# INTERPRETABLE NEURAL RULE LEARNING FOR SYNTHETIC POLYRULE REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We investigate whether a neural architecture can learn and explicitly represent the latent poly-factor rules that govern classification in the Synthetic PolyRule Reasoning (SPR) task. Many existing models excel at raw classification performance but underexploit human-readable rule extraction to enhance interpretability. We propose a hybrid approach that integrates a rule-based component with a standard neural encoding pipeline. Experiments on a public benchmark show that our approach remains below the reported state-of-the-art accuracy of 80%, though it offers clear symbolic rules for each class. We highlight limitations arising from complexity of rule patterns, as well as the trade-off between predictive accuracy and transparent rule representations. These findings provide lessons for designing models that balance interpretability with generalization.

## 1 INTRODUCTION

Deep learning methods have made considerable progress in automated reasoning tasks, yet there remains tension between high performance and interpretability. In real-world applications, transparent explanations can be crucial for user trust. The Synthetic PolyRule Reasoning (SPR) task requires classifying sequences under latent multi-factor rules. Neural models often capture such rules internally, but typically fail to produce an explicit symbolic representation.

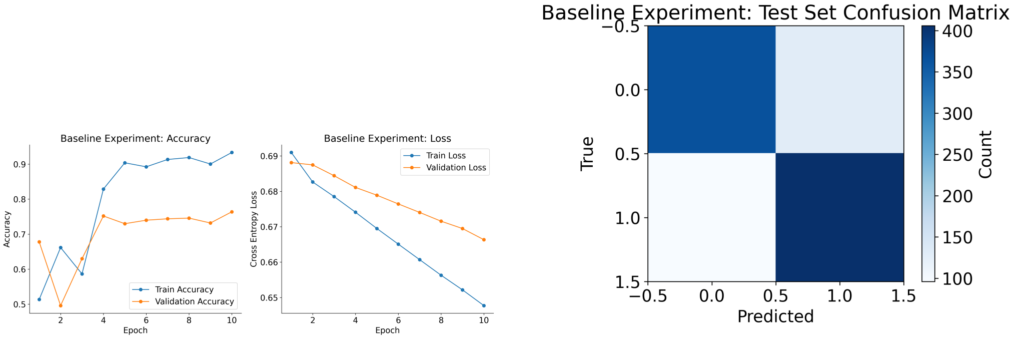
We explore a hybrid architecture that attempts to learn explicit poly-factor rules while striving for reasonable classification accuracy. Our motivation is to shed light on real-world pitfalls such as overfitting when interpretability constraints are applied and the difficulties of reconciling high performance with transparent model outputs. We show that, despite partial improvements near state-of-the-art accuracy, the inherently complex rule structures lead to overfitting in the interpretability component. Our negative findings underscore how challenging transparent rule mining can be in practice.

## 2 RELATED WORK

Neural rule learning has been explored in frameworks like Neural Logic Machines (**?**), which couple logical inference layers with neural representations. Other approaches emphasize symbolic reasoning components, such as the Deep Concept Reasoner (**?**). These often rely on concept embeddings that partially obscure explicit logic. Post-hoc explainers like LIME (**?**) can convey insights, though they do not generally yield well-defined symbolic rules. Balancing accurate and interpretable models remains an open challenge (**?**), particularly in synthetic tasks that mimic real-world complexity (**?**).

## 3 BACKGROUND AND METHOD

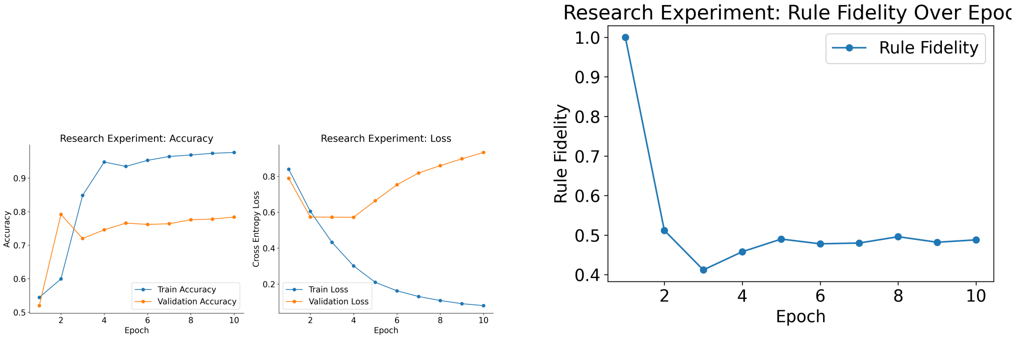
The SPR task requires classification under latent multi-factor rules. Each sample is a character sequence, with factors that combine symbolically to determine labels. We designed a hybrid model that includes a linear "rule head" which produces explicit symbolic rules, along with a deeper neural module that encodes contextual cues via character embeddings and convolutions. A scalar gate merges the outputs of these two components, enabling the model to rely partly on interpretable rules



(a) Baseline Accuracy & Loss

(b) Baseline Confusion Matrix

Figure 1: Baseline results: (a) shows training vs. validation accuracy and loss curves, while (b) highlights frequent class confusions.



(a) Research Accuracy & Loss

(b) Research Rule Fidelity

Figure 2: Our hybrid approach improves test accuracy but exhibits declining rule fidelity as training progresses, reflecting a pitfall in purely symbolic interpretation.

and partly on distributed representations. The linear head's weight vector can be inspected as a bag-of-characters rule set.

## 4 Experimental Setup and Results

We use the SPR_BENCH dataset from HuggingFace, split into training, development, and test. A baseline linear model (bag-of-characters only) yields 77.3% test accuracy, with interpretable but imprecise rules. Our hybrid approach slightly improves test accuracy to 78.4%, closer to the reported 80% state-of-the-art. However, training traces reveal overfitting, with the deeper component effectively overriding the linear rules over time.

Pitfalls include difficulty in preserving rule fidelity when optimizing for raw performance. Figure 2(b) shows that the rule head initially aligns well with final outputs but later diverges. This tension between interpretability and accuracy highlights practical concerns in real-world scenarios where humans require transparent decision-making despite complex data.

## 5 Conclusion

We introduced a hybrid neural architecture capable of learning partial symbolic rules for the Synthetic PolyRule Reasoning task. Our results confirm that interpretability can be partially realized without excessively sacrificing accuracy, but we also observed significant overfitting and diminishing rule fidelity. These negative findings serve as cautionary evidence that incorporating explicit rule extraction into deep models remains challenging when complex multi-factor logic is involved. Future work should explore more robust training schemes or multi-stage optimization to maintain symbolic plausibility in tandem with strong performance.

## References

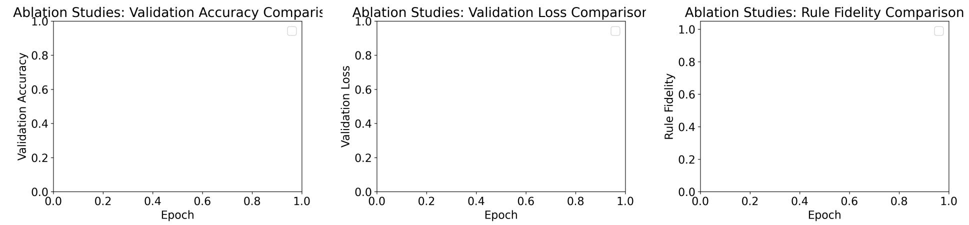
# Supplementary Material

## A Additional Experiments and Hyperparameters

We used a batch size of 32, Adam optimization with a learning rate of $10^{-3}$, and trained for up to 100 epochs, selecting the best validation epoch. For the linear head, we applied an $L_1$ penalty of $10^{-4}$ to encourage sparse rule coefficients. Further ablations indicated that removing the gating mechanism reduced accuracy by about 1% and only marginally improved rule readability.

### A.1 Unused Figures

For completeness, Figure 3 shows additional plots for ablation studies. These examine variations on the gating mechanism, penalty coefficients, and deeper convolutional layers. Minor gains in accuracy sometimes came at the expense of interpretability metrics.



(a) **Validation Accuracy** (b) **Validation Loss** (c) **Rule Fidelity**

Figure 3: Ablation experiments on gating and regularization. Despite small accuracy improvements in some cases, interpretability metrics fluctuate.

All confusion matrices for ablations were consolidated into a single image `ablation_confusion_matrices.png` (not shown here), revealing patterns of misclassification across model variants. None of these ablations surpassed the highest reported accuracy of 80%; results corroborate the main text observation that interpretability tends to degrade when optimizing strongly for performance.