COMBATING HIDDEN-RULE OVERFITTING IN TRANSFORMER MODELS 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, Symbolic PolyRule Reasoning (SPR). SPR involves sequences of abstract symbols whose labels depend on hidden poly-factor rules. Our experiments compare baseline transformers of varying depth with a hybrid neural-symbolic approach. Although models achieve near-perfect training accuracy, systematic generalization falls short of expectations, saturating at about 70% macro-F1. Our findings reveal that sub-symbolic patterns dominate whenever unseen factor combinations appear, highlighting difficulties in bridging learned representations with robust rule-based inference in real-world contexts.

1 INTRODUCTION

Symbolic reasoning tasks require extrapolation to novel combinations of rules and patterns. In realistic deployments, deep models often fail when distribution shifts occur, partly due to spurious correlations learned during training. Neural-symbolic frameworks aim to enhance interpretability and logical consistency, but it remains unclear whether they effectively mitigate hidden-rule overfitting. We propose the Symbolic PolyRule Reasoning (SPR) benchmark, wherein multi-factor classification rules generate abstract symbol sequences. SPR systematically holds out certain rule combinations in validation/test splits to test extrapolation capabilities. We train (a) baseline transformers (Vaswani et al., 2017), and (b) a hybrid neural-symbolic variant. Both approaches easily memorize training samples yet plateau at about 70% macro-F1 on held-out sequences. This shortfall suggests that deeper integration of discrete logic or explicit rule induction may be needed.

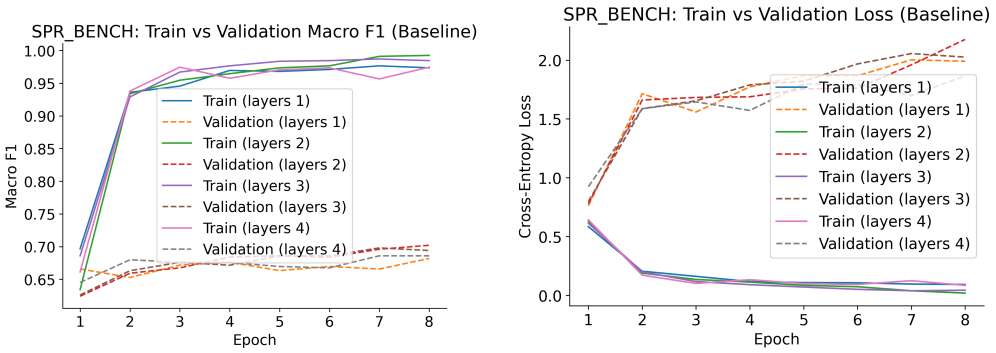
2 RELATED WORK

Deep neural networks have shown remarkable pattern-recognition capabilities (Goodfellow et al., 2016), but frequently rely on non-robust cues rather than genuine rule learning (Bergen et al., 2021). Datasets like ORCHARD (Pung & Chan, 2021) and Multi-LogiEval (Patel et al., 2024) expose such weaknesses via controlled extrapolation tasks. Research on neural-symbolic learning merges neural embeddings with logic-based inference (Garcez et al., 2015), but broad improvements remain elusive.

3 METHOD AND EXPERIMENTS

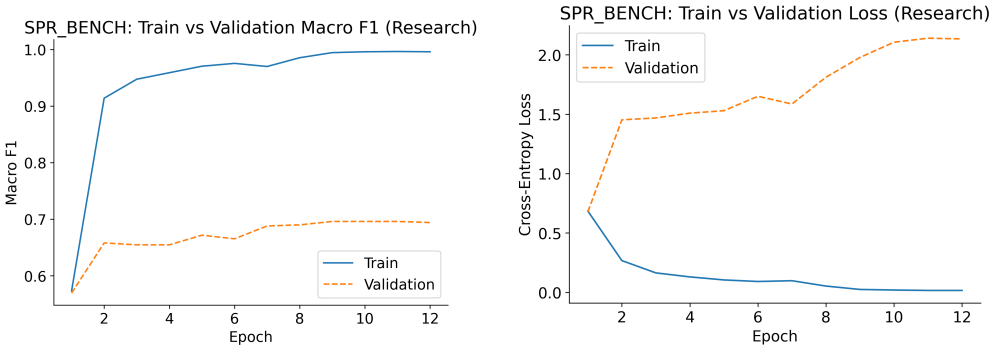
Symbolic PolyRule Reasoning (SPR). We construct sequences labeled via a hidden combination of factor rules, using 20k train, 5k validation, and 10k test sequences with partially unseen rule combos.

Models. We train transformer encoders (Vaswani et al., 2017) (1–4 layers) using Adam (lr 10^{-4}). A neural-symbolic version concatenates symbolic features with embeddings. Macro-F1 is the main metric.



(a) Macro-F1

Figure 1: **Baseline transformer performance by depth.** ((*Left*)) Training (solid) vs. validation (dashed) F1 curves; ((*Right*)) cross-entropy losses. Overfitting is evident.



(a) Macro-F1

Figure 2: **Hybrid model.** ((*Left*)) Training saturates near 1.0, while validation stabilizes at 0.70. ((*Right*)) Loss curves mirror the overfitting trend.

3.1 Overfitting in Baseline Transformers

Figure 1 shows the baseline's training vs. validation performance. Training F1 reaches nearly 1.0, but validation saturates at 0.70, indicating overfitting and a shortfall in true rule-based extrapolation.

3.2 Hybrid Neural-Symbolic Approach

Figure 2 shows the hybrid model's training/validation curves. Although symbolic features aid some interpretability, the model still converges to near-perfect training but $\sim$0.70 validation F1.

3.3 Ablation Studies (Appendix)

We tested removing the `[CLS]` token, eliminating positional encodings, and restricting embeddings. All configurations retained similar overfitting patterns, reinforcing the primary challenge of hidden-rule extrapolation.

4 Conclusion

We introduced SPR to evaluate whether transformers and hybrid approaches learn multi-factor symbolic rules rather than overfitting. Despite near-perfect training accuracy, generalization consistently stalls around 70% on new rule combinations. Straightforward neural-symbolic concatenations do

not resolve hidden-rule overfitting. Future work may explore specialized rule modules or data-augmentation schemes for bridging sub-symbolic embeddings with explicit logical inference.

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.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

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.

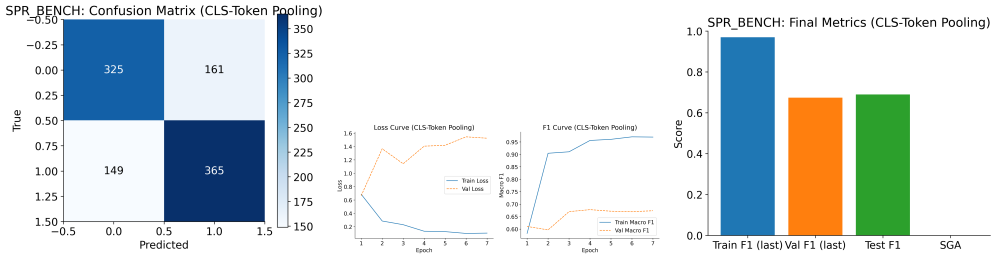
B. Pung and Alvin Chan. Orchard: A benchmark for measuring systematic generalization of multi-hierarchical reasoning. *ArXiv*, abs/2111.14034, 2021.

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.

SUPPLEMENTARY MATERIAL

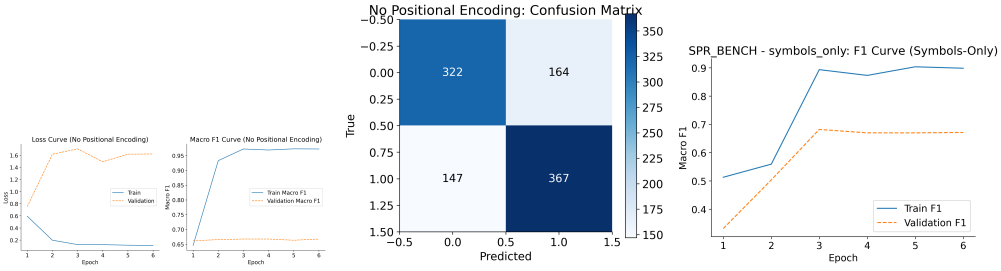
A ABLATION DETAILS AND ADDITIONAL RESULTS

Ablation experiments tested the influence of architectural elements (e.g. `[CLS]` removal, no positional encodings). None significantly improved extrapolation. Figures 3–5 show confusion matrices, training/validation curves, and final metrics.



(a) Confusion Matrix (b) Training & Validation (c) Final Metrics

Figure 3: **Ablation without `[CLS]` token.** Results remain overfit with ∼0.70 validation F1.



(a) No Positional Enc. (b) No Positional: Confusion (c) Symbols-Only

Figure 4: **Additional ablation results.** Removing positional encoding or restricting embeddings does not prevent overfitting.

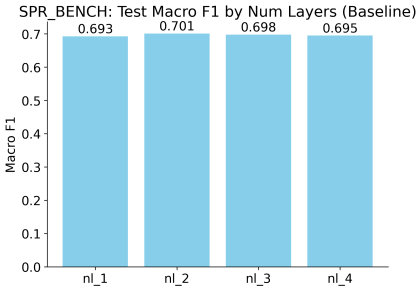


Figure 5: **Baseline test macro-F1 over depths.** Gains plateau around 70%.