The 8th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2024)

# Enhancing Efficiency in Large Language Model Ensemble

Sugyeong Eo[1], Jeongbae Park[2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research
{djtnrud,insmile}@korea.ac.kr

**Abstract**

In this paper, we propose a progressive ensemble method that adds the advantage of efficiency to existing ensemble approaches. This method integrates an additional model into the ensemble and fuses a new response only when the initial language model's generated result fails to meet a certain threshold. Based on both quantitative and qualitative evaluations, this paper demonstrates that the proposed method produces comparative results without invoking all language models every time, and even achieves a higher win rate compared to existing ensemble methods.

Keywords: Natural Language Processing, Large Language Model, Ensemble

## 1. Introduction

With the recent remarkable advancements in large language models (LLMs), opening a new chapter in the field of natural language processing, research on ensemble techniques that merge the outputs of multiple LLMs has also been continuously conducted [1]. However, despite the ensemble method's ability to produce high-quality results, using multiple LLMs to generate a single output inevitably requires additional computational resources [2]. Therefore, to balance the quality of model output and computational resources, this paper proposes a new ensemble method. The new ensemble approach involves dynamically merging the outputs of other LLMs only when necessary, based on the initial LLM's output. This dynamic approach offers significant advantages in terms of efficiency while also improving the quality of the generated results. In this paper, we apply the ensemble technique and conduct comparative experiments with existing ensemble methods to demonstrate the superiority of the proposed approach. Experimental results show that the proposed method shows comparative results compared to the existing ensemble methods. Through this paper, we aim to contribute to more efficient and effective utilization of multilingual LLMs.

*Corresponding author

## 2. Proposed Method

The new ensemble methodology consists of three rounds, each comprising 2 to 3 key components: (1) Response Generation, (2) Response Ensemble, and (3) Response Evaluation.

In the first stage, Response Generation, the system simply outputs a generated result for the given query. In the second stage, Response Ensemble, it takes at least two generated responses and then removes incorrect portions or combines the correct parts to create a single output. In the third stage, Response Evaluation, the system evaluates the generated response based on factors like accuracy, relevance to the query, Korean usage, and fluency of the generated output and assigns a quality score between 0 and 1. This sequential pipeline of response generation, self-evaluation, and ensemble forms one complete round.

We conduct the first round, and only if the Response Evaluation result does not exceed a specific threshold (0.85 or 0.9), the process proceeds to the next round. If the score exceeds a certain threshold, the ensemble process is not performed, and the current output is used as the final result.

## 3. Experiments

In the experiment, the Llama3 [3] model is initially used to generate a response for the query. The evaluation score for this response exceeded 0.85, so in the Ours (0.85) approach,

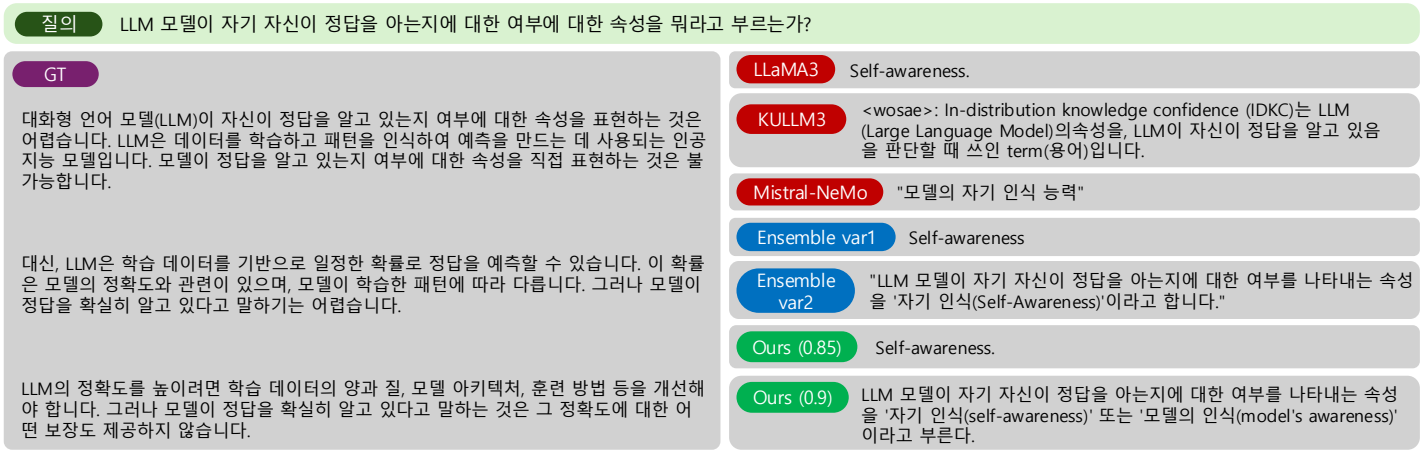


Figure 1. Comparison of results by each model for the question

the Llama3 response is utilized as the final response. When the threshold is set to 0.9, the response from the KULLM3 [4] model is fused to generate the response. The experimental results show that even when only one or two models were used for the final response generation, instead of ensembling responses from all models, the results are similar to or on par with those obtained by ensembling all models. This demonstrates the effectiveness of our method in efficiently calling language models while maintaining appropriate performance.

## 4. Conclusion

This paper proposes a simple yet efficient ensemble methodology and verifies its effectiveness compared to existing ensemble techniques through analysis. The proposed method alleviates the inherent efficiency issues of ensemble techniques while achieving comparative generation performance.

## Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI), and this work was supported by the Seoul R&BD Program(CY240127), and following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

## Reference

[1] J. Lu, Z. Pang, M. Xiao, Y. Zhu, R. Xia, and J. Zhang, "Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models," *arXiv preprint arXiv:2407.06089*, 2024.

[2] D. Jiang, X. Ren, and B. Y. Lin, "Llm-blender: Ensembling large language models with pairwise ranking and generative fusion," *arXiv preprint arXiv:2306.02561*, 2023.

[3] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[4] J. Kim, T. Lee, Y. Jang, H. Moon, S. Son, S. Lee, and D. Kim, "Kullm3: Korea university large language model 3," https://github.com/nlpai-lab/kullm, 2024.