Resourceful LayerSkip: Fast Yet Accurate Early Exit Inference

ACL Anonymous Submission

Abstract

We present Resourceful LayerSkip, an innovative approach that accelerates the inference process of large language models (LLMs) while preserving high accuracy. This method strategically leverages feedforward layers, aiming to refine tokens into the desired representation space. Our empirical evaluations across diverse datasets and model sizes demonstrate that Resourceful LayerSkip consistently surpasses traditional early exit methods in terms of accuracy, stability, and utility. Specifically, our approach reduces inference time by approximately 40% compared to full inference processes to reach the performance benchmark of 50% accuracy. This efficiency gain significantly lowers computational costs in LLM deployment, highlighting its potential for scalable and sustainable use of advanced language models.

1 Introduction

Large Language Models (LLMs) have become integral to natural language processing (NLP), enabling significant advances in tasks such as text generation and translation. However, the deployment of these models in real-world applications often encounters challenges related to their substantial computational requirements and slow inference speeds. These issues not only impact scalability but also pose environmental sustainability concerns due to the high energy consumption associated with extensive computational processes (Miao et al., 2023).

To address these challenges, various early exit strategies have been proposed (Panda et al., 2016; Teerapittayanon et al., 2016; Schuster et al., 2022), aiming to optimize the inference process by allowing models to terminate the generation process early, once a satisfactory output is reached. While these methods, such as LayerSkip (Elhoushi et al., 2024), have shown promise in reducing energy consumption and computational time, these methods often struggle to maintain high inference accuracy and stability. In response, we have developed a novel method called Resourceful LayerSkip, which builds upon traditional early exit strategies by leveraging later feedforward layers. This paper explores the effectiveness of Resourceful LayerSkip in optimizing LLM inference, improving inference accuracy and stability, focusing on its application to various NLP tasks and datasets.

2 Methods

2.1 Resourceful LayerSkip Architecture

The Resourceful LayerSkip approach modifies the traditional early exit inference mechanism, focusing on optimizing the balance between computational efficiency and predictive accuracy. The core idea is to strategically bypass the attention layers after a certain depth while retaining all subsequent feedforward layers. This approach is grounded in the efficiency of feedforward layers, as shown in a recent study, there is a potential of utilizing feedforward layers only to mimic the effect of attention mechanisms (Bozic et al., 2024).

Attention Layers: These layers gather and integrate relevant context from the input sequence. Multi-head attention captures dependencies between tokens by computing attention scores that weigh each token's contribution (Vaswani et al., 2023).

Feedforward Layers: Following the attention layers, the position-wise feedforward neural networks (FFNs) are crucial for refining token representations. This structure introduces non-linearity, allowing the model to learn complex patterns. The FFNs are applied identically to each token, independently refining their representations based on the contextual information aggregated by the attention layers.

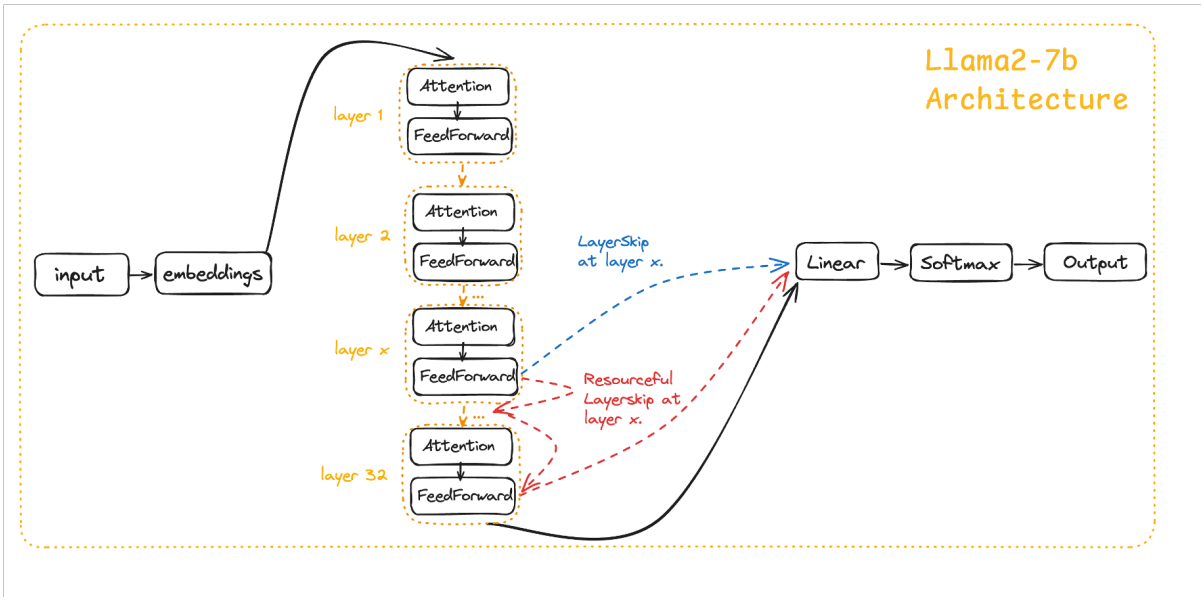

Figure 1: Model Architecture: Early Exit Inference

2.2 Inference Architecture Modification

In the Resourceful LayerSkip approach, an intermediate layer l is chosen where the attention mechanism is deemed to have gathered sufficient contextual information. From layer l onward, the model bypasses the attention layers via residual connections, allowing the feedforward layers to continue processing and refining the token representations. The output at any layer l using Resourceful LayerSkip, denoted as $\hat{y}_{res_l}$, is computed as:

$$\hat{y}_{res_l} = f(FFN_L(FFN_{L-1}(...FFN_{l+1}(h_l)...)))$$

where h_l is the hidden state at layer l, FFN_i represents the feedforward network of layer i, and f is the classification head. This formulation ensures that the output remains close to full inference, as the refined token representations continue to be refined through the deeper feedforward layers.

2.3 Computation Cost Analysis

The computational cost savings are directly tied to the reduced number of attention operations. In practical scenarios, this can translate into substantial reductions in GPU hours and energy consumption. Let's denote the total computational cost as C_{total} and the cost per layer as C_{layer}. Without layer skipping, the cost is:

$$C_{\text{total}} = L \times C_{\text{attention}} + L \times C_{\text{feedforward}}$$

With Resourceful LayerSkip, the effective computational cost becomes:

$$C_{\text{RL}} = (L - k) \times C_{\text{attention}} + L \times C_{\text{feedforward}}$$

where L represents the total number of layers and k represents the number of skipped attention layers.

In transformer models, attention layers have a time complexity of $O(n^2 d)$, where n is the sequence length and d is the dimensionality of the model. This quadratic dependency on the sequence length makes attention layers computationally expensive, especially for long inputs. In contrast, feedforward layers have a time complexity of $O(nd^2)$, which, although dependent on the model dimensionality, is linear with respect to the sequence length.

In many practical scenarios having large input size, the sequence length n tends to be much larger than the model dimensionality d. Given that $C_{\text{attention}} > C_{\text{feedforward}}$, bypassing attention layers results in significant cost savings.

2.4 Empirical Layer Selection

Layer selection is crucial for the success of Resourceful LayerSkip because it balances the trade-off between computational efficiency and model performance. Empirical evidence suggests that exiting at around 50% of the total layers can effectively balance computational efficiency and model performance. For instance, in the 32-layer Llama2-7B

model (Touvron et al., 2023), Resourceful LayerSkip typically achieves 50% accuracy by exiting at the 16th layer, compared to the 29th layer required by direct early exit method, demonstrating significant efficiency gains. However, the empirical layer selection result depends on the type of model and dataset we use.

3 Experimental Setup

Our experiments were conducted on the Nvidia DGX platform, equipped with eight H100 GPUs, each with 80 GB of memory. The dataset used in our study encompasses a diverse range of tasks, including GPT4-QA for question answering (Peng et al., 2023), Truthfulness for evaluating factual correctness (Lin et al., 2022), Human Ethics for ethical decision-making (Hendrycks et al., 2020), Mathematics for arithmetic problem-solving (Brown et al., 2020), Multi-Language for testing capabilities across different languages (Bandarkar et al., 2023), and Translation for assessing translation accuracy (Čihař, 2024).

To assess the effectiveness of Resourceful LayerSkip, we compared it against traditional direct early exit, which serves as a baseline and a simplified variant of LayerSkip. For consistency, we will refer to this baseline as LayerSkip in subsequent sections. Performance evaluation was carried out using key metrics such as token-wise utility, which measures efficiency on a per-token basis, and layer-wise accuracy, which captures the accuracy achieved at different layers of the model.

Our primary experiments were conducted using the Llama2-7B model, and we extended our evaluations to larger models within the Llama2 family to test scalability and performance across model sizes.

4 Results and Discussion

This section provides a detailed analysis of the performance of the Resourceful LayerSkip method compared to LayerSkip across various datasets and model sizes.

4.1 Performance Across Different Datasets

To comprehensively evaluate the effectiveness of Resourceful LayerSkip, we conducted experiments on a diverse range of datasets, each representing distinct natural language processing tasks. The results, depicted in Figure 2, highlight the method's performance across these datasets, demonstrating improvements in computational efficiency and accuracy across layers.

As shown in the Appendix 3, the GPT4-QA dataset, which involves factual question answering and suggestion generation tasks, and the Truthfulness dataset, designed to test the model's ability to provide factually correct responses, exhibit similar performance trends. Resourceful LayerSkip demonstrates a significant improvement over traditional LayerSkip, achieving around 50% accuracy by the 15th layer compared to only 15% for LayerSkip. By the 28th layer, Resourceful LayerSkip surpasses 90% accuracy, effectively reducing the computational cost while maintaining high prediction accuracy earlier in the inference process.

On the Mathematics dataset, composed of arithmetic and algebraic tasks, Resourceful LayerSkip shows notable efficiency. It achieves 50% accuracy by the 14th layer, whereas traditional LayerSkip requires nearly all 32 layers to reach a comparable level of accuracy.

In the Human Ethics dataset, both early exit methods perform effectively, likely due to the structured nature of the dataset, a hypothesis further confirmed by fine-tuning the model on structured ethical tasks, which enhanced early-exit performance.

Resourceful LayerSkip was evaluated on the Belebele multi-language dataset (Bandarkar et al., 2023), which covers a range of diverse languages, and on translation tasks. For languages well-represented in the original pre-training set of Llama2 (Touvron et al., 2023), such as English, it achieved higher accuracy compared to less-represented languages. Similarly, in translation tasks, well-represented languages exhibited better performance. Resourceful LayerSkip reached 50% accuracy with approximately 16 layers, while traditional LayerSkip required 28 layers to achieve comparable results.

4.2 Performance Across Different Model Sizes

Experiments were conducted across three model sizes: Llama2-7B, Llama2-13B, and Llama2-30B, with a focus on understanding how model size affects the efficiency of early exit methods.

The 32-layer Llama2-7B model serves as a baseline for evaluating early exit strategies. Resourceful LayerSkip achieves 50% accuracy by the 13th layer, significantly reducing the computational cost compared to the traditional method, which requires up to 29 layers to achieve similar accuracy. This approach results in approximately 40% savings in

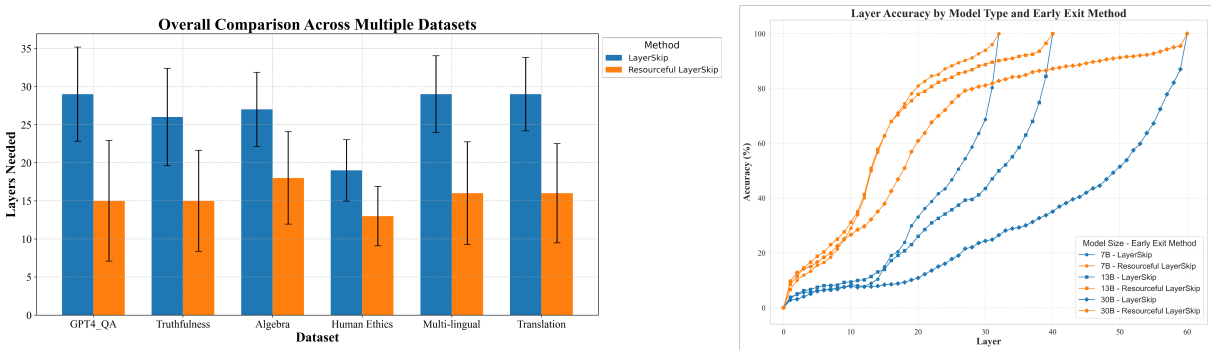

Figure 2: Performance Across Different Datasets and Different Model Size.

computational cost relative to full inference.

Scaling up to the Llama2-13B model with 40 layers, Resourceful LayerSkip continues to demonstrate efficiency. It achieves 75% accuracy using only 15 layers, while traditional LayerSkip requires 35 layers to reach the same performance.

For Llama2-30B with 60 layers, Resourceful LayerSkip still maintains its advantage. The method achieves high accuracy of 85% using only 25 layers, whereas traditional LayerSkip requires nearly all 60 layers. This efficiency is critical for applications involving large-scale LLMs, where computational resources are a limiting factor.

4.3 Overall Evaluation

Our extensive evaluation across diverse datasets, including GPT4-QA, Truthfulness, Human Ethics, Mathematics, Multiple Languages, and Translation, demonstrates that the Resourceful LayerSkip method consistently outperforms the traditional LayerSkip approach in both accuracy and efficiency. Typically, the method achieves a plateau accuracy of around 85% while requiring only 60% of the inference time compared to a full inference process, resulting in significant time, energy, and computational cost savings. This makes Resourceful LayerSkip a highly effective solution for real-world inference tasks.

Additionally, our findings suggest that larger models, which inherently possess greater redundancy, derive even more substantial benefits from Resourceful LayerSkip. It further indicates its potential for scaling to larger model architectures.

5 Conclusion

In this paper, we introduced the Resourceful LayerSkip method that offers a novel solution to optimize inference in large language models by selectively skipping feedforward layers while preserving accuracy. Our experiments show that this method achieves target performance using about 50% of model layers, leading to a 40% reduction in inference time while maintaining an accuracy plateau of 85% across various tasks.

Evaluations across diverse datasets, such as GPT-4 QA, Translation, and Mathematics, confirm its superiority over conventional LayerSkip methods, demonstrating both accuracy and efficiency gains. Larger models, like Meta's 405B parameter model (Dubey et al., 2024), are expected to benefit even further due to increased redundancy.

Future work will focus on extending the method to other model architectures, exploring more complex datasets, and conducting detailed energy consumption analysis. Resourceful LayerSkip addresses critical challenges in AI, offering a promising path for more efficient and scalable AI deployments.

Acknowledgments

This paper is based on our previous work under the supervision of Anonym. The idea of connecting later feedforward layers, which formed the foundation for the experimental design and direction of this paper, was provided by Anonym. The analysis of computational cost was advised by Anonym. The LLM experiments were supported and guided by Anonym.

Limitations

The Resourceful LayerSkip method demonstrates notable improvements in efficiency and accuracy, but several limitations need to be addressed.

Our experiments focused on models with up to 30B parameters, limiting the exploration of larger models. Additionally, the method was tested primarily on the LLaMA architecture, and its applicability to other model architectures requires further study.

The datasets used were relatively structured, and future work should evaluate the method on more diverse and complex datasets to ensure broader applicability.

While preliminary reductions in computational time were observed, a detailed analysis of energy consumption and environmental impact is needed to quantify the method's benefits in real-world deployments.

Finally, optimizing model design based on insights from the LayerSkip approach and evaluating its practical use in industry applications are key areas for future research.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Vukasin Bozic, Danilo Dordevic, Daniele Coppola, Joseph Thommes, and Sidak Pal Singh. 2024. Rethinking attention: Exploring shallow feed-forward neural networks as an alternative to attention layers in transformers. *Preprint*, arXiv:2311.10642.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed A Aly, Beidi Chen, and Carole-Jean Wu. 2024. Layerskip: Enabling early exit inference and self-speculative decoding. *Preprint*, arXiv:2404.16710.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, and Zhihao Jia. 2023. Towards efficient generative large language model serving: A survey from algorithms to systems. *arXiv preprint arXiv:2312.15234*.

Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy. 2016. Conditional deep learning for energy-efficient and enhanced pattern recognition. In *2016 design, automation & test in europe conference & exhibition (DATE)*, pages 475–480. IEEE.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Michal Čihař. 2024. Weblate.

Appendices

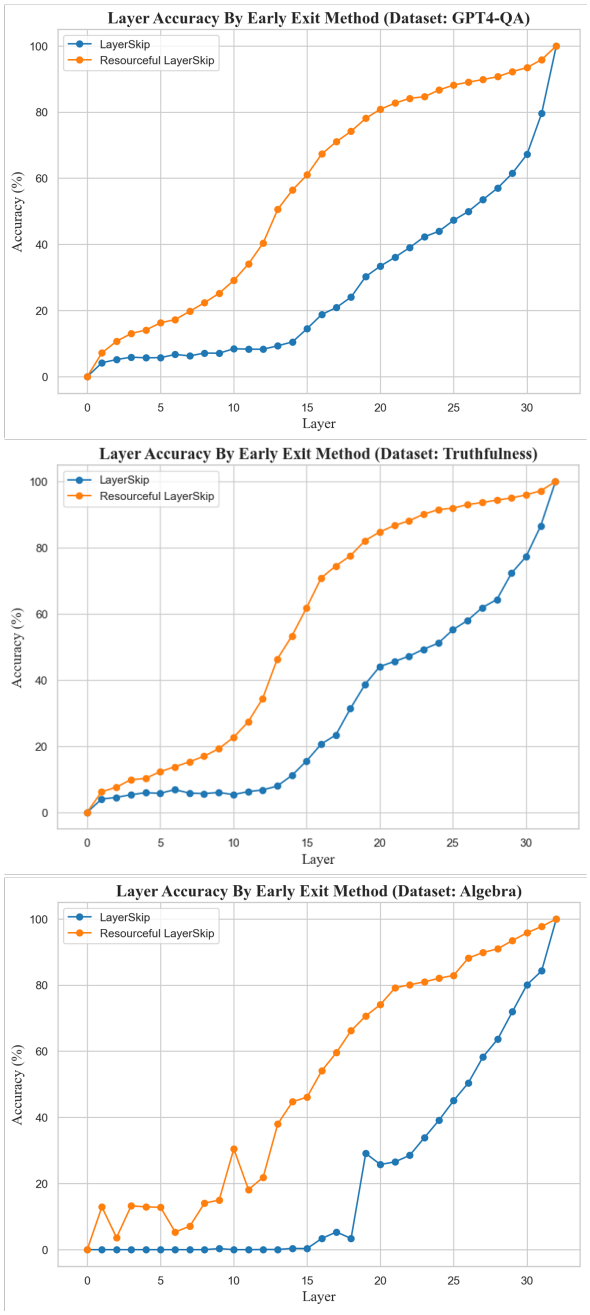

Figure 3: Detailed Performance Analysis By Layer.