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 is especially relevant for real-world scenarios such as product code validation or policy compliance checks, where diverse constraints operate simultaneously.

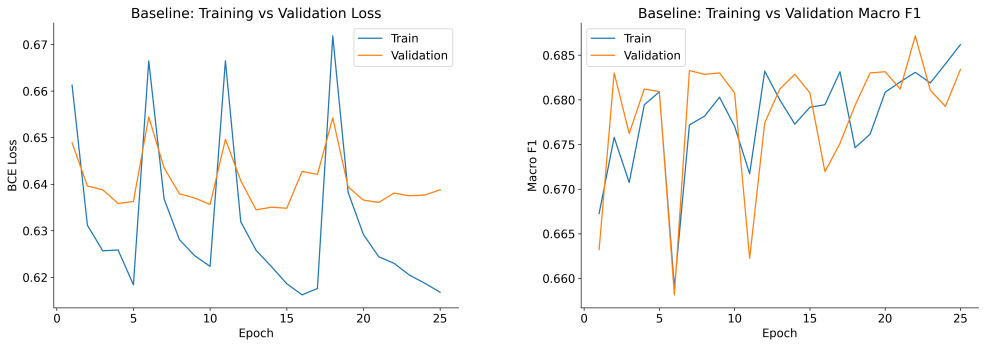
We propose an empirical evaluation on a newly developed SPR_BENCH dataset that fuses color attributes, shape frequency checks, and positional constraints. A rule-based baseline attains about 70% accuracy, representing a non-trivial standard. We explored gating-based recurrent architectures and lightweight Transformers, yet none systematically outperformed this heuristic. Our results include negative and inconclusive findings, illustrating the subtleties of multi-factor logical reasoning and revealing a tendency for models to overlook rarer rule compositions.

2 RELATED WORK

Methods that combine symbolic reasoning with deep networks have gained traction in various tasks, including neuro-symbolic rule learning (Cingillioglu & Russo, 2021), pipelines of parsing and symbolic reasoning (Li et al., 2020), and specialized benchmarks (Wang & Song, 2024; Xie et al., 2025). Several works emphasize fuzziness, robust classification (Lin & Zhang, 2024), or multi-step inference (Patel et al., 2024; Bortolotti et al., 2024). While prior studies highlight the promise of neural methods for logic-based tasks, bridging multiple interlocking rules under a single classification objective, as we do, remains challenging. We build on standard RNN-based encoders (Cho et al., 2014) and Transformers, using Adam-like optimization (Kingma & Ba, 2014), to test how well they manage multi-factor symbolic tasks.

3 METHOD

We define SPR as a binary classification problem on symbolic sequences. Each example involves attributes (color, shape, code) and a label that indicates acceptability based on a combination of logical predicates. Conjunctive rules capture requirements such as "must contain a red symbol"



(a) Baseline: Training vs Validation Loss (b) Baseline Train vs Validation Macro-F1

Figure 1: **Baseline GRU training curves on SPR_BENCH.** The training curves decline steadily, but the validation curves show considerable instability across epochs. The growing gap suggests potential overfitting and reveals the challenge in learning multi-factor constraints.

or "strict ordering of certain shapes based on position." We first train a GRU model (Cho et al., 2014) with a feedforward classifier on the final hidden state. We next test a lightweight Transformer, hoping that its attentional mechanism would track longer contextual constraints. Both approaches employ binary cross-entropy loss, optionally with class weighting to address label imbalance.

4 Experiments

We employ SPR_BENCH, splitting data into training, development, and test sets. A straightforward rule-based approach yields around 70% accuracy, demonstrating the dataset's complexity.

Figure 1 highlights the training and validation metrics for our GRU baseline. Training runs indicate a steady drop in binary cross-entropy (BCE) loss and a rise in Macro-F1, but the validation performance suffers from large fluctuations. Despite data augmentation and hyperparameter tuning, the GRU model plateaus at about 0.69 Macro-F1, barely matching the rule-based baseline. Below, we examine a Transformer variant.

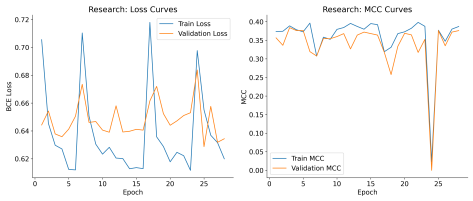
As shown in Figure 2, the Transformer converges more smoothly but fails to exceed the baseline's accuracy. While the training process exhibits fewer oscillations relative to the GRU, its Matthews correlation coefficient on the validation set remains low, around 0.40. The confusion matrix similarly indicates that both positive and negative classes are frequently misclassified, underscoring the complexity of SPR_BENCH.

5 Conclusion

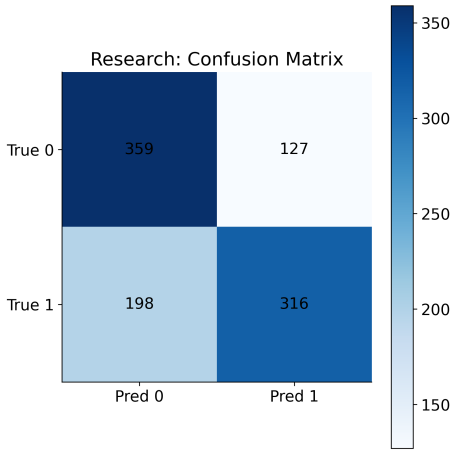
We investigated neural architectures for symbolic sequence classification, uncovering that standard models often fail to surpass a carefully designed rule-based baseline. Our findings indicate that multi-factor predicates can be unintuitive for data-driven learning, and highlight the importance of explicitly encoding logical constraints or devising specialized hybrid architectures. Future directions include refined neuro-symbolic models with improved interpretability and domain-awareness, which may address the pitfalls observed here.

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.



(a) Loss and MCC



(b) Confusion Matrix

Figure 2: **Lightweight Transformer results.** (a) Training loss trends downward while MCC remains between 0.38–0.40 on the validation set. (b) The confusion matrix indicates significant misclassifications for both classes, reflecting the difficulty of concurrently satisfying multiple constraints.

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.

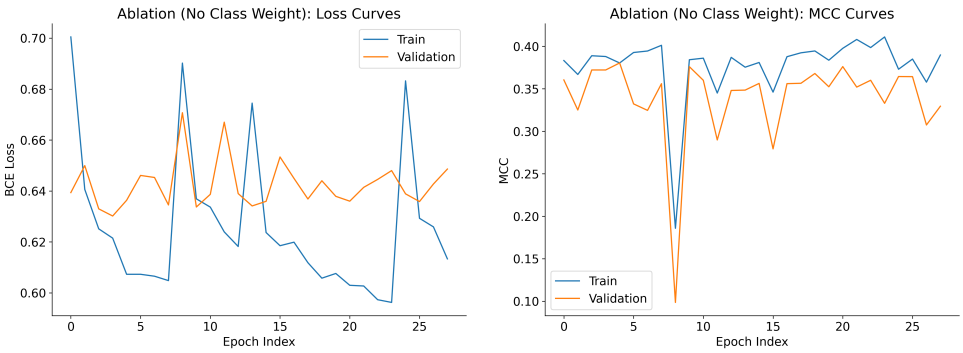
Zhuohan Xie, Dhruv Sahnan, Debopriyo Banerjee, Georgi N. Georgiev, Rushil Thareja, Hachem Madmoun, Jinyan Su, Aaryamonvikram Singh, Yuxia Wang, Rui Xing, Fajri Koto, Haonan Li, Ivan Koychev, Tanmoy Chakraborty, Salem Lahlou, Veselin Stoyanov, and Preslav Nakov. Finchain: A symbolic benchmark for verifiable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

SUPPLEMENTARY MATERIAL

In this appendix, we provide additional details and ablation studies. While these ablations offer insights into the training dynamics under modified parameter settings, none succeeded in surpassing or meaningfully improving upon the baseline results reported in the main text.

ABLATION: REMOVING CLASS WEIGHTING

We studied the effect of removing class weighting. Figure 3 shows that loss improvements become more erratic, while MCC remains nearly unchanged. These findings suggest that balancing labels can slightly stabilize training but does not resolve the fundamental challenges of compositional rule learning.

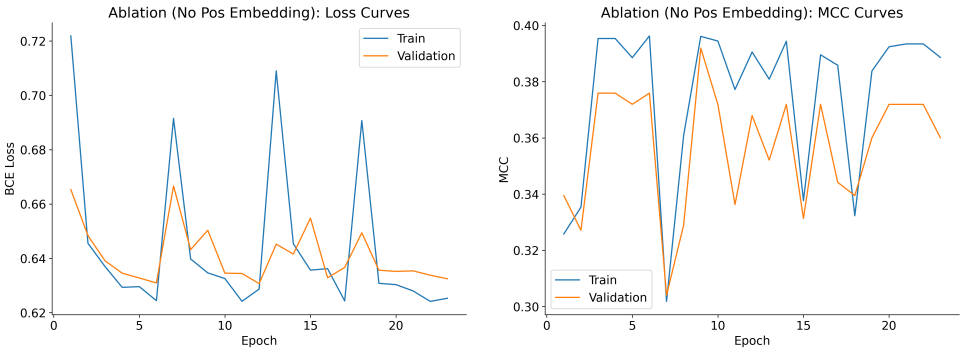


(a) No Class Weight: Loss Curves (b) No Class Weight: MCC

Figure 3: **Impact of removing class weighting.** The training/validation loss curves fluctuate more, and the MCC shows no substantial improvement.

ABLATION: REMOVING POSITIONAL EMBEDDINGS

We also removed positional embeddings from the Transformer to test their importance in capturing ordering constraints. As illustrated in Figure 4, discarding positional information further limits sequence-level reasoning, reducing both loss stability and MCC scores. This underscores the significance of explicit positional cues for multi-factor symbolic tasks.



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

Figure 4: **Effect of discarding positional embeddings.** Training and validation loss exhibit sharper fluctuations, and MCC degrades, emphasizing the need for positional cues.

Discussion of Removed Figures

We originally generated additional figures depicting confusion bar plots and fine-grained metrics for various ablations (e.g., no weight decay or fixed sinusoidal embeddings). However, these plots showed negligible differences or repeated similar trends to those already presented. For brevity and clarity, such figures have been removed to keep the focus on the most illustrative ablation results. All key observations are captured by the figures included above.