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.

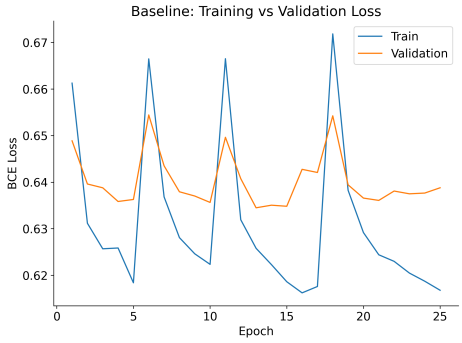
We propose an empirical evaluation on a newly developed SPR_BENCH dataset that fuses color attributes, shape frequency checks, and positional constraints. A rule-based baseline attains about 70% accuracy, representing a non-trivial standard. We explored gating-based recurrent architectures and lightweight 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

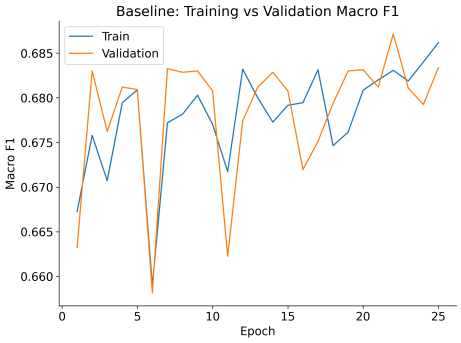
Methods that combine symbolic reasoning with deep networks have gained traction across diverse tasks, including neuro-symbolic rule learning (Cingillioglu & Russo, 2021), pipelines that integrate grammar parsing and symbolic components (Li et al., 2020), and specialized benchmarks (Wang & Song, 2024; Xie et al., 2025). Several works emphasize fuzziness or robust classification (Lin & Zhang, 2024) and multi-step inference (Patel et al., 2024; Bortolotti et al., 2024). However, bridging multiple interlocking rules under a single classification objective, as we do here, remains challenging. We build on standard RNN-based encoders (Cho et al., 2014) and Transformers, using Adam 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 indicating acceptability based on a combination of logical predicates. Conjunctive rules capture requirements such as “must contain a red symbol” or “strict ordering of certain shapes based on position.” We train a GRU (Cho et al., 2014) with a feedforward

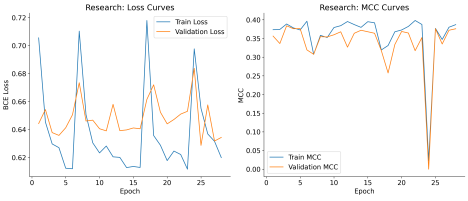


(a) Baseline: Training vs Validation Loss

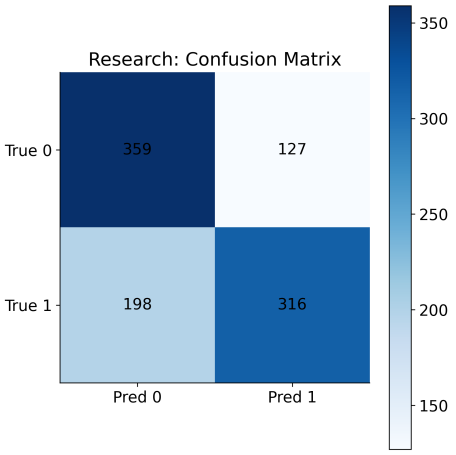


(b) Baseline Train vs Validation Macro-F1

Figure 1: **Baseline GRU training curves on SPR_BENCH.** Training loss decreases steadily, while validation loss and Macro-F1 fluctuate significantly, suggesting overfitting and difficulty in learning multi-factor constraints.



(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 validation. (b) The confusion matrix shows high misclassification for both classes, highlighting the challenge of satisfying multiple constraints.

classifier on the final hidden state. We then test a lightweight Transformer, hoping its attentional mechanism could track longer contextual constraints. Both approaches employ binary cross-entropy (BCE) loss, with optional 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 $\approx 70\%$ accuracy, suggesting non-trivial complexity.

Figure 1 shows the baseline GRU's learning dynamics. Although training improves predictably, validation metrics oscillate, hinting at overfitting. The GRU plateaus at about 0.69 Macro-F1, closely matching the rule-based accuracy. We next examined a Transformer variant.

Figure 2 reveals smoother training with the Transformer, but validation MCC hovers around 0.40. Both methods fail to outperform the rule-based baseline, underscoring the difficulty of compositional logic in SPR.

5 CONCLUSION

We examined neural architectures for a multi-factor symbolic reasoning task, finding that even carefully tuned RNNs and Transformers struggle to surpass a rule-based baseline. These negative results underscore pitfalls in capturing compositional constraints from data alone. Future efforts may require hybrid neuro-symbolic strategies or explicit logical encodings to overcome these challenges.

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

HYPERPARAMETER DETAILS

For the GRU model, we used a hidden dimension of 128, a single recurrent layer, and a dropout rate of 0.3. The Transformer model employed 4 attention heads and 2 encoder layers with a hidden dimension of 128. We trained for 25 epochs using a batch size of 64 and the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 1×10^{-3}. Weight decay and class weighting were toggled for ablations.

ADDITIONAL FIGURES

Figures 3, 4, 5, 6, and 7 provide complementary views of our experiments that were not included in the main text.

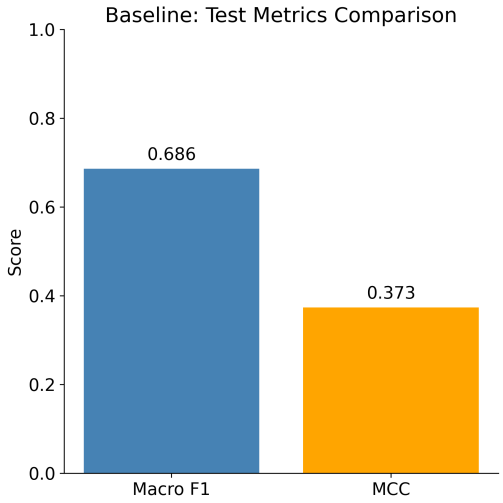


Figure 3: **Baseline final test performance.** This bar chart shows accuracy, Macro-F1, and MCC for the rule-based baseline and the GRU.

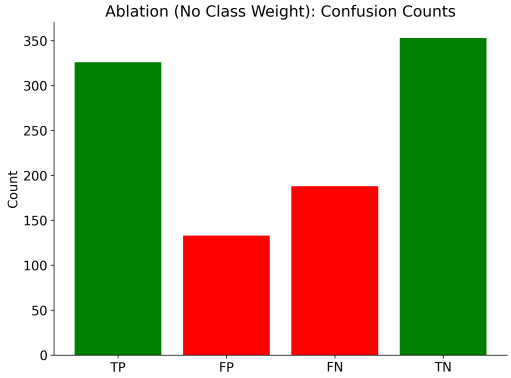


Figure 4: **No-Class-Weight Ablation: Confusion bars.** Distribution of predicted vs. true labels, illustrating persistent misclassifications.

Ablation Results

Removing Class Weighting. Eliminating class weighting often yielded more volatile training curves, but MCC remained comparable. Detailed confusion bars are shown in Figure 4.

Removing Positional Embeddings. We explored dropping positional embeddings, which hindered the model's ability to capture ordering constraints, reducing validation MCC (Figure 5).

Fixed Sinusoidal Embeddings. Using fixed sinusoidal embeddings (Figure 6) did not yield a measurable gain over learned embeddings, with loss curves showing similar patterns of fluctuation.

No Weight Decay. When removing weight decay, we observed slightly faster initial convergence but more pronounced overfitting. Figure 7 illustrates the remaining class-level confusion.

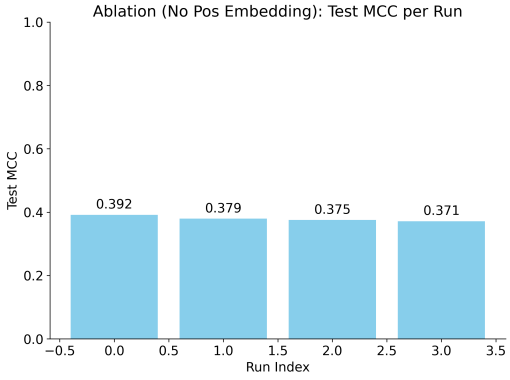


Figure 5: **No-Positional-Embeddings Ablation: Test MCC.** Dropping positional embeddings worsens classification consistency.

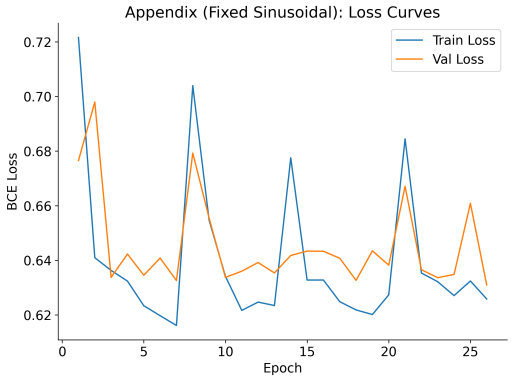


Figure 6: **Fixed Sinusoidal Embeddings: Loss Curves.** Loss trends with fixed sinusoidal positional embeddings.

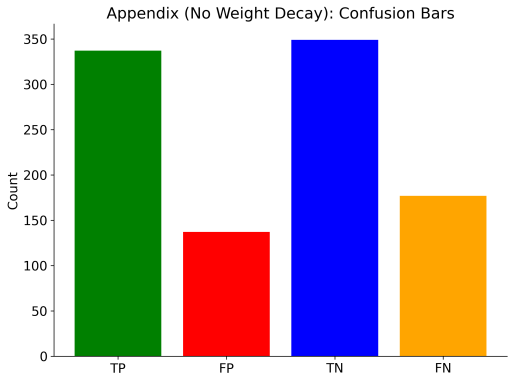


Figure 7: **No-Weight-Decay Ablation: Confusion Bars.** Visualizing misclassification distributions when weight decay is removed.