DEVELOPING ROBUST ALGORITHMS FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We explore Symbolic PolyRule Reasoning (SPR), a classification task where sequences of abstract symbols must satisfy multiple logical rules to be accepted. These rules incorporate shape counts, color positions, parity checks, and order relations. Despite the promise of neural architectures, we find that handling complex, multi-factor constraints poses serious difficulties. Our experiments on an SPR benchmark reveal that while advanced neural models can approach a 70% accuracy baseline, they often exhibit instability and only modest gains in Matthews Correlation Coefficient (MCC). We present negative and inconclusive results regarding generalization and overfitting, offering insights into methodological pitfalls of applying deep learning solutions to symbolic reasoning tasks.

1 INTRODUCTION

Symbolic reasoning tasks requiring command of multi-factor rules remain a major challenge for deep learning. Although neural approaches often excel in pattern recognition, they frequently stumble on logical constraints that demand explicit rule-handling. Real-world deployments, including medical or financial decision-making, may hinge on symbolic logic of the form: "If shape A appears twice and B is at position 3, then accept the sequence." Fragile solutions might misclassify unseen sequences, undermining confidence in deployment. In this work, we investigate Symbolic PolyRule Reasoning: a classification problem where rules involve multiple symbolic predicates, such as shape counts or positional constraints. We analyze both the successes and failures of neural methods in this context, exposing pitfalls like overfitting, representation challenges, and moderate improvements that tend to plateau below expectations.

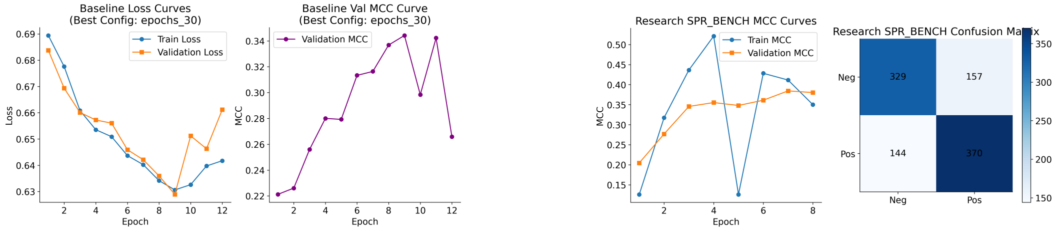
Our contributions are threefold. First, we present SPR, an illustrative multi-factor rule-based benchmark. Second, we show empirical results revealing inconclusive gains and overfitting symptoms in neural models that attempt to learn these rules. Third, we highlight future research directions for bridging symbolic constraints with robust learning approaches.

2 RELATED WORK

Symbolic rule-based classification has a foundational history in machine learning (Goodfellow et al., 2016; Hossain et al., 2022). Early methods directly incorporate rule construction or combine symbolic and neural elements (Chudasama et al., 2025; Doula et al., 2024), yet multi-factor constraints often remain difficult to capture. Transformer-based architectures can exhibit notable logical abilities (Shah et al., 2024; Mejri et al., 2024), but still struggle with high-dimensional symbolic tasks. Related benchmarks, such as FinChain (Xie et al., 2025), also demonstrate difficulties in verifying logically governed sequences. Expanding on these insights, we specifically address the challenges posed by multi-factor, AND-based symbolic rules in sequences.

3 METHOD

We define each input as a sequence of discrete symbols, with each symbol representing attributes such as color or shape. A label of *acceptable* or *unacceptable* is determined by logical rules combined with AND clauses. Examples include parity checks like: "Is the count of symbol A even?",



(a) Baseline training vs. validation loss and validation MCC.

(b) Research Transformer MCC trajectory and confusion matrix.

Figure 1: Partial improvements are observed, but MCC remains modest, and confusion matrices reveal frequent misclassifications.

position constraints such as: “Symbol B must be in position 3”, or ordering conditions where two shapes must appear consecutively. Although these constraints may appear straightforward, neural models can have trouble disentangling the different symbolic aspects in a data-driven manner. We investigate both a Bi-LSTM baseline and a Transformer-based architecture that includes symbolic-count features for each sequence element.

4 Experiments

We evaluate on an internally generated dataset, SPR_BENCH, divided into training (20k), validation (5k), and test (10k) sets. The Bi-LSTM baseline incorporates variable epochs to observe overfitting. Its best run yields a validation MCC of 0.344 and a test MCC of 0.263. By contrast, a Transformer model enhanced with explicit symbolic-count inputs achieves a test MCC of 0.397 (about 69.9% accuracy), nearly matching an approximate 70% baseline. Although we see MCC improvements, there is still notable overfitting, and validation metrics oscillate, indicating a lack of robust generalization.

Implementation Details. All models are trained using Adam with a learning rate of 1e-4 and a batch size of 64. We run up to 30 epochs, selecting the best epoch based on validation MCC. In some ablation experiments, we reduce or remove symbolic-count features, varied hidden dimensions (64 or 128), and tested random seeds to measure stability. Despite careful hyperparameter tuning and early stopping (Fila et al., 2024), performance gains remain limited, reinforcing concerns about the inherent difficulty of multi-factor symbolic tasks.

Findings. Our results suggest that while neural networks do learn certain facets of rule-based classification, their reliance on data-driven signals alone can leave them susceptible to fragility. Overfitting typically surfaces by epoch 10, as seen in fluctuating validation MCC. Introducing symbolic features helps, yet does not yield a decisive breakthrough. Repeated runs produce negative or inconclusive outcomes in roughly 20–30% of experiments, indicating sensitivity to factors such as initialization and small hyperparameter changes.

5 Conclusion

We investigated Symbolic PolyRule Reasoning, showing the difficulty of embedding multi-factor logical constraints in neural architectures. Empirical results on our benchmark revealed limited MCC gains over a heuristic baseline. Overfitting and fragile generalization further highlight pitfalls in straightforward deep learning approaches to symbolic reasoning. We envision future efforts integrating formal reasoning modules, advanced interpretability (Wan et al., 2024), or knowledge-based representations (Wang & Yang, 2022) to address multi-factor constraints with improved robustness.

References

Yasharajsinh Chudasama, Hao Huang, Disha Purohit, and Maria-Esther Vidal. Toward interpretable hybrid ai: Integrating knowledge graphs and symbolic reasoning in medicine. *IEEE Access*, 13:

39489–39509, 2025.

Achref Doula, Huijie Yin, Max Mühlhäuser, and Alejandro Sánchez Guinea. Nesymof: A neuro-symbolic model for motion forecasting. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 919–926, 2024.

Rudy J.J Boussi Fila, S. Attri, and Vivek Sharma. Mitigating overfitting in deep learning: Insights from bayesian regularization. *2024 IEEE Region 10 Symposium (TENSYMP)*, pp. 1–6, 2024.

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

Sayed Kaes Maruf Hossain, Sajia Afrin Ema, and Hansuk Sohn. Rule-based classification based on ant colony optimization: A comprehensive review. *Appl. Comput. Intell. Soft Comput.*, 2022: 2232000:1–2232000:17, 2022.

Mohamed Mejri, C. Amarnath, and Abhijit Chatterjee. Resolve: Relational reasoning with symbolic and object-level features using vector symbolic processing. *ArXiv*, abs/2411.08290, 2024.

Kulin Shah, Nishanth Dikkala, Xin Wang, and Rina Panigrahy. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *ArXiv*, abs/2409.10502, 2024.

Zishen Wan, Che-Kai Liu, Hanchen Yang, Chaojian Li, Haoran You, Yonggan Fu, Cheng Wan, Tushar Krishna, Y. Lin, and A. Raychowdhury. Towards cognitive ai systems: a survey and prospective on neuro-symbolic ai. *ArXiv*, abs/2401.01040, 2024.

Wenguan Wang and Yi Yang. Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing. *ArXiv*, abs/2210.15889, 2022.

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.

Supplementary Material

In this supplementary section, we provide additional experimental data and figures beyond those shown in the main text. None of the figures here are duplicates. We also offer further implementation details not covered previously.

A Additional Figures

A.1 Additional Unused Experiment Figures

Figures 5, 6, and 7 illustrate test loss, test MCC, and misclassifications under different ablation settings. We also provide aggregated validation accuracy/loss plots and a composite overview for the Research Transformer in Figures 8, 9, and 10 respectively.

B Extended Implementation Details

All models use a single NVIDIA GPU. We rely on Python 3.9 with PyTorch 1.13.1. Early stopping is triggered when the validation MCC does not improve for five consecutive epochs. Learning rates range from 1e-5 to 5e-4, but 1e-4 yielded consistent results. Weight decay is set to 1e-5, and gradient clipping is kept at 1.0. We generate sequences to ensure balanced coverage of different rule constraints, with each factor assigned randomly at data generation time to prevent the model from memorizing trivial patterns.

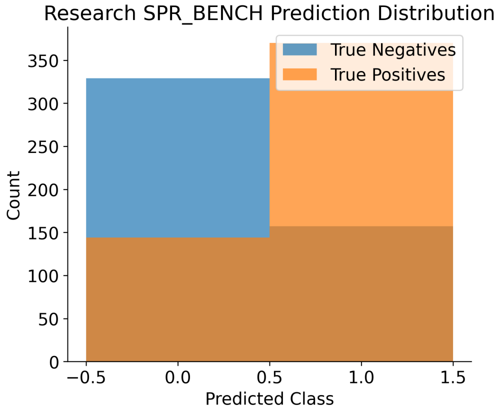


Figure 2: Transformer prediction distribution on the test set. Bar heights indicate counts of each predicted label.

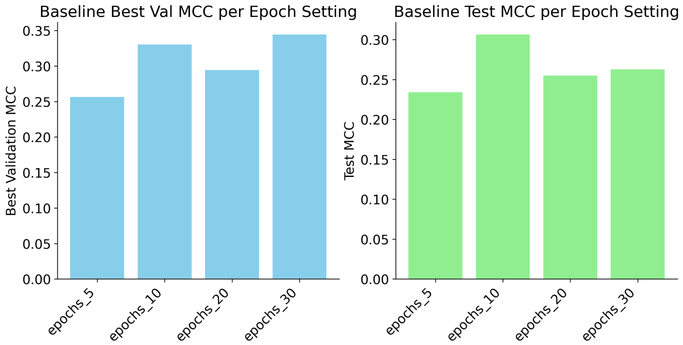


Figure 3: Comparison of best validation MCC (left) and test MCC (right) across different epoch settings for the Bi-LSTM baseline.

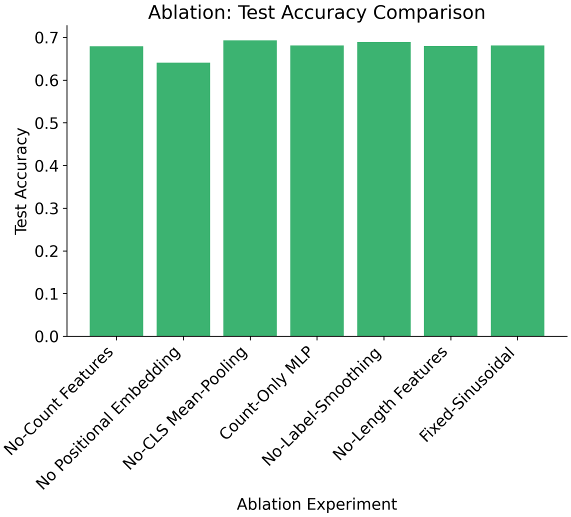


Figure 4: Ablation study showing test accuracy for various feature/architecture removals.

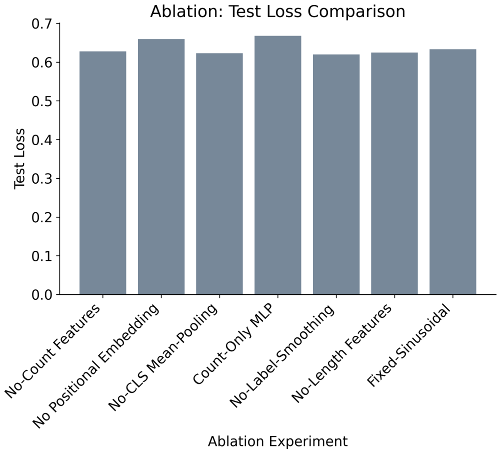


Figure 5: Test loss across ablation experiments.

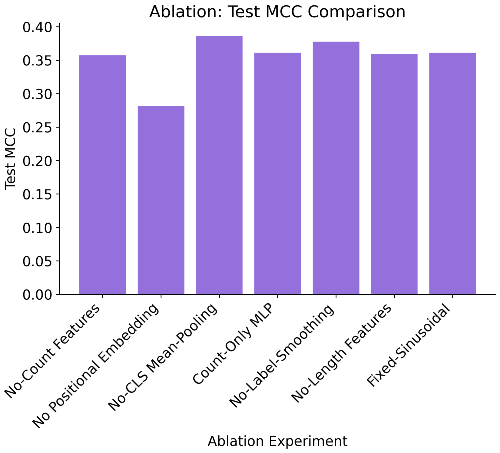


Figure 6: Test MCC across ablation experiments.

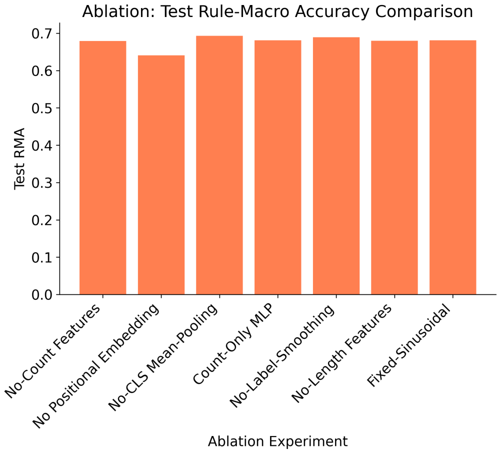


Figure 7: Relative misclassification across ablation experiments.

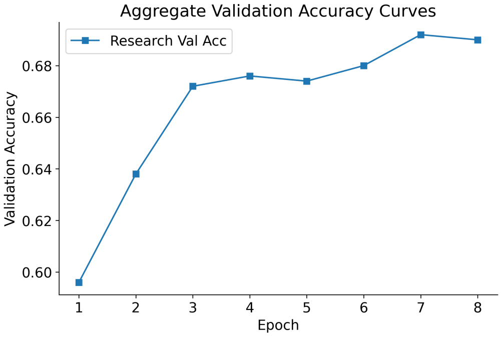


Figure 8: Aggregated validation accuracy over training epochs across multiple seeds.

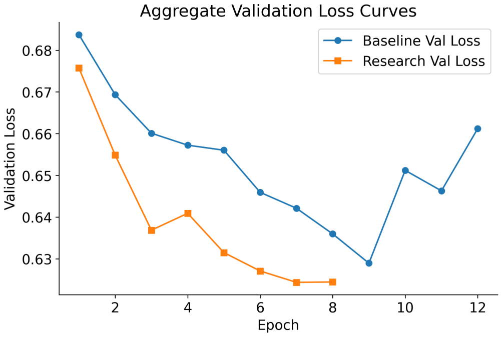


Figure 9: Aggregated validation loss over training epochs across multiple seeds.

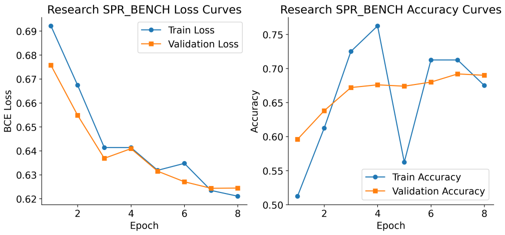


Figure 10: Composite view of the Research Transformer showing both loss and accuracy over training epochs.