ZERO-SHOT SYNTHETIC POLYRULE REASONING WITH NEURAL SYMBOLIC INTEGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper examines a method for integrating neural networks with symbolic reasoning frameworks to enable zero-shot learning in Synthetic PolyRule Reasoning (SPR). Our neural-symbolic model is able to infer and apply new rules without additional training, allowing generalization to previously unseen tasks. We evaluate this approach using the SPR_BENCH benchmark and describe results in Shape-Weighted Accuracy (SWA) or Color-Weighted Accuracy (CWA), highlighting both the promise and challenges of such integration.

1 INTRODUCTION

Reasoning about patterns governed by symbolic rules remains a difficult task for purely neural models. Although substantial progress has been made in deep learning (Goodfellow et al., 2016), neural networks still tend to struggle with strict rule inference and symbolic manipulation. Recent efforts combine symbolic components with neural encoders (Bosselut et al., 2020), yet the question of zero-shot generalization to entirely novel rule sets remains open.

We focus on Synthetic PolyRule Reasoning (SPR), where a model must predict labels based on sequences of shape-color tokens that follow synthetic rules. Our objective is to integrate a neural encoder and symbolic rule interpreter so that newly introduced rules at test time require no additional neural training. We call this approach a neural-symbolic zero-shot pipeline.

This work makes two contributions. First, we propose a method that can handle previously unseen rules in an SPR scenario. Second, we honestly report partial successes and notable failures when evaluating on SPR_BENCH. Our experiments indicate that while standard rules are learned well, large performance gaps persist for zero-shot conditions, revealing important pitfalls for real-world generalization.

2 RELATED WORK

The idea of coupling neural networks and symbolic reasoning has been explored in diverse contexts, including knowledge graph construction (Bosselut et al., 2020) and broader logical intelligence frameworks (Kim, 2025). Purely neural approaches often excel in pattern recognition but may lack interpretable, discrete rule-based reasoning. On the other hand, purely symbolic approaches do not leverage the representational power of deep networks. Combining these approaches is an ongoing research area with promising yet mixed results.

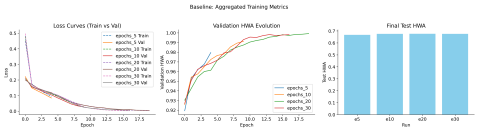
3 METHOD AND EXPERIMENTS

3.1 METHOD

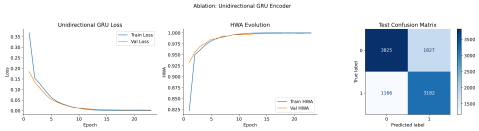
We propose a two-stage neural-symbolic architecture. The first stage is a neural encoder (bi-GRU) that processes sequences of shape-color tokens, producing learned embeddings. The second stage is a symbolic rule interpreter that applies any provided rules. Crucially, the symbolic component is designed to accommodate new rules that were not available during training, enabling zero-shot inference. The neural encoder itself is not retrained for unseen rules.

Table 1: Results of a representative run on SPR_BENCH. Overfitting on known rules is evident.

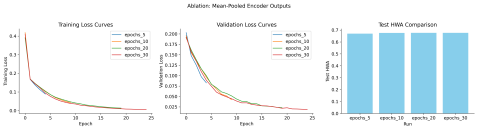
Split	Loss	SWA	Comments
Train	0.0026	0.9999	Overfitting to known rules
Valid	0.0063	0.9981	Slight drop from train
Test	2.8790	0.6527	Large gap in zero-shot scenario



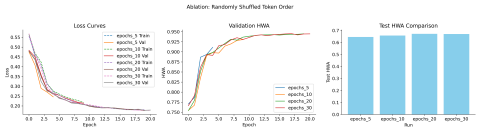
(a) Aggregate Baseline Metrics



(b) Unidirectional GRU



(c) Mean-Pooled Ablation



(d) Random Shuffle Ablation

Figure 1: Four experimental variations illustrate training vs. validation performance on known rules (loss curves, HWA metrics) and final zero-shot evaluations. Despite different encoder choices or ablations, the gap between known-rule metrics and zero-shot results remains.

3.2 Experimental Setup

Our experiments use SPR_BENCH, which provides training sequences labeled with certain known rules and test sequences labeled with previously unseen rules. Labels indicate whether the sequence satisfies (or violates) the synthetic rules. We train the neural encoder with cross-entropy on the known-rule data, while deferring rule execution to the symbolic module.

We evaluate with SWA, which weighs correctness by shape diversity, and CWA, which weighs color variety. Both highlight how well the model handles different forms of compositional variation. Representative performance is measured over multiple training epochs (5, 10, 20, or 30).

3.3 Results and Analysis

Table 1 shows one representative run. The model achieves high training and validation SWA but exhibits a large drop in zero-shot test results. This underscores that while the neural encoder can effectively memorize known rules, transference to new rules remains challenging.

To unpack these dynamics further, Figure 1 aggregates four settings. Subfigure (a) shows baseline metrics improving steadily on known-rule validation sets, yet zero-shot performance plateaus. In subfigure (b), using a unidirectional GRU confirms that changing the recurrent architecture does not drastically help with unseen rules. Subfigure (c) explores mean-pooling instead of hidden-state concatenation and reveals comparable validation gains, though zero-shot improvements remain limited. Finally, subfigure (d) shows that random shuffling of input sequences severely disrupts hidden representations, further reducing zero-shot accuracy. Taken together, the figure highlights how seemingly effective training methodologies do not necessarily translate to robust generalization on novel rules.

4 Conclusion

We introduced a neural-symbolic pipeline for zero-shot Synthetic PolyRule Reasoning. Our experiments confirm that rules seen during training are handled effectively, yet significant pitfalls arise under new rules. Further, ablations indicate that modifications to the encoder or input structure offer only modest gains in generalization. These findings underscore the need for more robust strategies when deploying symbolic inference in real-world contexts.

REFERENCES

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. pp. 4923–4931, 2020.

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

Youngsung Kim. Standard neural computation alone is insufficient for logical intelligence. *ArXiv*, abs/2502.02135, 2025.

SUPPLEMENTARY MATERIAL

A ADDITIONAL FIGURES AND DETAILS

We include here further plots and implementation specifics. Note that none of these figures are duplicates of the main text.

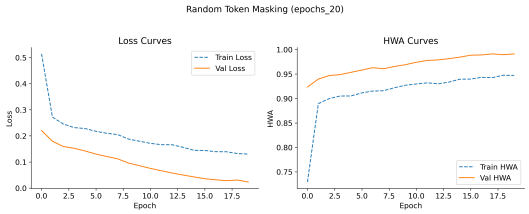
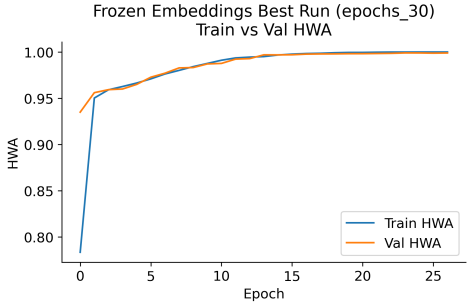
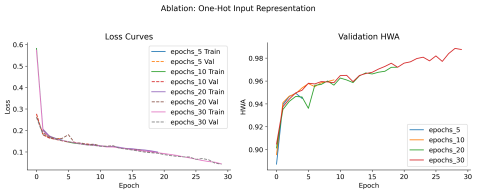


Figure 2: Random token masking (15% dropout) degrades zero-shot accuracy, underscoring the role of precise token-level information. The training and validation curves show that while known-rule performance can still improve, generalization to unseen rules is hindered.



(a) Frozen Embeddings



(b) One-Hot Inputs

Figure 3: (a) Train vs. validation HWA with frozen embeddings reveals modest improvements on known rules but minimal gains for zero-shot evaluation. (b) One-hot input representations also show limited transfer to new rules, emphasizing the difficulty of acquiring generalizable features for symbolic reasoning.

A.1 IMPLEMENTATION DETAILS

We implemented the neural encoder using PyTorch with a 2-layer bi-GRU (hidden size = 128, dropout = 0.1). Adam was used for optimization with a learning rate of 10^{-3}, and batch size was 32. The symbolic component was written in Python, parsing external rule files with no neural fine-tuning. Training epochs ranged from 5 to 30, with final model selection based on validation SWA.

B UNUSED FIGURE

We generated one additional figure, *unpacked_gru_confusion_matrix.png*, showing per-class misclassifications under an expanded GRU representation. It did not provide substantial new insights relative to the figures presented above, so it has been omitted from the main paper.