Early-Exit Routing via Intermediate RWKV Confidence Scores and GLUE Benchmark Performance Relative to Full-Depth Processing

Assignee Research

June 11, 2026

Abstract

Early Exiting (EE) is a promising technique for speeding up inference by adaptively allocating compute resources to data points based on their difficulty. The approach enables predictions to exit at earlier layers for simpler samples while reserving more computation for challenging ones. In this study, we first present a novel perspective on the EE approach, showing that larger models deployed with EE can achieve higher performance than smaller models while maintaining similar computational costs. As existing EE approaches rely on confidence estimation at each exit point, we further study the

1 Introduction

This paper examines: Performance Control in Early Exiting to Deploy Large Models at the Same Cost of Smaller Ones. Research question: To what extent does early-exit routing based on intermediate RWKV layer confidence scores affect GLUE benchmark performance compared to full-depth processing?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

10 papers retrieved. 12 claims extracted; 10 independently verified. Quality review score: 8.1/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
PCEE offers a simple yet computationally efficient approach that provides better control over performance than standard	✓	0.35
PCEE allows us to scale up model sizes to yield performance gain while reducing the computational cost.	✓	0.26
PCEE provides control over accuracy for any model returning a confidence score and a classification decision at each exi	✓	0.24
PCEE requires selecting one single threshold for all layers, unlike existing early exit methods that require learning a	✓	0.24
The threshold in PCEE is a simple accuracy lower bound—based on the target accuracy level chosen by the user—rather than	✓	0.26
Larger models require a reduced amount of computation to attain a certain accuracy level by exiting at very early layers	✓	0.28
Early Exit Neural Networks enable dynamic resource allocation during model inference, reducing computational demands by	✓	0.31
In experiments, H = 150 and 50 bins were used for the reliability diagrams.	×	0.07
Exit layers are implemented as fully-connected layers that output logits for a softmax layer.	×	0.05
Accuracy thresholds offer a simple approach to determine the earliest exit point that guarantees at least the desired ac	✓	0.27
Confidence estimates can present inconsistent behavior throughout layers, hence requiring the selection of a different t	✓	0.23
The decision to exit early is made if the confidence measure ci(x) at a given layer i exceeds a threshold δ $\in$[0, 1].	✓	0.17

References

- http://arxiv.org/abs/2407.11087v3
- http://arxiv.org/abs/2502.14620v1

- http://arxiv.org/abs/2412.19325v1