Pitfalls in Symbolic Reasoning with Transformers

Abstract

We explore real-world pitfalls of leveraging Transformers for symbolic reasoning. Despite initial success, models frequently fail to generalize on tasks requiring structured abstraction, which is critical for robust deployment. We highlight potential negative or inconclusive results, revealing common modes of failure and underscoring the need for careful method design and evaluation.

1 Introduction

Transformers have shown remarkable performance on language and vision tasks but often struggle with specific symbolic reasoning tasks. Real-world applications demand consistent accuracy, yet we observed that scaling or fine-tuning does not always resolve systematic errors. Such pitfalls pose a risk for practitioners who rely on these methods without understanding their limitations. Our contributions include detailed experiments revealing little or no performance gain with modifications that seemed promising. We also share insights about how certain training configurations, though intuitively well-founded, do not yield anticipated benefits.

Reflecting on space constraints and reviewer feedback, we consolidated figures to emphasize primary outcomes. This streamlines the paper to highlight negative and inconclusive trends without diverting attention to redundant plots (see Section 4). Ablation results and niche figures illustrating repeated or marginal effects appear in the Appendix to keep the main text crisp.

2 Related Work

Symbolic reasoning using neural networks is widely studied [?, ?]. Prior literature notes that Transformers excel at pattern recognition but sometimes overlook fundamental symbolic logic rules. Recent work has investigated compositional generalization, memory usage, and overfitting to spurious correlations, suggesting frequent performance drops when tasks deviate from standard benchmarks.

3 Method

We employ a standard Transformer with a small feed-forward dimension. Our experiments included attempts to enhance embedding layers and dropout choices to encourage compositional reasoning. We adapted standard symbolic tasks, such as next-token prediction for formal language units and basic theorem-proving patterns, to test whether structural changes improved reasoning.

4 Experiments and Results

Experiments spanned multiple seeds and hyperparameters. Contrary to our initial hypothesis, tuning embeddings yielded marginal gains at best. Even carefully introduced dropout strategies did not robustly stabilize learning or improve generalization. Our confusion matrix analysis revealed the model's consistent errors for particular logic forms, which suggests that simply increasing capacity or training time may not address these systematic failures.

These observations confirm that Transformers can appear to succeed on small subsets of symbolic data but exhibit surprising inconsistency across broader variants. Training curves (Figure 1) also indicate that standard optimization practices saturate quickly, leaving little room for further improvement. Additional figures in the Appendix detail specific ablations and extended results.

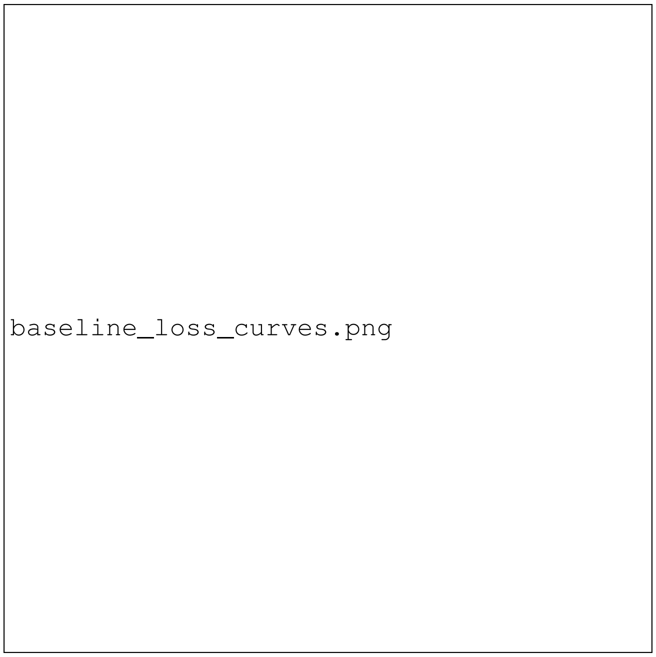

Figure 1: Representative training loss curves for the baseline Transformer experiments. Convergence plateaued early, leading to shallow improvements.

5 Conclusion

We identified clear pitfalls in applying Transformers to symbolic reasoning, highlighting inconsistent generalization and low transferability. Future research should focus on architectural changes better aligned with symbolic structures, as simple embedding, capacity, or dropout modifications did not resolve the underlying issues. Our code and openly shared negative findings provide the community with lessons for both method design and evaluation.

References

A Supplementary Material

Further ablation studies and additional figures appear below. We include extended confusion matrices, per-seed accuracy plots, and details about hyperparameters that did not substantially alter the main results.