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 poly-factor 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

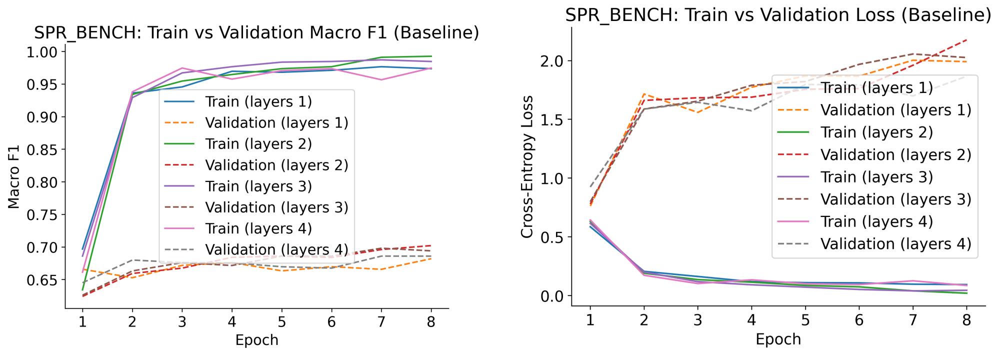
Symbolic reasoning tasks challenge neural networks to extrapolate beyond familiar patterns. In practical settings, lacking robust rule-grounded generalization often causes unreliability when new patterns arise. Researchers attempt to embed symbolic modules into neural networks to enhance interpretability 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.

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 the hallmark overfitting for baseline transformers. Training macro-F1 climbs to near 1.0 quickly, while validation saturates near 0.70. The loss curves show a similar discrepancy, confirming that the models memorize training data but fail to systematically capture underlying rules.

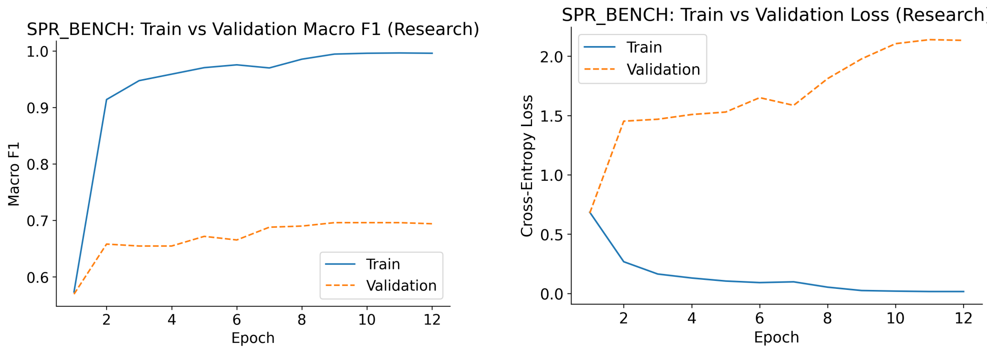
Hybrid Neural-Symbolic. Figure 2 shows analogous overfitting for the hybrid model, despite its additional symbolic features. Although early training remains stable, our symbolic approach does not substantially boost systematic generalization on hidden factor rules, leaving validation performance at ∼0.70 macro-F1.

Ablation: Removing [CLS]. Figure 3 demonstrates the impact of removing the `[CLS]` token. We observe only negligible differences in the ultimate overfitting pattern, suggesting that the root issue is not simply the presence or absence of specialized input tokens.

Confusion Matrices. We compare confusion matrices for baseline vs. hybrid in Section A. Both approaches frequently misclassify sequences with unseen factor combinations, indicating partial reliance on memorized bigrams.

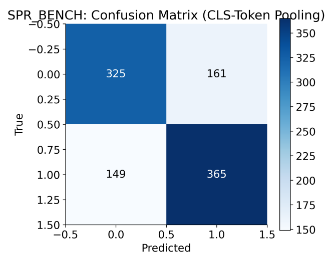
4 Conclusion

We introduced the Symbolic PolyRule Reasoning challenge to investigate how well transformer-based and hybrid neural-symbolic models generalize to previously unseen factor combinations. Despite near-perfect training performance, all approaches capped at around 70% macro-F1 on validation/test sets, demonstrating limited rule-based inference under distribution shifts. Our results

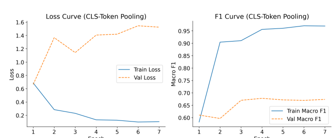


(a) Macro-F1

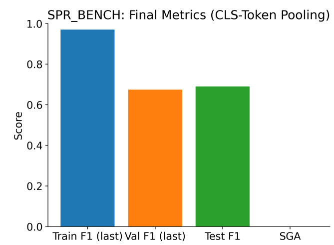
Figure 2: **Neural-symbolic hybrid model.** (*(Left)*) Training F1 saturates at nearly 1.0, while validation plateaus at 0.70. (*(Right)*) Loss curves highlight analogous overfitting.



(a) Confusion Matrix



(b) Training & Validation Curves



(c) Final Metrics

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

underscore the need for more explicit logical reasoning mechanisms or robust data augmentation strategies that help models extrapolate 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.

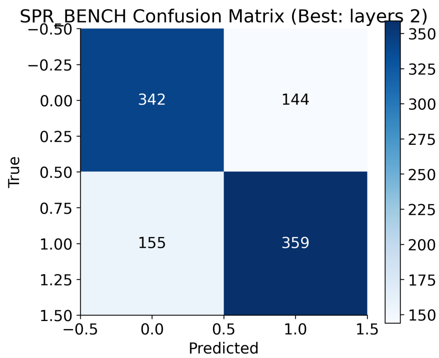
B. Pung and Alvin Chan. Orchard: A benchmark for measuring systematic generalization of multi-hierarchical reasoning. *ArXiv*, abs/2111.14034, 2021.

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.

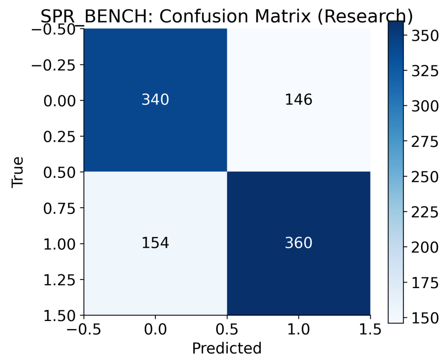
Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, M. Lee, and W. Hsu. Faithful logical reasoning via symbolic chain-of-thought. pp. 13326–13365, 2024.

SUPPLEMENTARY MATERIAL

A EXPANDED CONFUSION MATRICES



(a) Baseline

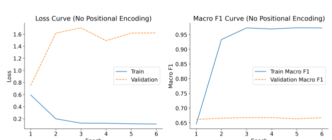


(b) Hybrid

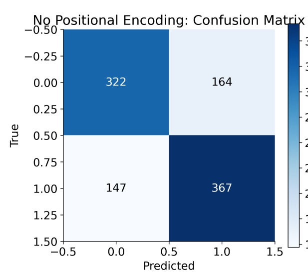
Figure 4: **Baseline vs. hybrid confusion matrices.** Both consistently struggle with sequences containing novel symbol combinations, misclassifying them at similar rates.

B FURTHER ABLATION STUDIES

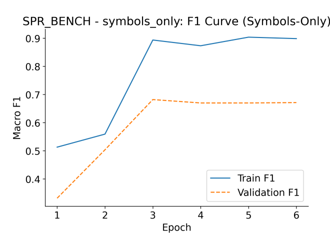
Figure 5 groups three additional ablations into one combined figure. We investigated removing positional encodings (subfigure *(a)*,*(b)*), analyzing confusion matrices without positional encodings (subfigure *(c)*), and restricting embeddings to symbols only (subfigure *(d)*). In all cases, the same overfitting pattern persists, reinforcing our conclusion that sub-symbolic pattern matching alone is insufficient for robust rule-based reasoning.



(a) No Positional Encoding: Curves



(b) No Positional Encoding: Confusion



(c) Symbols-Only: Curves

Figure 5: **Additional Ablation Results.** (*(a)*,*(b)*) Training saturates quickly without positional encodings, while validation stalls. (*(c)*) Confusion matrix shows persistent misclassifications for novel sequences. (*(d)*) Restricting embeddings to symbol-only also leads to little improvement.