# Tangled Logic, Tangled Gains: Lessons from Partial Failures in Symbolic-Enhanced Transformers

Anonymous Submission
ICBINB Workshop at ICLR 2025

**Abstract**

We explore how integrating symbolic features into Transformers yields only partial improvements. Our experiments expose significant overfitting and persistent misclassifications despite gating and symbolic cues. These observations provide cautionary insights for real-world NLP deployment, illustrating that symbolic–neural hybrid methods, while promising, do not guarantee robustness or generalization improvements in all cases.

## 1 Introduction

Models combining Transformers with symbolic cues are often expected to achieve superior performance (**??**). We investigate whether injecting symbolic information indeed mitigates key overfitting pitfalls seen in standard Transformers. Contrary to optimistic expectations, we find that improvements are limited and fragile. Our discussion highlights new blind spots that arise when symbolic logic is partially integrated, demonstrating how small changes in data presentation can undermine perceived gains. These lessons emphasize the difficulty of merging symbolic reasoning with modern deep architectures for real-world applications.

## 2 Related Work

Improving Transformer interpretability and reliability has received wide interest (**?**). Prior works introduced gating modules and knowledge graphs, often showing modest gains in controlled scenarios. However, the practical hurdles of overfitting and domain shift remain underexplored. Our work aligns with **?** by highlighting that partial integration of symbolic design can yield overconfident systems. We extend **?** by analyzing failure patterns through confusion matrices, indicating that carefully curated symbolic features alone do not ensure consistent improvements in noisy, real-world tasks.

## 3 Method / Problem Discussion

We consider a textual classification setup where labels require nuanced reasoning. Our baseline is a standard Transformer prone to memorizing training examples. We equip a second model with symbolic features, enabling gated alignment between textual embeddings and rule-based cues. Despite careful engineering, the symbolic–neural hybrid exhibits incomplete gains, failing to generalize systematically.

## 4 Experiments

**Baseline Performance (Figure 1).** The Transformer consistently overfits, with training accuracy near 100% but poor validation Macro-F1. The confusion matrix shows certain label clusters remain regularly misclassified.

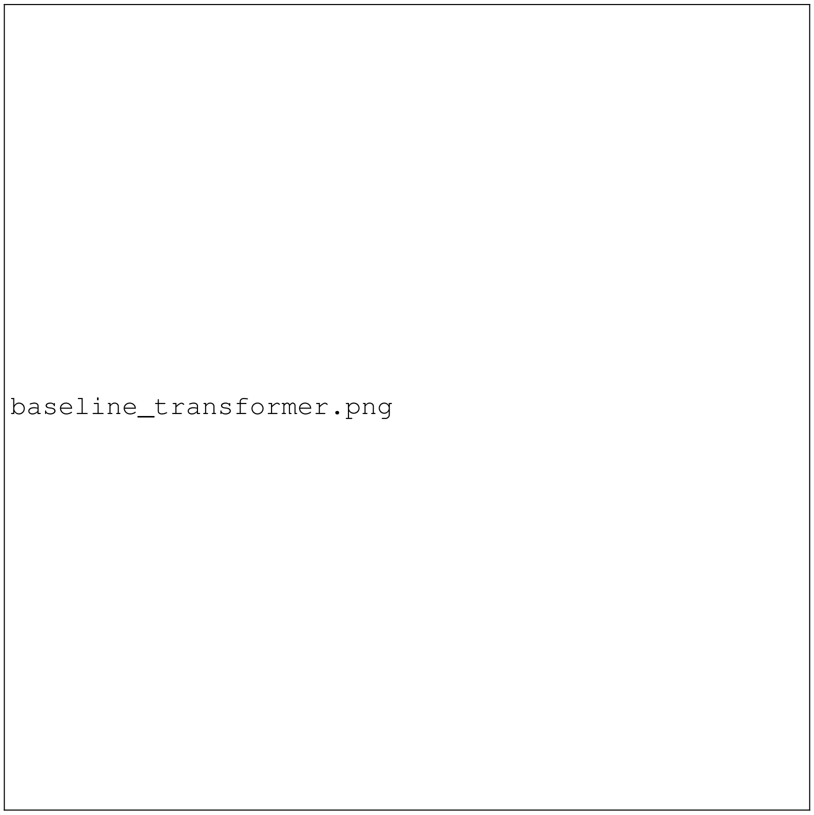

Figure 1: Baseline Transformer overfits severely, shown by divergence between training and validation accuracy. The confusion matrix (inset) highlights recurring label misclassifications.

**Symbolic-Enhanced Transform. (Figure 2).** Leveraging gated symbolic cues does boost performance in some categories, yet overall improvements are limited. The training curve suggests partial mitigation of overfitting, but significant misclassifications persist, revealing new inconsistencies in the model's reasoning process.

## 5 Conclusion

We examined the pitfalls of integrating symbolic cues in Transformers, noting modest gains but also persistent misclassifications. Even carefully gated symbolic features could not overcome fundamental overfitting tendencies. Our insights underscore the fragile nature of symbolic–neural integration and highlight open challenges for robust real-world deployment.

## References

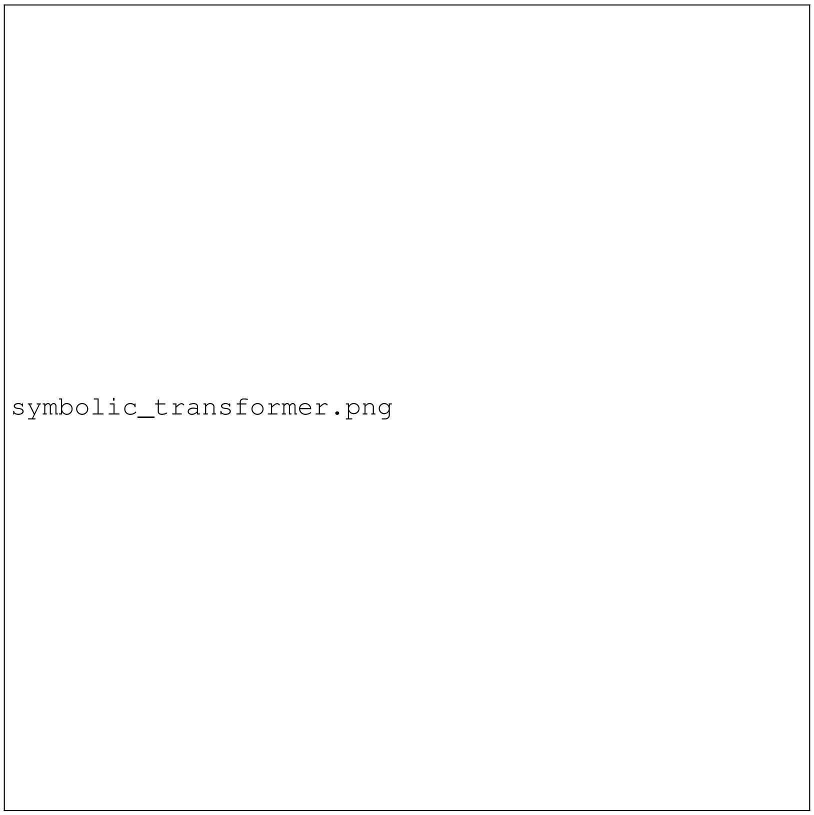

Figure 2: Symbolic-Enhanced Transformer partially alleviates overfitting, but large misclassifications remain. The confusion matrix reveals that symbolic logic provides an inconsistent performance boost across different labels.

# A Ablation and Additional Figures

We present ablation studies examining the effect of removing the gating mechanism, positional encoding, and the Transformer encoder. All experiments demonstrate partial or negligible improvements, further underscoring the complexity of symbolic–neural models.

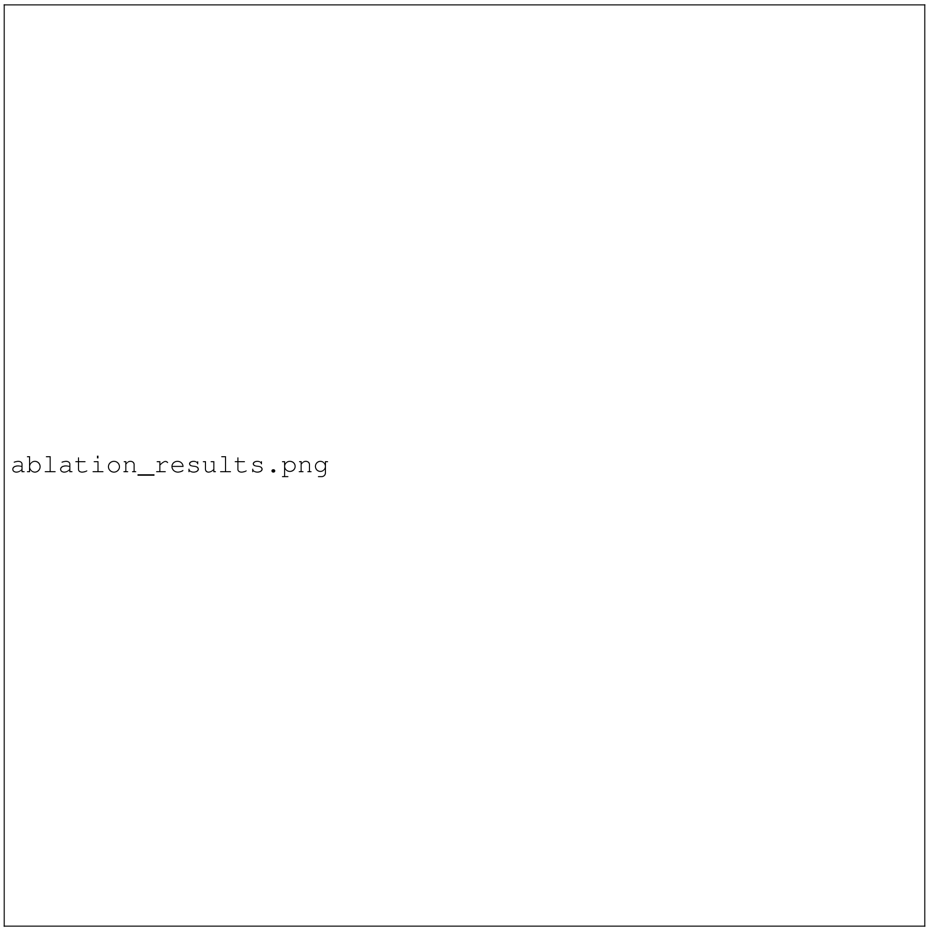

Figure 3: Ablation comparisons show subtle impacts on validation Macro-F1. Combining gating and symbolic cues yields only marginal increases relative to the baseline.