# Domain-Aware Router: Bridging Domain-Specific Models and Model Ensembles for Efficient and Scalable Performance

Gyuho Shim°, Imatitikua Aiyanyo†*
Department of Computer Science and Engineering, Korea University°, Human-Inspired AI Research†
{gjshim, titi}@korea.ac.kr

**Abstract**

Recent advancements in large language models (LLMs) have shown that domain-specific models and model ensemble techniques are both popular approaches for optimizing performance across diverse tasks. Domain-specific models excel in their respective areas but may fall short when tasked with generalizing across multiple domains. On the other hand, model ensembles attempt to combine various models' strengths, but they can be computationally expensive. This paper introduces a novel **Domain-Aware Router** framework designed to efficiently bridge the gap between these two approaches by directing each query to the most suitable model with the corresponding domain-expertise. Our approach maximizes the specialized capabilities of domain-specific models while enhancing the overall efficiency and scalability of ensemble methods. Through experiments on both in-domain and out-of-domain queries, we demonstrate the effectiveness of this approach in achieving decent performance with efficient use.

Keywords: Domain-Aware Router, Model Ensemble, Domain-Speicific LLM

## 1. Introduction

The growing prominence of large language models (LLMs) in natural language processing has spurred significant attention to two popular methodologies: domain-specific models [1] and model ensemble [2] techniques. Domain-specific models are fine-tuned on specialized data to excel in narrow tasks but often struggle when handling queries beyond their expertise [3]. In contrast, model ensemble techniques combine multiple models to handle a broader set of tasks, blending the strengths of each model for improved performance. However, this approach is not without its challenges; model ensembles can introduce substantial computational overhead and inefficiencies due to unnecessary evaluations by models that may not be well-suited to a particular query.

This paper proposes a novel solution to these challenges by introducing the **Domain-Aware Router**, which acts as an intelligent intermediary between domain-specific models and ensemble techniques. The Domain-Aware Router efficiently classifies each query according to its domain and directs it to the appropriate model within a pre-trained ensemble. By doing so, it optimizes system efficiency while maintaining high performance. This method reduces the need for costly ensemble evaluations by ensuring that each query is handled by the most relevant domain-tuned model.

### 1.1 Why Combine Domain-Specific Models and Ensembles?

The combination of domain-specific models and model ensembles addresses several key issues:

- **Efficiency**: Domain-specific models are optimized for their respective tasks but struggle with queries outside their domain, while ensembles introduce redundancy when models process queries they are not suited for. The Domain-Aware Router allows for efficient use of computational resources by routing queries only to relevant models.
- **Performance**: Domain-specific models offer superior accuracy for specialized tasks, while ensembles provide flexibility across multiple domains. Combining them enables high performance without sacrificing the specialized expertise of domain-specific models.
- **Scalability**: A router-based approach scales well as more domain-specific models are added, minimizing the burden of unnecessary inference by large model ensembles.

*Corresponding author

Table 1. KMMLU Category Classification Performance

| Model | Epoch | Training Loss | Validation Loss | Accuracy | F1 |
|---|---|---|---|---|---|
| **KoElectra** | 5 | 0.8313 | 1.112213 | 0.662496 | 0.658747 |

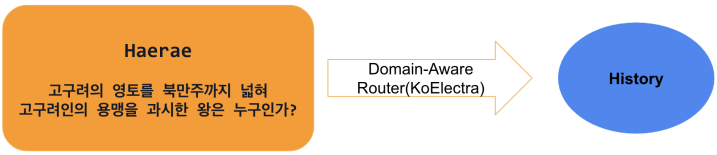


Figure 1. Evaluation on Haerae query

Thus, the Domain-Aware Router framework provides an efficient mechanism for blending the strengths of both approaches, ensuring that each model is leveraged to its fullest potential while optimizing overall system performance.

## 2. Contributions

This paper makes the following contributions:

1. **Introducing a Simple and Efficient Routing Model**: We propose a lightweight domain-aware router that classifies queries according to domain and routes them to the appropriate model. This system is both straightforward and computationally efficient.
2. **Exploring an Untapped Domain-Based Routing Approach**: While routing techniques have been applied to model ensembles, domain-based routing itself has not been adequately explored. This paper provides new insights into how domain classification can optimize model selection.

## 3. Related Work

Research on model ensembles, such as RouteLLM [4], has primarily focused on mixing outputs from multiple models or introducing post-inference routing techniques. These approaches attempt to improve accuracy by blending results from various models but do not address the inefficiencies caused by sending all queries through all models. Similarly, recent advancements in domain-specific LLMs have focused on tuning models for specific tasks, but these models often fail when queries fall outside their trained domain. The current literature has not sufficiently explored the potential of a lightweight routing system that can selectively choose domain-specific models, as our Domain-Aware Router does.

## 4. Domain-Aware Router

Our Domain-Aware Router is designed to efficiently classify input queries according to their domain. The router utilizes KoElectra [5] as its backbone and is fine-tuned on the 45 domain categories from KMMLU [6], a benchmark dataset for Korean language tasks.

### 4.1 Methodology

The router model is lightweight, making it scalable and efficient. It processes a given query to classify it into one of the 45 predefined domains. Based on this classification, the router then directs the query to the expert model fine-tuned for that specific domain. This targeted approach minimizes overhead and maximizes the effectiveness of both the domain-specific models and the overall ensemble.

## 5. Experiments

We evaluated the Domain-Aware Router using both in-domain and out-of-domain queries to test its classification accuracy and generalizability.

1. KMMLU Evaluation: The model was first tested on KMMLU, a dataset of questions spanning 45 domains. In table 1, the Domain-Aware Router showed high accuracy in classifying the categories of the given queries. significantly improving the efficiency of handling domain-specific tasks.
2. Out-of-Domain Test with Haerae [7]: To further test the model's performance, we introduced queries from Haerae, which lies outside the KMMLU training categories. As seen in figure 1, the model demonstrated strong handling capabilities with the out-of-domain queries, maintaining acceptable accuracy.

The experiments confirmed that the Domain-Aware Router efficiently directs queries to the appropriate categories. The router showed a strong ability to generalize to unseen domains while minimizing computational overhead.

## 6. Future Work

Future research will focus on enhancing the system by developing an end-to-end mapping solution where queries are automatically routed not only by domain but also by model-specific characteristics. This would further optimize the process by ensuring that the best-suited expert model within the

ensemble is selected based on both domain and task-specific performance metrics.

## 7. Conclusion

This paper presents a novel Domain-Aware Router framework that effectively bridges the gap between domain-specific models and model ensemble techniques. By efficiently classifying queries according to domain and routing them to the most appropriate model, the router ensures both high performance and computational efficiency. This approach optimizes the use of domain-specific expertise while avoiding the inefficiencies typically associated with ensemble methods. As a result, the proposed framework represents a significant advancement in the field of LLM-based ensemble methodologies, offering a scalable solution for diverse domain-specific tasks.


## 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). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).


## Reference


[1] C. Jeong, "Domain-specialized llm: Financial fine-tuning and utilization method using mistral 7b," *Journal of Intelligence and Information Systems*, Vol. 30, No. 1, p. 93–120, Mar. 2024. [Online]. Available: http://dx.doi.org/10.13088/jiis.2024.30.1.093

[2] 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," 2024. [Online]. Available: https://arxiv.org/abs/2407.06089

[3] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," 2024. [Online]. Available: https://arxiv.org/abs/2308.08747

[4] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "Routellm: Learning to route llms with preference data," 2024. [Online]. Available: https://arxiv.org/abs/2406.18665

[5] J. Park, "Koelectra: Pretrained electra model for korean," https://github.com/monologg/KoELECTRA, 2020.

[6] G. Son, H. Lee, S. Kim, S. Kim, N. Muennighoff, T. Choi, C. Park, K. M. Yoo, and S. Biderman, "Kmmlu: Measuring massive multitask language understanding in korean," 2024. [Online]. Available: https://arxiv.org/abs/2402.11548

[7] G. Son, H. Lee, S. Kim, H. Kim, J. Lee, J. W. Yeom, J. Jung, J. W. Kim, and S. Kim, "Hae-rae bench: Evaluation of korean knowledge in language models," 2024. [Online]. Available: https://arxiv.org/abs/2309.02706