Dear Dr. Rossi and Reviewers,

Thank you for the opportunity to revise and resubmit our manuscript titled "Benchmarking Neuro-Symbolic Description Logic Reasoners: Existing Challenges and A Way Forward." We sincerely appreciate the constructive feedback provided by the reviewers. The comments and suggestions were invaluable in refining our work, and we believe the revised manuscript is significantly improved as a result.

Response to Reviewers:

Reviewer 1:

- **Comments:** (a) "The authors are more focused on ontological reasoning, but neuro-symbolic reasoning goes well beyond that." ; (b) "Several papers in the past have dealt with the issue of developing benchmarks for neuro-symbolic learning and reasoning, event considering what is today referred to as multimodalities. Pioneering work dealing with different reasoning approaches aboud. However, the authors have not mentioned them or have not have access to them." ; (c) "Have you considered diversity in a wider scope? For instance, language diversity impacts the benchmarking development and the resulting experiments."
  Response: To address the reviewer's concern regarding existing literature in the broader neuro-symbolic AI domain, we have modified the title of our paper from "Benchmarking Neuro-Symbolic Reasoners: Existing Challenges and A Way Forward" to "Benchmarking Neuro-Symbolic Description Logic Reasoners: Existing Challenges and A Way Forward" to better reflect its focus on ontological reasoning within the realm of neuro-symbolic description logic reasoners. This adjustment aligns with the core content of our work and helps clarify its scope, addressing concerns.

- **Comment:** "Success metrics and key performance indicators. Here, the authors should consider whether or not the systems will be sound."
  Response: We have incorporated this suggestion- Page 9, Line 11. "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.

Reviewer 2:

- **Comment:** "One issue that is not addressed directly, but could be mentioned, is that the framework should support (to the extent possible) purely neural approaches, e.g., large language models. Neural approaches may not perform well on all tasks, but including them as a baseline should be possible.".
  Response: We have incorporated this suggestion on Page 8, Lines 34 to 39, discussing the inclusion of purely neural approaches as baselines.

- **Comment:** "The five categories given do not match up with those given in reference 9. Kautz has 6 categories and uses a different notation in places."
  Response: We have added the sixth category on Page 2, lines 18 and 19, and adjusted our taxonomy accordingly ensuring consistency with Kautz's updated categorization.

- **Comment:** "Page 5, lines 37-43: This information should be included in Table 1 rather than in the text after the table".
  Response: This comment was regarding Lines 44 to 51 on page 5. To ensure clarity, We have included the information in a paragraph, as adding another column to the table would disrupt its formatting.

- **Comment:** "Page 7, line 12: This seems to be an additional desideratum (i.e., a "living" benchmark) that could be included in the list on page 6.".
  Response: We agree with the reviewer. We moved this to the desiderata. See Page 9, Lines 24 to 32.

- **Comment:** "Page 6, lines 13-20: Regarding controlled inconsistencies: It might be helpful to have an example here, as some readers may not understand the importance of this issue."
  Response: Examples of controlled inconsistencies have been added on Page 6, starting from line 48.

Reviewer 3:

- **Comment:** "It could have put forth a desired benchmark, even to a narrow scope, that would be a big contribution making their desiderata easy to check for completeness and others to agree/ enhance."
  Response: Thank you for the suggestion. We have added another section (Page 9, Section 4) proposing a potential methodology to develop such a benchmark.

In addition to the changes made in response to the reviewers' comments, we have further improved the manuscript by breaking the content in the

desiderata in Section 3 into bullet points for enhanced readability. We believe these changes will make the content more accessible and easier to follow for readers.

We believe these revisions have significantly enhanced the manuscript, aligning it more closely with the journal's standards and the reviewers' expectations. We appreciate the thoughtful feedback provided and look forward to the possibility of our work being published in *NAI (Article in Special Issue)*.

Thank you for your consideration.

Sincerely,

Gunjan Singh