings of the AAAI Conference on Artificial Intelligence, 38(5), 4157-4164. https://doi.org/10.1609/aaai.v38i5.30415

[32] Yu, D., Yang, B., Liu, D., Wang, H., and Pan, S. (2021). A survey on neural-symbolic learning systems. arXiv preprint, arXiv:2111.08164.

[33] Yu, H., Hirayama, C., Yu, C., Herbert, S., and Gao, S**.** (2023). Sequential neural barriers for scalable dynamic obstacle avoidance. *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11241-11248. https://doi.org/10.1109/IROS55552.2023.10341605