Response to reviewers' comments

Dear Editor-in-Chief,

The authors would like to thank the reviewers for their valuable feedback and comments. Below is a list of changes made in response to the reviewers' suggestions. For each comment, we have provided a justification and indicate the corresponding changes in the manuscript (highlighted in yellow).

Review #1

This paper surveys neuro-symbolic AI (NSAI) architectures and explores their potential in the context of generative AI. It revisits and extends Kautz's taxonomy, and attempts to map various generative AI methods to these paradigms. The paper also proposes evaluation criteria for NSAI approaches and includes a case study on 4D printing as an illustrative application domain. The authors address a timely and important question: how to bring structure to the fragmented space of neuro-symbolic AI, and how to connect it to the rapid developments in generative AI. This framing is worthwhile, as surveys that attempt to systematize the field can orient researchers who otherwise encounter disparate terminology and frameworks.

Thank you for this positive overall assessment and for recognizing the motivation behind the paper.

There are however several weaknesses and points for improvement:

- While the paper in the introduction claims to define and extend NSAI architectures, I do not see this reflected in the paper. The analysis is qualitative and descriptive, rather than providing deeper theoretical or empirical insights.

Thank you for this comment. We would like to clarify that the paper is intentionally positioned as a theoretical and conceptual contribution, rather than an empirical one. The objective stated in the introduction — *to define and extend NSAI architectures* — is addressed in a theoretical manner in Section 3, where we formalize and structure the different classes of NSAI architectures. This includes the extension of the fibring architecture and the introduction of more granular subclassifications within the compiled architectures, as well as mathematical formulations specifying how neural and symbolic components interact.
The analysis is therefore qualitative by design, as the contribution of the paper is to provide a unified theoretical framework and a structured taxonomy grounded in prior empirical studies from the literature, rather than introducing new experiments or datasets. Our goal is to establish solid theoretical foundations and integration mechanisms for NSAI architectures, which we consider a necessary step before large-scale empirical benchmarking.

- The introduction of "fibring" as a new category is not well justified. As presented, it seems to resemble Symbolic[Neuro] with multiple neural components instead of one. Adding new categories without strong conceptual differentiation risks diluting the power of the classification.

Thank you for raising this important point. We acknowledge that the fibring architecture may appear, at first glance, similar to a Symbolic[Neuro] structure with multiple neural components. However, its conceptual basis is fundamentally different. Fibring is not defined by the *number* of neural modules but by the presence of a symbolic aggregation function—the fibring operator—which governs how multiple neural models interact under logical constraints. Unlike

Symbolic[Neuro], where the symbolic component orchestrates or filters a single neural subsystem, fibring establishes a logic-aware coordination mechanism that merges heterogeneous neural outputs while enforcing symbolic consistency across them. This makes fibring an ensemble integration paradigm, not a nested one: the symbolic layer does not simply consume neural features but actively mediates and constrains cross-network communication. Thus, fibring is introduced as a distinct category because it formalizes a type of symbolic–neural interaction—inter-model constraint propagation—that is not captured by existing architectures. Our intention is not to multiply categories arbitrarily, but to reflect an integration pattern already emerging in multi-agent, mixture-of-experts, and distributed neuro-symbolic systems, which cannot be adequately described by existing nested or sequential models.

- The section attempting to classify generative AI methods into NSAI paradigms is the most novel idea in the paper, but it falls short in execution. Several parallels feel superficial or misleading. Here are a few:

> -- The claim that "XAI fails in the nested paradigm" oversimplifies a vast subfield and cannot be reduced to one category.
>
> We thank the reviewer for this important remark. Our intention was not to dismiss the entire field of XAI, which indeed encompasses a wide variety of approaches, but to point out the limitations of a specific subset of methods when applied to nested neuro-symbolic architectures. In particular, many post-hoc XAI techniques that treat the neural component as a black box and explain it in isolation do not account for the tight interactions between the symbolic and subsymbolic layers in the nested paradigm. To avoid overgeneralization, we revise the sentence to make this scope explicit.
>
> -- GANs are placed in the cooperative paradigm, yet they involve two neural networks without a symbolic component; it is unclear how this qualifies as neuro-symbolic.
>
> We thank the reviewer for this helpful observation. Our intention was not to claim that standard GANs (with two purely neural components) are neuro-symbolic, but to use the GAN setup as an intuition for cooperative interaction. In the revised manuscript, we clarify this point by removing the direct reference to generic GANs and instead describing *GAN-inspired* cooperative training schemes in which at least one component is explicitly tied to symbolic rules or logic-based constraints.
>
> -- Knowledge distillation is mapped to the compiled paradigm, though it is simply neural-to-neural compression; by the same logic, one could have mapped it to cooperative.
>
> We thank the reviewer for this comment. We agree that standard knowledge distillation is a neural-to-neural compression technique and does not, by itself, constitute a neuro-symbolic architecture. In our taxonomy, however, the compiled paradigm refers to settings where the knowledge of one component is compiled into another to produce a single standalone model at inference time. Knowledge distillation fits this definition because the teacher–student interaction is only used during training, and the student ultimately operates independently. To avoid confusion, we revise the text to explicitly state this scope and clarify that knowledge distillation is included as an instance of compilation rather than as a neuro-symbolic example per se.
>
> -- Fine-tuning and pre-training are treated as compiled neuro-symbolic methods, but these are standard neural training regimes without symbolic constraints.

We thank the reviewer for this insightful comment. We agree that standard pre-training and fine-tuning procedures are purely neural training regimes and are not neuro-symbolic by themselves. In our taxonomy, the compiled paradigm is intended to cover settings in which symbolic knowledge (e.g., logical constraints or structured relations) is explicitly integrated into the training process and thereby compiled into the resulting neural model.
Accordingly, we have revised the manuscript to clarify that pre-training, fine-tuning, distillation, and transfer learning are considered part of the compiled paradigm only when they are augmented with symbolic constraints or objectives (for instance, through logic-informed loss functions or neuron-level mechanisms).

-- Data augmentation is also listed under the compiled paradigm, but the example provided (synthetic data with logical rules) is conceptually distinct from typical augmentation. Overall, these mappings appear forced, and the section does not achieve the promised contribution of establishing solid connections between NSAI and generative AI.

We thank the reviewer for this helpful observation. Our intention was not to claim that standard data augmentation (e.g., geometric or noise-based transformations) belongs to the compiled paradigm, but rather to focus on a specific form of *symbolically guided* augmentation, where synthetic data are generated under explicit logical rules. In the revised manuscript, we clarify this by framing our example as *rule-constrained synthetic data generation* and by making explicit that it is this variant—where symbolic constraints are used to generate logically valid labeled examples—that we map to the compiled paradigm. In this setting, symbolic knowledge is first compiled into the data distribution and then into the neural model through training.

- The evaluation criteria introduced (generalization, scalability, interpretability, etc.) are reasonable dimensions, but the way they are applied in the large comparison table is not transparent. It is unclear on what basis the authors judged one architecture stronger than another on each criterion. Without explicit methodology, the evaluation reads as subjective opinion rather than reproducible analysis.

Thank you for pointing this out. Our intention was not to present the comparison as a subjective ranking, but we agree that the methodology behind the scores in Tables 2–4 was not described with sufficient explicitness. As explained in Section 5.2, each main criterion (e.g., generalization, scalability, interpretability) is decomposed into three sub-criteria, and each architecture is assigned a score on a four-point scale (0–3) according to how many sub-criteria are met: 0 = none, 1 = one, 2 = two, 3 = three sub-criteria satisfied. These judgments are grounded in (i) reported empirical results from the literature associated with each architecture (Table 1), and (ii) an analysis of the design principles of the architectures, especially when empirical evidence is scarce.

This paper makes a commendable attempt to structure the field of neuro-symbolic AI and to link it with generative AI. However, the claimed contributions are not convincingly realized: the taxonomy offers little beyond existing classifications, the NSAI vs. GenAI parallels are often superficial or inaccurate, and the evaluation lacks methodological grounding. The case study also overlaps with other published work by the same authors. I recommend that the authors substantially revise the paper. In particular, they should: - sharpen the taxonomy and avoid

unnecessary category inflation, - present a more rigorous and defensible mapping between NSAI and generative AI methods, - clarify how evaluation judgments were made

Thank you for this overall assessment and for the concrete suggestions. As detailed in our responses to the specific points above, we have substantially revised the manuscript along the three main axes you mention.
Regarding the concern about the case study overlapping with our previous work, we agree that this needed clarification. The 4D printing case study indeed builds on our earlier domain-specific work, but in this paper it serves a different purpose: it is used as an *integrative testbed* to instantiate and compare the proposed NSAI architectures in a concrete application scenario.

Review #2

This paper attempts to define and contextualize current Neuro-symbolic Artificial Intelligence (NSAI) architectures. Given the rise of practices combining neural and symbolic components in diverse applications, this paper represents a significant effort in structuring them into defined categories and assessing their effectiveness based on several evaluation metrics. The paper moderately discusses the comparative analysis of the respective strengths and limitations of different architectures. In the end, it walks us through the potential use of NSAI architectures for 4D printing, which was interesting.

Thank you for this summary. We would like to clarify that the comparative strengths and limitations of the architectures are already captured in the evaluation tables, where each paradigm is scored in a literature-grounded way along multiple criteria. In the revised version, we have additionally made this comparative analysis more explicit in Section 5.3, where we now discuss in prose how the different architectures contrast with one another in terms of their main advantages and limitations.

The major concern, however, is that these NSAI architectures might not always fall into the set descriptions they provide; the reality may be more nuanced. In Section 4.2.2 they mention that "In-context learning, such as few-shot learning and CoT reasoning, aligns with the Symbolic[Neuro] approach by leveraging NNs for context-aware predictions, while symbolic systems facilitate higher-order reasoning." However, in many CoT-based neurosymbolic systems, the NN is not just a passive subcomponent but an active generator of reasoning traces. While the symbolic rules constrain or guide the reasoning, the NN often has flexibility in generating candidate reasoning paths. Thus, defining all CoT-based systems as a Symbolic[Neuro] approach does not seem logically sound, based on their definition, which places the NN as a subcomponent within a predominantly symbolic system.

Thank you for this thoughtful observation. We fully agree that CoT-based systems do not all fall under a single NSAI architecture and that the reality is more nuanced than the initial formulation suggested. This is precisely why we have revised Section 4.2.2 to soften the claim and explicitly state that in-context mechanisms such as few-shot learning and CoT can instantiate different nested paradigms depending on how symbolic structure and neural generation are combined. As clarified in the revised text, only *those* CoT settings in which an explicit symbolic framework (e.g., rules, templates, or knowledge bases) orchestrates the reasoning process correspond to the Symbolic[Neuro] paradigm. In contrast, many CoT variants operate differently: the neural model actively generates reasoning traces, while symbolic components merely constrain or

guide the search process, which aligns more closely with Neuro[Symbolic] or even purely neural reasoning.

In Section 4.2.1 they mention that "Techniques like RAG, GraphRAG, and Seq2Seq models (including LLMs, e.g., GPT [70]) align with this method due to their reliance on neural encodings of symbolic data (e.g., text or structured information) to perform complex transformations before outputting results in symbolic form." This, however, has a caveat regarding the generalization of RAG-based models. Saying the results are always “decoded back into symbolic output” might be slightly overstated (based on the description of Sequential architectures in Section 3.1). In some cases, the output remains natural language explanations that approximate symbolic reasoning; they are not guaranteed to be well-formed symbolic structures.

Thank you for this clarification. We agree that the original phrasing could be interpreted as implying that Sequential architectures always decode into formally well-formed symbolic structures, which is not universally the case for RAG-based and general Seq2Seq/LLM pipelines. Our intention was to follow Kautz’s Symbolic → Neuro → Symbolic view in which both inputs and outputs are symbolic in the discrete sense (e.g., text tokens). In practice, many RAG/LLM systems indeed return natural-language sequences that approximate symbolic reasoning without guaranteeing formal validity (e.g., logical well-formedness or strict KG structure). In the revised manuscript, we have modified Section 4.2.1 accordingly by softening the claim and explicitly distinguishing between Sequential systems that decode into structured symbolic forms (e.g., semantic parsing or RAG-Logic) and those whose outputs remain natural language and are only implicitly symbolic.

They provide an extensive evaluation of the NSAI architectures based on robustness, scalability, and explainability. However, they do not discuss the cost-effectiveness and runtime complexity of these systems, particularly the practicality of using multi-agent systems at scale.

Thank you for this comment. We would like to clarify that cost-effectiveness and runtime complexity are already encompassed in our Scalability criterion, specifically through the sub-criteria on hardware efficiency and complexity management.

Review #3

- Summary: The paper builds upon, reviews and extends a taxonomy of neuro-symbolic AI architectures and relates generative-AI techniques onto these architectural patterns. It introduces a qualitative evaluation framework based on seven criteria (generalization, scalability, data efficiency, reasoning, robustness, transferability, and interpretability) to compare these architectures qualitatively. Finally, it outlines 4D printing as an illustrative application domain.

Thank you for this summary.

- Novelty: Needs to highlighted much better with respect to referenced existing literature in the introduction (a bullet ppoint list might help here)

Thank you for this suggestion. We agree that the novelty of the paper needed to be made more explicit with respect to existing NSAI literature, including Kautz’s taxonomy. In the revised manuscript, we have strengthened the Introduction by clearly highlighting the novel aspects of our work.

- Literature: The literature contains some ArXiv papers, that should be replaced by the actual publication, if not published yet due to still being under review preferably not metnioned here.

Thank you for this remark. We agree with the concern and have revised the bibliography accordingly: a large number of references that were previously cited as arXiv preprints have been replaced by their corresponding peer-reviewed journal or conference publications whenever available.

Title: Is not very suited for the paper. The paper is primarily about a taxonomy and evaluation of NSAI architectures, with generative AI and 4D printing as applications or perspectives.Choose a title that reflects this better

Thank you for the suggestion. We agree that the original title did not fully reflect the main contribution of the paper. In the revised manuscript, we have updated the title accordingly to better match the scope and novelty of the work.

- Abstract: The abstract lacks a clear statement of novelty (same is true for introduction) and it seems almost a bit overloaded, due to studying diverse NSAI architectures, highlighting their unique approaches, examines the alignment of contemporary AI techniques such as RAG, GNN, RL, and multiagent systems, evaluating these architectures against comprehensive set of criteria, comparing respective strengths and limitations. The abstract would also improve from a motivation (preferably the first two sentences in the abstract, eg. mentioning open challenges like robotics/embodied AI, life-long learning, ...). The Abstract does not mention the section on 4D printing. Finally the abstract would benefit from a closing sentence on the impact of the paper. The claim "Notably, the Neuro → Symbolic ← Neuro model consistently outperforms its counterparts across all evaluation metrics." is s strong claim, here only based on qualitative scoring instead of experimental benchmarking and therefore carefully should be rephrased and revised.

Thank you for this detailed feedback. We agree that the initial abstract did not sufficiently foreground the novelty, motivation, and impact, and that it appeared overly dense. In the revised manuscript, we have streamlined the abstract and added an explicit novelty statement highlighting our contributions beyond prior NSAI surveys and Kautz's taxonomy (clarification/formalization/extension of architectures, rigorous NSAI–GenAI mapping, and a qualitative literature-grounded evaluation framework). We also strengthened the opening motivation to emphasize why a unified NSAI taxonomy is needed in the context of fast-moving generative and embodied AI. The abstract now explicitly mentions the 4D-printing application to reflect that component of the paper. Finally, we revised the claim about the Neuro → Symbolic ← Neuro paradigm to a qualitative, literature-based wording rather than an empirical superiority statement, and we added a closing sentence stating the paper's expected impact on guiding reproducible architectural choices and future NSAI research.

- Introduction: See above: Please clarify novelty for this paper better. The paper uses well-stablished taxonomy/catgorizations for neurosymbolic AI.

Thank you for this comment. We agree that the originality needed to be stated more clearly given that we build on a well-established NSAI taxonomy. In the revised manuscript, we have reworked the Introduction to explicitly highlight our novelty and added a concise contribution list that clarifies how our work extends and operationalizes prior classifications.

- General Comments: Section 6 on 4D printing has no previous relationship to the topic. It is not well placed in this paper. There needs to be mentioning, or further use cases and application scenarios discussed. Please either omit the section, or revise the document to present the motivation for exactly this application better.

Thank you for this comment. We respectfully disagree that Section 6 is unrelated, but we agree that its motivation and role in the paper needed to be made clearer. In the revised manuscript, we explicitly position 4D printing as an illustrative application testbed for the taxonomy, because it is a domain where purely neural approaches face well-known limits (data scarcity, black-box behavior, and difficulty reasoning over structured material–geometry–function knowledge), making it a natural fit for neuro-symbolic architectures. We therefore strengthen the linkage earlier in the paper (Introduction and Abstract) and clarify that the goal of Section 6 is not to introduce a new topic, but to instantiate each NSAI paradigm in a concrete, multidisciplinary workflow.

GANs and MoEs are not simply not neuro-symbolic. Please revise the whole section building relationships here or omit it.

Thank you for this comment. We thank the reviewer for this helpful observation. Our intention was not to present standard GANs (with two purely neural components) as neuro-symbolic, but to use the GAN setting as an intuition for cooperative interaction. In the revised manuscript, we clarify this by removing the claim about generic GANs and instead describing GAN-inspired cooperative schemes, where at least one component is explicitly grounded in symbolic rules or logic-based constraints. Concerning Mixture-of-Experts (MoE), we also clarified the discussion: while MoE is a powerful neural scaling strategy, standard MoE architectures rely on a learned neural router and are not neuro-symbolic by default. We now state that MoE only aligns with a fibring-like paradigm in variants where expert selection or aggregation is explicitly mediated by symbolic rules or constraints. These revisions tighten the conceptual relationships and avoid overextending the taxonomy.

Physics-informed learning and probabilistic programming and their impact cold be discussed more.

Thank you for this suggestion. Our primary objective in this paper is to provide a clear taxonomy and classification of NSAI architectures and to map contemporary generative-AI methods onto these paradigms. Physics-informed learning is already covered in the manuscript as a key instance of the compiled paradigm.

- Overall Presentation/Evaluation: The Table 2 and 3 are neither self-explanatory, nor are they explained well enough in the text. It is not clear what is really assessed. This needs to be carefull revised.

Thank you for this comment. We agree that Tables 2 and 3 were not sufficiently self-explanatory and that the manuscript did not clearly state what was being assessed. In the revised version, we carefully reworked the presentation.