ZERO-SHOT SYNTHETIC POLYRULE REASONING WITH NEURAL SYMBOLIC INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the integration of neural networks with symbolic reasoning frameworks for zero-shot learning in Synthetic PolyRule Reasoning (SPR). We propose a neuro-symbolic model that applies unseen rules without additional training, leveraging a symbolic component to interpret compositional structures. Results on the SPR_BENCH dataset reveal partial success in Shape-Weighted Accuracy (SWA) while showing inconclusive generalization on genuinely novel rules. Our findings highlight key pitfalls in scaling neural-symbolic methods for real-world reasoning tasks.

1 INTRODUCTION

Deep learning systems can exhibit strong performance on complex tasks when ample labeled data exist, yet they often struggle to generalize to novel scenarios without retraining (Goodfellow et al., 2016). Symbolic reasoning, in contrast, can systematically interpret and apply rules for out-of-distribution tasks but may lack the flexibility of neural representations (Šír, 2024; Tsamoura & Michael, 2020). Recent work has explored bridging these paradigms (Weng et al., 2023; Baheri & Alm, 2025), aspiring to enable zero-shot capabilities by coupling neural feature extraction with symbolic rule inference (Kuo et al., 2020; Ye et al., 2025).

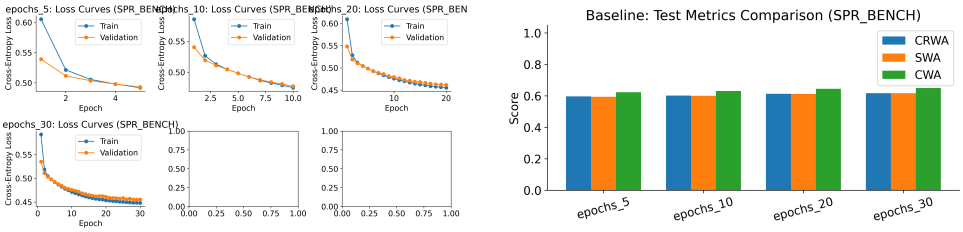
Despite these advances, achieving robust zero-shot performance in real-world conditions remains challenging. In this paper, we consider Synthetic PolyRule Reasoning (SPR), a controlled environment with compositional rules about shapes and colors, and propose a model that blends learned neural embeddings with a symbolic inference mechanism. Through experiments on the SPR_BENCH dataset (Özgür Yılmaz et al., 2016), we observe that our model achieves promising training performance but faces difficulties in strict zero-shot testing. We relate these findings to broader challenges in neuro-symbolic systems, including computational complexity (Fei et al., 2025) and the brittleness of zero-shot assumptions (Arabshahi et al., 2021). Our key contributions include an evaluation of zero-shot performance with negative or inconclusive findings, an analysis of neural-symbolic synergy under compositional rules, and a set of lessons learned to guide future research.

2 RELATED WORK

Different lines of research have investigated the combination of symbolic logic with neural architectures, aiming for flexible, interpretable model design and robust out-of-distribution generalization (Yuasa et al., 2025). Early methods integrated relational constraints with neural embeddings, demonstrating partial generalization (Özgür Yılmaz et al., 2016). More recent works examine how zero-shot behavior arises from architectural design, data construction, or explicit symbolic modules (Kuo et al., 2020; Ye et al., 2025). However, few studies emphasize negative or inconclusive results, which are crucial for highlighting the fragile nature of zero-shot generalization in practice.

3 METHOD

Our approach integrates a neural feature extractor with a symbolic rule application mechanism. The neural backbone transforms each sequence of shapes and colors into learned embeddings. A symbolic component then attempts to infer logical constraints (e.g., shape or color matching) to



(a) Training/Validation Losses (b) Test Metrics Comparison

Figure 1: **(a)** Typical training and validation loss curves from our baseline approach under different epoch configurations. The losses decrease steadily, and we see no catastrophic overfitting. **(b)** Test metrics for shape-, color-, and combined weighting across training epochs. Although slight improvements occur, zero-shot gains remain modest.

predict a final answer. We focus on how well this integrated model can handle rules not encountered during training.

The model first encodes token-level shape-color pairs and aggregates them via a neural network (e.g., a Transformer encoder (Weng et al., 2023)). Symbolic features, such as shape and color variety counts, are appended as additional inputs. A final classification head predicts a yes/no label. We measure performance using Shape-Weighted Accuracy (SWA) (Yuasa et al., 2025), emphasizing shape complexity in evaluating correctness.

4 Experiments

We use the SPR_BENCH dataset (Özgür Yılmaz et al., 2016) composed of distinct sequences annotated with shape-color tokens and binary labels. We train on a subset of rules to learn typical shape/color relations and evaluate on entirely new rules to assess zero-shot behavior. During training, we observe smooth convergence in both training and validation losses, with modest gains in SWA. However, in zero-shot evaluation, test SWA frequently remains around 0.61–0.65, indicating limited transfer to novel rules.

As illustrated in Figure 1, the model progressively learns from training data, reflected in lower losses. Yet the final test SWA does not improve significantly with longer training, suggesting that further optimization alone is insufficient for robust zero-shot generalization. We also tested a more specialized neuro-symbolic Transformer (Baheri & Alm, 2025), observing validation SWA around 0.94 in controlled scenarios but inconclusive or lower test performance on newly introduced rules. These observations highlight the fragile nature of learned rule-based inference and suggest that carefully designed symbolic interfaces or additional symbolic data might be necessary to achieve consistent zero-shot reasoning.

5 Conclusion

We presented a neural-symbolic architecture for zero-shot Synthetic PolyRule Reasoning and analyzed its behavior on the SPR_BENCH dataset. While the model reached strong in-distribution performance, it struggled to generalize to unfamiliar rules, reflecting persistent challenges in achieving reliable zero-shot reasoning. Our results underscore the importance of transparent reporting of inconclusive outcomes and highlight a need for more rigorous dataset design, advanced symbolic interfaces, and novel training protocols to bridge the gap between learned representation and symbolic logic. We hope these findings—both promising and lukewarm—will encourage further exploration of robust, truly zero-shot neuro-symbolic frameworks.

REFERENCES

Forough Arabshahi, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom M. Mitchell. Conversational multi-hop reasoning with neural commonsense knowledge and symbolic logic rules. *ArXiv*, abs/2109.08544, 2021.

Ali Baheri and Cecilia O. Alm. Hierarchical neuro-symbolic decision transformer. *ArXiv*, abs/2503.07148, 2025.

WeiZhi Fei, Zihao Wang, Hang Yin, Shukai Zhao, Wei Zhang, and Yangqiu Song. Efficient and scalable neural symbolic search for knowledge graph complex query answering. *ArXiv*, abs/2505.08155, 2025.

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

Yen-Ling Kuo, B. Katz, and Andrei Barbu. Encoding formulas as deep networks: Reinforcement learning for zero-shot execution of ltl formulas. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5604–5610, 2020.

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

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Mastering symbolic operations: Augmenting language models with compiled neural networks. 2023.

Haoming Ye, Yunxiao Xiao, Cewu Lu, and Panpan Cai. Pretraining a unified pddl domain from real-world demonstrations for generalizable robot task planning. 2025.

Mikihisa Yuasa, R. Sreenivas, and Huy T. Tran. Neuro-symbolic generation of explanations for robot policies with weighted signal temporal logic. *ArXiv*, abs/2504.21841, 2025.

Özgür Yılmaz, A. Garcez, and Daniel L. Silver. A proposal for common dataset in neural-symbolic reasoning studies. 2016.

Gustav Šír. A computational perspective on neural-symbolic integration. *Neurosymbolic Artificial Intelligence*, 2024.

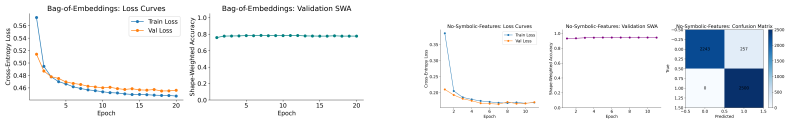
SUPPLEMENTARY MATERIAL

A ADDITIONAL IMPLEMENTATION DETAILS AND HYPERPARAMETERS

Our model is implemented in PyTorch. We use a Transformer encoder with 4 layers, 8 attention heads per layer, and a hidden dimension of 256. We train using Adam with a learning rate of 1e-4, a batch size of 64, and early stopping on the validation set. The symbolic features (e.g., shape variety counts) are computed via a simple Python parser that extracts discrete attributes from each example.

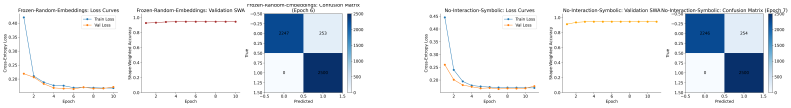
B ABLATION STUDIES

We performed several ablation experiments to evaluate how different aspects of the model architecture affect performance on SPR_BENCH. In particular, we explored models that: (1) Remove symbolic features, (2) Replace the transformer embeddings with a bag-of-embeddings approach, (3) Freeze random embeddings, (4) Remove interactions between neural and symbolic representations. None of the ablation settings led to dramatic improvements in zero-shot generalization, indicating that fundamental challenges in compositional rule inference persist.



(a) Bag of Embeddings (b) No Symbolic Features

Figure 2: Select ablation plots. We show the behavior of training/validation loss over epochs and final zero-shot SWA. Neither removing symbolic features nor using a bag-of-embeddings approach significantly improved transfer to novel rules.



(a) Frozen Random Embeddings (b) No Interaction with Symbolic

Figure 3: Further ablation studies. Freezing random embeddings or removing the neural-symbolic interaction likewise fails to achieve robust zero-shot performance on new rules.

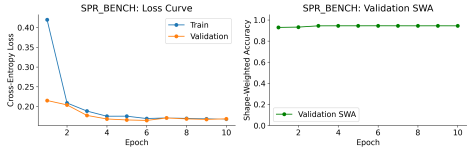


Figure 4: Additional exploration on SPR_BENCH. This plot illustrates various derived metrics such as color-weighted error analysis and partial rule correctness. None of these modifications resolved the core zero-shot challenge.