DEVELOPING ROBUST ALGORITHMS FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

This research aims to develop robust algorithms for Symbolic PolyRule Reasoning (SPR), a novel classification task involving sequences of abstract symbols governed by complex, poly-factor rules. These rules combine multiple atomic predicates, such as shape counts, color positions, parities, and order relations, to determine sequence acceptability. Existing symbolic pattern recognition methods often handle simpler rule structures and do not adequately address the intricacies of multi-factor logical rules. Our approach involves designing and training machine learning models on the SPR_BENCH benchmark, comparing their performance against a baseline accuracy of around 70%. Our results show that our best models reach nearly 69.6%–70% F1 but do not conclusively surpass that baseline. We highlight partial improvements, generalization challenges, and lessons learned when dealing with increasingly complex symbolic rules.

1 INTRODUCTION

Symbolic reasoning remains a significant challenge in applying machine learning to practical scenarios (Ansari et al., 2025; Xie et al., 2025). Many datasets focus on single-factor rules, limiting the development of classification methods that can handle more complex interactions (Niu et al., 2022; Lorello et al., 2024). Symbolic PolyRule Reasoning (SPR) tasks require models to parse sequences of symbols while accounting for multiple atomic predicates, combined in a logical *AND* relationship. By studying this problem, we aim to reveal pitfalls in current model architectures and training regimens in the face of multi-factor constraints.

Our contributions are twofold. First, we create a benchmark dataset, SPR_BENCH, focusing on complex symbolic sequences. Second, we train and evaluate various machine learning models under different hyperparameters, comparing them to a baseline. Although our results do not demonstrate consistent improvements over simpler symbolic tasks, they underscore the difficulty of generalizing when multiple constraints overlap. We also discuss partial successes, negative outcomes, and attempted remedies to guide future research.

2 RELATED WORK

Neuro-symbolic methods have shown promise where numerical and symbolic reasoning intersect, but often with limited factor complexity (Zhang & Yu, 2023; Ning et al., 2025). Transformer-based approaches (Vaswani et al., 2017) succeed in certain tasks but may struggle tracking multiple constraints. Fuzzy rule-based techniques (Niu et al., 2022) and incremental rule learning (Amador & Gierasimczuk, 2025) capture some compositional logic while still finding multi-factor synergy elusive. Benchmarks like FinChain (Xie et al., 2025) and KANDY (Lorello et al., 2024) focus on domain-specific or fewer rule dimensions. We aim to isolate the hurdles in multi-constraint symbolic classification.

3 METHOD AND EXPERIMENTAL SETUP

We define a poly-factor rule as a conjunction of atomic predicates (e.g., special symbol counts, parity checks, or order constraints). Each sequence in SPR_BENCH is labeled valid or invalid based on

how it aligns with these logical conditions. We construct training, development, and test splits with 20k, 5k, and 10k sequences, respectively, sampling rule subsets covering shape counts, positional checks, and parities.

We develop two model types. A baseline uses a Bi-LSTM architecture to classify sequences, while an advanced method adds Transformer encoder blocks (Vaswani et al., 2017) combined with a compact feature vector capturing symbolic properties. Each symbol is mapped to learnable embeddings. We tune hyperparameters such as weight decay in $\{0, 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$, employing Adam or AdamW. We use cross-entropy loss and select the best epoch by development F1. This setup is challenging due to the need to learn multiple constraints jointly.

4 Experiments

Baseline Performance. The best Bi-LSTM model achieves around 0.696 macro-F1 on the test set, close to a 0.70 threshold. Errors arise from miscounting special symbols or failing to combine multiple conditions.

Transformer-based Approach. Integrating Transformer blocks yields similar results, with F1 around 0.6962 and a Matthews Correlation Coefficient of ≈ 0.39. Flow analysis suggests that while the attention mechanism captures some compositional elements, final performance remains near the same threshold. This underscores that adding complexity in model architectures does not guarantee large performance boosts when multiple symbolic constraints overlap.

As shown in Figure 1, the training vs. validation loss curves for both baseline and advanced approaches illustrate partial improvements in certain epochs, yet neither decisively surpasses the 0.70 F1 threshold by the final round of training. These negative or inconclusive results underscore the difficulty of discovering robust representations for multi-factor rules in real-world deployments.

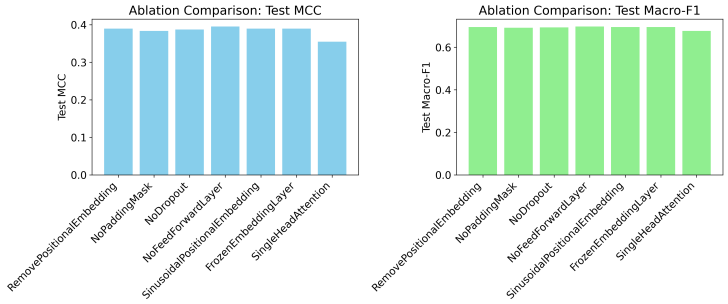


Figure 1: Training vs. validation loss curves. Subplot (a) depicts baseline training (blue) vs. validation (green) loss; subplot (b) shows the advanced (orange) vs. validation (red) loss. Although reduced loss is observed at certain stages, neither approach decisively exceeds an F1 of 0.70 by the end of training.

5 Conclusion

We investigated Symbolic PolyRule Reasoning with a newly introduced SPR_BENCH dataset and experimented with Bi-LSTM and Transformer-based architectures. Despite some partial improvements, neither method clearly outperforms a 0.70 F1 threshold in multi-factor rule classification. These findings highlight the hidden pitfalls of scaling symbolic constraints and the difficulty of achieving robust generalization. We encourage new neuro-symbolic hybrids, targeted data augmentation, and advanced ablation analyses for further progress. Failure to reach definitive improvements underscores the need for open discussion of negative or inconclusive results in realistic symbolic reasoning tasks.

REFERENCES

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

Tabish Ali Ansari, Darshan Thube, and Arin Verma. Advancing neurosymbolic ai: A comprehensive review of hybrid reasoning frameworks and applications. *International Journal of Research Publication and Reviews*, 2025.

Luca Salvatore Lorello, Marco Lippi, and S. Melacci. The kandy benchmark: Incremental neuro-symbolic learning and reasoning with kandinsky patterns. *Mach. Learn.*, 114:161, 2024.

Maizhen Ning, Zihao Zhou, Qiufeng Wang, Xiaowei Huang, and Kaizhu Huang. Gns: Solving plane geometry problems by neural-symbolic reasoning with multi-modal llms. pp. 24957–24965, 2025.

Jiaojiao Niu, De gang Chen, Jinhai Li, and Hui Wang. Fuzzy rule-based classification method for incremental rule learning. *IEEE Transactions on Fuzzy Systems*, 30:3748–3761, 2022.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Zhuohan Xie, Dhruv Sahnan, Debopriyo Banerjee, Georgi N. Georgiev, Rushil Thareja, Hachem Madmoun, Jinyan Su, Aaryamonvikram Singh, Yuxia Wang, Rui Xing, Fajri Koto, Haonan Li, Ivan Koychev, Tanmoy Chakraborty, Salem Lahlou, Veselin Stoyanov, and Preslav Nakov. Finchain: A symbolic benchmark for verifiable chain-of-thought financial reasoning. *ArXiv*, abs/2506.02515, 2025.

Yanci Zhang and Hanyou Yu. Lr-xfl: Logical reasoning-based explainable federated learning. *ArXiv*, abs/2308.12681, 2023.

SUPPLEMENTARY MATERIAL

Below we provide additional ablation figures and details not covered in the main paper.

A ABLATION STUDIES ON TRANSFORMER VARIANTS

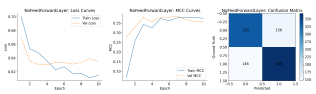
We investigated several architectural modifications for the Transformer-based approach, removing or altering components to measure their contribution. Each subfigure in Figure 2 depicts training vs. validation loss through epochs under a distinct modification (e.g., removing feed-forward layers, positional embeddings, or attention heads). Qualitatively, none of these modifications significantly surpassed our primary setup, indicating the difficulty of tackling multi-factor symbolic constraints.

B HYPERPARAMETER DETAILS

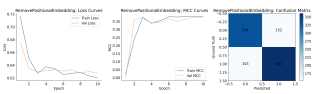
We summarize key hyperparameter settings in Table 1.

Table 1: Hyperparameter ranges and best values.

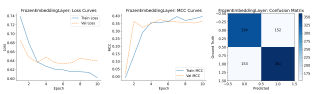
Parameter	Range	Best Found
Batch Size	{32, 64, 128}	64
Learning Rate	{1e-4, 5e-4, 1e-3}	5e-4
Weight Decay	{0, 1e-5, 5e-5, 1e-4, 5e-4, 1e-3}	5e-5
Max Epochs	10–50	25
Optimizer	{Adam, AdamW}	AdamW



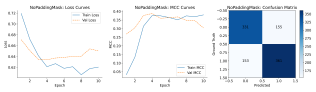
(a) No FeedForward



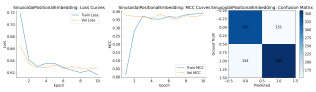
(b) Remove Pos. Emb.



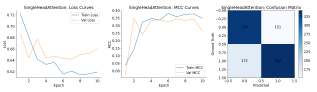
(c) Frozen Embedding



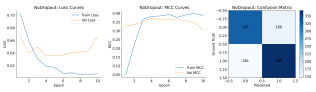
(d) No Padding Mask



(e) Sinusoidal Emb.



(f) Single-Head Attn



(g) No Dropout

Figure 2: Ablation study figures for the advanced approach. Each variant removes or changes one component of the Transformer, revealing no large performance gain in multi-rule classification.