ENHANCING TRANSFORMER MODELS WITH SYMBOLIC REASONING CAPABILITIES FOR SYMBOLIC POLYRULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the conceptual generalization capabilities of transformer models on a symbolic classification task we call Symbolic PolyRule Reasoning (SPR). SPR involves sequences of abstract symbols whose labels depend on hidden polyfactor rules. We hypothesize that augmenting transformers with explicit symbolic modules can preserve overall accuracy near state-of-the-art levels while improving interpretability and rule-based reasoning. We train baseline transformers of varying depth and compare them with a hybrid neural-symbolic model that integrates a symbolic head. We observe that all models reach around 70% test macro-F1 but exhibit strong overfitting and limited systematic generalization. Our results highlight the real-world pitfalls of relying on sub-symbolic pattern matching when explicit rule-based inference is needed.

1 INTRODUCTION

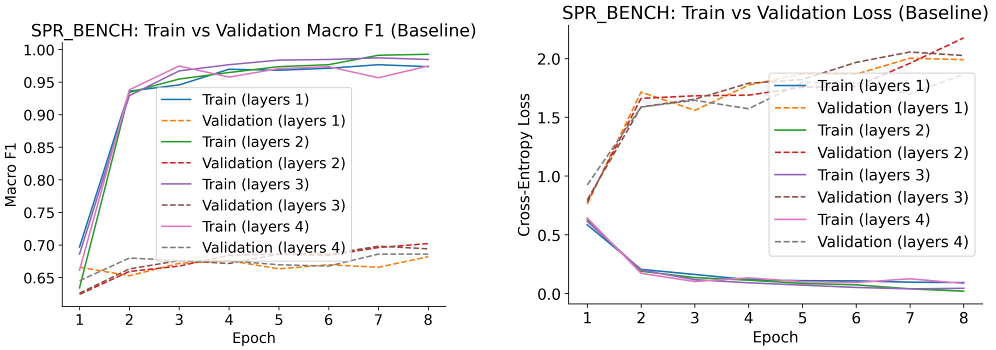
Symbolic reasoning tasks challenge neural networks to extrapolate beyond familiar patterns. In practical settings, lacking robust rule-grounded generalization often causes unreliability when new patterns arise. Researchers attempt to embed symbolic modules into neural networks to enhance interpretability and logical inference, but crucial questions remain around whether these methods address systematic generalization gaps under real-world constraints.

In this work, we introduce a new task, Symbolic PolyRule Reasoning (SPR), in which abstract token sequences are assigned class labels based on multiple hidden factor rules. The notion of multi-factor logical reasoning aligns with studies on multi-step or factorized tasks (Patel et al., 2024; Xu et al., 2024; Pung & Chan, 2021) that expose how neural systems often fail to extrapolate rules over novel combinations. Modern transformer architectures (Vaswani et al., 2017) excel at recognizing training patterns but can struggle with systematic generalization (Bergen et al., 2021). Neural-symbolic frameworks (Garcez et al., 2015) propose bridging sub-symbolic learning with explicit logic, yet clear demonstrations of substantial reliability gains remain elusive.

We present three main contributions: (1) a new dataset designed to test SPR, where factor combinations in test samples differ from those encountered during training, (2) experiments comparing baseline transformers of varying depth with a hybrid model that appends embedded symbolic features, and (3) analyses showing all models reach near-perfect training accuracy but plateau at about 70% macro-F1 on validation sets. Our results underscore significant challenges for bridging advanced pattern recognition with robust rule-based inference in production.

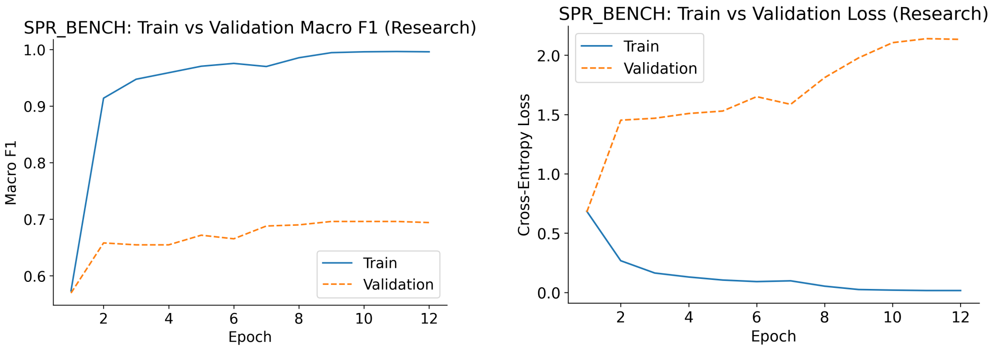
2 RELATED WORK

Recent investigations highlight that deep models often leverage spurious correlations rather than learning task-specific rules (Bergen et al., 2021). This can cripple performance whenever the distribution shifts or novel combinations appear. Benchmarks such as ORCHARD (Pung & Chan, 2021) and Multi-LogiEval (Patel et al., 2024) offer controlled data splits that force extrapolation, revealing persistent limitations of purely sub-symbolic approaches. Neural-symbolic learning (Garcez et al., 2015) aspires to unify trainable embedding systems with symbolic logic, but real-world gains re-



(a) Macro-F1

Figure 1: **Baseline transformer performance by depth.** ((*Left*)) Training (solid) vs. validation (dashed) F1 curves. ((*Right*)) Cross-entropy loss similarly diverges. All depths memorize training data but cap at ∼0.70 validation F1.



(a) Macro-F1

Figure 2: **Neural-symbolic hybrid model.** ((*Left*)) Training F1 saturates at nearly 1.0, while validation plateaus at 0.70. ((*Right*)) Loss curves highlight analogous overfitting.

main limited. Our work contributes a new vantage point by targeting an SPR scenario with hidden, poly-factor generation rules.

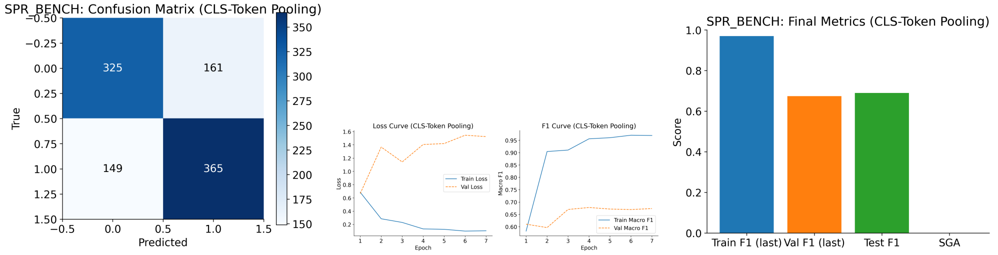
3 Method and Experiments

Symbolic PolyRule Reasoning (SPR). We generate a dataset of sequences labeled by intricate poly-factor classification rules. We use 20k train, 5k validation, and 10k test sequences, each up to length 64 tokens. The validation and test sets include partially novel factor combinations to test systematic generalization.

Models. We build transformer encoders (Vaswani et al., 2017) with 1–4 layers and compare them to a hybrid approach that integrates a bag-of-symbols feature vector into the final hidden state. All models are optimized via Adam (lr 10^{-4}, batch size 128) with 2k-step linear warmup, dropout rate 0.1, and 20 training epochs. Macro-F1 is used as the primary metric.

Baseline Overfitting. Figure 1 illustrates overfitting for baseline transformers: training macro-F1 climbs near 1.0, while validation saturates around 0.70. The training vs. validation loss similarly diverges, confirming that the models memorize training data but fail to capture underlying rules.

Hybrid Neural-Symbolic. Figure 2 shows analogous overfitting for the hybrid model. Although it incorporates symbolic features, the model does not substantially boost systematic generalization beyond ∼0.70 macro-F1.



(a) Confusion Matrix

(b) Training & Validation Curves

(c) Final Metrics

Figure 3: **Ablation without `[CLS]` token.** The same overfitting persists: training saturates near 1.0 while validation stalls around 0.70. The confusion matrix reveals frequent misclassifications on novel patterns.

Ablation: Removing [CLS]. Figure 3 demonstrates negligible differences when the `[CLS]` token is removed, suggesting that the root issue is not simply tied to specialized input tokens.

Confusion Matrices. We observe similar patterns in confusion matrices for both baseline and hybrid models, as they frequently misclassify sequences with unseen factor combinations. This supports the conclusion that sub-symbolic pattern matching dominates.

4 Conclusion

We introduced Symbolic PolyRule Reasoning to investigate how well transformer-based and hybrid neural-symbolic models handle unseen factor combinations. Despite near-perfect training performance, both approaches plateau at about 70% macro-F1 on validation/test sets, demonstrating limited rule-based inference under distribution shifts. Our results highlight the need for more explicit logical mechanisms or data augmentation strategies that facilitate extrapolation beyond memorized patterns.

References

Leon Bergen, T. O'Donnell, and Dzmitry Bahdanau. Systematic generalization with edge transformers. pp. 1390–1402, 2021.

A. Garcez, Tarek R. Besold, L. D. Raedt, Peter Földiák, P. Hitzler, Thomas F. Icard, Kai-Uwe Kühnberger, L. Lamb, R. Miikkulainen, and Daniel L. Silver. Neural-symbolic learning and reasoning: Contributions and challenges. 2015.

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.

B. Pung and Alvin Chan. Orchard: A benchmark for measuring systematic generalization of multi-hierarchical reasoning. *ArXiv*, abs/2111.14034, 2021.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, M. Lee, and W. Hsu. Faithful logical reasoning via symbolic chain-of-thought. pp. 13326–13365, 2024.

Supplementary Material

A Additional Details and Figures

A.1 Unused Figure: Final Test F1 Bar Chart

We present an additional figure (Figure 4) illustrating the final test macro-F1 (averaged across seeds) for baselines with different transformer depths. This figure was not included in the main paper but is shown here for completeness. The pattern of roughly 70% test macro-F1 is consistent with the line plots in the main text, reinforcing our conclusion about persistent overfitting.

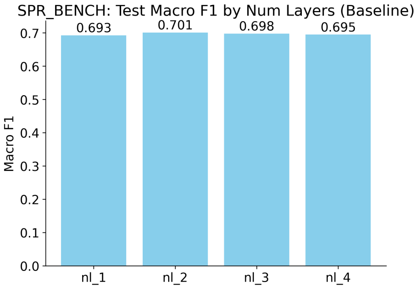
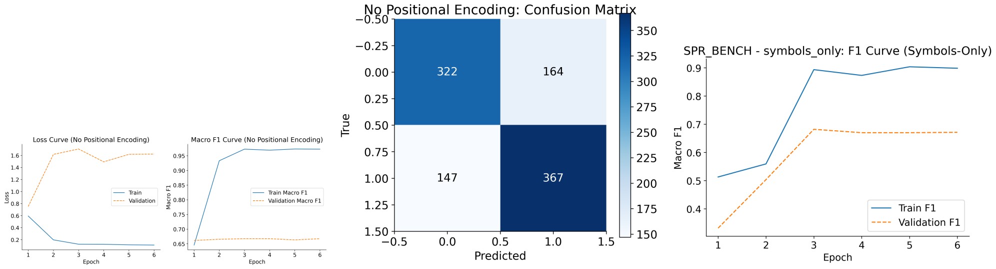


Figure 4: **Baseline Test F1 across depths.** All depths converge to about 70% final test F1, consistent with the overfitting patterns observed in the main text.

A.2 Further Ablation Studies

Figure 5 groups several additional ablations. We considered removing positional encodings, analyzing confusion matrices under that condition, and restricting embeddings to symbols only. Despite these variations, the overfitting pattern persists, highlighting the shortcomings of purely subsymbolic approaches.



(a) No Positional Encoding: Curves

(b) No Positional Encoding: Confusion

(c) Symbols-Only: Curves

Figure 5: **Additional Ablation Results.** ((*a*),(*b*)) Training saturates near 1.0 without positional encodings, while validation stalls at 0.70. ((*c*)) Confusion matrix shows persistent errors for novel sequences. ((*d*)) Restricting embeddings to symbols-only offers no significant improvement.