On the Unacknowledged Similarity Between "AI Flow: Perspectives, Scenarios, and Approaches" and EE-LLM: A Case of Uncredited Early-Exit Design in Large Language Models

CN-honestAI

Abstract

In this paper, we analyze the remarkable similarity between the Early Exiting with Scalable Branches (EESB) algorithm proposed in AI Flow: Perspectives, Scenarios, and Approaches and the Early-Exit LLM (EE-LLM) architecture published at ICML 2024. Both methods introduce early exit mechanisms in Transformer-based language models, enabling dynamic inference by inserting auxiliary modules and exits at intermediate layers. We argue that EESB replicates the core technical ideas of EE-LLM without proper citation or acknowledgment, despite the clear precedence and openness of EE-LLM's design. Through a detailed comparison of their mechanisms, structural designs, and training strategies, we highlight the lack of novelty in EESB and call for greater transparency and accountability in AI research publishing.

CONTENTS

I Introduction 3

II Technical Analysis: EESB vs EE-LLM 3

II-A Architectural Parallels 3

II-B Training and Inference Strategy 4

II-C Absence of Prior Art Acknowledgment 4

III Discussion and Implications 5

IV Conclusion 5

References 5

I. Introduction

The rapid development of large language models (LLMs) has necessitated innovations in model efficiency and scalability, especially for edge and resource-constrained deployments. One promising direction is early exit, which enables a model to produce outputs from intermediate layers, thereby reducing computation and latency when full-depth inference is unnecessary.

A notable implementation of this strategy is EE-LLM [1], which introduces a modular, scalable, and configurable early-exit architecture built upon GPT-style Transformers. EE-LLM supports user-defined exit layers, flexible branch architectures, and parameter sharing options, all within a unified framework. Its release, accompanied by thorough experiments and open-source code, has paved the way for further research in this area.

In contrast, the paper "AI Flow: Perspectives, Scenarios, and Approaches" recently released introduces a seemingly "new" algorithm termed EESB [2] (Early Exiting with Scalable Branches). Despite being described as an original contribution (e.g., via promotional channels such as the 机器之心 public account), EESB exhibits substantial overlap with the already-published EE-LLM framework—both conceptually and technically. Yet, EESB fails to cite or acknowledge EE-LLM or the broader body of early-exit research, thereby raising concerns of academic misappropriation.

This paper aims to:

1) Examine the technical parallels between EESB and EE-LLM;
2) Analyze the originality claims of the AI Flow paper;
3) Argue for the importance of appropriate attribution in LLM research.

II. Technical Analysis: EESB vs EE-LLM

A. Architectural Parallels

EESB proposes inserting branch networks between early exit points and the shared LM head. These branches are implemented as decomposed Transformer blocks with compressed linear layers, supporting scalable parameter counts under hardware constraints.

Similarly, EE-LLM introduces early exits at arbitrary Transformer layers, optionally equipped with full Transformer blocks or MLP modules. Each exit can share or untie embedding matrices and includes mechanisms to manage parameter count and inference-time memory usage.

Both approaches:

1) Allow flexible placement of exit points;

2) Add learnable computational modules at exits;
3) Support scalable model configurations under parameter budgets;
4) Enable per-exit training or shared fine-tuning;
5) Emphasize hardware efficiency during inference.

These similarities are not superficial—they reflect deep alignment in the design philosophy, structural formulation, and deployment motivation. EESB's use of decomposed layers and SVD initialization further mirrors common strategies in lightweight Transformer compression, without crediting existing early-exit or compression literature.

See Figure 1 and Figure 2 for the architecture similarity. One can notice that EESB shares similar design with EE-LLM with basically color difference.

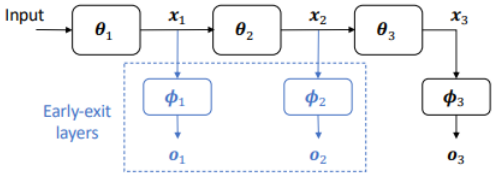

Figure 1: The model architecture of an early-exit LLM. Additional components compared to a standard LLM are highlighted in blue. Each θ_i represents a sequence of Transformer layers in the backbone of the LLM, with some additional modules in θ_1 for input processing. Each ϕ_i represents an early or final-exit layer that converts hidden states x_i into output o_i, e.g. logits for next-token prediction.

Fig. 1: EE-LLM architecture from paper [1].

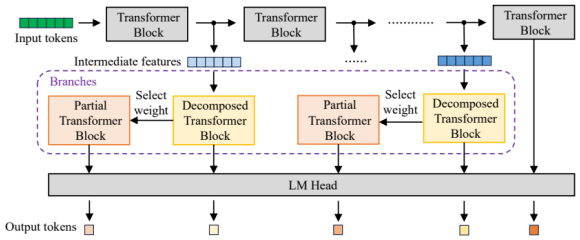

Fig. 2: EESB architecture from paper [2].

B. Training and Inference Strategy

EESB outlines three training approaches: freezing the main model, LoRA-based fine-tuning, and full end-to-end training. These match the flexibility offered in EE-LLM, which also allows separate training of early exits or global model updates. Furthermore, EESB emphasizes inference-time benefits such as reduced memory and compute footprints—claims already demonstrated and benchmarked in EE-LLM.

C. Absence of Prior Art Acknowledgment

Perhaps the most troubling aspect is the lack of proper citation. EESB's authors describe their approach as a novel implementation strategy for familial models, yet omit references to:

EE-LLM [1],

Pioneering early-exit works in NLP and vision (e.g., DeeBERT [3], FastBERT [4]),

Parameter sharing and modular training practices in LLMs.

This omission not only misleads readers about the novelty of the work but undermines the collaborative spirit of scientific progress.

III. Discussion and Implications

The case of EESB exemplifies a concerning trend in competitive AI publishing, where ideas are repackaged with minimal modification and presented as novel contributions. While scientific inspiration and iteration are essential, ethical research demands proper attribution—especially when similar methods exist in the literature.

Given that EE-LLM was:

1) Publicly released before the AI Flow paper
2) Published at a top-tier conference with peer review
3) Technically comprehensive and openly accessible

EESB's failure to cite it constitutes a breach of academic norms. It also raises questions about the review process and self-promotion practices in industrial research publications.

IV. Conclusion

Through a systematic comparison of EESB and EE-LLM, we have demonstrated that the former lacks significant novelty and likely derives from the latter without proper acknowledgment. We urge the authors of AI Flow: Perspectives, Scenarios, and Approaches to publicly clarify the inspiration and sources behind EESB and update their citations to reflect the existing body of work. Furthermore, we call upon the AI research community—journals, conferences, and media alike—to uphold stricter standards of transparency, citation, and accountability, particularly as competition intensifies in the era of large language models.

References

[1] Y. Chen, X. Pan, Y. Li, B. Ding, and J. Zhou, "Ee-llm: Large-scale training and inference of early-exit large language models with 3d parallelism," *arXiv preprint arXiv:2312.04916*, 2023.

[2] H. An, S. Huang, S. Huang, R. Li, Y. Liang, J. Shao, Z. Wang, C. Yuan, C. Zhang, H. Zhang *et al.*, "Ai flow: Perspectives, scenarios, and approaches," *arXiv preprint arXiv:2506.12479*, 2025.

[3] J. Xin, R. Tang, J. Lee, Y. Yu, and J. Lin, "Deebert: Dynamic early exiting for accelerating bert inference," *arXiv preprint arXiv:2004.12993*, 2020.

[4] W. Liu, P. Zhou, Z. Zhao, Z. Wang, H. Deng, and Q. Ju, "Fastbert: a self-distilling bert with adaptive inference time," *arXiv preprint arXiv:2004.02178*, 2020.