Ambiguity-Aware Optimization for Robust Neuro-Symbolic Learning

Dayana I. Agudo[1], **Christian F. Barreto**[1], **Igor C. Silveira**[1], **Denis D. Mauá**[1]

[1]Instituto de Matemática, Estatística e Ciência da Computação
Universidade de São Paulo – 05508-090 – São Paulo – SP – Brazil

{dagudo,cbarreto,igorcs,ddm}@ime.usp.br

Abstract. *Reasoning shortcuts are a well-known phenomenon in neuro-symbolic models, where high predictive performance is achieved through semantically incorrect concepts. While existing approaches mitigate reasoning shortcuts, they do not explicitly model the structural properties of the task that give rise to this behavior. This work defines structural task ambiguity as the multiplicity of concept configurations compatible with a symbolic output. Moreover, we propose an ambiguity-aware optimization and the RSR metric to quantify shortcut reliance. Experiments on MNIST tasks show that modeling ambiguity improves conceptual alignment maintaining competitive predictive performance, highlighting its importance for robust neuro-symbolic learning.*

1. Introduction

Neuro-symbolic (NeSy) Artificial Intelligence aims to combine the effective learning capabilities of neural networks with the reliable and interpretable nature of structured symbolic reasoning, enabling models that are more robust, interpretable, and capable of generalizing through prior knowledge [Colelough and Regli 2024, Yang et al. 2025, Acharya and Song 2026]. A central assumption in this paradigm is that symbolic constraints guide learning toward semantically meaningful concept representations. However, recent studies have shown that NeSy models can achieve high predictive performance while relying on concept assignments that deviate from the intended semantics [Marconato et al. 2023, Yang et al. 2024, Stein et al. 2025]. This phenomenon, known as *reasoning shortcuts* (RS), presents a major challenge to the reliability and interpretability of NeSy systems.

Reasoning shortcuts are not unique to NeSy models. In deep learning, Geirhos et al. [2020] describe shortcuts as decision rules that perform well on standard benchmarks but fail under distribution shifts. In the neuro-symbolic setting, Li et al. [2024] introduce *shortcut satisfaction*, where models exploit alternative concept configurations that satisfy logical constraints without being semantically correct. Similarly, Stein et al. [2025] describe *symbol hallucination*, where incorrectly learned concepts are guided by symbolic rules to produce correct outputs. From a probabilistic perspective, Marconato et al. [2023] show that RS can emerge as unintended optima consistent with prior knowledge, where models maximize likelihood while deviating from the (typically unknown) ground-truth concept distribution.

Prior work has shown that the emergence of RS is strongly influenced by both the structure of prior knowledge and the distribution of the data [Marconato et al. 2023,

Yang et al. 2024, Stein et al. 2025]. In this work, we argue that these factors are intrinsically connected through the *task ambiguity*, defined as the multiplicity of valid concept-level explanations for a given label. Intuitively, we argue that the larger the number of concept configurations that satisfy the symbolic constraints, the higher the chance that a RS will occur. As a consequence, pruning ambiguous data can be beficial to mitigate RS.

To illustrate this, consider the distant learning benchmark MNIST-Addition [Manhaeve et al. 2018], where the task is to learn how to map a pair of handwritten digit images to its sum (Figure 1). Prior knowledge is provided based on rules of arithmetic (e.g., $2+3=5$), and the data consists of pairs of MNIST images with their corresponding sum value. While the mapping from concepts to labels is fixed, multiple concept configurations may satisfy the same label constraint. In particular, different combinations of digit interpretations can produce identical sums, yet only one corresponds to the ground-truth semantic decomposition.

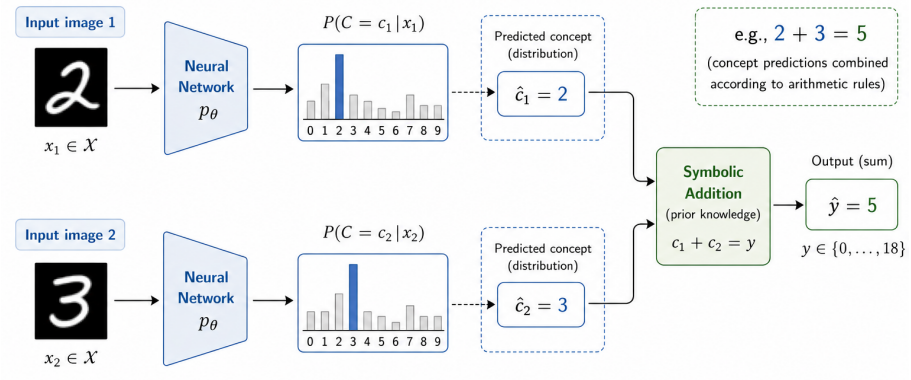


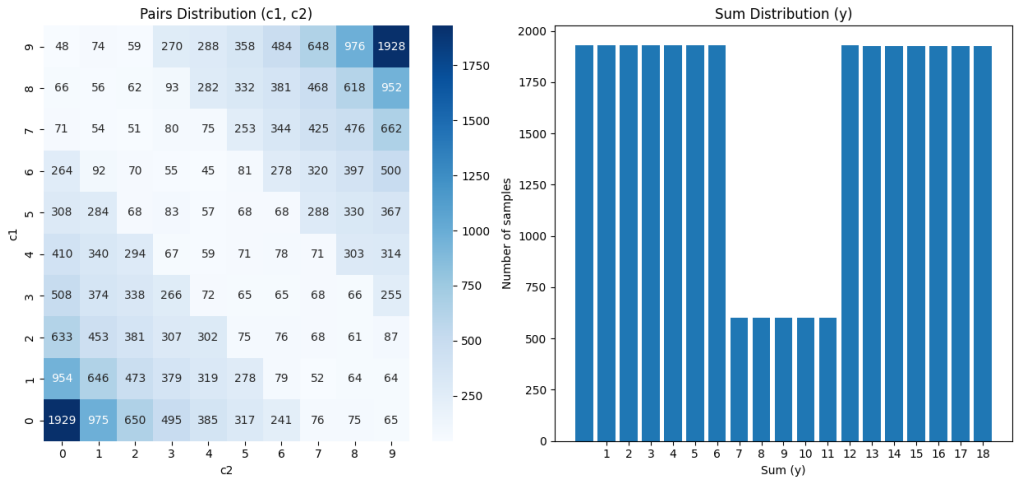
Figure 1. Illustration of the MNIST-Addition task. Two input images are processed by neural networks to predict digit concepts, which are then combined through symbolic addition to produce the final output.

This variability in the number of valid concept configurations induces different levels of structural ambiguity across the dataset, which in turn affects the likelihood of learning reasoning shortcuts. In practice, this phenomenon has a direct impact on learning dynamics. As shown in Figure 2, controlling the distribution of ambiguity levels leads to substantially different behaviors in terms of both predictive performance and reasoning shortcut metrics, even when label accuracy remains consistently high. These results highlight the importance of explicitly accounting for structural ambiguity during training.

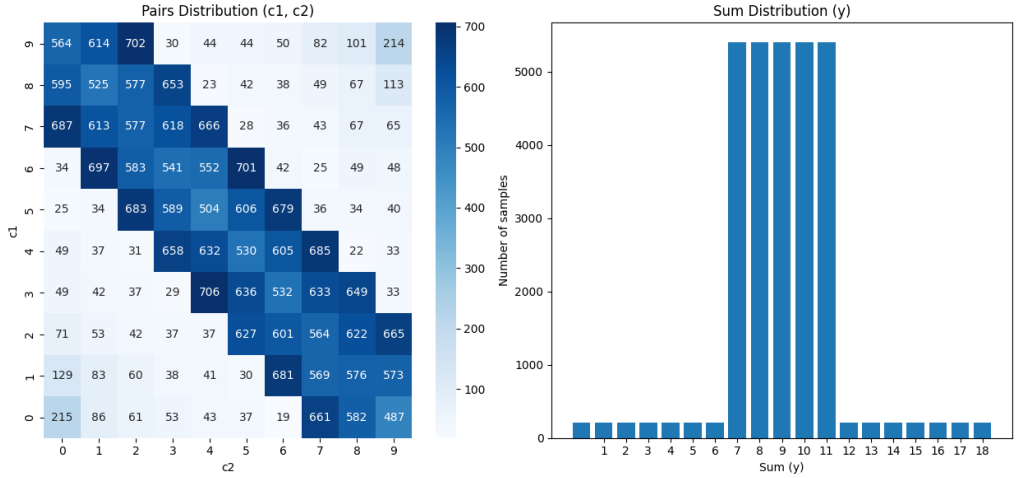
Despite recent progress in understanding RS, existing mitigation strategies primarily focus on modifying supervision signals, symbolic rules, or model architectures, without explicitly accounting for the structural ambiguity of the concept space. As a result, they do not directly control the conditions under which shortcut solutions emerge.

In this work, we address this gap by explicitly modeling task ambiguity and incorporating it into the learning process. Our contributions are threefold:

- We formalize **task ambiguity** in neuro-symbolic learning as the cardinality of concept configurations consistent with a given symbolic output, and show its role in enabling reasoning shortcuts.



(a) Datasets with a higher proportion of low-ambiguity labels (e.g., sums in $[0\text{-}6] \cup [12\text{-}18]$) lead to high accuracy both in labels and concepts (0.98 and 0.99 respectively)



(b) Datasets dominated by high-ambiguity labels (e.g., sums in $[7\text{-}11]$) result in significantly degraded concept accuracy (0.002), even when label accuracy remains relatively high (0.78)

Figure 2. Effect of dataset ambiguity on MNIST-Addition. While predictive accuracy remains relatively high across settings, concept accuracy and reasoning shortcut behavior vary significantly, highlighting the impact of structural ambiguity.

- We introduce the **Reasoning Shortcut Ratio (RSR)**, a metric that quantifies the extent to which correct predictions rely on semantically incorrect concept assignments.
- We propose an **Ambiguity-Aware Loss (AAL)** that incorporates structural ambiguity into the training objective, biasing learning toward semantically meaningful solutions.

We evaluate our approach on MNIST-Addition [Manhaeve et al. 2018] and MNIST-Max, two benchmark tasks with controlled ambiguity levels. Our results demonstrate that traditional accuracy metrics can mask significant differences in reasoning behavior, while explicitly accounting for ambiguity leads to improved semantic alignment and reduced reasoning shortcuts.

2. Related Work

In this section, we discuss Neuro-symbolic learning frameworks, then we move to the causes of RS, metrics for RS, strategies to mitigate it, and then situate our work in light of this literature.

Neuro-symbolic learning frameworks. Neuro-symbolic (NeSy) learning integrates neural perception with symbolic reasoning through probabilistic logic programming. Frameworks such as DeepProbLog [Manhaeve et al. 2018, Manhaeve et al. 2021], NeurASP [Yang et al. 2020], DPASP [Geh et al. 2023, Geh et al. 2024], and Scallop [Huang et al. 2021, Li et al. 2023] enable joint reasoning over neural predictions and logical constraints using techniques such as weighted model counting and differentiable inference. These systems support both two-stage training pipelines and end-to-end learning via distant supervision, where gradients are propagated through the symbolic component. In this work, we build on these frameworks to study how structural properties of the task influence learning behavior under ambiguous supervision.

Causes of reasoning shortcuts. The emergence of reasoning shortcuts (RS) has been attributed to multiple interacting factors. Marconato et al. [Marconato et al. 2023] identify four main sources: (i) the structure of prior knowledge, which constrains the solution space; (ii) dataset bias, which can bias learning toward shortcuts; (iii) the learning objective, which may not enforce semantic consistency; and (iv) the architecture of the concept extractor, which determines the expressivity of concept mappings. Subsequent work [Yang et al. 2024, Stein et al. 2025] further highlights the role of data complexity and rule complexity, showing that RS can persist even in well-specified symbolic systems.

These findings suggest that ambiguity arises from the structure of the underlying concept space, where multiple concept-level configurations remain compatible with a single label. While this structural mapping plays a central role in connecting data representation and knowledge complexity, prior frameworks have typically treated these factors independently, leaving this unifying aspect implicitly assumed rather than explicitly modeled or incorporated into the training objective.

Metrics for reasoning shortcuts. Several metrics have been proposed to analyze RS from different perspectives. Structural metrics, such as deterministic optima [Marconato et al. 2023, Bortolotti et al. 2024], quantify the number of concept configurations consistent with symbolic constraints, capturing the intrinsic ambiguity of the task. However, they do not reflect the behavior of learned models.

Representation-based metrics, such as concept collapse [Bortolotti et al. 2024], measure whether learned representations merge distinct semantic concepts, but do not directly capture whether correct predictions rely on semantically valid reasoning paths. Behavioral approaches, such as model comparison frameworks (e.g., BEARS [Marconato et al. 2024]), infer inconsistencies in predicted concepts across models, but remain indirect and computationally expensive.

Significantly, none of these metrics explicitly quantify the extent to which individual model predictions rely on semantically incorrect but symbolically valid concept configurations, which is the key notion captured by reasoning shortcuts in this work.

Mitigation strategies. Existing approaches to mitigating RS can be broadly catego-

rized into four groups. First, supervision-based methods, such as concept supervision [Marconato et al. 2023], enforce alignment between predicted and ground-truth concepts, while multi-task learning leverages auxiliary tasks to reduce ambiguity in representations. Second, knowledge-based approaches, including knowledge editing and abductive weak supervision [Yang et al. 2024], modify or augment the symbolic component to constrain the solution space. Third, regularization-based techniques, such as entropy minimization [Marconato et al. 2023] and representation smoothing [Yang et al. 2024], encourage more structured latent representations. Finally, architectural approaches, such as disentangled concept models [Marconato et al. 2023], restrict the hypothesis space through inductive biases.

Despite their effectiveness, these methods do not explicitly model the structure of the concept space that gives rise to ambiguity. As a result, they address the symptoms of reasoning shortcuts rather than their underlying structural cause.

Positioning of this work. In contrast to prior approaches, which mitigate reasoning shortcuts through supervision, architectural constraints, or symbolic refinement, we propose to explicitly model *task ambiguity* as a structural property of neuro-symbolic problems. Our approach directly quantifies the multiplicity of concept-level explanations consistent with a label and incorporates this structure into the learning objective, providing a principled mechanism to reduce reasoning shortcuts without requiring additional supervision.

3. Methodology

In this section, we formalize the notion of structural ambiguity in neuro-symbolic tasks, introduce the evaluation metrics used to quantify reasoning shortcuts, and present an ambiguity-aware learning objective designed to mitigate their emergence.

3.1. Structural Ambiguity in Neuro-Symbolic Tasks

We consider neuro-symbolic (NeSy) models as functions composed of a neural perception module and a symbolic reasoning operator. Formally, we define a NeSy model as $\mathbf{y} = \beta(\mathbf{c}), \mathbf{c} \sim p_\theta(\mathbf{c} \mid \mathbf{x})$, where $p_\theta(\mathbf{c} \mid \mathbf{x})$ is a neural network that maps inputs $\mathbf{x} \in \mathbb{R}^n$ to a distribution over latent concept vectors $\mathbf{c} \in \mathcal{C}$ (e.g., $\mathcal{C} = \{0, \dots, 9\}^2$ for MNIST-Addition), and $\beta : \mathcal{C} \to \mathcal{Y}$ is a deterministic symbolic function that maps concepts to labels $\mathbf{y}$. This induces a distribution over labels given inputs $p_\theta(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{c} \in \mathcal{C}} \mathbb{I}[\beta(\mathbf{c}) = \mathbf{y}]\, p_\theta(\mathbf{c} \mid \mathbf{x})$.

A key characteristic of NeSy tasks is that multiple concept configurations may lead to the same label. This induces, for each label $\mathbf{y}$, the preimage of β as $\mathcal{C}_\mathbf{y} = \beta^{-1}(\mathbf{y}) = \{\mathbf{c} \in \mathcal{C} \mid \beta(\mathbf{c}) = \mathbf{y}\}$.

We define *structural ambiguity* of a label $\mathbf{y}$ as the cardinality of this set as $A(\mathbf{y}) = |\mathcal{C}_\mathbf{y}|$. Intuitively, labels with larger $A(\mathbf{y})$ admit a greater number of concept-level explanations consistent with the symbolic rules, making the learning problem more underdetermined at the semantic level. In practice, for the MNIST-based tasks considered in this work, $A(\mathbf{y})$ is computed analytically from the symbolic composition rules defining each task.

3.2. Evaluation Metrics

We evaluate models at three levels: label prediction, concept prediction, and reasoning consistency. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ be the dataset, and let $\hat{y}_i$ and $\hat{\mathbf{c}}_i$ denote the predicted

label and concept configuration, respectively.

Label accuracy.

$$\mathrm{Acc}^{Y} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i] \qquad (1)$$

Concept accuracy.

$$\mathrm{Acc}^{C} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{\mathbf{c}}_i = \mathbf{c}_i] \qquad (2)$$

Reasoning Shortcut Ratio (RSR). We define the set of correctly classified samples as $\mathcal{A} = \{i \mid \hat{y}_i = y_i\}$ and the subset of samples where correct prediction is not supported by semantically correct concepts as $\mathcal{A}_{RS} = \{i \in \mathcal{A} \mid \hat{\mathbf{c}}_i \neq \mathbf{c}_i\}$. Thus, the Reasoning Shortcut Ratio, which quantifies the proportion of correct predictions unsupported by correct concepts, is defined as:

$$RSR = \frac{|\mathcal{A}_{RS}|}{|\mathcal{A}|} \qquad (3)$$

3.3. Ambiguity-Aware Loss Function

Standard neuro-symbolic models learn via maximum likelihood [Marconato et al. 2023]. However, such objectives treat all samples uniformly, ignoring the structural ambiguity of the underlying task.

To address this limitation, we propose an ambiguity-aware reweighting scheme. Let $p_\theta(y_i \mid x_i)$ denote the predicted probability of the correct label. We define the loss function as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w(y_i) \cdot \log p_\theta(y_i \mid x_i), \qquad (4)$$

where the ambiguity-based weighting function is defined as $w(y_i) = \log\left(1 + \frac{\alpha}{A(y_i)}\right)$. Here, $\alpha > 0$ is a hyperparameter that controls the influence of ambiguity on the learning process.

3.4. Experimental Setup

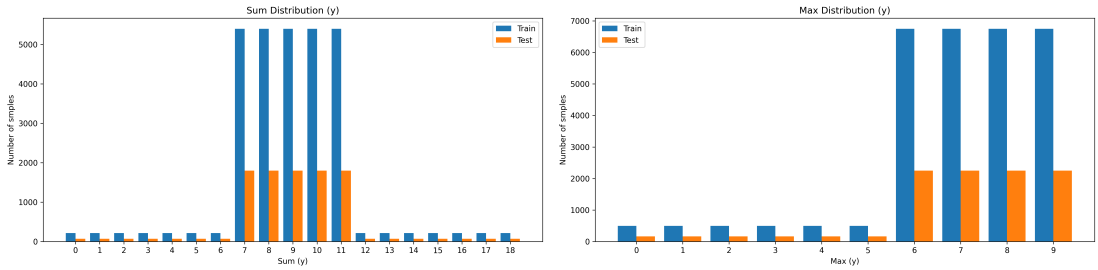
We evaluate our approach on two neuro-symbolic benchmark tasks: *MNIST-Addition* and *MNIST-Max*, both constructed from the MNIST dataset [LeCun et al. 2010].

In MNIST-Addition [Manhaeve et al. 2018], the goal is to predict the sum of two handwritten digits, while in MNIST-Max, the task is to predict the maximum value between two digits. Both tasks are implemented using the Scallop framework [Huang et al. 2021], which enables differentiable integration of neural perception and symbolic reasoning.

To investigate the effect of structural ambiguity on learning, we construct ambiguity-controlled versions of the datasets by modifying the distribution of labels while preserving the underlying symbolic reasoning rules. Specifically, we increase the frequency of labels associated with higher ambiguity $A(\mathbf{y})$. For MNIST-Addition, we increase the proportion of sum labels 7–11, whereas for MNIST-Max, we oversample maximum labels 6–9, corresponding to the highest ambiguity values in each task. Figure 3

illustrates the resulting label distributions. This design allows us to examine whether increasing the frequency of highly ambiguous examples affects label prediction, concept prediction, and the occurrence of reasoning shortcuts, without modifying the underlying symbolic reasoning task.

As baseline, we use a standard Scallop configuration with Binary Cross-Entropy (BCE) loss. We compare this baseline against our proposed Ambiguity-Aware Loss (AAL). We report label accuracy (Acc^{Y}), concept accuracy (Acc^{C}), and the Reasoning Shortcut Ratio (RSR) to evaluate both predictive performance and semantic alignment. To assess statistical robustness, each experiment was repeated five times with different random seeds, and results are reported as mean $\pm$ standard deviation.



(a) Synthetic MNIST-Addition dataset contains a higher proportion of 7-11 sum labels

(b) Synthetic MNIST-Max dataset contains a higher proportion of 6-9 max labels

Figure 3. High-ambiguity datasets generated with a higher proportion of ambiguous labels.

4. Results and Discussion

This section compares AAL and the BCE baseline on two neuro-symbolic benchmark tasks: MNIST-Addition and MNIST-Max, focusing on predictive performance and semantic alignment.

4.1. MNIST-Addition Results

Figure 4 compares the performance of BCE and AAL on the ambiguity-controlled MNIST-Addition dataset. Overall, AAL consistently outperforms the BCE baseline across the evaluated metrics.

In terms of label accuracy, AAL achieves consistently high performance, reaching approximately 97% across runs, while the BCE baseline reaches approximately 76%. A more pronounced difference is observed at the concept level. AAL achieves substantially higher concept accuracy, indicating that its predicted concepts are more consistent with the ground-truth concepts. In contrast, the BCE baseline exhibits considerably lower concept accuracy, suggesting that its predictions are less aligned with the intended concept semantics.

The difference in concept quality is also reflected in the Reasoning Shortcut Ratio (RSR). AAL consistently produces near-zero RSR values across the five runs, whereas BCE exhibits substantially higher RSR values. This indicates that, although BCE can achieve relatively high label accuracy, it is more frequently associated with concept configurations that produce correct labels while deviating from the intended semantics.

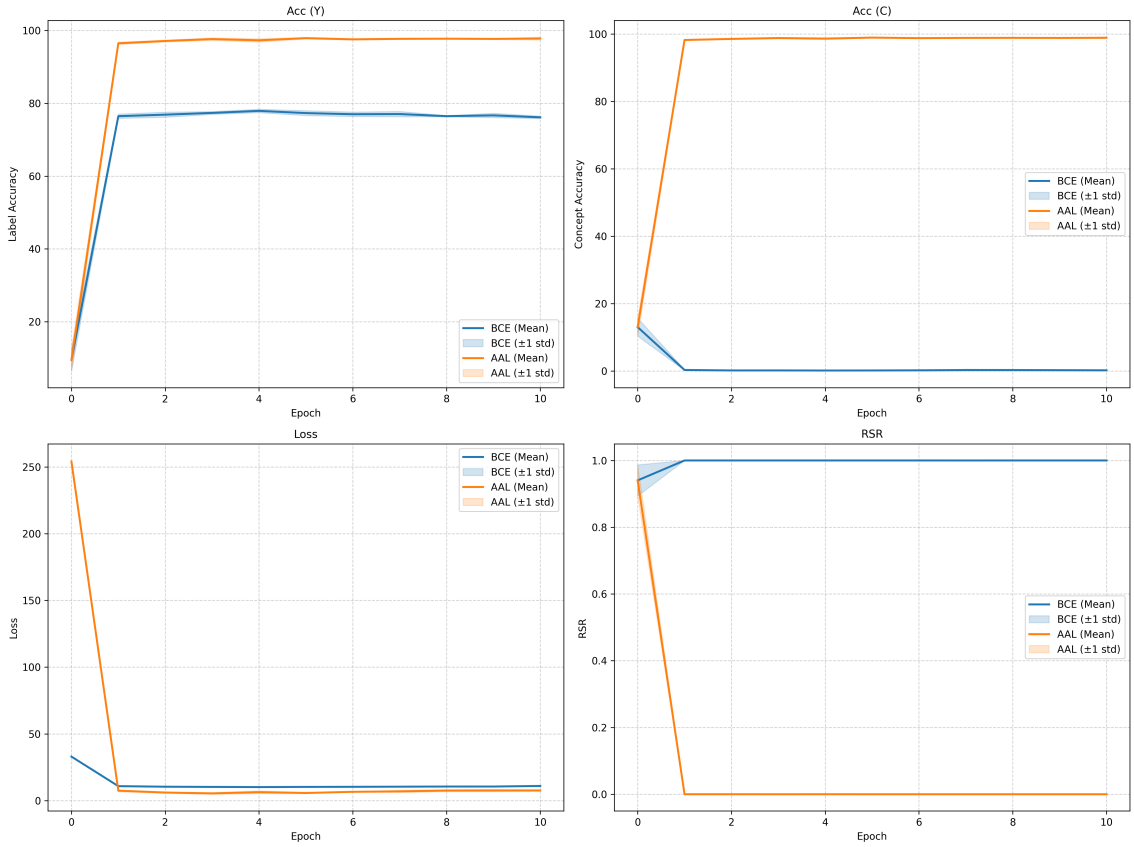


Figure 4. Test performance comparison on MNIST-Addition using BCE and AAL.

4.2. MNIST-Max Results

Figure 5 reports the results for the MNIST-Max task, which involves a higher degree of structural ambiguity than MNIST-Addition.

In this setting, AAL and the BCE baseline exhibit very similar performance across the evaluated runs. Both methods achieve approximately 98% label accuracy and 98% concept accuracy, indicating that AAL can match the predictive and concept-level performance of the BCE baseline in this task.

The RSR results show a similar pattern. For both methods, the reasoning shortcut ratio remains close to zero throughout evaluation, providing no evidence of significant shortcut behavior under the current experimental conditions. Importantly, this should not be interpreted as evidence that reasoning shortcuts are absent from the task, but rather that they are not manifested or detected under the current experimental conditions.

These results suggest that AAL does not necessarily provide a performance advantage in tasks where no significant reasoning shortcut behavior is detected. Instead, it can achieve performance comparable to the BCE baseline while maintaining similarly low levels of detected shortcut behavior.

4.3. Discussion

Overall, the results suggest that explicitly accounting for structural ambiguity can improve the semantic alignment of neuro-symbolic models without necessarily compromising pre-

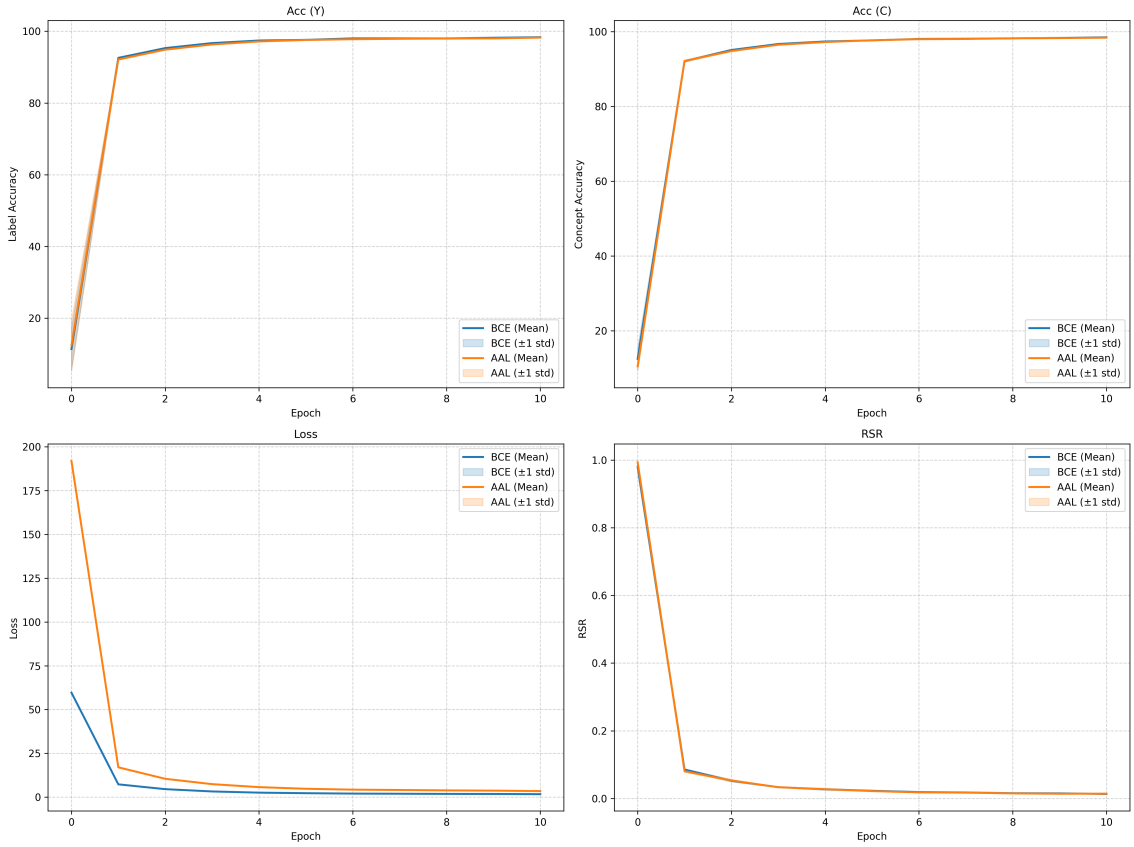


Figure 5. Test performance comparison on MNIST-Max using BCE and AAL.

dictive performance. This effect is particularly clear in MNIST-Addition, where AAL achieves substantially higher concept accuracy and lower RSR than the BCE baseline, while maintaining high label accuracy.

These results also highlight the limitations of evaluating neuro-symbolic models solely based on label accuracy. A model may achieve a high Acc^Y while relying on concept predictions that do not correspond to their intended semantics. In our experiments, this behavior is reflected by the BCE baseline on MNIST-Addition, which achieves relatively high label accuracy despite substantially lower concept accuracy and higher RSR. Thus, evaluating Acc^C and RSR alongside Acc^Y provides a more informative assessment of whether the model is solving the task using the intended concepts.

Interestingly, the advantage of AAL is not observed to the same extent in MNIST-Max. Despite the higher ambiguity associated with the selected labels in this task, AAL and BCE achieve comparable label and concept accuracy, and both exhibit low RSR. This suggests that the presence of highly ambiguous examples alone does not necessarily lead to detectable reasoning shortcuts under the experimental conditions considered here. Rather, the effect of ambiguity may depend on its interaction with the symbolic structure, the data distribution, and the learning objective.

Altogether, these findings provide preliminary evidence that ambiguity is a relevant factor in the learning dynamics of neuro-symbolic models, but they do not establish a direct causal relationship between ambiguity and reasoning shortcuts. Further exper-

iments varying the degree and distribution of ambiguity independently across a broader range of datasets and symbolic structures are therefore needed to determine under which conditions ambiguity-aware learning is most effective.

5. Conclusion

In this work, we investigated the role of structural ambiguity in neuro-symbolic learning and its relationship with reasoning shortcuts. We argued that ambiguity in the concept mapping is a fundamental property of neuro-symbolic tasks that directly influences the emergence of semantically inconsistent yet prediction-correct solutions.

To address this issue, we proposed an ambiguity-aware learning framework that explicitly incorporates the cardinality of valid concept configurations into the optimization process. Additionally, we introduced the Reasoning Shortcut Ratio (RSR) as a metric to quantify the extent to which correct predictions rely on semantically incorrect concept assignments.

Our experiments on MNIST-Addition and MNIST-Max provide preliminary evidence that accounting for structural ambiguity can improve conceptalignment without compromising predictive performance. In particular, on MNIST-Addition, AAL substantially improves concept accuracy and reduces the RSR while maintaining high label accuracy compared with the BCE baseline. This result highlights that a model can achieve relatively high predictive performance while still learning concepts that are inconsistent with their intended semantics. In contrast, on MNIST-Max, AAL and BCE achieve comparable label and concept accuracy, with both methods exhibiting low RSR. This indicates that the presence of highly ambiguous examples alone does not necessarily lead to detectable reasoning shortcuts.

Overall, these results suggest that structural ambiguity is a relevant factor in the learning dynamics of neuro-symbolic models, but that its effect on reasoning shortcuts depends on its interaction with the symbolic structure, data distribution, and learning objective. Therefore, ambiguity should not be regarded as a sufficient condition for the emergence of reasoning shortcuts. Rather, it provides a principled property of the task that can be explicitly considered when designing and evaluating neuro-symbolic learning systems. More generally, our results reinforce the importance of evaluating neuro-symbolic models beyond label accuracy by jointly considering concept quality and reasoning-shortcut behavior.

Limitations. This work has several limitations. First, the notion of structural ambiguity is currently defined at the level of symbolic task structure and computed analytically for the considered benchmarks, which may limit its applicability to more complex neuro-symbolic systems where ambiguity is not explicitly tractable. Second, the evaluation relies on synthetic MNIST-based tasks, which, although controlled, may not fully capture the complexity of real-world reasoning scenarios. Finally, the proposed RSR metric depends on the availability of ground-truth concept annotations, which may not always be accessible in practical settings.

Future work. Future research will focus on extending the notion of structural ambiguity to more expressive neuro-symbolic domains, where concept spaces are continuous or partially observable. Another promising direction is the development of approximate or

weakly supervised variants of ambiguity estimation that do not require full knowledge of the symbolic configuration space. Additionally, integrating ambiguity-aware learning with large-scale neural architectures and evaluating its impact in open-world settings represents an important step toward improving the robustness and interpretability of neuro-symbolic systems.

Acknowledgements

This work was partially supported by São Paulo Research Foundation (FAPESP) grants no. 2022/02937-9, 2025/17392-6, 2025/17393-2 and 2026/15768-1; CNPq grant no. 301420/2026-2, CAPES Finance Code 001.

References

Acharya, K. and Song, H. (2026). A comprehensive review of neuro-symbolic ai for robustness, uncertainty quantification, and intervenability. *Arabian Journal for Science and Engineering*, 51(1):35–67.

Bortolotti, S., Marconato, E., Carraro, T., Morettin, P., van Krieken, E., Vergari, A., Teso, S., and Passerini, A. (2024). A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Colelough, B. C. and Regli, W. (2024). Neuro-symbolic ai in 2024: A systematic review. In *Proceedings of the First International Workshop on Logical Foundations of Neuro-Symbolic AI*, volume 3819 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org.

Geh, R. L., Gonçalves, J., Silveira, I. C., Mauá, D. D., and Cozman, F. G. (2023). dpasp: A comprehensive differentiable probabilistic answer set programming environment for neurosymbolic learning and reasoning.

Geh, R. L., Gonçalves, J., Silveira, I. C., Mauá, D. D., and Cozman, F. G. (2024). dPASP: A Probabilistic Logic Programming Environment For Neurosymbolic Learning and Reasoning. In *Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning*, pages 731–742.

Huang, J., Li, Z., Chen, B., Samel, K., Naik, M., Song, L., and Si, X. (2021). Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, page 25134–25145. Curran Associates, Inc.

LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Li, Z., Huang, J., and Naik, M. (2023). Scallop: A language for neurosymbolic programming. *Proceedings of the ACM on Programming Languages*, 7(4):1463–1487.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., and De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T., and De Raedt, L. (2021). Neural probabilistic logic programming in deepproblog. *Artificial Intelligence*, 298:103504.

Marconato, E., Bortolotti, S., van Krieken, E., Vergari, A., Passerini, A., and Teso, S. (2024). BEARS make neuro-symbolic models aware of their reasoning shortcuts. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*. JMLR.org.

Marconato, E., Teso, S., Vergari, A., and Passerini, A. (2023). Not all neuro-symbolic concepts are created equal: analysis and mitigation of reasoning shortcuts. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Stein, A., Velingker, N., Naik, M., and Wong, E. (2025). Neuro-symbolic programming in the age of foundation models: Pitfalls and opportunities. *Proceedings of Machine Learning Research*, pages 1–18.

Yang, X.-W., Shao, J.-J., Guo, L.-Z., Zhang, B.-W., Zhou, Z., Jia, L.-H., Dai, W.-Z., and Li, Y.-F. (2025). Neuro-symbolic artificial intelligence: towards improving the reasoning abilities of large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 10770–10778.

Yang, X.-W., Wei, W.-D., Shao, J.-J., Li, Y.-F., and Zhou, Z.-H. (2024). Analysis for abductive learning and neural-symbolic reasoning shortcuts. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56524–56541. JMLR.org.

Yang, Z., Ishay, A., and Lee, J. (2020). Neurasp: Embracing neural networks into answer set programming. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 1755–1762.