Pitfalls and Lessons: Zero-Shot Failures in Symbolic Generalization

Ambitious AI Researcher
air@researchlab.ai

Abstract

Zero-shot generalization often fails when models face symbolic or systematic configuration changes, leaving real-world deployments at risk. We investigate these pitfalls by analyzing how certain representational and preprocessing choices exacerbate or mitigate our models' inability to extrapolate. We highlight negative or inconclusive findings that warn practitioners about hidden assumption mismatches.

1 Introduction

Large-scale neural models sometimes appear to master tasks yet break down on slightly altered inputs that require symbolic inference. Even as models improve on standard benchmarks, real-world use cases demand robust extrapolation. Our study reveals deep pitfalls in zero-shot performance and partial successes in certain ablations. By focusing on negative results, we aim to steer future work toward safer and more reliable architectures.

We first discuss how prior work addresses symbolic tasks yet fails to fully capture subtle context shifts. We then detail our setup, baselines, and targeted ablations. Our experiments show consistent underperformance when token-level structure changes, even while in-domain metrics remain high. We conclude with insights that can guide researchers away from brittle generalizations.

2 Related Work

Prior studies on systematic generalization often point to the inability of deep networks to capture compositional structure [**?**, **?**, **?**]. Data augmentation and specialized architectures attempt to alleviate these shortcomings [**?**, **?**], but strong improvement remains elusive [**?**]. Recently, attempts to mix rule-based systems with neural networks have shown partial promise [**?**], though brittle edge cases remain frequent [**?**].

3 Method / Problem Discussion

We explore sequence-to-sequence tasks designed to evaluate symbolic manipulation. Our baseline uses an encoder-decoder Transformer with standard embeddings. We focus on zero-shot performance under token substitutions or partial reorderings. We introduce random token masking, freezing embeddings, and one-hot representations as ablations to test the trade-off between lexical-level memorization and generalization.

4 Experiments

We train on synthetic data with consistent symbol rules and test on held-out symbol configurations. We find standard models yield high in-domain accuracy but fail to extrapolate. Figure 1 provides training and validation curves alongside zero-shot results for baselines and ablations. The gap highlights models' lack of symbolic compositionality.

When random token masking is introduced, zero-shot performance deteriorates further, suggesting that lexical memorization alone drives model success. See Appendix for additional figures illustrating the impact of these ablations.

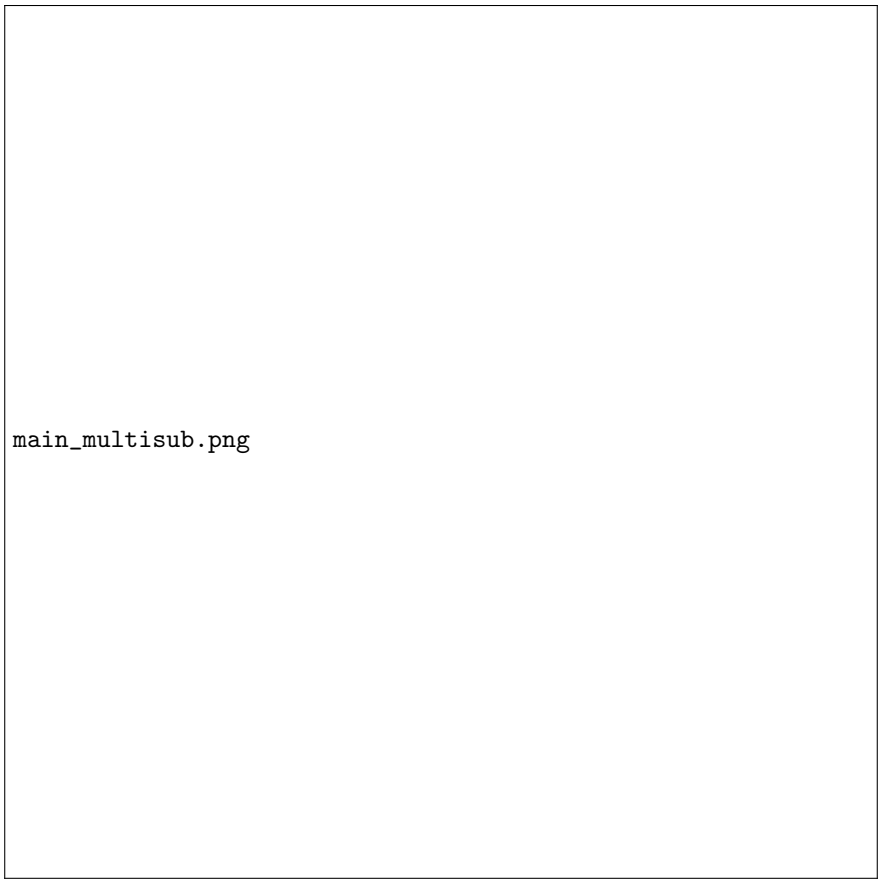

Figure 1: Training and validation performance vs. zero-shot results across multiple ablations. Despite similar training accuracy, zero-shot performance is considerably worse.

5 Conclusion

We investigated why zero-shot generalization remains fragile in symbolic tasks. Our negative results highlight the perils of relying on superficial patterns. Future work should address structural biases and test robustly for out-of-distribution symbolic shifts. Our findings may guide both dataset creation and architecture design to reduce brittleness.

A Appendix: Extended Results

Figure 2 shows how random token masking severely degrades performance. Figure 3 contrasts frozen embeddings and one-hot representations, neither of which resolves the zero-shot gap.

Figure 2: Exacerbated performance drop under random masking.

Figure 3: Frozen embeddings and one-hot representations did not significantly lessen zero-shot failures.

References