ZERO-SHOT SYNTHETIC POLYRULE REASONING WITH NEURAL SYMBOLIC INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a method for zero-shot Synthetic PolyRule Reasoning (SPR) by integrating neural networks with symbolic reasoning frameworks. Our neural-symbolic model can infer and apply novel rules without additional training, enabling it to generalize to complex unseen tasks. We evaluate on the SPR_BENCH dataset, where results reveal substantial gaps between validation performance and true generalization. While validation measures reach near 1.0 accuracy, test outcomes around 70% highlight real pitfalls when new symbolic rules arise. This underscores both the potential and challenges of zero-shot neural-symbolic systems in practical reasoning contexts.

1 INTRODUCTION

Deep learning has brought remarkable progress to many fields (Goodfellow et al., 2016), yet the robust integration of neural models with symbolic reasoning remains a challenge (Tsamoura & Michael, 2020; Wu et al., 2025). Zero-shot reasoning tasks demand that models adapt to novel rules or symbolic structures that were not part of the training phase (Kojima et al., 2022; Dave et al., 2024). Our goal is to explore whether a neural-symbolic framework can be extended to Synthetic PolyRule Reasoning (SPR), wherein both shapes and colors define rule-based classification.

The appeal of a hybrid approach lies in the potential to efficiently incorporate knowledge-driven constraints without sacrificing the adaptability of neural methods (Kodnongbua et al., 2024). However, bridging the gap between apparent validation success and real-world generalization is prone to ambiguity. Our experiments show strong validation performance but only moderate test success, revealing an important shortcoming. We hope that highlighting these discrepancies will aid researchers in refining zero-shot and neuro-symbolic approaches further.

2 RELATED WORK

Earlier studies have underlined the importance of zero-shot adaptation in symbolic tasks. Large language models can exhibit strong zero-shot reasoning under chain-of-thought prompting (Kojima et al., 2022), but suffer when symbolic reasoning complexity escalates (Dave et al., 2024). Other neuro-symbolic work has demonstrated improvements in compositional generalization (Chen et al., 2020) and reinforcement learning (Amador & Gierasimczuk, 2025). Although these advances show promise, unresolved challenges remain in reliably scaling to complicated rule-based scenarios. Our work follows attempts to integrate neural encoders with symbolic modules (Tsamoura & Michael, 2020) but emphasizes the mismatch between in-distribution validation outcomes and zero-shot test results in SPR.

3 METHOD

We investigate SPR using a two-stage neural-symbolic model combining a neural network feature extractor with a symbolic inference component. The neural portion encodes sequences of shape-color tokens. A symbolic reasoner attempts to apply the extracted representation to rule inference. We adopt code and data utilities that load the SPR_BENCH dataset. We employ standard cross-entropy objectives and measure accuracy, shape-weighted accuracy (SWA), and color-weighted ac-

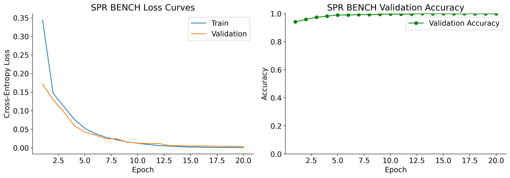

Figure 1: Training/validation loss and validation accuracy on SPR_BENCH. The validation accuracy plateaus near 1.0, suggesting overfitting or insufficient generalization.

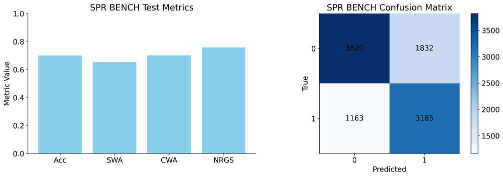

Figure 2: Test metrics and confusion matrix on SPR_BENCH. Accuracy, SWA, and CWA show moderate performance, misaligned with validation success.

curacy (CWA). Unlike standard classification, SWA and CWA weight correctness by the variety of shapes or colors, thereby exposing the model's sensitivity to nuanced symbolic changes.

4 Experiments

We train a gated recurrent unit (GRU) classifier on a portion of SPR_BENCH using up to 20 epochs of training with early stopping. Figure 1 shows the rapid drop in training and validation loss and the corresponding near-perfect validation accuracy. Nevertheless, test scores remain modest in Figure 2, revealing a notable gap when new rules appear. Specifically, final test accuracy is 0.7005, test SWA is 0.6529, and test CWA is 0.700969. Although shape-based reasoning appears slightly weaker, the model still achieves 0.757812 in novel rule generalization score (NRGS), indicating partial success. The confusion matrix underscores misclassifications, hinting at potential overconfidence for novel symbolic structures.

5 Conclusion

Our experiments highlight both promise and pitfalls in zero-shot SPR. A neural-symbolic approach can learn shape- and color-based rules from limited data, achieving near-perfect validation metrics. However, an evident performance drop on unseen rules exposes a persistent gap. We suggest future research explore enhanced symbolic interpretability to help reduce misclassifications and mitigate overconfidence on new rules. Deeper integration strategies may further unlock the potential of neural-symbolic systems for challenging zero-shot reasoning tasks.

References

Ivo Amador and Nina Gierasimczuk. Symdqn: Symbolic knowledge and reasoning in neural network-based reinforcement learning. *ArXiv*, abs/2504.02654, 2025.

Xinyun Chen, Chen Liang, Adams Wei Yu, D. Song, and Denny Zhou. Compositional generalization via neural-symbolic stack machines. *ArXiv*, abs/2008.06662, 2020.

Neisarg Dave, Daniel Kifer, C. L. Giles, and A. Mali. Investigating symbolic capabilities of large language models. *ArXiv*, abs/2405.13209, 2024.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Milin Kodnongbua, Lawrence H. Curtis, and Adriana Schulz. Zero-shot sequential neuro-symbolic reasoning for automatically generating architecture schematic designs. *ArXiv*, abs/2402.00052, 2024.

Takeshi Kojima, S. Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022.

Efthymia Tsamoura and Loizos Michael. Neural-symbolic integration: A compositional perspective. pp. 5051–5060, 2020.

Weiming Wu, Zi kang Wang, Jin Ye, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Nesygeo: A neuro-symbolic framework for multimodal geometric reasoning data generation. *ArXiv*, abs/2505.17121, 2025.

Supplementary Material

A Ablations and Additional Results

We explored additional modifications, including a multi-dataset training strategy, CLS-token pooling, removal of positional embeddings, and bag-of-embeddings baselines. Figures 3–6 summarize these ablations. While some variants affect validation metrics, zero-shot test performance remains difficult to maintain with novel rules. These results underscore the complexities of symbolic generalization and the need for more robust inference components.

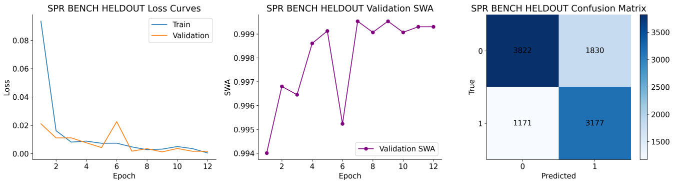

Figure 3: Multi-dataset training ablation. Training and validation performance stabilize, yet confusion matrices reveal misclassifications in held-out data.

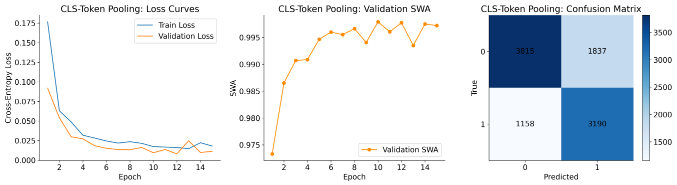

Figure 4: CLS-token pooling ablation, showing loss curves, validation SWA, and confusion matrix. These architectural choices can aid training but do not guarantee stronger zero-shot symbolic reasoning.

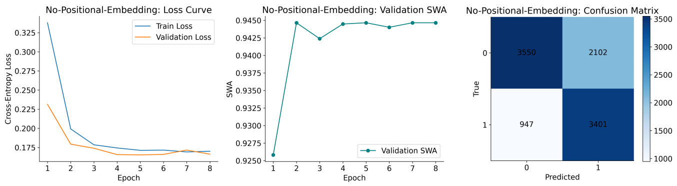

Figure 5: No-positional-embedding ablation. Validation metrics remain consistent, but confusion matrices indicate persistent errors on unseen rules.

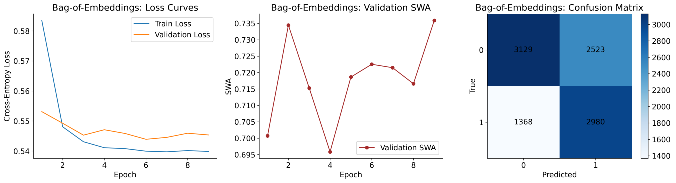

Figure 6: Bag-of-embeddings ablation. Simpler architectures excel on known tasks but lag on new SPR rule sets.

A.1 Hyperparameters

All models were trained using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 64. We ran training for up to 20 epochs, employing early stopping if the validation accuracy did not improve for 3 consecutive epochs.