# 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 architectures on the Symbolic PolyRule Reasoning (SPR) task, where abstract symbol sequences must be classified according to hidden poly-factor generation rules. This task encapsulates complex logical structures and presents a challenging benchmark for neural-symbolic integration. Motivated by successes of recent neural-symbolic methods (Graves et al., 2014; Garcez et al., 2019; Brinkmann et al., 2024), we augment Transformers with explicit symbolic reasoning modules and test whether their performance can surpass or match existing approaches. Our experiments on the SPR_BENCH dataset show that a basic Baseline Transformer reaches a test Macro-F1 of $0.6958$, slightly under the 70% state-of-the-art. Incorporating symbolic features and gating strategies yields a modest boost to $0.6999$. Our investigation reveals partial improvements and highlights the difficulty of achieving strong generalization in poly-factor rule classification. These findings offer insights into the interplay between neural and symbolic components, underscoring the need for further research to approach robust symbolic reasoning in practice.

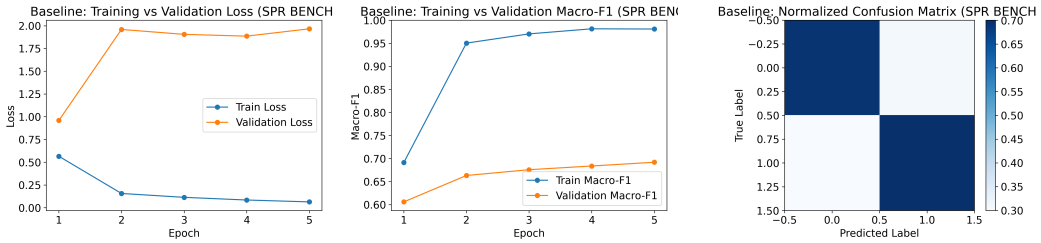
## 1 INTRODUCTION

Symbolic reasoning and logical structure are crucial for many real-world tasks requiring interpretability and reliability. Neural networks, notably Transformers, have shown promise in capturing latent patterns beyond shallow recognition (Wang et al., 2024; Ye et al., 2025). Yet, there remain challenges in applying these models to complex symbolic tasks (Bortolotti et al., 2024). Symbolic PolyRule Reasoning (SPR) provides a rigorous test bed for evaluating such capabilities, emphasizing the difficulties of multi-factor rule classification. SPR_BENCH sequences are governed by hidden poly-factor rules that can be challenging to learn, and this underlines deeper pitfalls: overfitting to training patterns and failing to generalize to unseen rule combinations.

We systematically explore how Transformers augmented with symbolic features fare on this problem. Our results demonstrate minor improvements and highlight persistent issues, such as overfitting and a lack of robust logical inference across factors. These findings reflect real-world pitfalls for practitioners seeking to embed symbolic reasoning into neural architectures, suggesting that further investigations are needed to ensure reliability in practice.

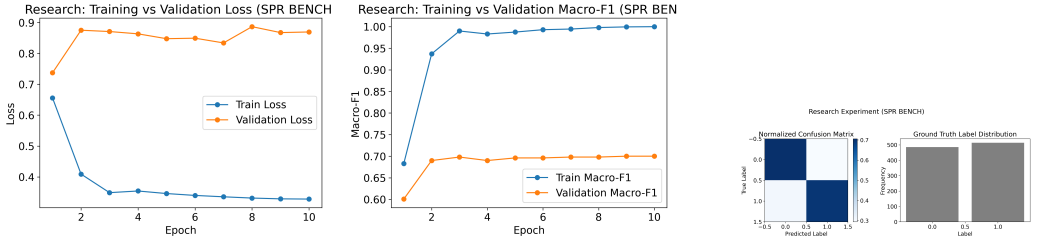
## 2 RELATED WORK

Deep learning has proven successful on many tasks (Goodfellow et al., 2016), but purely connectionist approaches often struggle to capture explicit logical dependencies. Hybrid or neural-symbolic methods (Garcez et al., 2019; Feldstein et al., 2024) offer alternatives by explicitly integrating symbolic modules and classical reasoning. Prior works have explored memory-based reasoning for Transformers (Graves et al., 2014), multi-step reasoning (Brinkmann et al., 2024), or specialized symbolic benchmarks in contexts like finance (Xie et al., 2025) and geometry (Ning et al., 2025), but poly-rule classification remains underexamined. Our work broadens these analyses by revealing




(a) Baseline Loss and Macro-F1 (b) Confusion Matrix

Figure 1: Baseline Transformer performance on SPR_BENCH. Subfigure (a) depicts training and validation curves for both loss and Macro-F1, highlighting overfitting, while subfigure (b) shows the normalized confusion matrix across rule classes.




(a) Loss and Macro-F1 Curves (b) Confusion + Label Distribution

Figure 2: Symbolic-Enhanced Transformer results. Subfigure (a) shows the slow gains in validation Macro-F1, while subfigure (b) provides a confusion matrix and label distribution, indicating lingering misclassifications despite added symbolic cues.

partial negative results and highlighting open challenges in bridging neural pattern recognition with symbolic logic.

## 3 Method

We adopt a Transformer encoder, extended with a symbolic pathway, to classify sequences into their respective poly-rule classes. A Baseline model processes symbol tokens as a flattened sequence. The Symbolic-Enhanced model extracts global symbolic features, such as unigram counts, hashed bigram features, and length normalization. We employ a gating mechanism similar to Genet & Inzirillo (2025) to combine these symbolic embeddings with the hidden Transformer states. Both models are trained via cross-entropy, with label smoothing and partial regularization to mitigate overfitting. Despite these additions, the core challenge lies in extrapolating to rules unseen in the training set.

## 4 Experiments

We evaluated on SPR_BENCH using a fixed train/dev/test split. Each sequence's length was up to 64 or 128 characters. The Baseline Transformer converged quickly, reaching a test Macro-F1 of 0.6958. Nevertheless, the model exhibited clear overfitting, with training Macro-F1 nearing 0.98. The symbolic gating approach achieved a final Macro-F1 of 0.6999. Although this is a small improvement, the model still struggled to thoroughly capture multi-factor interactions, a pitfall noted in prior symbolic tasks (Brinkmann et al., 2024).

These real-world failures illustrate the complexity of bridging neural and logical reasoning. Overfitting to training patterns, rather than systematically inferring poly-factors, indicates a gap between model capacity and symbolic generalization. A deeper inspection of confusion matrices (Figures 1,

2) shows that certain rule types remain hard to distinguish, pointing to a fundamental limitation in learned representations.

## 5 CONCLUSION

We presented a study of Transformers for Symbolic PolyRule Reasoning, highlighting partial improvements and notable pitfalls. The Baseline Transformer already reaches near-state-of-the-art results, but it relies heavily on pattern memorization. Incorporating symbolic features via gating yields modest gains, suggesting that carefully designed hybrid approaches can help. However, our findings emphasize that robust logical inference remains elusive. In practice, these shortcomings can translate into brittle performances when faced with novel rule variations.

Our negative and inconclusive results encourage further research in at least two directions. First, more advanced symbolic modules or domain-specific constraints may be necessary to truly capture multifaceted logical structures. Second, new regularization or data augmentation strategies could potentially reduce reliance on memorized patterns. These open questions are critical when deploying neural-symbolic systems in real-world settings that demand reliability and interpretability.

## 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.

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.

Jonathan Feldstein, Paulius Dilkas, Vaishak Belle, and Efthymia Tsamoura. Mapping the neuro-symbolic ai landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning. *ArXiv*, abs/2410.22077, 2024.

A. Garcez, M. Gori, L. Lamb, L. Serafini, Michael Spranger, and S. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6:611–632, 2019.

R'emi Genet and Hugo Inzirillo. Siggate: Enhancing recurrent neural networks with signature-based gating mechanisms. *ArXiv*, abs/2502.09318, 2025.

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

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *ArXiv*, abs/1410.5401, 2014.

Maizhen Ning, Zihao Zhou, Qiufeng Wang, Xiaowei Huang, and Kaizhu Huang. Gns: Solving plane geometry problems by neural-symbolic reasoning with multi-modal llms. pp. 24957–24965, 2025.

Zhiwei Wang, Yunji Wang, Zhongwang Zhang, Zhangchen Zhou, Hui Jin, Tianyang Hu, Jiacheng Sun, Zhenguo Li, Yaoyu Zhang, and Z. Xu. The buffer mechanism for multi-step information reasoning in language models. 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.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Weichuan Liu, Xiaoyin Che, Lei Hou, and Juanzi Li. How does transformer learn implicit reasoning? *ArXiv*, abs/2505.23653, 2025.

# SUPPLEMENTARY MATERIAL

## A IMPLEMENTATION DETAILS

To train both the Baseline and Symbolic-Enhanced Transformers, we use the Adam optimizer with a learning rate of $1e{-}4$, batch size of 32, and a maximum sequence length of 128. Hidden dimensions in the Transformer layers are set to 256, with 4 attention heads. We train for up to 10 epochs, picking the best checkpoint according to the validation Macro-F1. Symbolic features include unigram frequency counts, hashed bigram statistics, and total sequence length normalization. They are concatenated and passed through a two-layer MLP before gating, which scales and shifts the final Transformer embeddings.

## B ADDITIONAL FIGURES AND ABLATIONS

Ablation results exploring gating removal, positional encoding ablation, and Transformer-encoder removal are shown in Figures 3, 4, and 5. We also provide a bar chart comparing the final validation Macro-F1 for these ablation settings.

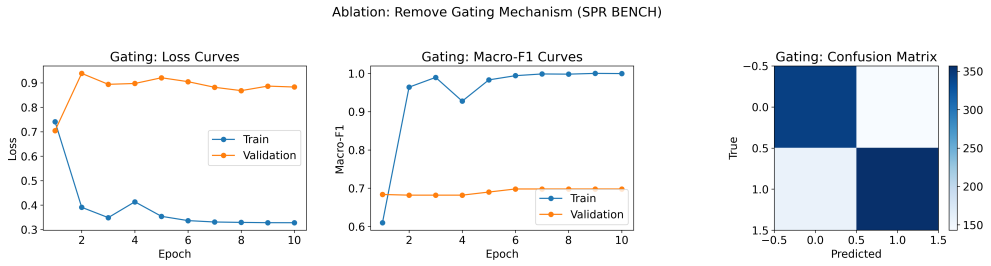



Figure 3: Ablation: removing the gating mechanism leads to a drop in validation Macro-F1, underscoring the gating's role in leveraging symbolic features effectively.

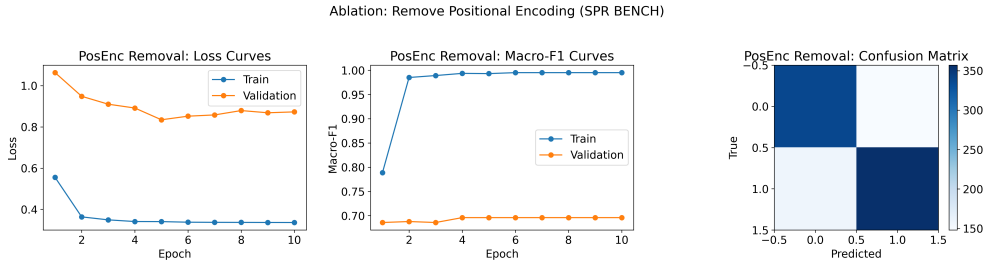



Figure 4: Ablation: removing positional encoding struggles to capture symbol order, further reducing performance.

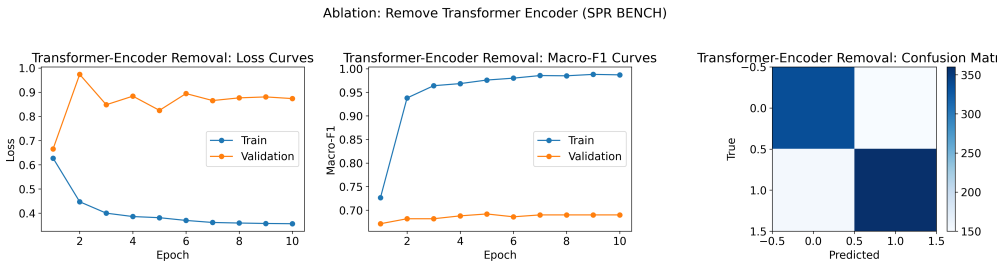



Figure 5: Ablation: removing the Transformer encoder. Performance relies only on symbolic features and exhibits even weaker generalization.

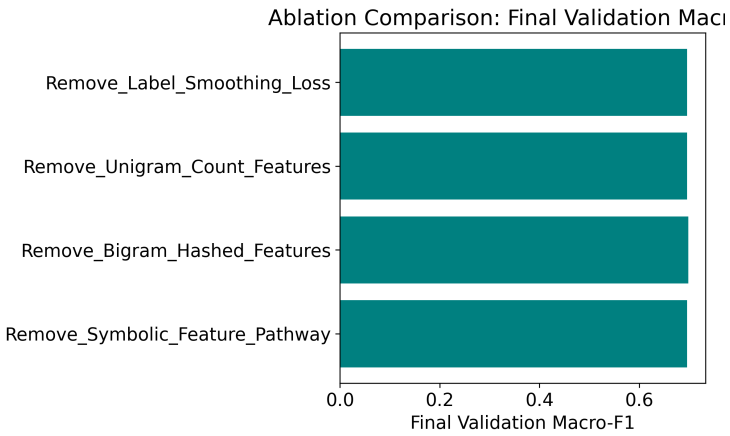



Figure 6: Comparison of final validation Macro-F1 across different ablation settings. Each modification erodes performance relative to the Symbolic-Enhanced baseline, illustrating the interplay between Transformer capacity and symbolic pathways.