Benchmarking Neuro-Symbolic Description Logic Reasoners: Existing Challenges and A Way Forward

Gunjan Singh [a,*], Riccardo Tommasini [b], Sumit Bhatia [c] and Raghava Mutharaju [a]
[a] *Knowledgeable Computing and Reasoning Lab, IIIT-Delhi, Delhi, India*
E-mails: gunjans@iiitd.ac.in, raghava.mutharaju@iiiitd.ac.in
[b] *LIRIS Lab, INSA, Lyon, France*
E-mail: riccardo.tommasini@insa-lyon.fr
[c] *Media and Data Science Research Lab, Adobe Inc., Delhi, India*
E-mail: sumit.bhatia@adobe.com

Abstract. Recently, there has been significant progress in the development of robust and highly scalable neuro-symbolic description logic reasoners. However, the field faces challenges arising from diverse design strategies and evaluation methods. We address the latter challenge by emphasizing the critical requirement for a comprehensive benchmark framework tailored to the unique evaluation needs of neuro-symbolic description logic reasoners. In this paper, we address barriers that must be overcome to facilitate the effective evaluation of these reasoners and outline a potential methodology for designing the benchmark framework. This work contributes towards a more systematic and principled evaluation framework for neuro-symbolic reasoning, highlighting the broader role of benchmarks in advancing the field.

Keywords: Benchmark, Neuro-Symbolic AI, Reasoning, Description Logics, Ontology, Neural Network

1. Introduction

Neuro-symbolic Artificial Intelligence (AI) [1, 2] is a promising field that aims to bridge the gap between traditional symbolic logic and modern neural network-based machine learning. The idea is to combine the strengths of both approaches while overcoming their weaknesses. The focus of this paper lies within the realm of neuro-symbolic reasoning. At its core, neuro-symbolic reasoning involves integrating symbolic reasoning, which relies on structured logic and formal knowledge representation, with neural network-based methods known for their capacity to process large-scale, unstructured data and learn complex patterns from it. This fusion holds the potential for developing systems with enhanced performance, explainability, and generalization abilities [3]. It's important to note that these approaches, unlike traditional reasoning methods, are not necessarily sound and complete. Instead, they strike a balance between approximating the precise reasoning capabilities of symbolic systems and harnessing the robust learning capabilities of machine learning techniques.

However, progress in this field faces significant challenges because neuro-symbolic reasoning is emerging, in contrast to other areas with extensive research and well-established benchmarks. For instance, several models (Graph

*Corresponding author. E-mail: gunjans@iiitd.ac.in.

Neural Networks (GNN) [4], Logic Tensor Networks [5]), methodologies (Inductive Logic Programming [6]) and innovative ideas (explainable AI [7], zero-shot learning [8]) enrich this field. As a result, existing works in this field exhibit diversity in techniques, and hence, different methods and criteria are used to evaluate the performance of neuro-symbolic reasoning systems (see Table 1). The lack of a standardized approach makes it difficult to compare these systems and make progress in the field. Furthermore, based on the reciprocal relationships between neural and symbolic components and how they benefit each other, neuro-symbolic reasoning systems, and in general neuro-symbolic AI systems, as discussed by Henry Kautz, can be categorized into one of the six distinct categories [9].

- Symbolic Neuro Symbolic: In this category, the input and output are represented symbolically, such as with words or sequences of words. These symbols are converted into vectors using methods like word2vec [10] and then fed into a neural network for processing.
- Symbolic[Neuro]: Symbolic solvers use neural models internally for some functions, as seen in systems like AlphaGo [11].
- Neuro|Symbolic: This category involves a refined integration of neural and symbolic approaches, where both systems collaborate to enhance specific tasks, such as in the case of Neuro-Symbolic Concept-Learner [12].
- Neuro:Symbolic $\rightarrow$ Neuro: These approaches take symbolic rules as input and compile them during training, effectively integrating symbolic knowledge into the structure of neural models, as demonstrated in Deep Learning For Symbolic Mathematics [13].
- $Neuro_{Symbolic}$: This category involves transforming symbolic rules into templates for structures within the neural network, such as Logic Tensor Network [14].
- Neuro[Symbolic]: Refers to the embedding of symbolic reasoning inside a neural engine, such as Graph Neural Networks (GNN) [4].

Each of these categories represents a unique approach to neuro-symbolic AI, adding an extra layer of diversity to the advancements in this field.

Drawing inspiration from Jim Gray's pioneering work [15] on domain-specific benchmarks for databases, our goal is to tackle the challenge of benchmarking neuro-symbolic reasoners. The primary purpose of such a benchmark is two-fold. Firstly, it serves as a tool to identify the performance bottlenecks, enabling targeted improvements in the systems where algorithms are still evolving. Secondly, benchmarks facilitate meaningful comparisons between various systems, offering insights into their relative strengths and weaknesses. While this paper does not put forth an alternative benchmark, we highlight the strong need for such benchmarks, including their features, and explain why they are essential for moving the field forward.

In Section 2, we delve into the recent advancements in neuro-symbolic reasoning, highlighting the challenges in evaluating and comparing the existing state-of-the-art neuro-symbolic reasoners. Subsequently, in Section 3, we address the barriers that must be overcome to facilitate the effective evaluation of neuro-symbolic reasoners. Finally, in Section 4, we outline a potential methodology for designing the benchmark.

2. Neuro-Symbolic Reasoning for Description Logics

In recent years, there have been significant advancements in developing neuro-symbolic reasoners for description logics (DLs) [16], a formal underpinning for the Web Ontology Language (OWL 2) [17]. While most of these works predominantly focus on classification and consistency checking [18–20], the other reasoning tasks, such as instance retrieval, query rewriting, materialization, abduction, and explanation generation, remain relatively unexplored. The intricacy of these tasks varies significantly, and delving into their complexities offers a promising avenue for further exploration.

Research in this domain takes an alternative approach to traditional reasoning tasks such as classification and consistency, breaking them into class subsumption, class membership, and satisfiability tasks. Various techniques are employed, such as geometric embeddings [21–24] that map ontological relationships to geometric spaces and emulating logical reasoning through machine learning [25–27]. A comprehensive overview and detailed insights into the state-of-the-art neuro-symbolic reasoning landscape are discussed in [19, 20]. Regarding other categories, a limited amount of work, such as that for e-commerce search [28], merges neuro-symbolic reasoning with query

rewriting. This involves a Knowledge Graph (KG) [29] enhanced neural network approach that integrates auxiliary knowledge from a product Knowledge Graph, enhancing semantic understanding of user queries and improving query reformulation.

The existing traditional benchmarks such as LUBM (Lehigh University Benchmark) [30], UOBM (University Ontology Benchmark) [31], and OWL2Bench [32] lack suitability for evaluating neuro-symbolic reasoners due to their narrow focus on conventional reasoning tasks. Traditional evaluations of reasoning systems often rely on metrics such as reasoning time, which may not align well with the evaluation requirements of neuro-symbolic reasoners. Although the ontologies of these benchmarks, along with those from the OWL Reasoner Evaluation (ORE) Competition [33], can serve as initial datasets for the proposed neuro-symbolic benchmark framework, these datasets fall short of addressing the distinct challenges posed by neuro-symbolic reasoning. To our knowledge, no benchmarks or evaluation frameworks have been designed to evaluate and compare neuro-symbolic reasoning systems. Most reasoner evaluations are performed on different publicly available ontologies, including but not restricted to SNOMED CT[1], Gene Ontology (GO)[2], and Galen[3], as well as other ontologies available in public repositories such as DBpedia [34], YAGO [35], Wikidata [36], Claros[4], NCBO Bioportal[5], and AgroPortal[6]. However, these offer a limited set of ontologies for evaluation, which does not cover the full spectrum of possible scenarios.

As discussed in Section 1, neuro-symbolic approaches encompass a range of evaluation methodologies and reasoning techniques. This diversity becomes evident in Table 1, highlighting the necessity for a dedicated benchmark to systematically and comprehensively assess the performance of neuro-symbolic reasoning systems. The table reveals the utilization of subsets of description logics, such as $\mathcal{ALC}$ and $\mathcal{EL}^{++}$, and various OWL 2 profiles like EL and RL [37]. Some works also incorporate RDF and RDFS into their reasoning techniques, underlining the diversity in the supported ontology languages and profiles, which implies that existing works handle different levels of complexity. Furthermore, the table showcases the variety of reasoning tasks undertaken, different datasets utilized, and the diverse metrics employed for evaluating each approach. The summary column in Table 1 highlights the differences in techniques used by each work. It is important to note that the paper does not aim to provide an exhaustive list of all the existing work. Instead, it emphasizes the variations in reasoning and evaluation approaches. The collective representation highlights the pressing need for a standardized benchmark to facilitate fair and consistent comparisons, thereby advancing the progress of neuro-symbolic reasoning research. The table reveals that similar works may differ significantly by employing distinct metrics and datasets to evaluate their contributions. For instance, consider the works of Makni et al. [27] and Ebrahimi et al. [26]. Both studies focus on RDFS entailment reasoning, aiming to replicate deductive reasoning processes. However, they adopt different metrics and datasets to assess the effectiveness and performance of their approaches. Such variations in evaluation criteria can lead to diverse insights and perspectives on the contributions within the field.

Paper	**Logic**	**Reasoning Task**	**Datasets Used**	**Metrics**	**Summary of Approaches Used**
ELEm [21]	$\mathcal{EL}^{++}$	Subsumption	GO	Hits@n, AUC, Mean Rank	To capture entity relationships, embeddings were created by representing Concepts as n-balls and the relations as translation vectors between the centers of each Concept ball. The embeddings were utilized to predict protein-protein interactions.

[1] https://bioportal.bioontology.org/ontologies/SNOMEDCT
[2] https://bioportal.bioontology.org/ontologies/GO
[3] https://bioportal.bioontology.org/ontologies/GALEN
[4] https://www.clarosnet.org
[5] https://bioportal.bioontology.org/
[6] http://agroportal.lirmm.fr/

EmEL^{++} [22]	$\mathcal{EL}^{++}$	Subsumption	SNOMED CT, Anatomy, GO, Galen	Hits@n, AUC, Median Rank, 90^{th} percentile rank	Extended ELEm with relation inclusion and role chains. Also introduced negative samples for training.
EmEL-V [23]	$\mathcal{EL}^{++}$	Subsumption	SNOMED CT, GO, Galen	Top@n, Median Rank, 90^{th} Percentile Rank	Extended EmEL^{++} to include many-to-many relationships
BoxEL [24]	$\mathcal{EL}^{++}$	Subsumption	Anatomy, GO, Galen	Hits@n, AUC, Mean Rank	To capture entity relationships, mapped concepts as boxes and deals with the limitations of n-ball [21–23] based embeddings.
Box^2EL [38]	$\mathcal{EL}^{++}$	Subsumption, Role Assertion and Deductive Reasoning	Anatomy, GO, Galen	Hits@n, AUC, Median, Mean Rank	Maps both concepts and roles as boxes, and models inter-concept relationships using a bumping mechanism.
Özçep et al. [39]	$\mathcal{ALC}$	Concept Membership	NA	NA	Embeds Concepts in the ontology as convex regions in vector spaces.
E2R [40]	$\mathcal{ALC}$	Concept Membership	LUBM	Hits@n, Mean Rank, MRR	Aiming to preserve the logical structure, proposed embeddings in the quantum space.
Makni and Hendler [27]	RDFS	Entailment Reasoning	LUBM and Scientist dataset created from DBpedia	Precision, Recall, and F1 Score	The evaluation focused on assessing noise tolerance by employing an encoder-decoder architecture to translate input RDF graph embeddings into corresponding inference graph embeddings.
Ebrahimi et al. [41]	RDFS	Query-based Classification	Created from Linked Data Cloud and Data Hub websites	Precision, Recall, and F1 score	Explored the capabilities of end-2-end memory networks. The model's capability for multi-hop reasoning is demonstrated. The use of normalized embeddings support transfer.
Ebrahimi et al. [26]	RDFS and $\mathcal{EL}^{+}$	Entailment Reasoning	Synthetic Data and LUBM	Exact Matching Accuracy	Utilized pointer networks for learning the sequential application of inference rules used in many deductive reasoning algorithms.
Hohenecker and Lukasiewicz [42]	OWL 2 RL	Entailment Reasoning	Claros, DBpedia, UMLS, and Synthetic Data	Accuracy	Developed a deep learning-based model called Recursive Reasoning Networks (RNN).

Eberhart et al. [25]	$\mathcal{EL}^+$	Ontology Completion (concept inclusions and existential restrictions)	Synthetic Data and SNOMED	Precision, Recall, and F1 Score	Showcases completion reasoning behavior using various LSTM neural networks to learn reasoning patterns, employing three distance measures to assess prediction accuracy.
Makni et al. [43]	RDFS	Explainable Entailment Reasoning	LUBM and real-world scholarly dataset	Accuracy	Built upon the previous work [27] for generating explanations for the derived conclusions by taking the RDF graph and inferred triples as input and the explanations as the target.
Hohenecker and Lukasiewicz [44]	RDF	Concept Membership and Relation Prediction	LUBM, UOBM, Claros, DBpedia	F1 Score and Accuracy	Proposed Relational Tensor Network (RTN). Embeddings of the individuals are computed by applying RTNs on the Directed Acyclic Graph representation of the ontology (including the inferences).
Farzana et al. [28]	RDF	Query pruning and complete query rewriting	Created from user search logs from eBay Inc.	Precision, Recall, and F Score, and Query Accuracy	Proposes a Knowledge Graph (KG) enhanced approach for query rewriting in e-commerce, leveraging RDF2Vec entity embeddings, entity types, category information, and entity frequency extracted from a product KG.
OWL2Vec* [45]	$\mathcal{SROIQ}$	Concept membership and concept subsumption	HeLis, FoodOn, GO	MRR and Hits@n	Ontologies are transformed into RDF graphs, and Word2Vec is utilized on the resulting paths. The training dataset comprises three documents: structural, lexical, and a combination of both, enhancing entity interrelation understanding compared to earlier Word2Vec methodologies.

Table 1: Overview of Variations in Neuro-Symbolic Reasoning and Evaluation Approaches

To further highlight the diversity in the current approaches, we classify the works mentioned in Table 1 into one of the six distinct categories discussed in Section 1. [45] involves converting the symbolic input, ontologies, and RDF graphs, to vectors (Symbolic Neuro Symbolic). [25–27, 41, 43, 44] take symbolic reasoning rules as input and compile them during training (Neuro:Symbolic] $\rightarrow$ Neuro), integrating symbolic knowledge into neural models. [21–24, 38–40] embed symbolic reasoning inside neural engines, representing symbolic information in geometric or vector spaces and employing neural methods for reasoning tasks (Neuro[Symbolic]). [28] falls into the category involving a refined integration of neural and symbolic approaches to enhance query rewriting (Neuro|Symbolic).

3. Desiderata for Benchmarking Neuro-Symbolic Reasoners

Creating an effective benchmark demands careful consideration of critical principles such as simplicity for accessibility, portability for impartial assessments across various approaches, scalability to accommodate diverse system sizes, and relevance to reflect practical challenges in benchmark scenarios [15]. However, the evaluation of neuro-symbolic reasoners presents its own set of distinctive challenges. Given the field's novelty, state-of-the-art solutions do not approach such challenges systematically. Therefore, we advocate below the issues that should be prioritized in constructing a fair neuro-symbolic reasoning benchmark.

1. **Diverse benchmark scenarios**
   To effectively evaluate neuro-symbolic reasoners, the benchmark must incorporate diverse scenarios that mirror the complexity and variety encountered in real-world applications. This approach ensures a thorough assessment of the reasoners' capabilities across different contexts. Key aspects to consider include:
   - *Variety of Ontologies*: The benchmark should encompass a range of ontologies differing in size, profile, and axiom types. This includes:
     * *Size and Complexity*: Include ontologies with varying sizes and complexities in both assertional knowledge (ABox) and terminological knowledge (TBox) to evaluate how reasoners handle different levels of detail and scope.
     * *OWL 2 Profiles*: Use ontologies that adhere to various OWL 2 profiles (such as EL, QL, RL, and DL) to test the reasoners' ability to handle different levels of expressiveness.
     * *Axiom Types*: Incorporate different types of axioms (such as subclass relations and property restrictions) and their combinations to assess how well the reasoners manage diverse logical constructs.
   - *Specific and Generic Reasoning Tasks*: Benchmark scenarios should include both specific reasoning tasks and generic information needs:
     * *Specific Reasoning Tasks*: Design tasks that test particular reasoning capabilities, such as classification, consistency checking, and instance retrieval. These tasks enable micro-benchmarking and provide insights into the strengths and limitations of individual reasoners.
     * *Generic Information Needs*: Include tasks that assess the reasoners' ability to handle complex and broader reasoning scenarios, such as evaluating how well the reasoners can address multi-step queries that involve integrating diverse information sources and applying both symbolic rules and neural network-derived insights. This includes testing the reasoners' ability to synthesize and leverage contextual information to generate coherent and relevant responses.
   - *Real-World Applicability*: Ensure that benchmark scenarios reflect real-world use cases:
     * *Real-World Ontologies*: Analyze and incorporate real-world ontologies and existing benchmarks to capture practical challenges and scenarios.
     * *Scalability and Realism*: Design scenarios that not only address current requirements but also scale beyond them to foster technological advancement and future-proof the evaluation process.
2. **Introducing controlled inconsistencies**
   Incorporating controlled inconsistencies into benchmark design presents a significant challenge but is essential for evaluating the robustness of neuro-symbolic reasoners. Controlled inconsistencies should be introduced in a deterministic manner to assess how well the systems handle and resolve contradictions. Key aspects to consider include:
   - *Types of Inconsistencies*:
     * *Structural Inconsistency*: Introduce structural inconsistencies by creating contradictions in the ontological hierarchy or relationships. For example: If the ontology specifies that the entities 'Male' and 'Female' are disjoint classes, add instances in the ABox that are classified as both 'Male' and 'Female'. This tests the system's ability to detect and resolve structural conflicts. Similarly, create inconsistencies by defining contradictory class hierarchies or property restrictions that violate the logical constraints of the ontology.

  - ∗ *Semantic Inconsistency*: Introduce semantic inconsistencies by modifying entity names or attributes to introduce ambiguity or slight deviations. For example: Change names or attributes in a way that creates near-identical but distinct instances, such as altering "John" to "Jonh" to test how well the system identifies and resolves semantic conflicts. Introduce synonyms or typographical errors that may lead to semantic ambiguities and test how the system manages these issues.
- *Reproducibility and Control*:
  - ∗ *Deterministic Generation*: Ensure that the process of generating inconsistencies is deterministic, allowing for consistent reproduction of test scenarios. This is crucial for evaluating the effectiveness of the system's handling of inconsistencies.
  - ∗ *Controlled Environment*: Design mechanisms to introduce inconsistencies in a controlled manner, avoiding randomness that could obscure the evaluation of specific reasoning capabilities.

Note that existing benchmarks may lack the capability to introduce generic inconsistencies effectively or in a contextually relevant manner. This highlights the need for novel approaches to benchmark design. Traditional generative AI models, such as Large Language Models, may not be well-suited for creating controlled inconsistencies. This underscores the unique requirements for benchmark design that effectively simulates real-world contradictions. Incorporating controlled inconsistencies into the benchmark will provide a deeper understanding of a reasoner's robustness and its ability to manage and resolve conflicts, reflecting the complexity of real-world scenarios where inconsistencies are prevalent.

3. **Input representation for benchmarking**

A critical aspect of benchmarking neuro-symbolic reasoners is the representation of input data. This involves ensuring that ontological knowledge, both ABox (assertional knowledge) and TBox (terminological knowledge), is formatted in a manner that various reasoning systems can effectively process. This flexibility ensures comprehensive and realistic evaluation conditions, enabling the assessment of reasoning systems across the spectrum of neuro-symbolic methodologies.

- *Ontology Formats*: The benchmark should support multiple ontology formats such as RDF/XML[7], Turtle[8], and Manchester OWL Syntax[9], among others. This ensures compatibility with a wide range of systems that may require specific formats.
- *Axiom Format*: While some neuro-symbolic systems may utilize embedding techniques to transform ontological entities and relationships into continuous vector spaces, others might directly process axioms in their logical form. For instance, certain systems might require axioms in a normalized form as per the $\mathcal{EL}^{++}$ profile. Additionally, some approaches may require axioms to be in triple format. Therefore, the benchmark should accommodate these varying requirements by providing tools for transforming and normalizing ontological data as needed.
- *Pre-embedded Entities*: Some systems may necessitate entities represented as embeddings, using models such as BERT [46] or other neural embeddings such as TransE [47]. The benchmark should offer pre-embedded entity representations, ensuring compatibility with these methods and enabling comprehensive evaluation across different representation techniques.
- *Dataset Splits*: The benchmark should facilitate the generation of dataset splits tailored to diverse testing needs, such as train-test-validation splits. This enables a thorough evaluation of a system's learning and generalization capabilities across different segments of data. Properly managed splits ensure that the performance metrics accurately reflect the system's ability to handle unseen data and prevent overfitting.
- *Domain Agnostic Datasets*: To assess a system's understanding of logical semantics independently of specific domain knowledge, the benchmark should have the capability to generate domain-agnostic datasets. This allows for evaluation focused on the system's ability to interpret and apply logical rules universally rather than relying on domain-specific information.

[7] https://www.w3.org/TR/rdf-syntax-grammar/
[8] https://www.w3.org/TR/turtle/
[9] https://www.w3.org/2007/OWL/wiki/ManchesterSyntax

4. **Assessment of the deductive capabilities of existing approaches**
In the trajectory towards developing a new generation of reasoners that effectively harness the potential of both neural networks and logical reasoning, a foundational requirement involves conducting an equitable assessment of state-of-the-art solutions. This assessment provides insights into the present capabilities of these approaches and illuminates the trajectory of the field's future development. Evaluating these aspects ensures that the reasoners can generalize beyond specific datasets and apply logical rules consistently across different domains. The key points of this assessment include:

   - *Soundness and Completeness*: Traditional deductive reasoners are sound and complete, meaning they produce correct and exhaustive inferences based on given axioms. Evaluating whether neuro-symbolic reasoners can achieve similar standards is critical.
   - *Generalization Capabilities*: Deductive reasoners should be able to work across any ontology from any domain. This includes verifying that the reasoners can generalize logical rules universally and not be confined to specific datasets or domains. For instance, a rule stating "if A is a subclass of B and B is a subclass of C, then A is a subclass of C" should apply universally, irrespective of the specific terms involved. This ensures the systems can apply logical rules consistently across different domains.
   - *Scalability and Efficiency*: Assessing the scalability and efficiency of these reasoners in handling large and complex ontologies is essential. Traditional deductive reasoners are designed to handle extensive datasets and intricate logical structures, which serve as a benchmark for emerging neuro-symbolic systems. Understanding how these models perform under varying degrees of complexity and scale can guide the development of more robust and versatile reasoning systems.
   - *Handling Noise and Inconsistent Data*: When evaluating the deductive capabilities of these reasoners, it is crucial to consider how well they handle noise and inconsistent data. Real-world applications often involve datasets with inaccuracies, ambiguities, and inconsistencies that can challenge the reasoning process. Assessing a system's ability to manage and mitigate the impact of such issues provides valuable insights into its robustness and practical applicability. This includes evaluating how well the system maintains soundness and completeness in the presence of noisy or conflicting data and its effectiveness in adapting to varying degrees of data quality and integrity.

   When assessing deductive reasoning capabilities and comparing them with conventional deductive reasoners, it is advantageous to also include neural-based approaches, such as large language models, in the evaluation framework. While neural methods may not always excel in every aspect of deductive reasoning, incorporating them as a baseline can offer valuable comparative insights. This approach not only underscores the benefits of neuro-symbolic methods, which integrate both neural and symbolic reasoning, but also provides a more comprehensive understanding of the strengths and potential synergies between different reasoning paradigms.
5. **Success metrics and key performance indicators**
In order to accurately measure the performance of neuro-symbolic reasoners, the benchmark must support a range of metrics and key performance indicators (KPIs) that capture various aspects of system performance. Standard metrics commonly used in evaluation include:

   - *Accuracy, Precision, Recall, and F1-Score*: These metrics provide insights into the classification performance of neural components, assessing how well the system identifies correct versus incorrect predictions. That is, giving insights into how accurately the system produces only correct inferences (soundness) and whether it generates all possible correct inferences based on the given axioms (completeness).
   - *Mean Reciprocal Rank (MRR) and Hits@K*: These are used to evaluate ranking tasks, measuring the position of correct answers in a ranked list of predictions. They help assess the effectiveness of the system in retrieving relevant information.
   - *Scalability Metrics*: Metrics such as processing time and memory usage that evaluate how well the system handles large and complex datasets.

   While these standard metrics are essential for evaluating traditional aspects of system performance, there remains a need for developing new metrics tailored to the unique characteristics of neuro-symbolic reasoning. Current benchmarks might not fully capture critical aspects such as:

- *Robustness to Noise and Inconsistent Data*: Evaluating how well the system manages inaccuracies, ambiguities, and inconsistencies in real-world data. This requires metrics that measure the impact of noisy or conflicting data on performance and the system's ability to maintain robustness.
- *Inference Generation Efficiency*: Metrics that assess the system's capability to generate all inferences in a single run, while ensuring system soundness, measuring computational efficiency and the number of iterations required.
- *Explanatory Capabilities*: Evaluating the quality and usefulness of explanations provided by the system for its inferences. This includes measuring the clarity and completeness of explanations, which is crucial for transparency and user trust.
- *Generalization Across Domains*: Metrics to assess how well the system transfers reasoning capabilities across different domains, ensuring consistent performance and applicability in varied contexts.
- *Embedding Quality*: For systems that use embeddings, metrics evaluate how well embeddings capture logical relationships and nuances. This includes assessing the embeddings' effectiveness in preserving logical structures and supporting accurate inferences.

The inclusion of these novel metrics, alongside traditional ones, ensures a comprehensive evaluation of neuro-symbolic reasoners. This approach provides deeper insights into the performance and limitations of current systems, guiding future improvements and research directions.

6. **Adaptability**
In the rapidly evolving field of neuro-symbolic reasoning, the benchmark's adaptability is crucial for ensuring its relevance and effectiveness. The benchmark should be designed to accommodate the following aspects:

- *Continuous Updates*: The benchmark should be capable of integrating new tasks and methodologies as they emerge. This involves regularly updating the benchmark to reflect the latest advancements and challenges in neuro-symbolic reasoning.
- *Ongoing Review and Feedback*: Regular reviews and updates based on the latest research and feedback from the community are essential to keep the benchmark aligned with current practices and real-world needs.

4. Proposed Methodology for Designing the Benchmark

In this section, we propose one of the possible methodologies to design a benchmark comprising of the objectives outlined in Section 3. The methodology involves the following key steps:

1. *Generating Diverse Benchmark Scenarios*: The initial step towards creating a benchmark involves curating datasets that cover a wide range of benchmarking scenarios. One approach is to start with a study of existing neuro-symbolic description logic reasoners, beginning with basic ontology profiles like RDFS and gradually progressing to more complex ones such as OWL 2 EL and OWL 2 DL. This process includes evaluating existing datasets across various models and systems, comparing results with traditional reasoning systems like Konclude [48], and identifying performance variations. This analysis can provide insights about the datasets that prove critical for existing systems. We can then focus on generating synthetic datasets that replicate these patterns at various scales and complexities, ensuring coverage of diverse ontology constructs and reasoning tasks.
2. *Introduction of Controlled Inconsistencies*: After generating the datasets, the next step is to introduce controlled inconsistencies, similar to those discussed in desiderata 2 of Section 3. This approach allows for the evaluation of how effectively the system handles and resolves these inconsistencies.
3. *Input formats*: One essential benchmarking feature could be support for various OWL and RDF serialization formats, such as RDF/XML [10] or OWL/XML [11]. This capability would allow the tool to handle ontologies in any input format and convert them into the required format, facilitating seamless integration and testing.

[10] https://www.w3.org/TR/rdf-syntax-grammar/
[11] https://www.w3.org/TR/owl-xmlsyntax/

Additionally, incorporating options for generating different dataset splits and profile-specific features, like generating axioms in normal form for the $\mathcal{EL}++$ profile, can further enhance the tool's versatility and effectiveness in benchmarking neuro-symbolic reasoning systems.

4. *Evaluation of deductive capabilities*: Traditional reasoners often struggle with inconsistent ontologies, underlining the importance of starting with consistent ontologies. Therefore, generating inconsistencies is kept as a separate step in the benchmarking process. Simultaneously, it's crucial to evaluate the features attributed to the neural aspect of the system, such as learning capabilities, repair abilities, and scalability. This involves assessing performances based on handling ontologic1al complexities, scalability, and overall performance compared to traditional reasoning systems. Evaluating performances after introducing controlled inconsistencies highlights the benefits and results obtained in the presence of inconsistencies, emphasizing deductive prowess and unique contributions of the neural aspect in neuro-symbolic reasoning, showcasing the system's overall capabilities comprehensively.
5. *Metric Design*: The existing standard learning metrics, such as accuracy, precision, and F1 score, only provide an overall idea of the efficacy of the systems. However, these systems need a thorough analysis, emphasizing areas well-supported by systems and areas needing improvement, such as handling different ontological constructs. These metrics should encompass not only deductive capabilities but also adaptability to diverse scenarios and overall efficacy in handling complex neuro-symbolic reasoning tasks.

This methodology outlines a foundational approach to benchmark design that can be adapted and expanded to include more expressive profiles. It provides a systematic starting point for addressing the challenges in neuro-symbolic reasoning, with the flexibility to evolve and incorporate additional complexity and features as the field progresses.

5. Conclusion

We highlighted the significant need for a comprehensive benchmark framework to tackle the challenges tied to evaluating neuro-symbolic description logic reasoning systems. Merging symbolic logic and neural network-based machine learning brings great promise, but the lack of common evaluation methods has held back progress in the field. By underlining the importance of creating benchmarks, our aim for the future is to establish a structured way of evaluating these systems that can drive the field forward.

Acknowledgement

Gunjan Singh and Raghava Mutharaju would like to acknowledge the partial support of the Infosys Centre for Artificial Intelligence (CAI), IIIT-Delhi, in this work.

References

[1] A.S. d'Avila Garcez, T.R. Besold, L.D. Raedt, P. Földiák, P. Hitzler, T. Icard, K. Kühnberger, L.C. Lamb, R. Miikkulainen and D.L. Silver, Neural-Symbolic Learning and Reasoning: Contributions and Challenges, in: *2015 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 22-25, 2015*, AAAI Press, 2015. http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10281.

[2] M.K. Sarker, L. Zhou, A. Eberhart and P. Hitzler, Neuro-Symbolic Artificial Intelligence, *AI Communications* **34**(3) (2021), 197–209.

[3] J. Ott, A. Ledaguenel, C. Hudelot and M. Hartwig, How to Think About Benchmarking Neurosymbolic AI?, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, A.S. d'Avila Garcez, T.R. Besold, M. Gori and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3432, CEUR-WS.org, 2023, pp. 248–254. https://ceur-ws.org/Vol-3432/paper22.pdf.

[4] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner and G. Monfardini, The Graph Neural Network Model, *IEEE Transactions on Neural Networks* **20**(1) (2009), 61–80. doi:10.1109/TNN.2008.2005605.

[5] S. Badreddine, A.S. d'Avila Garcez, L. Serafini and M. Spranger, Logic Tensor Networks, *Artif. Intell.* **303** (2022), 103649. doi:10.1016/j.artint.2021.103649.

[6] P. Sen, B.W.S.R. de Carvalho, R. Riegel and A.G. Gray, Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 8212–8219. https://ojs.aaai.org/index.php/AAAI/article/view/20795.

[7] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao and J. Zhu, Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges, in: *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, J. Tang, M. Kan, D. Zhao, S. Li and H. Zan, eds, Lecture Notes in Computer Science, Vol. 11839, Springer, 2019, pp. 563–574. doi:10.1007/978-3-030-32236-6_51.

[8] J. Chen, F. Lécué, Y. Geng, J.Z. Pan and H. Chen, Ontology-guided Semantic Composition for Zero-shot Learning, in: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020*, D. Calvanese, E. Erdem and M. Thielscher, eds, 2020, pp. 850–854. doi:10.24963/kr.2020/87.

[9] H.A. Kautz, The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture, *AI Magazine* **43**(1) (2022), 93–104. doi:10.1609/aimag.v43i1.19122.

[10] T. Mikolov, K. Chen, G.S. Corrado and J. Dean, Efficient Estimation of Word Representations in Vector Space, in: *International Conference on Learning Representations*, 2013. https://api.semanticscholar.org/CorpusID:5959482.

[11] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T.P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, Mastering the game of Go with deep neural networks and tree search, *Nat.* **529**(7587) (2016), 484–489. doi:10.1038/nature16961.

[12] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum and J. Wu, The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision, in: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. https://openreview.net/forum?id=rJgMlhRctm.

[13] G. Lample and F. Charton, Deep Learning For Symbolic Mathematics, in: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. https://openreview.net/forum?id=S1eZYeHFDS.

[14] S. Badreddine, A. d'Avila Garcez, L. Serafini and M. Spranger, Logic Tensor Networks, *Artificial Intelligence* **303** (2022), 103649. doi:https://doi.org/10.1016/j.artint.2021.103649. https://www.sciencedirect.com/science/article/pii/S0004370221002009.

[15] J. Gray (ed.), *The Benchmark Handbook for Database and Transaction Systems (2nd Edition)*, Morgan Kaufmann, 1993. ISBN 1-55860-292-5.

[16] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi and P.F. Patel-Schneider (eds), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2003. ISBN 0-521-78176-0.

[17] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P.F. Patel-Schneider and U. Sattler, OWL 2: The next step for OWL, *Journal of Web Semantics* **6**(4) (2008), 309–322. doi:10.1016/j.websem.2008.05.001.

[18] P. Hitzler, M. Krötzsch and S. Rudolph, *Foundations of Semantic Web Technologies*, Chapman and Hall/CRC Press, 2010. ISBN 9781420090505. http://www.semantic-web-book.org/.

[19] B. Makni, M. Ebrahimi, D. Gromann and A. Eberhart, Neuro-Symbolic Semantic Reasoning, in: *Neuro-Symbolic Artificial Intelligence: The State of the Art*, P. Hitzler and M.K. Sarker, eds, Frontiers in Artificial Intelligence and Applications, Vol. 342, IOS Press, 2021, pp. 253–279. ISBN 978-1-64368-244-0. doi:10.3233/FAIA210358.

[20] G. Singh, S. Bhatia and R. Mutharaju, Neuro-Symbolic RDF and Description Logic Reasoners: The State-Of-The-Art and Challenges, in: *Compendium of Neurosymbolic Artificial Intelligence*, P. Hitzler and M.K. Sarker, eds, Frontiers in Artificial Intelligence and Applications, Vol. 369, IOS Press, 2023, pp. 29–63. ISBN 978-1-64368-407-9. doi:10.3233/FAIA230134.

[21] M. Kulmanov, W. Liu-Wei, Y. Yan and R. Hoehndorf, EL Embeddings: Geometric Construction of Models for the Description Logic $\mathcal{EL}^{++}$, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, ed., ijcai.org, 2019, pp. 6103–6109. doi:10.24963/ijcai.2019/845.

[22] S. Mondal, S. Bhatia and R. Mutharaju, EmEL^{++}: Embeddings for $\mathcal{EL}^{++}$ Description Logic, in: *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021*, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, Vol. 2846, CEUR-WS.org, 2021. http://ceur-ws.org/Vol-2846/paper19.pdf.

[23] B. Mohapatra, S. Bhatia, R. Mutharaju and G. Srinivasaraghavan, Why Settle for Just One? Extending $\mathcal{EL}^{++}$ Ontology Embeddings with Many-to-Many Relationships, *CoRR* **abs/2110.10555** (2021). https://arxiv.org/abs/2110.10555.

[24] B. Xiong, N. Potyka, T. Tran, M. Nayyeri and S. Staab, Faithful Embeddings for $\mathcal{EL}^{++}$ Knowledge Bases, in: *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, U. Sattler, A. Hogan, C.M. Keet, V. Presutti, J.P.A. Almeida, H. Takeda, P. Monnin, G. Pirrò and C. d'Amato, eds, Lecture Notes in Computer Science, Vol. 13489, Springer, 2022, pp. 22–38. doi:10.1007/978-3-031-19433-7_2.

[25] A. Eberhart, M. Ebrahimi, L. Zhou, C. Shimizu and P. Hitzler, Completion Reasoning Emulation for the Description Logic $\mathcal{EL}^{+}$, in: *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020, Palo Alto, CA, USA, March 23-25, 2020, Volume I*, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, Vol. 2600, CEUR-WS.org, 2020. http://ceur-ws.org/Vol-2600/paper5.pdf.

[26] M. Ebrahimi, A. Eberhart and P. Hitzler, On the Capabilities of Pointer Networks for Deep Deductive Reasoning, *CoRR* **abs/2106.09225** (2021). https://arxiv.org/abs/2106.09225.

[27] B. Makni and J.A. Hendler, Deep learning for noise-tolerant RDFS reasoning, *Semantic Web* **10**(5) (2019), 823–862. doi:10.3233/SW-190363.

[28] S. Farzana, Q. Zhou and P. Ristoski, Knowledge Graph-Enhanced Neural Query Rewriting, in: *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, ACM, 2023, pp. 911–919. doi:10.1145/3543873.3587678.

[29] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs, *ACM Computing Surveys* **54**(4) (2022), 71:1–71:37. doi:10.1145/3447772.

[30] Y. Guo, Z. Pan and J. Heflin, LUBM: A Benchmark for OWL Knowledge Base Systems, *Journal of Web Semantics.* 3(2–3) (2005), 158–182.

[31] L. Ma, Y. Yang, G. Qiu Z .and Xie, Y. Pan and S. Liu, Towards a Complete OWL Ontology Benchmark, in: *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 125–139.

[32] G. Singh, S. Bhatia and R. Mutharaju, OWL2Bench: A Benchmark for OWL 2 Reasoners, in: *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, J.Z. Pan, V.A.M. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Lecture Notes in Computer Science, Vol. 12507, Springer, 2020, pp. 81–96. doi:10.1007/978-3-030-62466-8_6.

[33] B. Parsia, N. Matentzoglu, R.S. Gonçalves, B. Glimm and A. Steigmiller, The OWL Reasoner Evaluation (ORE) 2015 Competition Report, *Journal of Automated Reasoning* **59**(4) (2017), 455–482. doi:10.1007/s10817-017-9406-8.

[34] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer and C. Bizer, DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, *Semantic Web Journal* **6** (2014). doi:10.3233/SW-140134.

[35] F.M. Suchanek, G. Kasneci and G. Weikum, YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, in: *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 697–706–. ISBN 9781595936547. doi:10.1145/1242572.1242667.

[36] D. Vrandečić, Wikidata: A New Platform for Collaborative Data Collection, in: *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, Association for Computing Machinery, New York, NY, USA, 2012, pp. 1063–1064–. ISBN 9781450312301. doi:10.1145/2187980.2188242.

[37] B. Motik, B.C. Grau, I. Horrocks, Z. Wu, A. Fokoue and C. Lutz, OWL 2 Web Ontology Language Profiles (Second Edition), 2012. https://www.w3.org/TR/owl2-profiles/.

[38] M. Jackermeier, J. Chen and I. Horrocks, Dual box embeddings for the description logic EL++, Association for Computing Machinery, 2024.

[39] Ö.L. Özçep, M. Leemhuis and D. Wolter, Cone Semantics for Logics with Negation, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, ed., ijcai.org, 2020, pp. 1820–1826. doi:10.24963/ijcai.2020/252.

[40] D. Garg, S. Ikbal, S.K. Srivastava, H. Vishwakarma, H.P. Karanam and L.V. Subramaniam, Quantum Embedding of Knowledge for Reasoning, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 5595–5605. https://proceedings.neurips.cc/paper/2019/hash/cb12d7f933e7d102c52231bf62b8a678-Abstract.html.

[41] M. Ebrahimi, M.K. Sarker, F. Bianchi, N. Xie, D. Doran and P. Hitzler, Reasoning over RDF Knowledge Bases using Deep Learning, *CoRR* **abs/1811.04132** (2018). http://arxiv.org/abs/1811.04132.

[42] P. Hohenecker and T. Lukasiewicz, Ontology Reasoning with Deep Neural Networks, *Journal of Artificial Intelligence Research* **68** (2020), 503–540. doi:10.1613/jair.1.11661.

[43] B. Makni, I. Abdelaziz and J.A. Hendler, Explainable Deep RDFS Reasoner, *CoRR* **abs/2002.03514** (2020). https://arxiv.org/abs/2002.03514.

[44] P. Hohenecker and T. Lukasiewicz, Deep Learning for Ontology Reasoning, *CoRR* **abs/1705.10342** (2017). http://arxiv.org/abs/1705.10342.

[45] J. Chen, P. Hu, E. Jiménez-Ruiz, O.M. Holter, D. Antonyrajah and I. Horrocks, OWL2Vec*: Embedding of OWL Ontologies, *Machine Learning* **110**(7) (2021), 1813–1845. doi:10.1007/s10994-021-05997-6.

[46] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

[47] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 2787–2795–.

[48] A. Steigmiller, T. Liebig and B. Glimm, Konclude: System description, *Journal of Web Semantics.* 27 (2014), 78–85.