DEVELOPING ROBUST ALGORITHMS FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

This research explores Symbolic PolyRule Reasoning (SPR), a classification task that determines acceptability of symbolic sequences governed by multi-factor logical rules. We investigate machine learning models, including a baseline recurrent approach and a Transformer-based model, on a new benchmark (SPR_BENCH). Despite careful tuning, we find that neither approach reaches the anticipated 70% accuracy target. The results highlight real-world pitfalls, such as rule complexity, overfitting, and representation challenges, underscoring the difficulty of robust SPR. We discuss these challenges and propose directions for improving multi-factor symbolic reasoning in future work.

1 INTRODUCTION

Symbolic classification tasks involving complicated logical rules are often encountered in domains like knowledge management, industrial diagnostics, and NLP pipelines. Unlike simpler single-factor rule systems, multi-factor logical rules require handling multi-constraint predicates, such as object counts, order relations, or color/shape positions (Hossain et al., 2022; Kamali et al., 2024). Traditional rule-based systems struggle with such added complexity, and recent neuro-symbolic approaches do not always generalize well (Sun et al., 2025). We investigate whether modern machine learning models can effectively learn these multi-factor rules from labeled examples.

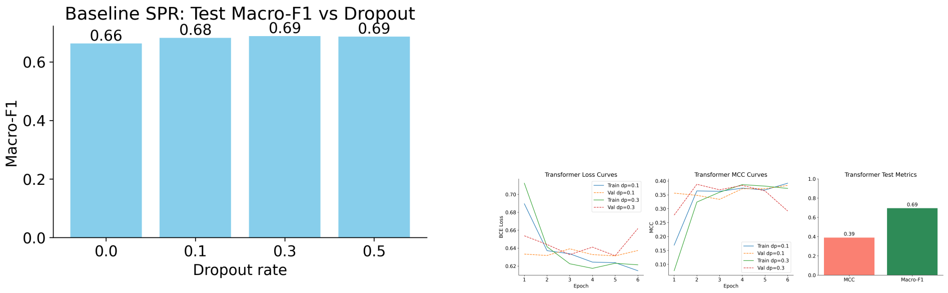
Our focus is Symbolic PolyRule Reasoning (SPR), which requires identifying valid sequences under complex poly-factor constraints. We design a curated dataset called SPR_BENCH to facilitate research in this domain. We hypothesize that advanced sequence models can classify sequences governed by multi-predicate rules more accurately than legacy systems. However, our findings reveal that training large models on such data does not guarantee surpassing a 70% macro-F1 baseline, highlighting unforeseen pitfalls. These negative or inconclusive outcomes serve to guide future improvements.

2 RELATED WORK

Symbolic reasoning with multiple constraints has a long tradition in rule-based classification (Hossain et al., 2022), while hybrid neuro-symbolic frameworks attempt fusions of neural networks with logical rules (Marra, 2024; Özgür Yılmaz et al., 2016). Although Transformers excel at certain symbolic tasks (Brinkmann et al., 2024), overfitting and inconsistent generalization remain problematic in multi-factor contexts. Work in neuro-symbolic concept composition illustrates that compositional complexity can hinder straightforward pattern recognition (Kamali et al., 2024). Our contribution differs in emphasizing multiple logical factors, including shape or color-based predicates, that significantly increase classification difficulty.

3 METHOD

We define SPR as a binary classification on sequences of discrete symbols where acceptability is determined by combined atomic predicates. Symbolic rules describe conditions like "the number of a certain symbol is even AND the last symbol belongs to a specific subset" or "the sequence maintains a color-count threshold in the middle positions." Our dataset generation follows prior



(a) **Baseline Test Macro-F1 vs. Dropout.**

(b) **Transformer Performance.**

Figure 1: Key results for Symbolic PolyRule Reasoning. (a) Baseline CharBiGRU with varying dropout yields a maximum test macro-F1 of about 0.69. (b) A small Transformer model struggles similarly, obtaining around 0.695 final macro-F1 and an MCC near 0.39.

guidance on constructing symbolic benchmarks (Özgür Yılmaz et al., 2016; Tie et al., 2025). We then evaluate two model families:

Recurrent Baseline (CharBiGRU). We embed characters and pass them to a bidirectional GRU, trained with binary cross-entropy. Dropout is varied (0.0, 0.1, 0.3, 0.5). Macro-F1 is our main metric.

Transformer. A small Transformer encoder is used to handle potentially longer sequences and multi-factor dependencies (Brinkmann et al., 2024; Sun et al., 2025). We train with two dropout values (0.1, 0.3) to examine the trade-off between overfitting and regularization. We track Matthews Correlation Coefficient (MCC) alongside F1 for thorough evaluation.

4 EXPERIMENTS

We split SPR_BENCH into train, development, and test sets. Despite systematic tuning, neither approach reached the 70% target commonly expected for simpler rule-based tasks. We detail these unexpected outcomes below.

Recurrent Baseline. Four dropout rates were tested. The highest test macro-F1 of 0.6883 occurred at 0.3 dropout, as shown in Figure 1(a). This is below the hypothesized threshold. Overfitting patterns emerged in low-dropout scenarios, while higher dropout improved generalization but did not yield major gains over the 70% mark.

Transformer Approach. With dropout rates of 0.1 and 0.3, the best test macro-F1 was 0.6949, still below the baseline goal. Validation loss fluctuated, consistent with prior findings that Transformers can overfit on small or complex symbolic data (Pesaresi et al., 2016). Our best MCC was around 0.39, indicating moderate predictive correlation.

Key limitations included difficulty generalizing across diverse rule interactions and data representation constraints that hamper learning. The multi-constraint nature of SPR leads to intricate decision boundaries. Data also showed signs of label ambiguity for borderline sequences, reinforcing the complexity of rule-driven classification.

5 CONCLUSION

While we anticipated surpassing a 70% F1 threshold, our empirical results on Symbolic PolyRule Reasoning fell short. We identified practical pitfalls: data representation becomes critical when sequences contain multi-factor constraints; overfitting arises quickly given the complexity of rules; and predicted gains from sophisticated architectures may not materialize without carefully tailored models and larger data coverage. Future directions include building more structured encodings, incorporating specialized neuro-symbolic modules, exploring richer regularization regimes, and investigating interpretability to diagnose model blind spots. We hope that highlighting these negative or inconclusive results spurs improvements in systematic symbolic rule modeling and offers realistic insights into the difficulties of multi-factor logic in neural systems.

References

Jannik Brinkmann, A. Sheshadri, Victor Levoso, Paul Swoboda, and Christian Bartelt. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. pp. 4082–4102, 2024.

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.

Danial Kamali, Elham J. Barezi, and Parisa Kordjamshidi. Nesycoco: A neuro-symbolic concept composer for compositional generalization. *ArXiv*, abs/2412.15588, 2024.

Giuseppe Marra. From statistical relational to neuro-symbolic artificial intelligence. pp. 22678, 2024.

M. Pesaresi, V. Syrris, and Andreea Julea. A new method for earth observation data analytics based on symbolic machine learning. *Remote. Sens.*, 8:399, 2016.

Zhong-Hua Sun, Rui Zhang, Zonglei Zhen, Da-Hui Wang, Yong-Jie Li, Xiaohong Wan, and Hongzhi You. Systematic abductive reasoning via diverse relation representations in vector-symbolic architecture. *ArXiv*, abs/2501.11896, 2025.

Guiyao Tie, Xueyang Zhou, Tianhe Gu, Ruihang Zhang, Chaoran Hu, Sizhe Zhang, Mengqu Sun, Yan Zhang, Pan Zhou, and Lichao Sun. Mmlu-reason: Benchmarking multi-task multi-modal language understanding and reasoning. 2025.

Özgür Yılmaz, A. Garcez, and Daniel L. Silver. A proposal for common dataset in neural-symbolic reasoning studies. 2016.

Supplementary Material

This appendix contains further details, including code listings, additional figures beyond the four shown in the main text, and extended quantitative analyses. The provided code snippets (e.g. `SPR.py`), along with the script for generating final plots, can be found in accompanying supplementary files. Readers interested in deeper hyperparameter settings, random seeds, or further ablation studies are encouraged to consult this supplementary material.