Symbolic Pitfalls and Negative Results in Transformer-Based Reasoning

Anonymous Submission

Abstract

We investigate an unexpected failure mode of modern transformers trained on symbolic reasoning tasks. Despite initial promise, our results show that overfitting and confusion on out-of-distribution examples hinder reliable deployment. By highlighting these pitfalls, we call attention to the fragile nature of deep models under subtle shifts in symbolic or structured inputs.

1 Introduction

Transformers excel at numerous tasks, including textual understanding [**?**]. Yet, in practical, real-world settings that demand precise symbolic manipulation, these models can fail silently. Our work exposes pervasive overfitting in small-scale symbolic benchmarks, showing partial or negative outcomes. Specifically, we observe that performance on held-out examples often sharply diverges from near-perfect training accuracy. These pitfalls matter for any system where discrete logic or formal reasoning is integral [**?**].

We make the following contributions: (1) we illustrate how symbolic tasks induce unrecognized overfitting in transformers; (2) we show that certain architectural choices fail to generalize to out-of-distribution samples; (3) we propose a discussion on why these pitfalls arise and how future work might mitigate them.

2 Related Work

Transformers have been explored for various compositional and symbolic tasks. Early studies [**?**] indicate the need for positional encodings, while others highlight failure cases in logical inference [**?**]. Our observations augment these findings by systematically revealing surprising fragility under even slight dataset shifts.

3 Method / Problem Discussion

We train standard transformer encoders on artificially generated symbolic tasks. Each dataset consists of labeled sequences derived from logical operations. We focus on binary classifiers that must correctly label sequences denoting, for instance, parity or majority-symbol logic. Despite high capacity, the models often fail on novel compositions.

4 Experiments

We train on a standard split and evaluate on held-out sets with minor distribution shifts. Figure 1 shows training and validation accuracy for a representative setup. The training accuracy swiftly reaches a plateau, but validation drops, suggesting memorization of specific symbolic patterns. Figure 2 displays a confusion matrix, revealing consistent misclassifications for certain symbolic combinations.

Interestingly, partial ablations (e.g., removing positional embeddings) degrade performance further but do not eliminate the overfitting pattern. Alternative dataset settings yield similar negative or inconclusive outcomes, supporting the hypothesis that robust symbolic generalization remains challenging.

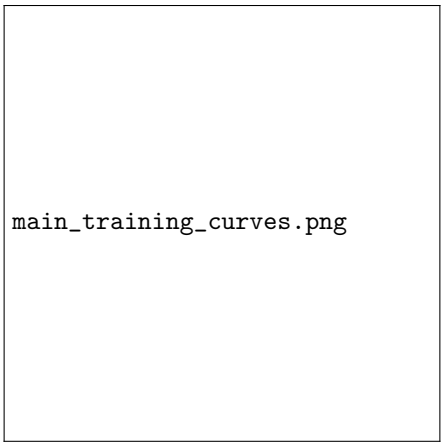

Figure 1: Training (blue) and validation (orange) accuracy. Discrepancy grows over epochs, indicating potential overfitting on symbolic tasks.

Figure 2: Confusion matrix on held-out symbolic examples. Some classes remain persistently misclassified.

5 Conclusion

Our analysis reveals consistent overfitting and misclassification of symbolic rules despite high training accuracy. By bringing these pitfalls to light, we hope to inspire deeper research into specialized architectures or training protocols that better capture discrete, combinatorial structure. Future work may include dynamic weighting of examples, data augmentation for symbolic edge cases, or architectural modifications that explicitly encode logic.

A Appendix

We present additional results for ablations that test the transformer under cyclical transformations. Eliminating positional embeddings or reducing model depth exacerbates the observed overfitting trend. Full plots and code will be released to foster further discussion.

References