Neuro-Symbolic Artificial Intelligence:
A Task-Directed Survey in the Black-Box Models Era

Giovanni Pio Delvecchio , **Lorenzo Molfetta** , **Gianluca Moro**

Department of Computer Science and Engineering, University of Bologna, Cesena Campus
Via dell'Università 50, I-47522 Cesena, Italy
{g.delvecchio, lorenzo.molfetta, gianluca.moro}@unibo.it

Abstract

The integration of symbolic computing with neural networks has intrigued researchers since the first theorizations of Artificial intelligence (AI). The ability of Neuro-Symbolic (NeSy) methods to infer or exploit behavioral schema has been widely considered as one of the possible proxies for human-level intelligence. However, the limited semantic generalizability and the challenges in declining complex domains with pre-defined patterns and rules hinder their practical implementation in real-world scenarios. The unprecedented results achieved by connectionist systems since the last AI breakthrough in 2017 have raised questions about the competitiveness of NeSy solutions, with particular emphasis on the Natural Language Processing and Computer Vision fields. This survey examines task-specific advancements in the NeSy domain to explore how incorporating symbolic systems can enhance explainability and reasoning capabilities. Our findings are meant to serve as a resource for researchers exploring explainable NeSy methodologies for real-life tasks and applications. Reproducibility details and in-depth comments on each surveyed research work are made available at https://github.com/disi-unibo-nlp/task-oriented-neuro-symbolic.git.

1 Introduction

"*Stacked neural layers is all we need*". This statement summarizes most of the current research efforts in the Artificial Intelligence (AI) field, especially in Natural Language Processing (NLP) and Computer Vision (CV). The advancements and achievements of the latest neural models are marvelous, but they conceal fundamental drawbacks regarding data efficiency and explainability. The need to resort to Neuro-Symbolic (NeSy) components naturally arises from the necessity for trustworthy and efficient solutions. The data type of "thoughts" in the connectionist approaches—also known as tensors—and the hidden unsupervised manipulations of such information constitute a physical barrier to hierarchical and abstract planning, whose achievement cannot be reached via mere input-output relationships [Marcus, 2018]. If we
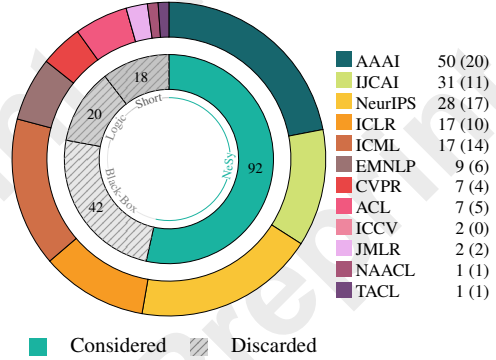

Figure 1: Distribution of NeSy peer-reviewed papers over the period 2017-2024. The innermost ring delineates the inclusion criteria. Greyscale slices denote studies excluded for relying on black-box methods, exclusively logical approaches, or brevity, while the colored slice comprises the surveyed ones. The outermost ring represents the number of research works from each selected venue, with exact counts reported in the legend. The number of papers considered for each track is enclosed in parentheses.

envision a not-too-distant future where data availability and quality become critical concerns, mainly due to the potential bias introduced by synthetic information, we urge finding and exploring research pathways with compositional abilities weakly related to the training set's dimension [Giunchiglia *et al.*, 2022]. We uphold that neuro-symbolic methods can provide an alternative to break the curse-of-dimensionality modern methodologies suffer from. While effective, sparsity enforcement regularization mechanisms [Bach, 2017] require strong assumptions about the nature of the target distribution, which is unrealistic for the complex domains NLP and CV models usually deal with [Bronstein *et al.*, 2021]. We argue that symbolic components can be the regularization for forcing more complex behavior into neural models. By combining data-driven insights with explainable, logic-based representations, we can achieve higher generalization capabilities [Besold *et al.*, 2021]. Other data-driven paradigms have also been considered [Lodi *et al.*, 2010; Domeniconi *et al.*, 2015].

As the keywords "reasoning" and "explainability" are gaining more attention in the community, we highlight a key notational misunderstanding between rule-guided and post-

inference forensic strategies. While both methods aim to provide a rationale behind predictions, they fundamentally differ in their approach. Rule-guided methods constrain the output to meet specific criteria, while post-inference approaches retrospectively reconstruct plausible interpretable reasoning pathways that led to the results. We further claim that these terms should not be used to advocate human-like faculties but as proxies for generalization behavior within a hierarchical backbone architecture governed by predefined rules or component-interaction schemas.

Our Contribution The lack of a clear definition of NeSy in the AI field, the inconsistency in evaluation benchmarks, and the multitude of uncoordinated study directions with no common comparison grounds primarily drive this survey. In this work, we conduct a task-directed literature analysis focusing on how NeSy methods apply and scale in diverse application contexts. We assess the methodological soundness of such approaches and compare them to black-box systems in real-world scenarios to identify the current limitations of these strategies. Departing from recent surveys that attempt to classify research works along taxonomical axes or network architectural variations, we propose a new methodological approach that disentangles these conceptual and practical inconsistencies by providing a comprehensive and critical overview of NeSy systems. We aim to deepen the understanding of these hybrid systems' advantages and inherent limitations, laying the groundwork for future research and bridging the gap between explainability and performance in AI.

2 Taxonomy

This section discusses the structure and systematic reproducible methodologies followed to define our task-directed taxonomy for NeSy methods. We further examine datasets, evaluation benchmarks, and their limitations in delineating fair comparisons across different approaches. As previously discussed, we acknowledge the nuanced usages and inconsistency in the declination of the term NeSy in NLP and CV literature, where it is often misapplied to methods claiming reasoning and interpretability without structured symbolic frameworks. In this paper, NeSy exclusively refers to approaches integrating neural networks with symbolic components such as solvers, logical rules, or state-action schemas.

Inclusion Criteria This survey aims to investigate NeSy systems from the last AI revolution in 2017. To identify relevant research systematically, we used the DBLP SPARQL endpoint to collect research papers published between 2017 and 2024 from general-purpose leading venues in those fields where such technological leap has taken hold more — Natural Language Processing (NLP) and Computer Vision (CV). A customized query was executed to retrieve publications related to NeSy methods using the keywords: "neuro-symbolic", "nesy", "rule-based", "probabilistic-logic", "probabilistic-reasoning", "logic-based", "soft-logic", "fuzzy-logic", "concept-learning", "inductive-logic programming". Figure 1 illustrates the number and distribution of the research works meeting our requirements. We meticulously analyzed the resulting 172 papers and categorized them to

identify works that deviated from the designated NeSy formulation, excluding non-interpretable methods, purely logical approaches, and short papers. This collection was thoroughly studied to derive a pool of research works focusing on NeSy approaches that integrate symbolic reasoning with neural networks. We further explored relevant related works to gain a precise picture of the NeSy research landscape.

Datasets and Benchmarks We analyzed real-world datasets and benchmarks to ensure an unbiased comparison across methods while focusing on practical applications. This study highlighted reproducibility challenges in tasks involving sampling operations, such as negative-example mining. We found substantial inconsistencies on the WN18RR[1] benchmark for different research works using the same black-box model baselines. We further spotlight an evaluation trend in the Nesy literature where synthetic datasets and toy tasks are specifically tailored for the features of the newly proposed method. For this reason, we consider them unsuitable for drawing conclusions and comparisons. In Figure 2, we label most of the identified NeSy tasks with their respective evaluation benchmark if they respect our fairness criteria. Missing values appear in image generation, where different instruction prompts significantly impact results; in Reinforcement Learning (RL), where each method is evaluated on custom tasks or games; and in causal effect estimation, due to outdated benchmarks with limited entries.

A Task-Oriented Formulation We organize our analysis by identifying three distinct methodological frameworks that traverse the neural-symbolic research landscape: (1) Rule Mining, (2) Rule Enforcement, and (3) Program Synthesis. Within each framework, we categorize techniques based on their theoretical grounding and the tasks they address, focusing on how constraints, symbolic structures, and interpretable programmatic constructs are respectively extracted, integrated, or synthesized. Our taxonomy in Figure 2 is designed to be read in two complementary ways. A top-down perspective highlights how these frameworks interconnect, emphasizing shared logical formalisms and learning paradigms. A bottom-up approach begins with a specific task, moving upward to identify the most relevant methodological families and techniques for that objective. This dual perspective preserves the field's theoretical coherence while promoting a task-directed formulation of NeSy solutions. Enabling bidirectional exploration between related methodologies allows researchers to identify analogous strategies to refine or expand existing approaches based on their objectives. In Figure 3 we show examples of the formal languages employed by each theoretical framework in the taxonomy.

In the following sections, we analyze each methodology in depth, illustrating its core principles and highlighting opportunities for advancements in neural-symbolic integration.

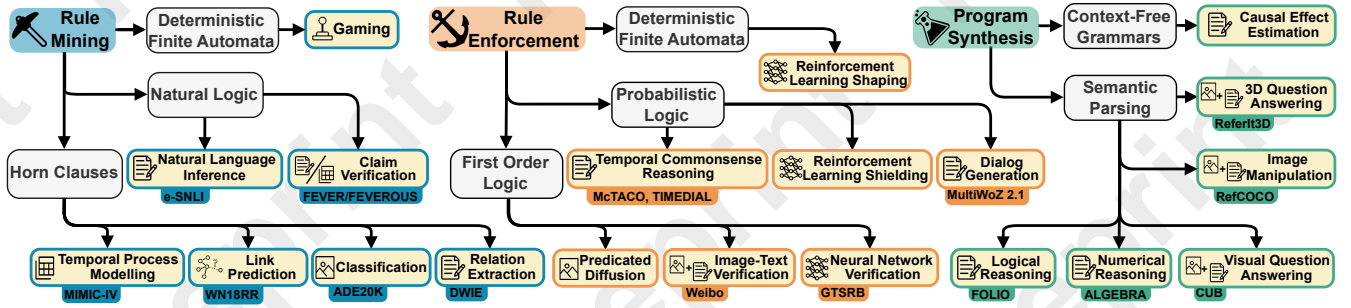[1]https://github.com/TimDettmers/ConvE

Figure 2: Task-Directed NeSy Taxonomy. We organize the most relevant task families by classifying each according to applicable NeSy techniques and grouping them into three macro-categories: (1) Rule Mining, (2) Rule Enforcement, and (3) Program Synthesis. We also label the most commonly used datasets to highlight their role in real-world benchmarking and model evaluation. Tasks are organized from bottom to top and from left to right within each category, mirroring the survey's conceptual progression.

3 Rule-Mining techniques ⚒

Neuro-Symbolic techniques employ rule-mining as a core methodology for model construction, focusing on extracting interpretable rules from diverse input data. The computational complexity of rule extraction directly correlates with the chosen formal specification language. Horn clause approaches exhibit the highest complexity, requiring matrix representations for ground atom truth values. Natural Logic frameworks show moderate complexity, combining linguistic preprocessing with question-answering for entity and operator extraction. Deterministic Finite Automata (DFAs) present the lowest complexity, leveraging operational traces with SAT-solving for rule learning.

3.1 Horn Clauses

Horn Clauses (HCs), a fundamental component of logic programming, provide a structured framework for rule-based reasoning. Defined as disjunctions of literals with at most one positive literal as the clause head, HCs enable logical inference across various domains.

Temporal point process modeling relies on HC-based reasoning to capture event dependencies, which is particularly crucial in fields such as medicine and autonomous driving, where sudden changes in conditions can have severe consequences. [Yang *et al.*, 2024] introduced a NeSy rule induction method that leverages a sequential covering algorithm and a custom attention mechanism to extract HCs. While demonstrating scalability, trustworthiness, and strong performance over existing models, its ability to capture increasingly complex rules remains an open research question.

The Inductive Logic Programming (ILP) paradigm automates the discovery of HC-based rules, supporting generalized reasoning and inference over structured data. In the context of extracting rules from structured data like tables and graphs, ILP seeks to identify missing connections between entities by uncovering patterns in existing relationships. DFORL [Gao *et al.*, 2024], a recent lightweight method for efficient rule extraction, introduces a depth-limited breadth-first search for neighborhood extraction. This propositionalization technique converts relational facts into vector representations suitable for neural network-based learning and applies syntactic constraints to reduce the rule search space.

By integrating auxiliary matrices and curriculum learning, DFORL uncovers hidden predicates and enhances the efficiency and progression of the training process. Although this approach demonstrates potential in real-world applications such as drug design [King *et al.*, 1995], its performance on the WN18RR link prediction benchmark remains inconsistent. This variability primarily stems from the dataset's design, which omits inverse relations in the test set, complicating direct comparisons with baseline methods and suggesting that methods like DFORL could benefit from incorporating negation in rule bodies to enhance the model's ability to handle higher-arity predicates. NCRL, a complementary approach by [Cheng *et al.*, 2023], focuses on rule compositionality, sampling alternative paths between connected nodes to construct Horn clauses. This method employs an iterative algorithm combining an RNN-based selection process with an attention mechanism, aiming to maximize the likelihood that head predicates can be reconstructed from sampled path predicates. By leveraging predicate embeddings, NCRL outperforms [Gao *et al.*, 2024] on WN18RR, mainly due to its higher ability to infer semantic relationships when inverse relations are missing. Its strong performance in low-data regimes and rapid convergence on GPU makes NCRL a compelling alternative for the same task.

Applied to tabular data, Inductive Logic Programming foundations are leveraged in Fold-SE [Wang and Gupta, 2024] to extract interpretable logical rules for supervised tabular classification. Designed to handle both categorical and numerical data with minimal preprocessing, it scales efficiently to large datasets while preserving rule coherence. Following a sequential covering strategy, Fold-SE iteratively refines rules by optimizing a modified Gini Impurity metric to enhance classification performance. Built upon Fold-SE, NeSyFOLD [Padalkar *et al.*, 2024] adopts ILP on extracted high level features for an image classification task. NeSyFOLD derives rules from binary activation masks produced by Convolutional Neural Networks (CNNs), assigning semantic labels to kernels post hoc. Its reliance on CNN architectures —extended and outperformed by modern transformer-based models— constrains its effectiveness, and the exclusion of exception predicates in semantic labeling further restricts its representational capacity.

In the textual domain, Document-level Relation Extraction (DocRE) is a supervised learning task that identifies relations $r(h, t)$ between entities, where r represents the relation, h is the head entity, and t is the tail entity. Both NeSy models and black-box approaches have been evaluated on DWIE, a widely recognized benchmark for DocRE. JMLR, a novel framework introduced in [Qi *et al.*, 2024], integrates DocRE with rule extraction through residual connections. This approach computes rule support as a weighted sum of atom supports, facilitating the implementation of soft-proof mechanisms that combine relations into HCs. [Jain *et al.*, 2024] adopted an alternative formulation, treating DocRE as a knowledge base link prediction task and utilizing the black-box model DocRE-CLiP to infer relational structures. FedNSL, introduced in [Xing *et al.*, 2024], extends NeSy methods to a federated learning setting for DocRE, achieving performance between that of DocRE-CLiP and JMLR. While its evaluation of real-world datasets remains limited to DWIE, the framework demonstrates the potential for privacy-preserving algorithms. By leveraging variational expectation maximization, it efficiently constrains the rule search space, reducing training rounds while maintaining competitive performance. A comparative evaluation of these methods —as shown in Table 1— indicates that JMLR significantly outperforms DocRE-CLiP on DWIE while achieving similar results on other benchmarks, highlighting the advantages of rule-based reasoning in capturing complex multi-sentence relationships. However, JMLR does not explicitly incorporate external knowledge, a feature more naturally integrated into knowledge-base-driven methods.

3.2 Natural Logic

Natural Language Inference (NLI) determines entailment relations between a premise and a hypothesis using formal operators such as equivalence ($\equiv$) and forward entailment ($\sqsubset$). [Feng *et al.*, 2022] proposed a transformer-based NeSy model that integrated GPT-2 with reinforcement learning for operator composition and introspective revision using WordNet. This approach outperforms baselines like BERT on most benchmarks but remains sensitive to linguistic noise, such as adverbs and prepositional phrase modifications in benchmark datasets. QA-NatVer [Aly *et al.*, 2023] extended this framework to claim verification, constructing proofs through sentence alignment, operator assignment, and a custom DFA. As shown in Table 1, QA-NatVer highlights the trade-off between NeSy models' explainability and performance, with a notable accuracy drop due to its reliance on a small training set. Its black-box counterpart, SFAVEL [Bazaga *et al.*, 2024], overcomes this limitation using self-supervision and distillation to remove the need for labeled data. TabVer [Aly and Vlachos, 2024] further extended NLI to tabular claim verification by incorporating numerical reasoning, recognizing equivalences, alongside pragmatic reasoning through deterministic rules. While achieving strong results on FEVER-OUS datasets, it struggles with exact numerical matching, underscoring ongoing challenges in adapting NeSy methods to less rigid reasoning tasks.

3.3 Deterministic Finite Automaton

Deep Q-learning has proven effective in reinforcement learning (RL), particularly in tasks that require sequential decision-making and structured exploration. Traditional black-box methods struggle with environments like the Atari classic game "Montezuma's Revenge", where success depends on complex interactions between objects and precise room navigation. A Deterministic Finite Automaton provides a structured way to model sequential dependencies by representing state transitions through a finite set of rules. This formalism can be leveraged to guide exploration and improve decision-making in complex RL environments. Leveraging this formalism, [Hasanbeig *et al.*, 2024] proposes a deep Q-learning framework augmented with DFA synthesis to enhance exploration efficiency and policy optimization. By structuring learned behaviors as state transitions, this approach enables more effective reasoning over sequential dependencies. Within this framework, the RL algorithm generates exploration traces that capture sequences of state-action pairs alongside reward estimates. These traces inform a DFA synthesis module, which formulates logical constraints and employs a SAT solver to construct a minimal-state DFA encoding the agent's learned behavior. By leveraging this structured representation, the deep learning module refines policy transitions, optimizing state-action mappings and iteratively improving decision-making within the RL process. Experiments demonstrate that the framework achieves faster convergence in tasks requiring sequential planning and object interactions, outperforming conventional models that fail to converge. Integrating expert-designed DFAs further enhances efficiency by eliminating the initial exploration phase and expediting training. Future research may extend this approach beyond gaming and investigate using more expressive automata, such as Pushdown Automata, to model hierarchical memory structures [Sipser, 1997].

4 Rule-Enforcement Techniques ✄

This section presents promising methodologies for enforcing constraints in the form of rules over systems for various application purposes. These constraints can be enforced through tailored networks or by regularization.

4.1 First-Order-Logic

Integrating First-Order Logic (FOL) rules into constrained image generation has been explored to enhance controllability and trustworthiness. [Sueyoshi and Matsubara, 2024] introduced a pioneering approach that extracts FOL constraints from text using dependency parsing and translates them into equations governing the intensity of attention maps. Rather than mining logical constraints from large datasets, this method employs a regularization strategy —a well-established practice in neuro-symbolic reasoning— as demonstrated by later works in this survey [Cai *et al.*, 2022; Pryor *et al.*, 2023]. In this approach, the logical structure is not inferred from data but explicitly imposed through a specialized loss function that guides the generation process. Notably, this method natively supports logical quantifiers, which

Natural Language Inference

The kid / does not love / table tennis
≡ / ≡ / ⊑
The kid / does not like / sports

Semantic Parsing

Subjects: All rectangles have four-sides
Predicates: All four-sided things are shapes
Objects: All rectangles are shapes

Horn Clauses / First Order Logic

All rectangles have four sides ⟹ FourSided(x) ← Rectangle(x)

/

All rectangles have four sides ⟹ ∀x FourSided(x) ← Rectangle(x)

Probabilistic Logic

I'm almost sure that all rectangles have four sides

⟱

0.9: FourSided(x) ← Rectangle(x)

Deterministic Finite Automaton

start → (👤) — run → (🚶)
idle

Context Free Grammars

Well formed parentheses: $\{\lambda, (), (()), ..., (^n)^n\}$
R1: S→(S)   Terminal symbols: {"(",")",λ}
R2: S→λ   Non terminal symbols: {S}
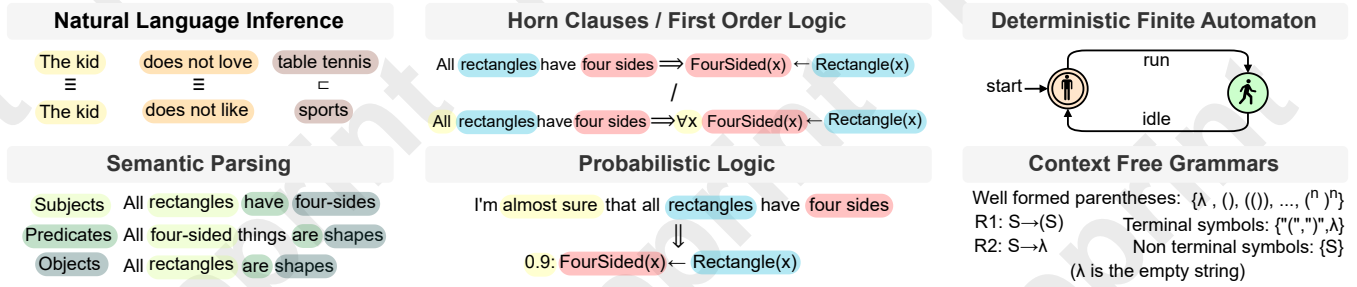(λ is the empty string)

Figure 3: Overview of the intermediate formal languages employed by each taxonomy method.

allow for formulating more precise and interpretable constraints. Qualitative analyses suggest that incorporating FOL constraints in image generation significantly enhances content fidelity, making this approach highly relevant for applications like chatbots and automated content creation. However, the absence of a counting mechanism prevents constraint enforcement on repeated entities (e.g., "a black dog and a white dog"), while a bias toward prototypical examples limits flexibility. Visual concept learning and semantic parsing could be incorporated to address these challenges [Mao *et al.*, 2019].

Reasoning over images is essential for detecting multimodal misinformation, where deception stems from the interplay of text and visuals. While traditional black-box approaches have achieved high accuracy in this domain, the lack of interpretability remains a major limitation. [Dong *et al.*, 2024] introduced a NeSy pipeline that enhances transparency by integrating logical reasoning into the classification process. Using two encoder models, the model extracts patterns from both modalities and employs a teacher-student network to estimate probabilities for three latent variables: image manipulation, cross-modal inconsistency, and image repurposing. A unique FOL rule classifies a sample as fake if any of the three variables hold true. Table 1 shows this method trailing its black-box counterpart by a few points, likely due to the more advanced vision models used by the competitor.

Ensuring neural network reliability is a major challenge, especially in fact-checking [Bussotti *et al.*, 2024], and media forensics, where errors can have serious consequences. Formal verification methods guarantee model behavior under specific conditions, ensuring compliance with predefined properties. [Xie *et al.*, 2022] introduced a structured verification framework based on the Neuro-Symbolic Assertion Language (NeSAL), a fragment of FOL designed for expressing and verifying neural network properties. NeSAL formalizes relationships between the inputs and outputs of a neural network under verification (NUV) and those of a specification network, which acts as a binary classifier to assess compliance with a given property. For instance, in an image classification task, the specification network outputs "true" whenever the NUV correctly identifies a specific class, such as the digit "2" in MNIST. This framework enables the formal specification and verification of key properties, such as correct classification of specific classes, stability of high-confidence predictions under certain data distributions, and functional equivalence between different architectures within a defined

margin of error. While NeSAL does not provide quantitative measures for violations, its qualitative evaluations of model confidence under distributional shifts demonstrate its potential for debugging and improving neural networks. This approach enhances model reliability by providing a structured verification method, making it valuable in high-risk tasks.

4.2 Probabilistic Logic

Probabilistic Logic (PL) integrates probability theory with formal logic to model uncertainty, enabling structured reasoning in noisy or incomplete data environments. SLEER [Cai *et al.*, 2022] addresses reporting bias in temporal event data expressed in text, where uncommon events are overrepresented compared to routine occurrences (e.g., "It took me an hour to get out of bed" versus "It takes me 15 seconds to get out of bed"). This imbalance poses challenges for temporal commonsense reasoning tasks, which rely on multiple-choice question formats derived from textual benchmarks. SLEER aims to improve model robustness in handling temporal information by incorporating structured reasoning principles. A language encoder supports multi-task learning, where classification heads predict specific temporal dimensions, each optimized using cross-entropy loss. The total loss function integrates these individual losses with a regularization term derived from predefined probabilistic soft logic (PSL) rules. Implemented with t-norms, these rules enforce logical consistency by connecting the outputs of the classification heads. Unlike traditional PSL models that use weighted rules, this approach relies on carefully crafted constraints for effectiveness. Despite its innovative design, the approach yields only marginal improvements on the McTACO [Zhou *et al.*, 2019] benchmark[2], underperforming on other multi-task encoders such as [Pereira *et al.*, 2020]. Expanding on this work, [Cai *et al.*, 2023] introduced LECTER, which integrates a context encoder, a logic induction module for temporal dependencies, and a DeepProbLog-based logic validator incorporating human knowledge. Optimized through regression and logic entailment losses, LECTER demonstrated strong zero-shot performance on the TIMEDIAL [Qin *et al.*, 2021] benchmark. As shown in Table 1, LECTER outperforms GPT-3.5 [Brown *et al.*, 2020] in 2-best accuracy, which checks if the top two ranked answers are correct, while GPT-3.5 is benchmarked with standard accuracy. However, its evaluation remains limited to TIMEDIAL, leaving its generalizability unverified.

---

[2]https://leaderboard.allenai.org/mctaco/submissions/public

Generating interpretable dialogues presents a significant challenge that can be addressed through reasoning over external KBs. [Yang *et al.*, 2022] introduced NS-Dial, a model that integrates external knowledge bases into the dialogue generation process to enhance interpretability. The framework processes input dialogues through an encoder, which extracts relevant features before a hypothesis generator formulates an initial assumption. A hierarchical reasoning engine then identifies supporting triples within the KB, constructing a structured reasoning chain represented as a single weighted HC. This inferred knowledge guides the decoder, which generates the next token based on belief scores derived from the reasoning process. Although this approach improves interpretability, black-box methods demonstrate higher performance in dialogue generation—as shown in Table 1. In the same context, [Pryor *et al.*, 2023] proposed a NeSy model to identify dialogue states and transitions in goal-oriented conversations (Dialog Structure Induction), especially in low-resource settings. The model combines a Variational Recurrent Neural Network (VRNN) with a symbolic loss based on PSL to enforce structural consistency in sequence prediction. Though effective in domain generalization and adaptation, the approach is surpassed by a simpler VRNN baseline as labeled data increases, revealing potential complexity limitations.

PL, like FOL, provides a framework for ensuring network safety. [Yang *et al.*, 2023] proposed a RL policy that enforces action safety through structured probabilistic reasoning. The policy uses annotated disjunctions for action probabilities, probabilistic facts for environmental conditions, and logic clauses for safety specifications, enabling queries on shielded policy probabilities and action safety metrics. The approach employs a PL-based strategy, where a shielded policy gradient optimizes rewards while enforcing safety constraints in situationally aware environments, and a safety gradient enhances action reliability in physically constrained settings like those affected by inertia. Empirical evaluations indicate that the collaboration between these components is essential to avoid performance degradation. With RL playing an increasing role in language modeling, we argue that integrating these techniques with instruction tuning may enhance response accuracy and safety in future real-world applications.

4.3 Deterministic Finite Automaton

[Jiang *et al.*, 2021] introduced a temporal-logic-based reward shaping technique to enhance Q-learning in scenarios where the behavior policy deviates from the optimal policy. Their approach incorporates a potential-based reward term derived from a DFA to encode temporal logic constraints, such as "following a human" or "staying in a corridor". In this framework, the agent iteratively refines a Q-table to estimate state-action values, typically updating it at episode endpoints. By integrating these rules into the reward function, the method encourages alignment with desired behaviors, accelerating convergence to the optimal policy without enforcing specific actions. Simulations across various tasks demonstrate that reward shaping is more effective than shielding in exploration-intensive scenarios, while shielding performs better when constraints precisely match task requirements.

5 Program Synthesis Techniques 🖋

Program Synthesis (PS) techniques generate structured programs in either general-purpose programming languages or domain-specific languages (DSLs) to guide system prediction or behavior. This approach shares similarities with methodologies such as NLI syntax, where words and operations are combined to generate a structured representation for each sample. However, PS introduces significant complexity, as generating a valid program often depends on semantic parsing or search algorithms like A*. Its greater expressive power distinguishes it from formal rule-based methods, which are typically constrained by the removal of quantifiers and face challenges in integrating domains such as mathematics.

5.1 Context-Free Grammars

PS has been applied to causal effect estimation, where traditional models often rely on strong assumptions that limit their flexibility. [Reddy and Balasubramanian, 2024] introduced NESTER, a model using a DSL defined by a context-free grammar. An A* algorithm with heuristic guidance selects grammar rules iteratively, generating differentiable programs acting as inductive biases. The DSL supports operations such as binary selection, addition, multiplication, and subset computation but excludes recursion to maintain simplicity. While evaluations showed competitive results on small benchmarks, scalability to more diverse datasets remains uncertain.

5.2 Semantic Parsing

LINC [Olausson *et al.*, 2023] approaches logical reasoning through semantic parsing, translating natural language inputs into FOL clauses processed by the Prover9 solver. To enhance robustness, the method employs majority voting across multiple generated programs. Comparisons with alternative reasoning paradigms, including naive prediction, scratchpad reasoning, and Chain-Of-Thought, show superior performance on synthetic datasets and competitive results on expert benchmarks. A key limitation of this approach is the occurrence of semantic parsing errors, which could be addressed by integrating context-free grammars and consistency-checking mechanisms to improve reliability. Unlike SATLM [Ye *et al.*, 2023], which encodes problems into SAT formulations and demonstrates strong benchmark performance, LINC relies on FOL-based reasoning, offering a distinct approach to explainability compared to SAT's declarative framework.

Beyond textual reasoning, NeSy methods have been explored for visual tasks, with NS-CL [Mao *et al.*, 2019] serving as a foundational model. By integrating a visual encoder, semantic parser, and program executor, it achieves near-perfect generalization on CLEVR [Johnson *et al.*, 2017] and strong performance on real-world datasets like VQS [Gan *et al.*, 2017], requiring only 10% of the training data. [Stammer *et al.*, 2021] extended this approach by incorporating user feedback and explanation loss, improving robustness against confounding factors through symbolic explanations and Slot Attention mechanisms. Building on NS-CL, FALCON [Mei *et al.*, 2022] introduced meta-learning concepts via hyperboxes, enhancing visual encoders for few-shot learning and biased settings, with competitive results on CLEVR, CUB

Method	Task	Benchmark	Evaluation Metric	Score	Black-Box Competitor	Δ Score
JMLR [2024]	⚒ Document Lv. Relation Extraction	DWIE [2021]	ACCURACY	77.9%	DocRE-CLiP [2024]	+10.8%
QA-NatVer [2023]	⚒ Claim Verification	FEVER (Dev) [2018]	ACCURACY	70.3%	SFAVEL [2024]	−20.0%
NSLM [2024]	⟆ Image-Text Verification	Weibo [2017]	F1-MACRO	84.4%	SAFE + Hami-m³d [2024]	−3.7%
SLEER [2022]	⟆ Temporal Commonsense Reasoning	McTACO [2019]	F1-MACRO	69.0%	ALICE [2020]	−10.5%
LECTER [2023]	⟆ Temporal Commonsense Reasoning	TIMEDIAL [2021]	ACCURACY	71.5%	GPT-3.5 [2020]	+6.5%*
NS-Dial [2022]	⟆ Goal-oriented Dialogue Generation	MultiWoZ 2.1 [2020]	BLEU	10.6	GALAXY [2022]	−9.4
NS3D [2023]	⟐ Visual Question-Answering	Sr3D [2020]	ACCURACY	67.0%	GPS [2024]	−10.5%

⚒ = Rule Mining Task; ⟆ = Rule Enforcement Task; ⟐ = Program Synthesis Task.
∗ LECTER [2023] is evaluated using the 2-best accuracy.

Table 1: NeSy Competitiveness. The Δ **Score** field reports NeSy methods' performance gain over the black-box competitors.

[Wah *et al.*, 2011], and GQA [Hudson and Manning, 2019]. NEUROSIM [Singh *et al.*, 2023] further explored image manipulation by integrating scene graph-based operations and tailored loss functions, though its real-world applicability remains unclear. GENOME [Chen *et al.*, 2024] shifted the focus to modular incremental learning, generating Python-based components for visual reasoning, image manipulation, and knowledge tagging. While outperforming [Singh *et al.*, 2023] in real-world tasks, its comparison with [Mei *et al.*, 2022] remains limited. In continual learning, COOL [Marconato *et al.*, 2023] tackled catastrophic forgetting, surpassing regularization-based approaches on CLEVR. NS3D [Hsu *et al.*, 2023], adopting the NS-CL formalism in the 3D setting, demonstrates strong generalization and zero-shot reasoning. However, as shown in Table 1, its performance on the ReferIt3D (SR3D) [Achlioptas *et al.*, 2020] benchmark falls short of the black-box GPS model [Jia *et al.*, 2024].

6 Discussion

Finally, we address open problems, NeSy application challenges, and comparisons with black-box methods.

Competing with Black-Box Methods NeSy models excel in tasks requiring explicit rule enforcement and structured reasoning, but often struggle in open-domain settings where black-box architectures benefit from large-scale, unstructured data. Table 1 highlights these differences, showing that performance inconsistencies across similar benchmarks make some NeSy solutions highly sensitive to their application frameworks. The Temporal Commonsense Reasoning task exemplifies how score discrepancies may arise from differing benchmarking strategies. While TIMEDIAL focuses on dialogue masking prediction, McTACO is designed for multiple-choice Question-Answering, resulting in varied performance and further highlighting the sensitivity of NeSy solutions to specific application frameworks. Performance drops are particularly evident in tasks such as claim verification on FEVER [Thorne *et al.*, 2018] and visual question-answering on Sr3D [Achlioptas *et al.*, 2020], where black-box models excel in leveraging semantically-reach unlabeled data. A key insight from QA-NatVer [Aly *et al.*, 2023] is the trade-off between interpretability and data efficiency. Multi-granular chunking and step-by-step scoring improve explainability but limit generalization versus black-box methods, creating a balance between symbolic clarity and neural adaptability as NeSy models vary in efficiency and expressiveness.

Future Applications As research in NeSy models evolves, certain areas appear to be approaching saturation. The extensive body of work following NS-CL [Mao *et al.*, 2019] suggests that image-based reasoning has reached a point of diminishing returns. This evolution demonstrates that well-structured synthetic benchmarks reflecting real-world scenarios can drive field-wide advances (e.g., CLEVR's impact on spatial reasoning). While reinforcement learning gains traction in robotics and NLP, its high-stakes applications remain limited. Enhancing trustworthiness through structured constraints with NLI—integrating domain knowledge via DFAs, shaping rewards to enforce desired behaviors, and using shielding to prevent unsafe actions—could enable deployment in critical domains like surgery, medical Q&A and gene discovery [Frisoni *et al.*, 2024; Domeniconi *et al.*, 2016], autonomous driving, and legal analysis. In this regard, the robust NLI methods proposed in [Feng *et al.*, 2022] would benefit from the rigorous evaluation on benchmarks featuring complex sentence structures with multi-phrase hypotheses, premises, and conclusions, while systematically analyzing alternation, negation, and independence operators—particularly valuable for legal text analysis. We argue that the design pattern introduced by NeSyFOLD [Padalkar *et al.*, 2024], which treats high-level neural features as semi-structured data for rule mining and partial interpretability, warrants further exploration. Investigating its application to models like transformers, autoencoders, and GNNs could enable explainability in black-box models.

Acknowledgments

Contribution Statement

All authors contributed equally to this work. For more details, visit https://disi-unibo-nlp.github.io.

References

[Achlioptas *et al.*, 2020] Panos Achlioptas, Ahmed Abdelreheem, et al. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, volume 12346. Springer, 2020.

[Aly and Vlachos, 2024] Rami Aly and Andreas Vlachos. Tabver: Tabular fact verification with natural logic. *TACL*, 12, 2024.

[Aly *et al.*, 2023] Rami Aly, Marek Strong, and Andreas Vlachos. Qa-natver: Question answering for natural logic-based fact verification. In *EMNLP*. ACL, 2023.

[Bach, 2017] Francis R. Bach. Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.*, 18, 2017.

[Bazaga *et al.*, 2024] Adrián Bazaga, Pietro Lio, and Gos Micklem. Unsupervised pretraining for fact verification by language model distillation. In *ICLR*, 2024.

[Besold *et al.*, 2021] Tarek R. Besold, Artur S. d'Avila Garcez, et al. Neural-symbolic learning and reasoning: A survey and interpretation. In *Neu.-Sym. AI: The State of the Art*, volume 342. IOS Press, 2021.

[Bronstein *et al.*, 2021] Michael M. Bronstein, Joan Bruna, et al. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, et al. Language models are few-shot learners. In Hugo Larochelle et al., editors, *NeurIPS*, 2020.

[Bussotti *et al.*, 2024] Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. In Yaser Al-Onaizan et al., editor, *EMNLP*, pages 12105–12122. ACL, 2024.

[Cai *et al.*, 2022] Bibo Cai, Xiao Ding, et al. Mitigating reporting bias in semi-supervised temporal commonsense inference with probabilistic soft logic. In *AAAI*, 2022.

[Cai *et al.*, 2023] Bibo Cai, Xiao Ding, et al. Self-supervised logic induction for explainable fuzzy temporal commonsense reasoning. In *AAAI*, 2023.

[Chen *et al.*, 2024] Zhenfang Chen, Rui Sun, et al. GENOME: generative neuro-symbolic visual reasoning by growing and reusing modules. In *ICLR*, 2024.

[Cheng *et al.*, 2023] Kewei Cheng, Nesreen K. Ahmed, and Yizhou Sun. Neural compositional rule learning for knowledge graph reasoning. In *ICLR*, 2023.

[Domeniconi *et al.*, 2015] Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, and Roberto Pasolini. Markov chain based method for in-domain and cross-domain sentiment classification. In Ana L. N. Fred, Jan L. G. Dietz, David Aveiro, Kecheng Liu, and Joaquim Filipe, editors, *KDIR 2015*, pages 127–137. SciTePress, 2015.

[Domeniconi *et al.*, 2016] Giacomo Domeniconi, Marco Masseroli, Gianluca Moro, and Pietro Pinoli. Cross-organism learning method to discover new gene functionalities. *Comput. Methods Programs Biomed.*, 126:20–34, 2016.

[Dong *et al.*, 2024] Yiqi Dong, Dongxiao He, et al. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In Michael J. Wooldridge et al., editors, *AAAI*, 2024.

[Eric *et al.*, 2020] Mihail Eric, Rahul Goel, et al. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In Nicoletta Calzolari et al., editors, *LREC*. European Language Resources Association, 2020.

[Feng *et al.*, 2022] Yufei Feng, Xiaoyu Yang, et al. Neuro-symbolic natural logic with introspective revision for natural language inference. *TACL*, 10, 2022.

[Frisoni *et al.*, 2024] Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In Lun-Wei Ku et al., editor, *ACL*, pages 9878–9919. ACL, 2024.

[Gan *et al.*, 2017] Chuang Gan, Yandong Li, et al. VQS: linking segmentations to questions and answers for supervised attention in VQA and question-focused semantic segmentation. In *ICCV*. IEEE Computer Society, 2017.

[Gao *et al.*, 2024] Kun Gao, Katsumi Inoue, et al. A differentiable first-order rule learner for inductive logic programming. *Artif. Intell.*, 331, 2024.

[Giunchiglia *et al.*, 2022] Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. Deep learning with logical constraints. In *IJCAI*. ijcai.org, 2022.

[Hasanbeig *et al.*, 2024] Hosein Hasanbeig, Natasha Yogananda Jeppu, et al. Symbolic task inference in deep reinforcement learning. *J. Artif. Intell. Res.*, 80, 2024.

[He *et al.*, 2022] Wanwei He, Yinpei Dai, et al. GALAXY: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *AAAI*, 2022.

[Hsu *et al.*, 2023] Joy Hsu, Jiayuan Mao, and Jiajun Wu. NS3D: neuro-symbolic grounding of 3d objects and relations. In *CVPR*. IEEE, 2023.

[Hudson and Manning, 2019] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*. Computer Vision Foundation / IEEE, 2019.

[Jain *et al.*, 2024] Monika Jain, Raghava Mutharaju, et al. Revisiting document-level relation extraction with context-guided link prediction. In Michael J. Wooldridge et al., editors, *AAAI*, 2024.

[Jia *et al.*, 2024] Baoxiong Jia, Yixin Chen, et al. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In Ales Leonardis et al., editors, *ECCV*, volume 15067. Springer, 2024.

[Jiang *et al.*, 2021] Yuqian Jiang, Suda Bharadwaj, et al. Temporal-logic-based reward shaping for continuing reinforcement learning tasks. In *AAAI*, 2021.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *ACM on MM*. ACM, 2017.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, et al. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. IEEE Computer Society, 2017.

[King *et al.*, 1995] Ross D. King, Michael J. E. Sternberg, and Ashwin Srinivasan. Relating chemical activity to structure: An examination of ILP successes. *New Gener. Comput.*, 13(3&4), 1995.

[Lodi *et al.*, 2010] Stefano Lodi, Gianluca Moro, and Claudio Sartori. Distributed data clustering in multi-dimensional peer-to-peer networks. In Heng Tao Shen and Athman Bouguettaya, editors, *(ADC 2010*, volume 104, pages 171–178. Australian Computer Society, 2010.

[Mao *et al.*, 2019] Jiayuan Mao, Chuang Gan, et al. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*, 2019.

[Marconato *et al.*, 2023] Emanuele Marconato, Gianpaolo Bontempo, et al. Neuro-symbolic continual learning: Knowledge, reasoning shortcuts and concept rehearsal. In *ICML*, volume 202. PMLR, 2023.

[Marcus, 2018] Gary Marcus. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631, 2018.

[Mei *et al.*, 2022] Lingjie Mei, Jiayuan Mao, et al. FALCON: fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. In *ICLR*, 2022.

[Olausson *et al.*, 2023] Theo Olausson, Alex Gu, et al. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *EMNLP*. ACL, 2023.

[Padalkar *et al.*, 2024] Parth Padalkar, Huaduo Wang, and Gopal Gupta. Nesyfold: A framework for interpretable image classification. In Michael J. Wooldridge et al., editors, *AAAI*, 2024.

[Pereira *et al.*, 2020] Lis Pereira, Xiaodong Liu, et al. Adversarial training for commonsense inference. In Spandana Gella et al., editors, *RepL4NLP@ACL*. ACL, 2020.

[Pryor *et al.*, 2023] Connor Pryor, Quan Yuan, et al. Using domain knowledge to guide dialog structure induction via neural probabilistic soft logic. In *ACL*. ACL, 2023.

[Qi *et al.*, 2024] Kunxun Qi, Jianfeng Du, and Hai Wan. End-to-end learning of logical rules for enhancing document-level relation extraction. In Lun-Wei Ku et al., editors, *ACL*. ACL, 2024.

[Qin *et al.*, 2021] Lianhui Qin, Aditya Gupta, et al. TIME-DIAL: temporal commonsense reasoning in dialog. In Chengqing Zong et al., editors, *IJCNLP*. ACL, 2021.

[Reddy and Balasubramanian, 2024] Abbavaram Gowtham Reddy and Vineeth N. Balasubramanian. NESTER: an adaptive neurosymbolic method for causal effect estimation. In Michael J. Wooldridge et al., editors, *AAAI*, 2024.

[Singh *et al.*, 2023] Harman Singh, Poorva Garg, et al. Image manipulation via multi-hop instructions - A new dataset and weakly-supervised neuro-symbolic approach. In Houda Bouamor et al., editors, *EMNLP*. ACL, 2023.

[Sipser, 1997] Michael Sipser. *Introduction to the theory of computation*. PWS Publishing Company, 1997.

[Stammer *et al.*, 2021] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *CVPR*. Computer Vision Foundation / IEEE, 2021.

[Sueyoshi and Matsubara, 2024] Kota Sueyoshi and Takashi Matsubara. Predicated diffusion: Predicate logic-based attention guidance for text-to-image diffusion models. In *CVPR*. IEEE, 2024.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, et al. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*. ACL, 2018.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang and Gupta, 2024] Huaduo Wang and Gopal Gupta. FOLD-SE: an efficient rule-based machine learning algorithm with scalable explainability. In Martin Gebser and Ilya Sergey, editors, *PADL*, volume 14512. Springer, 2024.

[Wang *et al.*, 2024] Bing Wang, Shengsheng Wang, et al. Harmfully manipulated images matter in multimodal misinformation detection. In Jianfei Cai et al., editors, *ACM Multimedia, 2024*. ACM, 2024.

[Xie *et al.*, 2022] Xuan Xie, Kristian Kersting, and Daniel Neider. Neuro-symbolic verification of deep neural networks. In *IJCAI*. ijcai.org, 2022.

[Xing *et al.*, 2024] Pengwei Xing, Songtao Lu, and Han Yu. Federated neuro-symbolic learning. In *ICML*, 2024.

[Yang *et al.*, 2022] Shiquan Yang, Rui Zhang, et al. An interpretable neuro-symbolic reasoning framework for task-oriented dialogue generation. In *ACL*. ACL, 2022.

[Yang *et al.*, 2023] Wen-Chi Yang, Giuseppe Marra, et al. Safe reinforcement learning via probabilistic logic shields. In *IJCAI*. ijcai.org, 2023.

[Yang *et al.*, 2024] Yang Yang, Chao Yang, et al. Neuro-symbolic temporal point processes. In *ICML*, 2024.

[Ye *et al.*, 2023] Xi Ye, Qiaochu Chen, et al. Satlm: Satisfiability-aided language models using declarative prompting. In Alice Oh et al., editors, *NeurIPS*, 2023.

[Zaporojets *et al.*, 2021] Klim Zaporojets, Johannes Deleu, et al. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4), 2021.

[Zhou *et al.*, 2019] Ben Zhou, Daniel Khashabi, et al. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Kentaro Inui et al., editors, *EMNLP*. ACL, 2019.