DEVELOPING ROBUST ALGORITHMS FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic PolyRule Reasoning (SPR) involves the classification of sequences of abstract symbols regulated by multi-factor logical rules. We focus on a novel benchmark, SPR_BENCH, where atomic predicates based on color, shape frequency, and positional constraints jointly determine whether a sequence is acceptable. Our study hypothesized that carefully designed architectures might surpass a rule-based baseline (70% accuracy). However, our experiments reveal challenges in generalizing across rule compositionality, with the best Macro-F1 score near 0.69 and a Matthews correlation coefficient around 0.38. These inconclusive findings highlight pitfalls in capturing intricate, logical constraints for real-world symbolic tasks.

1 INTRODUCTION

Symbolic reasoning tasks rely on discrete logical rules for classification or prediction (Cingillioglu & Russo, 2021; Li et al., 2020; Bortolotti et al., 2024). Although deep learning excels in language and vision domains, handling multi-factor and compositional rules remains a challenge (Lin & Zhang, 2024; Patel et al., 2024; Vats et al., 2025). We address Symbolic PolyRule Reasoning (SPR), where sequences must satisfy multiple interacting predicates to be deemed acceptable. This problem relates to real-world settings, such as validating product codes or ensuring policy compliance, where diverse constraints operate simultaneously.

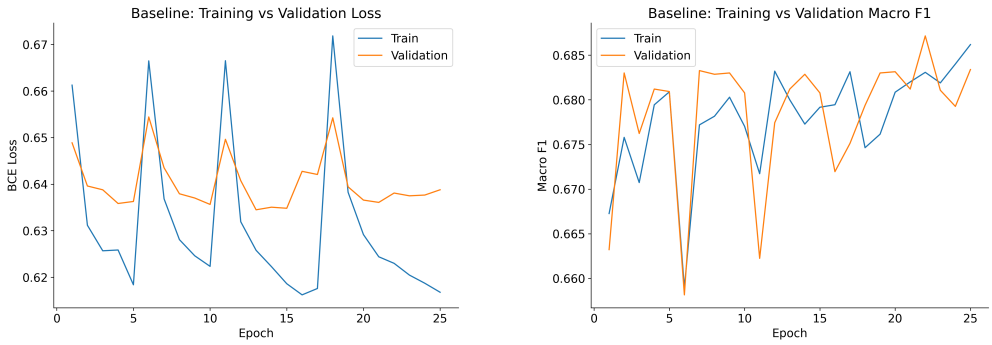
Our key contribution is an empirical evaluation on a newly developed SPR_BENCH dataset, incorporating color attributes, shape frequency checks, and positional constraints. A rule-based baseline attains about 70% accuracy, representing a non-trivial standard. We explored gating-based recurrent architectures and lightweight Transformers, yet none consistently outperformed the baseline. Our analysis includes negative or inconclusive results, underscoring the subtleties of multi-factor logical reasoning and the tendency for models to overlook rarer rule compositions.

2 RELATED WORK

Methods that combine symbolic reasoning with deep networks have grown in popularity, exemplified by rule-based neuro-symbolic systems (Cingillioglu & Russo, 2021), integrated pipelines of parsing and symbolic reasoning (Li et al., 2020), and specialized reasoning benchmarks (Wang & Song, 2024; Xie et al., 2025). Others emphasize fuzziness and robust classification (Lin & Zhang, 2024) or multi-step reasoning (Patel et al., 2024; Bortolotti et al., 2024). While such work shows promise, bridging multiple interlocking rules in a single classification setting, as we do here, is less studied. We build on RNN-based encoders (Cho et al., 2014) and Transformers, optimizing with Adam-like algorithms (Kingma & Ba, 2014). The present findings highlight pitfalls in applying these approaches to multi-factor symbolic tasks and offer cautionary insights for real-world systems.

3 METHOD

We frame SPR as a binary classification problem over symbolic sequences. Each example can involve attributes (e.g., color, shape, code) and a label indicating overall acceptability. We include



(a) Baseline Loss vs Validation Loss

(b) Baseline Train vs Validation Macro-F1

Figure 1: **Baseline GRU training curves on SPR_BENCH.** Our model's training metrics (blue) and validation metrics (red) demonstrate a growing gap between training and validation, indicating challenges in learning across multi-factor logical constraints. Notable instability in the validation curves highlights how subtle rule conditions may be overlooked.

conjunctive predicates such as "must contain a red symbol" or "the pairwise ordering of triangles must be strictly ascending." To capture these interactions, we first employed a Gated Recurrent Unit (GRU) model (Cho et al., 2014), passing the final hidden state through a binary classifier. We then tested a lightweight Transformer with position embeddings, hypothesizing better learning of long-range constraints. Both models used binary cross-entropy loss with optional class weighting to mitigate label imbalance.

4 Experiments

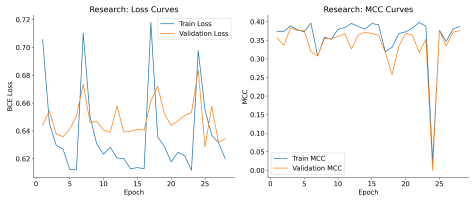
We used SPR_BENCH, splitting data into training, development, and test sets. A simple rule-based heuristic achieves about 70% accuracy.

As shown in Figure 1, the GRU baseline achieves around 0.69 Macro-F1 and an MCC near 0.38, barely meeting the rule-based benchmark. The gap between training and validation performance suggests the model has difficulty generalizing beyond straightforward patterns. Architectural tuning yielded limited improvements, pointing to inherent challenges in capturing the combinatorial variety of rules.

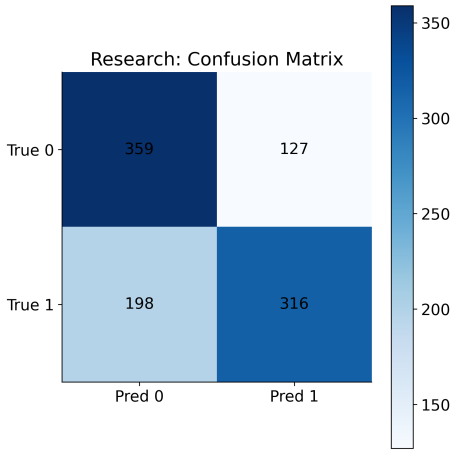
Figure 2 illustrates that while the Transformer also converges on the training set, its ability to capture multiple constraints simultaneously remains limited. The confusion matrix further reveals that both positive and negative classes are misclassified at non-negligible rates. These results underscore the complex interplay among color relationships, shape frequencies, and positional constraints, which even advanced sequence models fail to learn adequately.

5 Conclusion

We evaluated neural architectures on a complex symbolic task governed by multiple logical constraints. While baseline methods achieve near 70% accuracy, standard sequence models struggle to surpass this figure, yielding inconclusive or negative results. Our findings illustrate pitfalls in applying data-driven models to multi-factor symbolic challenges and emphasize careful attention to domain constraints, data coverage, and architecture design. We hope these results spur deeper investigation into specialized, interpretable, or hybrid approaches that can robustly tackle real-world symbolic reasoning.



(a) Loss and MCC



(b) Confusion Matrix

Figure 2: **Lightweight Transformer results.** (Left) Although training loss is relatively smooth, the MCC remains low (0.38–0.40). (Right) Confusion matrix showing misclassifications in both classes, underscoring the model's difficulties with intricate rule interactions.

REFERENCES

Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. 2024.

Kyunghyun Cho, B. V. Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. pp. 1724–1734, 2014.

Nuri Cingillioglu and A. Russo. pix2rule: End-to-end neuro-symbolic rule learning. pp. 15–56, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Qing Li, Siyuan Huang, Yining Hong, Yixin Chen, Y. Wu, and Song-Chun Zhu. Closed loop neural-symbolic learning via integrating neural perception, grammar parsing, and symbolic reasoning. *ArXiv*, abs/2006.06649, 2020.

Guo Lin and Yongfeng Zhang. Fuzzy neural logic reasoning for robust classification. *ACM Transactions on Knowledge Discovery from Data*, 19:1 – 29, 2024.

Nisarg Patel, Mohith Kulkarni, Mihir Parmar, Aashna Budhiraja, Mutsumi Nakamura, Neeraj Varshney, and Chitta Baral. Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models. *ArXiv*, abs/2406.17169, 2024.

Shaurya Vats, Sai Phani Chatti, Aravind Devanand, Sandeep Krishnan, and Rohit Karanth Kota. Empowering llms for mathematical reasoning and optimization: A multi-agent symbolic regression system. *Systems and Control Transactions*, 2025.

Weiqi Wang and Yangqiu Song. Mars: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *ArXiv*, abs/2406.02106, 2024.

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.

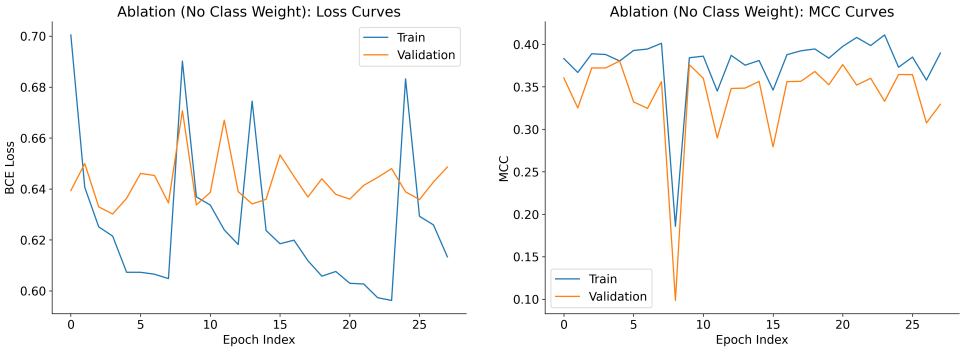
SUPPLEMENTARY MATERIAL

HYPERPARAMETERS AND IMPLEMENTATION DETAILS

All models were implemented in PyTorch (v1.13). For the GRU, we used 2 layers of 256 hidden units each, initialized with Xavier uniform. The Transformer used 2 encoder layers, 4 attention heads, and 128-dimensional embeddings, with a dropout rate of 0.1. We optimized all models using Adam (Kingma & Ba, 2014), an initial learning rate of 1×10^{-3}, and a mini-batch size of 64. A weight decay of 1×10^{-5} was tested but did not yield systematic improvements.

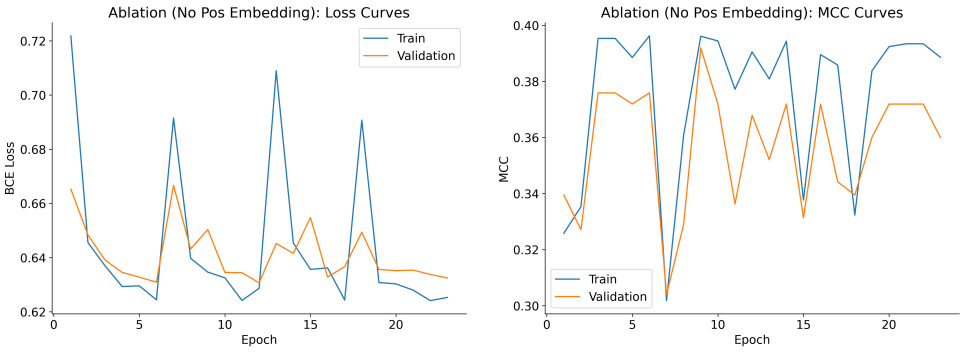
ADDITIONAL ABLATION STUDIES

We conducted ablations removing class weighting and positional embeddings, as well as variations in weight decay. While each factor sometimes improved training stability, no single configuration substantially outperformed the baseline. Below are illustrative figures for two representative ablations:



(a) No Class Weight: Loss Curves (b) No Class Weight: MCC

Figure 3: Removing class weighting led to small fluctuations in training loss without significant gains in MCC.



(a) No Pos Embedding: Loss Curves (b) No Pos Embedding: MCC

Figure 4: Discarding positional embeddings weakened sequence-level reasoning, further reducing MCC.

We found that eliminating positional cues or class weights often compromised model performance, though not dramatically. Figures that did not offer further insight into model behaviors, such as additional bar plots of minimal differences, were omitted for brevity.