

Router-S: Conditional Compute for Language Models

Contents

1 Introduction

2 Method

3 Results

References

1

2

3

4

Abstract

Sparse routing reduces the compute a dense model spends on tokens that are trivially predictable. We evaluate Router-S against a dense baseline on an identical token budget. The headline result is that Router-S reaches

Our evaluation harness holds the data order fixed across runs so that differences in loss cannot be attributed to shuffling. Each configuration is trained three times and we report the median, since seed variance

Introduction

a validation loss of 3.41 while consuming five times fewer FLOPs than the dense baseline it replaces. That saving is not uniform across the corpus; it concentrates on passages the router marks as low entropy.

exceeds the gap we are measuring here. Prior work on conditional computation reports similar savings but measures them on synthetic corpora. We show that the router learns the skipping policy unaided.

Method

Sparse routing reduces the compute a dense model spends on tokens that are trivially predictable. We evaluate Router-S against a dense baseline on an identical token budget. The headline result is that Router-S reaches a validation loss of 3.41 while consuming five times fewer FLOPs than the dense baseline it replaces. That saving is not uniform across the corpus; it concentrates on passages the router marks as low entropy.

Our evaluation harness holds the data order fixed across runs so that differences in loss cannot be attributed to shuffling. Each configuration is trained three times and we report the median, since seed variance exceeds the gap we are measuring here. Prior work on conditional computation reports similar savings but measures them on synthetic corpora. We show that the router learns the skipping policy unaided.

Results

Sparse routing reduces the compute a dense model spends on tokens that are trivially predictable. We evaluate Router-S against a dense baseline on an identical token budget. The headline result is that Router-S reaches a validation loss of 3.41 while consuming

Our evaluation harness holds the data order fixed across runs so that differences in loss cannot be attributed to shuffling. Each configuration is trained three times and we report the median, since seed variance exceeds the gap we are measuring here.

References

- [1] A. Researcher. Conditional computation. 2024.
- [2] B. Author. Sparse routing at scale. 2025.
- [3] C. Scientist. Entropy and capacity. 2026.