

# Reviewer 1

## Comment 1:

Kautz categorization of neurosymbolic frameworks is a large part of the paper and one of the main contributions. It would be good to clarify this in the title of the paper (now it only mentions evaluation strategies and dataset overview).

### Response:

- We have revised the title from “A Scoping Review of Neurosymbolic Reasoning over Ontologies and Knowledge Graphs: Datasets, Evaluation Practices, and Open Challenges” to “A Scoping Review of Neurosymbolic Reasoning over Ontologies and Knowledge Graphs: System Categorization, Datasets, Evaluation Practices, and Open Challenges.” We use the broader term “system categorization” because the revised manuscript complements Kautz’s high-level integration types with a more detailed characterization of the neural components, symbolic components, knowledge representations, and information flows of the reviewed systems.

## Comment 2:

In Section 'Introduction' or 'Background' more detailed information on knowledge graphs and ontologies should be added: what is a knowledge graph (formally), what is an ontology, what are Description Logics (syntax and semantics), ABox / TBox axioms, OWL / RDF, which tasks support ontology and knowledge graph reasoning (e.g., link prediction, subsumption prediction, query answering etc.), maybe brief mentioning of symbolic reasoners (ELK, HermiT etc.). Without proper introduction of all the terminology it would be difficult for a beginner to understand the rest of the paper. This also applies to datasets mentioned in the introductory part: what is FB15k, WN18 etc., which data do they contain?

### Response:

- We have substantially expanded the Introduction and Background sections. The revised manuscript now defines knowledge graphs, relational triples, ontologies, Description Logics, TBox, ABox and RBox axioms, RDF, OWL, and representative symbolic reasoners such as ELK and HermiT. We also introduce the main deductive and predictive tasks considered in the review, including classification, consistency checking, subsumption, instance checking, ontology completion, link prediction, knowledge graph completion, and logical query answering.
- We additionally provide introductory descriptions of frequently used benchmark families such as FB15k/FB15k-237, WN18/WN18RR, YAGO-derived benchmarks, LUBM, ORE, and OWL2Bench. These additions are intended to make the remainder of the review accessible without assuming extensive prior knowledge of Semantic Web or Description Logic terminology.

### Comment 3:

In Section 'Introduction' some prior surveys on neurosymbolic systems operating with knowledge graphs and ontologies are mentioned, yet this list is not complete. E.g. consider the works below:

- DeLong, Lauren Nicole, Ramon Fernández Mir, and Jacques D. Fleuriot. "Neurosymbolic AI for reasoning over knowledge graphs: A survey." IEEE Transactions on Neural Networks and Learning Systems 36.5 (2024): 7822-7842.
- Seeliger, Arne, Matthias Pfaff, and Helmut Krcmar. "Semantic web technologies for explainable machine learning models: A literature review." PROFILES/SEMEX@ ISWC 2465 (2019): 1-16.
- Steinmetz, Nadine, and Kai-Uwe Sattler. "What is in the KGQA benchmark datasets? Survey on challenges in datasets for question answering on knowledge graphs." Journal on Data Semantics 10.3 (2021): 241-265.
- Dai, Yuanfei, et al. "A survey on knowledge graph embedding: Approaches, applications and benchmarks." Electronics 9.5 (2020): 750.
- Zamazal, Ondřej. "A survey of ontology benchmarks for semantic web ontology tools." international Journal on Semantic Web and information Systems (iJSWiS) 16.1 (2020): 47-68.

It should be stated explicitly what novelty this survey brings compared to other similar works.

#### **Response:**

- We have expanded the related work discussion to include the surveys by DeLong et al., Seeliger et al., Steinmetz and Sattler, Dai et al., and Zamazal. The revised Introduction now distinguishes their respective scopes, including neurosymbolic reasoning over knowledge graphs, semantic technologies for explainable machine learning, knowledge graph question-answering benchmarks, knowledge graph embedding methods, and ontology benchmarks for Semantic Web tools.
- We have also made the novelty of the present review more explicit. In contrast to surveys focused on a single method family, task, or benchmark tradition, our review jointly analyzes: (1) the architectural organization of neurosymbolic systems operating over both knowledge graphs and ontologies; (2) the datasets used to assess them; (3) the mapping between reasoning tasks and evaluation metrics; and (4) limitations and recommendations for comprehensive neurosymbolic evaluation.

### Comment 4:

In Section 'Methods', subsection 'Stage 1: Identifying the research questions' RQ 1 should be more specific: the objective is not just to list the works with focus on reasoning over ontologies and knowledge graphs, yet to categorize selected works according to Kautz neurosymbolic taxonomy.

#### **Response:**

- We agree that the previous formulation of RQ1 was too broad and did not reflect the system-categorization objective of the review. We have revised RQ1 to ask which systems have been developed for predictive or deductive reasoning over ontologies and knowledge graphs and how their knowledge representations, neural components, symbolic components, and integration mechanisms can be characterized. This formulation makes the

architectural analysis explicit while avoiding the implication that the objective is only to enumerate existing systems.

#### Comment 5:

'Methods' ('Stage 3: Study Selection', 'Stage 4: Charting the data' and 'Stage 5: Collating, summarizing, and reporting the results' subsections) and the beginning of 'Results' sections (everything up until 'Neurosymbolic systems (RQ 1)') need to be reorganized since there are some parts that repeat each other. I would suggest to put everything from 'Results' until 'Neurosymbolic systems (RQ 1)' to the 'Methods' section and start 'Results' with Kautz categorization. I also suggest to move the entire 'Methods' section to the Appendix since this is not the most important part of the paper.

#### Response:

- We agree that the previous manuscript contained unnecessary repetition between the Methods section and the beginning of the Results section. We have reorganized these sections accordingly.
- The study-selection procedure, screening limits, eligibility criteria, numbers of records retained at each stage, and the PRISMA-ScR flow diagram are now presented together under Stage 3 of the Methods section. The benchmark-dataset identification and filtering procedure has been consolidated under Stage 4 together with the data-charting process. We removed the repeated "Literature Search" subsection and the preliminary list of notable benchmark families from the Results section, which now begins directly with the analysis of neurosymbolic systems under RQ1.
- We retained a concise Methods section in the main manuscript rather than moving it entirely to the Appendix because the search, screening, and charting procedures are central to the transparency and reproducibility of a scoping review.

'Stage 3: Study Selection' from 'Methods' does not include the information how many papers were selected at the end and after each stage of relevant work selection. This part can be merged with paragraphs 1-2 of 'Literature Search' from 'Results' which also reveal the process of article collection.

#### Response:

- We have merged the study-selection information previously reported at the beginning of the Results section into Stage 3 of the Methods section. Stage 3 now reports the number of records retrieved, the number remaining after duplicate removal, the number retained after title and abstract screening, and the final number of included studies.

'Stage 4: Charting the data' from 'Methods', paragraph 3 (about benchmark datasets) needs to be joined with paragraph 3 of 'Literature Search' from 'Results' which includes information about benchmark datasets as well.

#### Response:

- We have consolidated the benchmark-dataset identification and filtering procedure under Stage 4 of the Methods section. This section now explains how the initial set of candidate

datasets was derived from the included studies, which exclusion criteria were applied, and how the final set of 83 datasets was obtained.

Paragraph 4 of 'Literature Search' from 'Results' about notable benchmark families can be omitted since benchmarks are discussed further in the paper.

**Response:**

- We removed the preliminary paragraph listing notable benchmark families from the beginning of the Results section because these datasets are discussed later in the manuscript.

Paragraph 2 of 'Results' section includes information that neurosymbolic systems were characterized across 4 dimensions including neural and symbolic components. It would be useful to add very brief information regarding these components (neural architectures, algorithms, symbolic reasoners, ontologies/KGs) for each work in Kautz categorization since some work descriptions are too high-level (e.g., for BEUrRE: 'combines neural uncertainty modeling with symbolic constraints that encourage global consistency').

**Response:**

- We agree that several system descriptions were previously too high-level. We revised the descriptions throughout the Kautz-based categorization to identify, where applicable, the underlying ontology or knowledge graph representation, the neural architecture or learning algorithm, the symbolic component or reasoning mechanism, and the way these components interact. For example, the revised description of BEUrRE now explains that it represents entities as probabilistic boxes and relations as affine transformations, while relational constraints are incorporated into the geometric model to promote globally consistent uncertain knowledge graph predictions.

Figure 1, 'Neural component' part needs also 'ontology embeddings' mentioning since many discussed works apply this form of ontology representation for underlying problem solving.

**Response:**

- We revised Figure 1 to include ontology embeddings among the neural components used in the reviewed neurosymbolic systems.

Figure 2 needs to be moved to 'Methods' section as well since it describes the workflow of article selection.

**Response:**

- We moved the PRISMA-ScR flow diagram to Stage 3 of the Methods section, where the study-selection and eligibility-assessment procedure is now described.

In Figure 2, 'Identification', the right rectangle: could you please clarify which 'automation tools' were used to assess records eligibility? In the 'Screening' part, could you please clarify which sort of 'retrieval' was applied (rectangles in the middle)?

#### Response:

- The box “Records marked as ineligible by automation tools: (n=0)” follows the standard PRISMA-ScR terminology and indicates that no automation tools were used to assess or exclude records for eligibility. In the PRISMA-ScR diagram, “reports sought for retrieval” refers to the full-text articles that were sought after title and abstract screening, while “reports not retrieved” refers to articles for which the full text could not be obtained.

#### Comment 6:

In 'Neurosymbolic systems (RQ 1)' subsection of 'Results' section, it would be good to unify theme names and their purposes. E.g., theme names in Type 1 systems usually reflect problems to solve ('predicting ontology relations', 'template-based symbolic output selection'), themes in Type 2 systems focus more on methods ('symbolic enrichment followed by neural prediction', 'neural parametrization and constraint-based filtering in symbolic inference'), and some themes in Type 4 systems mix both methods and problems to solve (e.g., 'logic-regularized neural representations for complex query answering'). Since Kautz categorization is about methods, it would be relevant to stick by methods while choosing theme names in all system types and organize the works descriptions accordingly.

In this part in general some important and famous works for knowledge graph embeddings are missed, e.g. TransE, DistMult, TransH, RotatE etc.

#### Response:

- We agree that the previous theme names did not consistently describe the same level of analysis: some referred primarily to application tasks, whereas others described neural-symbolic integration mechanisms. We therefore revised the terminology throughout the Kautz-based categorization so that the integration mechanisms consistently reflect the dominant architectural or methodological pattern. For example, “template-based symbolic output selection” was reformulated as “neural classification over symbolic templates,” “predicting ontology relations, especially subsumptions” as “ontology-informed representation learning”, and “logic-regularized neural representations for complex query answering” as “logic-regularized representation learning.” We also added an introductory explanation that the integration mechanisms derived inductively from the reviewed corpus and should be interpreted as descriptive rather than mutually exclusive groupings.
- We further agree that TransE, TransH, DistMult, and RotatE are foundational knowledge graph embedding methods. We added a brief description of these approaches to the Background section to contextualize the embedding components used by several reviewed systems. However, we did not add them as standalone systems to the Kautz categorization because they are primarily subsymbolic embedding models and do not, by themselves, contain an explicit symbolic reasoning or constraint component.

In Type 1 systems, theme 3 'predicting ontology relations, especially subsumptions' includes one box-embedding approach, yet other geometric ontology embedding methods are operating like this: they encode individuals and concepts as geometric regions and interpret logical operators as operations over geometric objects. Despite that, some geometric ontology embedding methods (e.g., ELEmbeddings, Box2EL) are included in Type 5 systems as per Kautz although they solve similar problems and represent ontologies in a similar way.

**Response:**

- The previous organization placed closely related geometric ontology embedding approaches in different Kautz types, although methods such as ELEmbeddings, EmEL++, and Box2EL share important methodological characteristics: they encode Description Logic concepts and relations geometrically and incorporate ontology semantics through constraints on the learned representations. We have therefore revised the discussion to make this methodological relationship explicit and to group these approaches together when discussing geometric ontology embeddings.
- We also clarify that the mapping of such methods to Kautz's high-level types is not always unambiguous, particularly for approaches in which symbolic semantics are directly embedded into differentiable representations. The revised manuscript therefore uses the Kautz categorization as a high-level description of the dominant integration pattern while additionally organizing closely related methods according to their concrete architectural and representational characteristics.

In Type 1 systems, why EmEL++ is not grouped together with Box2EL, ELEmbeddings and other similar methods? It basically extends ELEmbeddings to include role hierarchy and role chains. As for ELK inferences, Box2EL authors use ELK inferences as ground truth to test for entailed knowledge as well. Also, why this work is not coupled with 'predicting ontology relations' work since it also evaluates ontology completion?

**Response:**

- Upon revisiting the architectural characteristics of EmEL++, we found that its previous placement separated it from closely related geometric ontology embedding approaches primarily on the basis of how the method was described and evaluated, rather than on a substantive difference in its neural-symbolic integration mechanism. EmEL++ extends the geometric embedding approach for EL++ by supporting additional constructs, including role hierarchies and role chains, and is methodologically closely related to ELEmbeddings, Box2EL, and other ontology embedding approaches.
- We therefore moved EmEL++ from the Type 1 grouping to the Type 5 theme on “ontology semantics embedded as differentiable constraints for representation learning”, where it is now discussed together with related geometric ontology embedding methods. We also revised the discussion so that the use of reasoner-generated entailments, such as ELK inferences, is not treated as a defining criterion for the Kautz assignment, since such inferences may also be used as evaluation ground truth by methods including Box2EL.

In Type 1 systems, EBR does not directly 'use <...> reasoner outputs as ground truth' yet rather use symbolic reasoners as SOTA methods to compare against.

**Response:**

- We revised the description of EBR to clarify that symbolic reasoners are used as state-of-the-art baselines for comparison, rather than as sources of ground-truth labels. The previous wording was therefore removed.

In Type 2 systems, I would place DPLogic and DiffLogic into Neuro|Symbolic type since these methods have two co-equal components, KG embeddings and MLN which exchange information via EM loop.

**Response:**

- We revisited the architectures of DPLogic and DiffLogic and concluded that their previous placement in Type 2 did not adequately capture the interaction between their neural and symbolic components. In both systems, knowledge graph embeddings and probabilistic logical reasoning remain distinct components and exchange information during iterative optimization, rather than following a predominantly symbolic-first pipeline with one-way neural guidance. We therefore reassigned DPLogic and DiffLogic to Type 3 (Neuro|Symbolic) and revised the Type 3 description to explicitly characterize this category in terms of distinct, co-equal neural and symbolic components with bidirectional information exchange.

In Type 5 systems for the second theme, it is argued that discussed methods 'support ranking or predicting axioms', yet they predict axioms via ranking (via rank calculation of axioms from the test set among similar ones).

**Response:**

- We revised the wording in the Type 5 discussion to reflect that these methods perform axiom prediction through ranking candidate axioms rather than treating ranking and prediction as separate operations. The revised text therefore refers to the ranking of candidate axioms.

**Comment 7:**

In 'Evaluation Methodology (RQ 2)' subsection of 'Results' section, please elaborate more on datasets which are used just in one work among selected ones; why they are not widely adopted? Also, add more information about OWL benchmark suites reported (LUBM, ORE, OWL2Bench), which ontologies are included there, how do they look like, in which works were they mentioned, for which tasks were they applied in these works? These datasets collections are not reported in dataset tables.

**Response:**

- We expanded the RQ 2.1 dataset analysis to provide additional context for both infrequently reused datasets and ontology-oriented benchmark suites.
- First, we clarified the organization of the dataset catalogue. Datasets used by at least two reported neurosymbolic systems are presented in the main manuscript, whereas datasets used by exactly one reported system are provided in the Appendix. We also added a discussion of these single-use datasets, noting that many are domain-specific, task-specific, or introduced or adapted for a particular evaluation setting. Their limited reuse within the reviewed corpus therefore reflects the specialization and fragmentation of current evaluation practices and limits direct cross-system comparison.
- Second, we expanded the discussion of LUBM, ORE, and OWL2Bench. The revised manuscript now describes their structure and intended evaluation setting and explicitly identifies their use within the reviewed corpus. LUBM is used by Poderoso for link prediction, while ORE and OWL2Bench are used by EmELvar for subsumption prediction. We also clarify that these benchmark suites are included in the Appendix table containing datasets used by exactly one reported neurosymbolic system. Finally, the revised discussion emphasizes that ontology-oriented benchmarks can support both deductive and predictive



evaluation rather than treating ontology and KG benchmarks as corresponding to strictly separate reasoning paradigms.

'The co-existence of KG-style benchmarks and OWL-centric resources suggests that current neurosymbolic evaluation spans two partially distinct traditions: one emphasizes performance on predictive KG tasks, whereas the other emphasizes deductive reasoning' - this claim might not be true: some of reported ontology embedding methods (e.g., ELEmbeddings, Box2EL etc.) try to predict new knowledge (e.g., missing subsumptions), not entailed axioms. Similar claim can be found in 'Conclusion' section, RQ 3.

**Response:**

- We agree that the previous wording implied an overly strict distinction between predictive knowledge graph evaluation and deductive ontology reasoning. As the reviewer notes, ontology-based resources are also used by embedding methods for predictive tasks such as subsumption prediction and ontology completion.
- We therefore revised the RQ 2.1 discussion to clarify that ontology-oriented resources can support both deductive and predictive evaluation settings and that the distinction concerns the representation, semantic structure, and task formulation rather than a strict predictive-versus-deductive divide. We also revised the corresponding statement in the RQ 3 conclusion, replacing the previous characterization of two separate evaluation traditions with a description of heterogeneous benchmark families that differ in representation, semantic richness, domain, and supported reasoning tasks.

For Tables 8 and 11 from the Appendix, it would be interesting to add ontology DL expressivity for ontology datasets to see which methods target certain DL expressivity for ontology reasoning tasks.

**Response:**

- We have added Description Logic expressivity information to the ontology datasets reported in Tables 8 and 11 of the Appendix, where this information is available and applicable.

**Comment 8:**

For 'Tasks and Metrics (RQ 2.2)' subsection of 'Results' section, it would be useful to specify whether reported metrics are task-specific or general. As for metrics reported beyond the 10 most common ones, it would be useful not just to list them, yet to mention what they measure, how they are calculated, what do they mean. It would be good to discuss here which limitations do reported metrics have, how they miss some aspects of methods evaluation etc., for now only critical analysis of benchmark sets is present.

**Response:**

- We agree that the previous presentation of metric frequencies lacked sufficient task context. We therefore now report both overall metric frequencies across the 63 included studies and task-conditioned frequencies for the most frequently represented task categories. For example, among the 28 studies evaluating link prediction, 22 report Hits@K and 17 report MRR; among the 10 studies evaluating subsumption prediction, 8 report Hits@K and 4 report MRR; and among the 13 studies evaluating complex query answering, 8 report MRR and 6 report Hits@K. We similarly report the corresponding frequencies for



multi-hop reasoning and knowledge graph question answering. This makes explicit both the task context and the denominator underlying the reported metric frequencies.

In the text accompanying Table 9, in G3 part comments about 'Membership' are missing although membership task is mentioned in the table.

**Response:**

- We have extended the discussion of G3 to explicitly include the membership (realization) task reported in Table 9. The revised text now explains that, in membership evaluation, neural components typically score or predict whether an individual belongs to a class, while ontology axioms or rules define the corresponding membership relation and may support entailment or consistency checking.

'Ontology-centric structure induction was also prominent.' - could you please clarify what does this sentence mean?

**Response:**

- By “ontology-centric structure induction,” we specifically referred to the G4 tasks of subsumption prediction and rule learning. In these approaches, neural models rank candidate axioms or rules, while symbolic semantics are used to evaluate their correctness, for example through ontology-coherence or constraint-violation checks. We have replaced the original sentence with this more explicit explanation.

G5 and G6 topics are mentioned briefly, it would be good to explain what they mean (although this part is contained in the Appendix entirely).

**Response:**

- We expanded the manuscript to clarify the scope of G5 and G6. G5, “search and planning with neural guidance” concerns the use of symbolic constraints to guide the correction, completion, or adaptation of knowledge under incompleteness, noise, or inconsistency. G6, “explainable inference,” concerns methods that provide human-interpretable reasoning evidence, such as rules, proof traces, or graph paths, and evaluate the coherence or faithfulness of these explanations.

**Comment 9:**

'Discussion' section, 'Limitations and future improvements (RQ 3)' subsection: 'First, the benchmark landscape is fragmented across two partially distinct traditions: widely adopted KG completion datasets <...> and ontology-centric OWL/DL benchmarks (e.g., LUBM, ORE and OWL2Bench).' - what about other ontologies or LQA datasets?

**Response:**

- We agree that the previous sentence presented the benchmark landscape too narrowly by contrasting knowledge graph completion datasets only with the LUBM, ORE, and OWL2Bench suites. We revised the sentence to acknowledge several partially overlapping benchmark traditions, including knowledge graph completion datasets, logical query

answering benchmarks, individual domain ontologies, and ontology benchmark suites. LUBM, ORE, and OWL2Bench are now presented as examples rather than as an exhaustive account of ontology-based evaluation resources.

As for Table 10 (benchmark limitations), were these limitations extracted from the results of Query 4 from the Appendix?

**Response:**

- Query 4 was specifically designed to identify literature discussing limitations of ontology- and knowledge graph-based benchmarks. However, the limitations summarized in Table 10 were not extracted exclusively from the results of Query 4. They were synthesized from the full set of included studies, including papers retrieved through Query 4 and limitations identified during the broader analysis of the reviewed datasets and evaluation practices. We have clarified this in the manuscript.

As for Figure 3, are these recommendations universal for both KGs and ontologies? Are there any specific metrics / evaluation strategies tailored for KGs or ontologies?

**Response:**

- The five dimensions presented in Figure 3 are intended as a general evaluation framework applicable to neurosymbolic systems operating over both knowledge graphs and ontologies. However, the specific metrics and evaluation protocols used within these dimensions depend on the underlying representation and task. For knowledge graph prediction tasks, ranking metrics such as MRR and Hits@K, filtered evaluation, and robustness to missing or noisy edges are particularly relevant. For ontology-oriented reasoning, evaluation may additionally consider logical soundness and completeness, consistency preservation, supported Description Logic expressivity, reasoning depth, and the correctness of inferred axioms or proof traces. We have clarified in the manuscript that the recommendations are general at the dimensional level but should be operationalized using representation- and task-specific measures.

In 'Evaluation metrics' part of the five-dimensional evaluation framework, why MRR/Hits@K, precision/recall/F1/accuracy are chosen among standard metrics?

**Response:**

- MRR and Hits@K were included because they are the standard metrics most frequently used for ranking-based tasks such as link prediction, knowledge graph completion, and axiom ranking. Precision, recall, F1-score, and accuracy were included because they are widely used for classification, retrieval, and answer-set evaluation. These metrics were therefore selected to reflect the dominant evaluation practices observed across the reviewed studies, rather than as a universal or exhaustive set of measures. We have clarified that they should be complemented, where appropriate, by task-specific measures of logical correctness, consistency, proof quality, robustness, explainability, and computational efficiency.

#### Comment 10:

Supporting Material, 'Full Search Queries' part: in Query 6 'Extended coverage', were "knowledge graph embedding" approach included and "link prediction" task? If no, why not?

#### Response:

- The exact terms “knowledge graph embedding” and “link prediction” were not included in Query 6. The extended-coverage query was designed to retrieve explicitly neurosymbolic or logic-aware approaches, using terms such as “ontology embedding,” “logical embedding,” “box embedding,” “neural theorem proving,” and “differentiable reasoning.” We did not use the broader terms “knowledge graph embedding” and “link prediction” because, without an explicit neural-symbolic or logic-related qualifier, they retrieve a very large literature on purely subsymbolic knowledge graph embedding methods that falls outside the scope of this review. Nevertheless, eligible neurosymbolic systems evaluated on link-prediction tasks could still be identified through the broader neurosymbolic system queries and were retained when they satisfied the inclusion criteria. We have clarified this rationale in the Supporting Material.

#### Minor comments:

- 1. The font in Figure 1 is too small, please enlarge it for easier readability. Response: We made the changes.
- 2. On page 15 there's a typo: 'ALCconstraints' -> 'ALC constraints'. Response: We made the changes.
- 3. On page 31 in the text accompanying Table 9, it might be better to renumber G-themes so that G1 appears first in the text, G2 second etc., not G2 first, then G3, then G4, and finally G1. Response: We made the changes.

## Reviewer 2

#### Comment 1:

Kautz Neuro-symbolic taxonomy adoption

While it is true that Kautz taxonomy is well known and has been adopted in several works, one of the most problematic aspects is that it does not provide very rigorous types definition. It stemmed out of a nice lecture, but the majority of NeSy systems, if analysed in their actual componency, can be classified by more than one type, and that's because this taxonomy lacks the level of detail that in this context would be necessary. Adopting it as the backbone to classify previous existing work unfortunately does not grant the immediate understandability of the classified approaches, which is, instead, very relevant as scope of a survey paper.

#### Response:

- We agree that Kautz's taxonomy provides a useful high-level characterization of neural-symbolic integration, but does not define sufficiently rigid or mutually exclusive architectural classes to fully characterize contemporary neurosymbolic systems. In

particular, multi-stage systems may exhibit characteristics associated with more than one Kautz type depending on where and how neural and symbolic components interact.

- We have therefore revised the role of Kautz's taxonomy in the manuscript. Rather than treating the six types as definitive and mutually exclusive system classes, we use them as high-level organizational categories. Systems are assigned to a primary type according to their dominant neural-symbolic integration mechanism, while their architecture is further characterized in terms of the underlying knowledge representation, neural component, symbolic component, and the interaction and information flow between these components.
- To improve the interpretability of the categorization, we have also substantially expanded the Type 1-6 subsections. Each subsection now begins with an explicit description of the architectural pattern represented by the corresponding type, followed by finer-grained integration mechanisms identified within the reviewed corpus. We additionally revisited the classification of systems for which the previous assignment did not adequately reflect their architecture; for example, DPLogic and DiffLogic are now grouped with systems exhibiting parallel and iterative neural-symbolic interaction, while EmEL++ is grouped with related ontology-embedding approaches.
- Finally, we complement the Kautz-based organization with a component-oriented architectural description and a schematic comparison of the neural and symbolic components and their information flows. The intention is therefore not to claim that each system belongs unambiguously to one Kautz type, but to retain Kautz's taxonomy as a recognizable high-level structure while providing the finer architectural detail required to understand and compare the reviewed approaches.

## Comment 2:

2. Lack of clarity: the scope of a survey paper is to provide a clear description of (a specific aspect of) the state of the art in a domain. The organisation of the paper is, in my opinion, poor, and it is very difficult to get an idea of the domain from this work.

Some examples (detailed in the followings):

### Response:

- We agree that the previous organization did not sufficiently distinguish the methodological procedure from the substantive analysis. We therefore reorganized the Methods and Results sections to remove repetition and improve the progression of the paper. The study-selection procedure, eligibility criteria, screening counts, and PRISMA-ScR diagram are now presented together in the Methods section. The dataset-identification and filtering procedure has likewise been consolidated under the data-charting stage. The repeated literature-search material was removed from the Results section, which now begins directly with the analysis of neurosymbolic systems. We also expanded the Background section and clarified the system, task, dataset, and metric terminology used throughout the manuscript.

Tables express in a more schematic way what it is said in the text, and the text briefly summarises what can be found in tables, but they do not add anything one to the other. This results in several tables with a column "Task" which is barely mentioned, and which is instead potentially fairly relevant, whose entries are keywords left to be interpreted by the reader.

### Response:

- We agree that the previous version did not sufficiently differentiate the role of the narrative discussion from that of the tables, and that the task labels in particular required

more explanation. We have therefore revised these sections so that the tables serve primarily as a structured reference providing the system-dataset-task-metric mapping, while the accompanying text focuses on synthesis and interpretation rather than reproducing the tabulated information.

- In the revised analysis of the six integration types, we now discuss how the identified architectural mechanisms relate to the tasks on which the systems are evaluated, highlighting recurring task patterns and differences across the integration types. We have also substantially expanded the dedicated RQ 2.2 analysis of tasks and metrics. The task categories are now explained in the text rather than being presented only as keywords in the tables, including clarification of ontology-oriented tasks such as subsumption and membership, predictive tasks such as link prediction and knowledge graph completion, and the less common task groups concerning “constrained search, repair and evolution”, and “explainable and auditable inference”.
- The tables have been retained because they provide the detailed per-system mapping that would be cumbersome to reproduce in prose, whereas the revised text is intended to explain the meaning, relationships, and broader patterns represented by those entries.

The reason for clustering papers in “Types of neurosymbolic reasoners” is left to a few lines sentences such as “overall, type X systems in our corpus...” while it should be clearly stated and highlighted at the very beginning, including what are the details of that specific cluster, architectural patterns, nuances based on their scope, etc. This information is mostly missing from the main paper, or sparse, or not detailed enough.

#### **Response:**

- We agree that the previous version did not make the rationale for clustering the reviewed systems sufficiently explicit at the beginning of the Results section. We have therefore substantially revised the RQ 1 analysis. The revised manuscript now states explicitly that systems are organized according to their dominant neural-symbolic integration mechanism, considering where symbolic knowledge enters the architecture, the respective roles of the neural and symbolic components, and the direction and timing of information flow between them. We also clarify that the Kautz categories are used as primary organizational clusters rather than as mutually exclusive architectural classes, since contemporary systems may combine characteristics of more than one paradigm.
- Within each Kautz type, the systems are further grouped according to recurring integration mechanisms identified in the reviewed corpus. Each Type 1-6 subsection now begins with a description of the defining architectural organization of the cluster, followed by the principal integration mechanisms and architectural variations observed within it, before the individual systems and their task/evaluation settings are discussed.

#### **Additional general remarks:**

It is not always clear how the clustering of “themes” is done. They start appearing on p. 11, and they are used then to introduce clusters of works considered, but it is not clarified if they are extracted bottom up, analysing the selected corpus, if they are regularities

#### **Response:**

- We have clarified how the integration mechanisms used to organize systems within the Kautz categories were derived. These mechanisms were not predefined. Rather, after assigning the included systems to the corresponding Kautz integration types, we compared

their neural components, symbolic components, and neural-symbolic interactions and identified recurring integration mechanisms inductively from the reviewed corpus. These mechanisms were then used to organize the systems within each Kautz type. We also clarify that they are descriptive rather than mutually exclusive, since some systems combine multiple forms of neural-symbolic interaction.

The paper would benefit of more clear figures, showing difference in architectures, possibly with the adoption of a more standard terminology, such as the Boxology approach [1]

[1]: van Bekkum, M., de Boer, M., van Harmelen, F., Meyer-Vitali, A., & Teije, A. T. (2021). Modular design patterns for hybrid learning and reasoning systems: a taxonomy, patterns and use cases. *Applied Intelligence*, 51(9), 6528-6546.

#### **Response:**

- We agree that the architectural differences between the integration types were not sufficiently apparent in the previous presentation. We therefore introduced component-oriented terminology inspired by the Boxology framework of van Bekkum et al. [1] to describe neural and symbolic components and their information flows more consistently.
- In addition, we added a new schematic figure comparing the six neural-symbolic integration architectures used in the review. The figure uses a consistent graphical representation to make the relative position of neural and symbolic components and the dominant direction of information flow explicit. Kautz's taxonomy remains the primary organizational framework of the review, while the Boxology approach is used as a complementary architectural vocabulary to improve the clarity and comparability of the different integration types.

#### **Detailed Review:**

- p. 2: "Knowledge graphs and ontologies provide structured and explicit representations of domain knowledge that are accessible to both neural and symbolic reasoning modules" -> what does it mean in this context "that are accessible"? **Response:** The sentence containing the term "accessible" has been removed during the revision of the Introduction. The revised text now directly defines knowledge graphs and ontologies and explains how these structured representations are used in the reasoning tasks considered in this review, avoiding the ambiguous use of "accessible."
- p. 2: "knowledge representation and reasoning (KRR) frameworks, particularly knowledge graphs (KGs) and ontologies" -> would you define knowledge graphs and ontologies as frameworks in this context? **Response:** We agree that referring to knowledge graphs and ontologies themselves as knowledge representation and reasoning (KRR) frameworks was imprecise. In the revised Introduction, we have removed this formulation and instead describe knowledge graphs and ontologies as two prominent forms of structured knowledge used in neurosymbolic systems. We then define each representation separately and distinguish the representations themselves from the reasoning mechanisms that operate over them.
- p. 2: "and YAGO-based resources" -> which resources are you referring to? re reference points at YAGO, but here are mentioned "YAGO-based resources" **Response:** We have replaced the generic reference to "YAGO-based resources" with the specific benchmark used in the reviewed corpus, YAGO3-10, and explicitly state that it is derived from YAGO3.



- p.4: Table 1 here is not necessary, one possible rework would be to describe in more detail the specific type in the corresponding part of the paper. **Response:** We retained Table 1 because it provides a concise background overview of the six Kautz integration types and serves as a reference point for the subsequent analysis. However, we agree that the table alone did not provide sufficient detail. We have therefore substantially expanded the corresponding Type 1-6 subsections in the Results section, where each type is now introduced with a more detailed description of its architectural organization and neural-symbolic interaction mechanism before the individual systems are discussed. We also added a schematic architectural figure to facilitate comparison across the six types.
- p. 4: the Background section is pretty short, and the same 4-5 papers appears repeated in different sentences before and after Table 1. **Response:** We agree that the previous Background section was too concise for readers who are not already familiar with knowledge representation and Semantic Web terminology. We have substantially expanded this section to define knowledge graphs, ontologies, Description Logics, TBox, ABox and RBox axioms, RDF, OWL, and representative symbolic reasoners. We also introduce the principal deductive and predictive tasks considered in the review and provide brief descriptions of frequently used knowledge graph and ontology benchmark families. These additions are intended to make the subsequent system and evaluation analysis understandable without assuming extensive prior familiarity with the field.
- p.6: RQ1 is very broad, what does it mean here to "reason over ontologies and knowledge graphs" ? **Response:** We agree that the previous formulation of RQ1 was too broad. We revised it to ask which systems have been developed for predictive or deductive reasoning over ontologies and knowledge graphs and how their knowledge representations, neural components, symbolic components, and integration mechanisms can be characterized. This revised formulation makes clear that the objective is not merely to enumerate systems, but to analyze their architectures and modes of neural-symbolic integration.
- p. 6: "The initial search retrieved 984 records from Google Scholar, 202 from the ACM Digital Library, and 81 from IEEE Xplore" -> which timespan did you consider? **Response:** No publication-year restriction was applied; the searches therefore covered all records indexed by the selected databases up to the respective search dates. Google Scholar was searched on 15 December 2025, while IEEE Xplore and the ACM Digital Library were searched on 25 February 2026. We have made this information explicit in the Methods section.
- p. 7: it is probably not necessary to declare both inclusion and exclusion criteria, at least not in this detail, especially if one criterion is e.g. "English only" which automatically excludes "Publications not written in English." **Response:** We agree that presenting separate inclusion and exclusion criteria resulted in unnecessary duplication, particularly where exclusion criteria were simply the inverse of the corresponding inclusion criteria. We have therefore consolidated these into a single, more concise set of eligibility criteria and removed redundant formulations such as separately specifying both English-language inclusion and non-English-language exclusion. The subsequent description of the principal reasons for exclusion during full-text screening has been retained to document how the eligibility criteria were applied in practice.
- p. 8: "(3) omission of required components such as ontologies..." -> omission in the sense that it was not included as material or that it was not used? Also, the definition of KG and ontology is per se ambiguous, what were the inclusion criteria here? **Response:** By "omission of required components," we intended to refer to the absence of an ontology or knowledge graph that was actually used within the neurosymbolic approach, rather than merely being absent from the material provided with the paper. We have revised this wording accordingly. We have also made the study-selection criterion more explicit: knowledge graphs were considered in scope when graph-structured entities and relations were used by the learning or reasoning process, while ontologies were considered in scope



when their classes, relations, axioms, or constraints provided explicit structure used for learning, inference, prediction, validation, or evaluation. Mere mention of an ontology or knowledge graph without an operational role in the approach was not sufficient for inclusion.

- p. 8: "The database search yielded 984 records from Google Scholar, 202 from the ACM Digital Library, and 81 from IEEE Xplore." -> this exact sentence is present on page 6. Maybe it would be useful to provide some more statistics? For example the percentage of journal papers? Of conference papers? Etc. **Response:** We removed the duplicated sentence from the Results section. The database-search counts and subsequent screening statistics are now reported once under Stage 3 of the Methods section, together with the number of records retained after duplicate removal, title and abstract screening, and full-text eligibility assessment.
- p.9: if "Neural Systems" and "Symbolic Systems" are "out of scope", why do you have them in the image? It would be enough a brief sentence in the text, but the image should highlight the main and most relevant aspects. **Response:** We agree that including purely neural and purely symbolic systems in the taxonomy figure gave unnecessary visual prominence to categories that are outside the scope of the review. We have therefore revised the figure to focus exclusively on the dimensions used to characterize the included neurosymbolic systems. The exclusion of purely neural and purely symbolic approaches is now stated briefly in the accompanying text.
- p. 9: "Datasets were excluded if they primarily contained image-based or lexical data" -> why? **Response:** We thank the reviewer for pointing out that the previous wording did not explain the rationale sufficiently and could be interpreted as excluding datasets solely on the basis of their data modality. This was not our intention. The exclusion was based on the scope of the review: datasets were excluded when their primary evaluation setting did not involve reasoning over an ontology or knowledge graph. We have revised the text accordingly and clarified that image-based or textual/lexical datasets were excluded only when they did not contain or use an operative structured knowledge representation for the reasoning task. Thus, structured lexical resources used as knowledge graphs, such as WordNet-derived benchmarks, remain within scope.
- p. 10: Fig. 2: what is the meaning of the arrows here? **Response:** The arrows in Figure 2 indicate the progression of records through the successive stages of the PRISMA-ScR study-selection process, from identification through screening and eligibility assessment to final inclusion. We have clarified this in the figure caption.
- p. 10: here the paper would benefit of some diagrammatic language such as Boxology **Response:** We agree that a more explicit diagrammatic representation facilitates comparison between the different neural-symbolic architectures. We have therefore introduced component-oriented terminology inspired by the Boxology framework of van Bekkum et al. and added a schematic representation of the six integration types. The diagrams distinguish neural and symbolic components and make their relative position and dominant information flow explicit. Kautz's taxonomy remains the high-level organizational framework, while the Boxology-inspired representation is used as a complementary architectural vocabulary to improve the clarity and comparability of the different integration patterns.
- p. 11: "In our corpus, one theme is template-based symbolic output selection, where HTL (Hereditary attentive Tree-LSTM) (Gomes et al., 2022) represents..." -> how come that it is one "theme", and then there is only one paper cited here? a big, well organised table with clusters and dimension is necessary for a survey paper **Response:** We agree that the term "theme" was potentially misleading, particularly where an identified architectural pattern was represented by only one system in the reviewed corpus. Our intention was not to imply that each grouping constituted a recurrent research theme, but rather to distinguish

integration mechanisms observed within each Kautz category. We have therefore replaced the term “theme” with “integration mechanism” throughout the RQ 1 analysis and explicitly clarified that an integration mechanism may be instantiated by one or several systems. The revised RQ 1 section also makes the basis of the clustering explicit by describing the defining architecture of each Kautz type and the corresponding neural-symbolic interaction mechanisms before discussing the individual systems.

- We have additionally reorganized the corresponding system tables according to these integration mechanisms, making the association between each architectural grouping and the individual reviewed systems explicit while retaining the dataset, task, and metric information.
- p. 11: “Table 2. Type 1 neurosymbolic reasoners.” -> are they only “reasoners”? **Response:** We agree that the term “reasoners” is too restrictive, since the reviewed approaches include not only explicit reasoning systems but also neurosymbolic models and frameworks for tasks such as representation learning, prediction, alignment, and question answering. We have therefore replaced “neurosymbolic reasoners” with the broader term “neurosymbolic systems” in the captions of the Type 1-6 tables and, where appropriate, throughout the corresponding discussion.
- p. 11 Table 2: if one of the criteria is the shape of the ontology - KG, why are these not described here? not even the domain? **Response:** Systems are assigned to the Kautz types primarily according to their dominant neural-symbolic integration mechanism, whereas the underlying knowledge representation and application domain are treated as complementary descriptive dimensions. To avoid duplicating dataset-level information across the six system tables, Table 2 and the corresponding Type 1-6 tables retain the system-dataset-task-metric mapping, while dataset characteristics and application domains are analyzed separately in RQ 2.1.
- p. 12: “A second theme is...” here again, a figure / diagram of the components of the architecture could be useful **Response:** We agree that the textual description alone can make the architectural differences between the identified integration mechanisms difficult to follow. We have therefore complemented the discussion with schematic diagrams illustrating the principal neural and symbolic components and their information flow for the integration mechanisms identified within the corresponding Kautz types.
- p. 12: “ontology evolution” -> have you defined ontology evolution before? **Response:** The term “ontology evolution” had not been explicitly defined before its first use. We have therefore added a brief definition at its first occurrence, clarifying that ontology evolution refers here to the controlled update or extension of an ontology to incorporate new or changed knowledge.
- p. 14: “themes” division here is chaotic, it would be clearer if divided in paragraphs **Response:** We agree that the previous presentation of the architectural themes made the discussion difficult to follow. We have therefore reorganized the Type 1-6 analysis so that the identified architectural patterns are presented in separate paragraphs. This clearer paragraph structure distinguishes the individual mechanisms and the systems associated with each of them, while making the progression of the architectural analysis easier to follow.
- p. 15: at the top of the table it would be good to have a short description of the neuro-symbolic architecture type, otherwise it is difficult to follow **Response:** We agree that the tables are easier to interpret when the defining architectural organization of each neurosymbolic type is visible directly within the table. We have therefore added a concise architectural description at the top of each Type 1-6 table, before the individual integration mechanisms and systems are listed. This allows the reader to immediately relate the systems and mechanisms in each table to the corresponding high-level neural-symbolic architecture.

- p. 15: "to satisfy ALC constraints" -> I think there is a missing space after ALC, but also: what if the reader does not have familiarity with logic language acronyms? a short description would be beneficial **Response:** We corrected the typographical error from "ALCconstraints" to "ALC constraints." We also expanded the Background section to introduce Description Logics and explain the role of logical constructors and expressivity, so that readers are not expected to interpret such acronyms without prior context.
- p. 16: given that a survey paper should be the reading to start from to have an idea of a domain, here what is provided is a list of works, organised (subjectively) according to a taxonomy, but no further explanation is provided. **Response:** We agree that the previous version was too close to a catalogue of individual works and did not provide sufficient architectural synthesis to serve as an accessible entry point to the field. We have therefore substantially expanded the analysis of RQ 1.
- First, we now explain explicitly the rationale used to organize the reviewed systems. Kautz's taxonomy is retained as a high-level organizational framework, while systems are assigned according to their dominant neural-symbolic integration mechanism, considering where symbolic knowledge enters the system, the respective roles of the neural and symbolic components, and the direction and timing of information flow. We also clarify that these categories are not treated as rigid or mutually exclusive architectural classes.
- Second, each Type 1-6 subsection now begins with an explanation of the defining architectural organization of that type and subsequently identifies finer-grained integration mechanisms observed within the reviewed systems. The discussion therefore focuses not only on which works belong to each group, but also on how their neural and symbolic components interact, where important architectural variations occur, and for which tasks the corresponding approaches are evaluated.
- Finally, we complement the textual analysis with schematic architectural figures and reorganized system tables that explicitly relate the individual systems to the identified integration mechanisms. These revisions are intended to provide the reader with a clearer conceptual overview of the main architectural approaches in the field before presenting the details of the individual works.
- p. 16: "Overall, Type 4 systems..." -> This is what the paper should provide, but not as a single sentence, but detailed grouping of different aspects of several works **Response:** We agree that the synthesis should not be limited to a brief concluding statement, but should explicitly compare and group the reviewed approaches according to the different ways in which neural and symbolic components interact. We have therefore substantially expanded the Type 4 analysis.
- The revised section now distinguishes five integration mechanisms through which symbolic knowledge shapes neural learning: (1) constraint-driven grounding and revision under description logic, (2) deduction-guided supervision and evaluation, (3) logic-regularized representation learning, (4) symbolic guidance of neural reasoning policies, and (5) symbolic knowledge injection into representation learning. Within each mechanism, we discuss the corresponding systems together and explain how symbolic knowledge is used, for example as training supervision, deductive filtering, logical regularization, rule-based guidance, or structured knowledge injected into representation learning.
- The ``Overall, Type 4 systems..." paragraph is therefore now used only to summarize this more detailed preceding analysis rather than serving as the primary synthesis itself.
- p. 18: "constraints for completion and approximate deduction." -> What do you mean here with "approximate deduction"? **Response:** We intended it to refer to approaches that encode ontology semantics as differentiable or geometric constraints and use the resulting learned representations to approximate consequences of the ontology, rather than performing exact symbolic deduction. To avoid this ambiguity, we have removed the expression "approximate deduction" from the description of this integration mechanism.

The revised text now refers to “ontology semantics embedded as differentiable constraints for representation learning” and explains more explicitly how the corresponding systems encode ontology semantics and use the learned representations for tasks such as ontology completion and candidate-axiom ranking.

- p. 22: “Similarly, the commonsense reasoner...” -> why is it a “commonsense reasoner”?  
**Response:** The term “commonsense reasoner” is taken directly from the title and terminology of the original work, *Neuralsymbolic Commonsense Reasoner with Relation Predictors*. To make this clearer, we have revised the manuscript to refer to the approach by its full designation at first mention rather than using the shortened expression “commonsense reasoner” without explanation.
- p. 24: “Overall, these findings suggest that neurosymbolic evaluation currently draws on...” -> this formulations should be avoided in survey papers, these are not findings, it should be the systematic analysis of a set of papers and organisation according to certain dimensions, clustering, and explicit description of the rationale  
**Response:** We agree that the term “findings” was not appropriate in this context and that the Results should more explicitly reflect the systematic analysis and organization of the reviewed literature. We have therefore revised this part of RQ 2.1 to describe the dataset landscape as a structured characterization of the included studies rather than as a set of findings. Specifically, the datasets are organized according to their type and domain, their reuse across the reviewed systems, the tasks for which they are employed, and the extent to which they provide explicit ontological semantics. We also make the rationale for the organization explicit by distinguishing recurrent datasets, used by at least two systems in the reviewed corpus, from datasets occurring in only one study. The corresponding discussion has been revised accordingly to emphasize this systematic mapping and comparison rather than presenting the observations as independent findings.
- p. 25: Table 8 could benefit of a rework to make it more readable, maybe adding one column per each work considered and some ticks, if they are all relevant, otherwise show only meaningful clusters  
**Response:** We agree that the previous version of Table 8 was difficult to read because the Task(s) column contained detailed study-level task mappings and citations. Rather than adding one column for each included work, which would result in an excessively wide table, we followed the reviewer’s alternative suggestion to present meaningful clusters. We therefore reorganized the individual tasks into recurring evaluation task families and added the number of distinct studies associated with each task family. The task families are non-exclusive, as individual studies may evaluate more than one task. Detailed system-dataset-task mappings remain available in the Type 1-6 system tables. This redesign substantially reduces the density of Table 8 while retaining the principal comparative information.
- p. 26-28: In some cases the column “Domain” it is the actual domain, in others it is the extended name of the Dataset, in others it seems to be the scope...?  
**Response:** We agree that the previous “Domain” column conflated several different characteristics, including application domain, resource type, and benchmark scope. We have therefore revised the column to “Resource type / domain” and standardized its entries throughout the table. Each dataset is now characterized consistently by the type of resource (e.g., knowledge graph benchmark, ontology, KGQA benchmark, or ontology-matching benchmark) together with its application domain or scope where applicable. This avoids using extended dataset descriptions and domain information interchangeably.
- p. 29: “ComplexWebQuestions” overlaps with next cell  
**Response:** We fixed this.
- p. 30: This list of metrics lacks a bit of context, e.g. Hits@K out of how many works focused on the same task? and is this related to which task specifically?  
**Response:** We agree that the previous presentation of metric frequencies lacked sufficient task context. We therefore now report both overall metric frequencies across the 63 included studies and

task-conditioned frequencies for the most frequently represented task categories. For example, among the 28 studies evaluating link prediction, 22 report Hits@K and 17 report MRR; among the 10 studies evaluating subsumption prediction, 8 report Hits@K and 4 report MRR; and among the 13 studies evaluating complex query answering, 8 report MRR and 6 report Hits@K. We similarly report the corresponding frequencies for multi-hop reasoning and knowledge graph question answering. This makes explicit both the task context and the denominator underlying the reported metric frequencies.

- p. 31: Table 9 starts at the end of the page, and then again at the beginning of the subsequent page, but lacking the “G1” specification? **Response:** Corrected. We adjusted the pagination of Table 9 so that the task-group headings are not separated from their corresponding rows across page breaks. We also checked the continuation of the remaining task groups after recompilation.
- p. 35: Fig. 3 actually helps a lot, but a. it should appear much earlier in the paper, not as a surprise towards the end, it is one of the main outcomes and it deserves to be highlighted. However, it could be formatted in a better way: it is not clear why the entries are on two columns, given that you have the space. It is necessary also to motivate the choice of the dimensions in advance, not after the limitations **Response:** We agree that the evaluation framework represents one of the main outcomes of the review and should be introduced and motivated earlier in the manuscript. We have therefore moved the framework to the Results section, where it is now presented in a dedicated subsection on the synthesis of evaluation dimensions.
- Before introducing the figure, we now explain the rationale for distinguishing the five dimensions—dataset characteristics, reasoning capabilities, evaluation protocols, evaluation metrics, and robustness and generalization—and clarify that they provide complementary perspectives on the evaluation of neurosymbolic systems over ontologies and knowledge graphs.
- We have also redesigned the figure to improve its readability. Rather than arranging the dimensions in two columns, the revised figure presents the five dimensions in a single vertical layout, with one dimension per row and its main evaluation aspects shown directly alongside it. This makes the structure of the framework more explicit and gives equal visual emphasis to each dimension.
- p. 37: “Our synthesis highlights three main limitations of...” -> this is another outcome, identifying some limitations allows to extend and refine the area of research, it deserves to be listed among the main contributions **Response:** We agree that the systematic identification of evaluation and benchmark limitations constitutes an important outcome of the review. We therefore revised the Introduction to make this contribution explicit. The contribution statement now highlights not only the system categorization, dataset analysis, and task-to-metric mapping, but also the synthesis of current benchmark limitations and the resulting recommendations for more comprehensive neurosymbolic evaluation.