Sensors and Materials, Vol. 38, No. 6 (2026) 3413–3420
MYU Tokyo

S & M 4510

Learning Marine Spatial Information Data Using a Korean Speech-based Large Language Model

Je Hyung Tak, Yun Soo Choi,* Min Sung Kim, and Chan Woo Lee

Department of Geoinformatics, University of Seoul, Seoul 02504, Korea

(Received November 24, 2025; accepted June 4, 2026)

Keywords: marine spatial information, spatial information, LLM , RAG, fine tuning

In previous studies, large language models (LLMs) were fine-tuned using Korean utterance data to improve performance in small-scale computing environments by applying the low-rank adaptation (LoRA) method. In addition, Gradient Checkpointing and Gradient Accumulation techniques were employed to address computational resource limitations, enabling efficient fine-tuning under constrained computing conditions. The purpose of this study is to develop a marine geospatial information LLM. To achieve this, the LLM previously fine-tuned on Korean utterance data was enhanced by integrating a retrieval-augmented generation (RAG) framework, a document-based inference approach, with marine geospatial information data. First, domain-specific terminology learning was conducted using the International Hydrographic Organization Dictionary (S-32), which provides standardized definitions. Additionally, data on current speed, current direction, wind speed, and wind direction were collected from the Badanuri marine information service provided by the Korea Hydrographic and Oceanographic Agency and incorporated into the RAG knowledge base. Subsequently, S-101 and S-102 datasets were preprocessed to extract bathymetric depth information and were also integrated into the RAG framework. In conclusion, this study demonstrates the feasibility of developing a marine geospatial information-specialized LLM in a resource-constrained environment and enhances the practical applicability of the proposed marine geospatial information LLM through RAG-based knowledge integration.

1. Introduction

1.1 Research background

In the field of marine spatial information, heterogeneous datasets—such as electronic navigational charts, bathymetric grids, and tidal and current data—are produced using standardized frameworks. However, conventional rule-based systems have limitations in supporting natural language question answering, document understanding, and knowledge organization required for operational decision-making. Recently, large language models (LLMs) have demonstrated strong potential for such knowledge-intensive tasks. However, ensuring the

*Corresponding author: e-mail: choiys@uos.ac.kr
https://doi.org/10.18494/SAM6189

ISSN 0914-4935
https://myukk.org/

accuracy and timeliness of domain-specific knowledge, as well as managing training and inference costs, remains a major challenge. Meanwhile, considering the environmental and economic costs of LLM development, efficient learning based on small-scale computing resources is an important research motivation.(1) Discussions on "Green AI" have urged that, in addition to performance, indicators of computational, energy, and carbon costs should be jointly considered and that efficient methods should be adopted.(1) In the preliminary stage of this study, gradient checkpointing was adopted; this classical technique reduces memory usage by not storing intermediate activations and recalculating them during backpropagation, and gradient accumulation expands the effective batch size by accumulating gradients from small minibatches, thereby enabling stable training under memory constraints.(1) These methods also contribute to reducing the carbon footprint in graphics processing unit (GPU)-limited environments. In the marine domain, the terminology and data standards of the International Hydrographic Organization (IHO) effectively serve as a common language. In this study, IHO S-32 (Hydrographic Dictionary), which provides official terms and definitions for hydrography and marine surveying, was used as a basic lexical resource to improve the consistency of the model with domain terminology. In Korea, the Badanuri Marine Information Service operated by the Korea Hydrographic and Oceanographic Agency (KHOA) provides time-series data—such as current speed, current direction, wind speed, and wind direction—via an application programming interface (API). These data are collected from observation networks, including tide gauges, ocean buoys, and scientific stations, as well as numerical prediction systems. As these marine dynamic variables are suitable for integration into a retrieval-augmented generation (RAG) knowledge base, they were selected as domain data. Finally, S-101 and S-102 datasets were incorporated into RAG-based training for accurate bathymetric verification. In this study, we used S-32, S-101, S-102, current speed, current direction, wind speed, and wind direction data, as summarized in Table 1.

1.2 Purpose

The objective of this study is to develop a domain-specific LLM that, on the basis of a Korean LLM trained in a low-resource environment, combines RAG with domain knowledge of marine spatial information (IHO standards and domestic observation/prediction data) so as to accurately and reliably perform marine terminology understanding, interpretation of numerical information such as water depth, current, and wind field, and document-based question answering. In summary, the goal is to advance low-resource and environmentally friendly learning techniques

Table 1
Data used.

Division	Data
IHO standard data	IHO S-32 Hydrographic Dictionary
IHO standard data	IHO S-101 Electronic Navigational Chart data
IHO standard data	IHO S-102 Bathymetric Surface data
KHOA Badanuri data	current speed
KHOA Badanuri data	current direction
KHOA Badanuri data	wind speed
KHOA Badanuri data	wind direction

for the ecosystem of structured and semi-structured marine data provided by the IHO standards (S-32, S-101, S-102) and the national marine observation infrastructure (Badanuri), while simultaneously leveraging the effectiveness of RAG for knowledge grounding and ensuring up-to-dateness. This study establishes a technical foundation for constructing an LLM specialized in marine spatial information. On the basis of the previously fine-tuned Korean utterance-based model, this study aims to enhance the practical applicability of the proposed marine geospatial information LLM in real-world operations by integrating domain knowledge such as IHO terminology, bathymetric grids, currents, and wind field data through RAG.

2. Related Work

The objective of this study is to develop an optimized LLM that can provide knowledge in the field of marine spatial information. First, a base model was constructed by fine-tuning Llama-3.1-8B-Instruct model using Korean utterance data,[2] and then data produced and accumulated in the marine spatial information domain was utilized to strengthen domain-specific reasoning ability and response reliability. As a starting point for this effort, we systematically review, as in previous studies, the development trends and application cases of domain-specific LLM reported in various fields, and summarize the necessity of such approaches (handling of specialized terminology, task-specific performance, and maintenance of knowledge recency) as well as their differences from general-purpose LLM. Through this discussion, we clarify the academic and practical significance of LLM specialization in the marine spatial information domain and present the design rationale of this study.

In this study, GeoChat, a grounded large vision-language model (VLM) for remote sensing, is proposed as a conversational multitask VLM for remote sensing imagery that covers images, regions, and coordinates.[3] By processing high-resolution patches and exchanging task tokens and coordinates in textual form, the model can indicate regions and provide supporting evidence.[3] The model integrates image, region, visual question answering, and grounding within a single architecture, providing coordinates in its responses as visual evidence. In addition, rapid fine-tuning is achieved through low-rank adaptation (LoRA), and raising the input resolution improves the recognition of small objects.[3,4] The significance of this work lies in presenting conversational, grounding, and multitask capabilities within a single model. In addition, recent geospatial LLM studies have emphasized the importance of domain-specific data organization and task-oriented model adaptation for Earth observation and geospatial applications.[5] Next, the "Less Is More for Alignment" (LIMA) study examined whether a model can achieve effective performance using only 1,000 high-quality examples instead of large-scale datasets.[6] The results showed that LIMA achieved performance comparable to existing models with just 1,000 examples, suggesting that the quality and diversity of data may be more important than sheer quantity. The paper on "FinGPT: Open-Source Financial Large Language Model"[7] presents a data-centric approach for financial domain-specific LLM, an automatic data curation pipeline, and lightweight LoRA-based adaptation. It argues that, for LLM in specific domains, performance is determined more by the automation of data collection, cleaning, and preprocessing (a data-centric approach) than by changes in model architecture.[8] To this end, it emphasizes that the roles of open data accessibility and quality management pipelines should be established in advance.

These previous studies confirmed the importance of data contributions to the performance and reliability of domain-specific LLM.[9] Related studies have also reported LLM- or AI-based applications in intelligent information systems, smart media, safety, land and geospatial analysis, engineering applications, and Korean-language document processing fields.[10–17] In particular, by comparing and analyzing development trends of LLM in the geospatial and legal domains, it has been found that systematic injection of domain knowledge and high-quality data management are essential. On the basis of this analysis, in this study, we aim to design and implement a conversational language model optimized for question-answering and reasoning tasks by performing supervised fine-tuning (SFT) using diverse high-quality data collected and refined in the marine spatial information domain, and simultaneously combining the RAG approach for knowledge grounding and enhancement of recency.

3. Research Method

3.1 Previous research

In this study, on the basis of a previously developed Korean utterance-based LLM, we aimed to build a more complete model by training it further on marine spatial information data. The Korean utterance-based LLM was designed to minimize computational resource consumption by applying gradient checkpointing and gradient accumulation, and was developed by fine-tuning the Llama-3.1-8B-Instruct model using the LoRA approach implemented with Unsloth, an open-source framework for efficient LLM fine-tuning. The training data for the Korean utterance model were constructed to enable the model to understand various questions in Korean and to generate appropriate answers,[2] and included general commonsense sentence generation data, book information summarization data, paper information summarization data, and document summarization data.

During model training, we observed that, under a constraint on the GPU computing power, the clock frequency varied depending on load and operating conditions, and that owing to the power limitation, the GPU did not operate at its maximum performance but was regulated to a certain level (Fig. 1).

In addition, Fig. 2 shows the GPU memory allocation size as a function of time, where the *x*-axis represents the elapsed GPU runtime and the *y*-axis represents the amount of memory used by the GPU in bytes. Owing to the computing power limitation applied in this study, it was

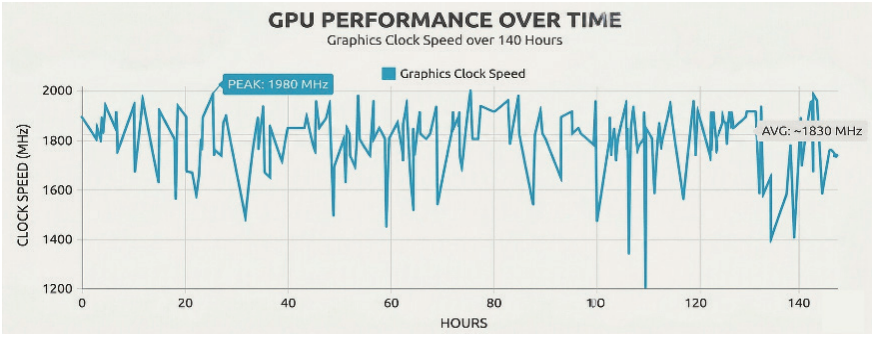

Fig. 1. (Color online) GPU graphics clock speed.

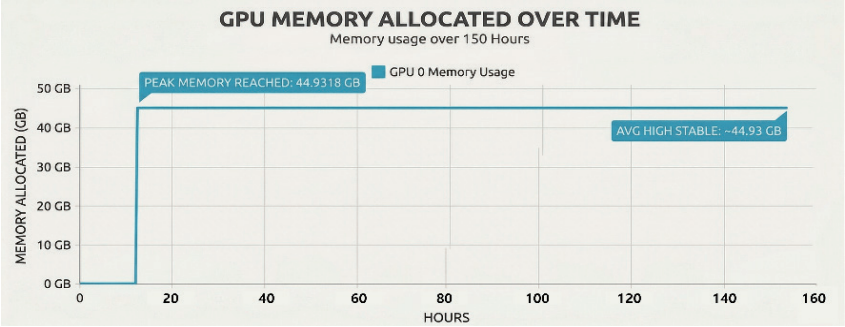

Fig. 2. (Color online) GPU memory allocation over time.

observed that once GPU memory was allocated for a specific task, it remained occupied continuously over time.

The performance of the developed model was evaluated by submitting it to the Open Ko-LLM LeaderBoard, a platform jointly operated by the National Information Society Agency. The benchmark results showed an average score of 43.33; in particular, the model achieved 61.17 on Ko-Winogrande, which evaluates logical reasoning ability, and 58.3 on Ko-GSM8k, which measures mathematical problem-solving ability, demonstrating competitive performance compared with other open-source models.

3.2 Marine spatial information data learning

In a Google Colab environment integrated with Google Drive and Hugging Face, we constructed an end-to-end pipeline that incorporates 4-bit quantization and LoRA-based fine-tuning,[4] while integrating a RAG framework to specialize a Llama 3.1 series language model for the ocean and meteorological domain. As the base model, we used NAPS-ai/naps-llama-3_1_instruct-v0.6.0 on Hugging Face, applied 4-bit NF4 quantization using bitsandbytes, and attached LoRA adapters with rank 16 and dropout 0.05 to the main projection modules[4] (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj), thereby constructing an efficient fine-tuning environment with minimized memory usage.

The domain data were organized along three axes (bathymetry data and meteorological and current time-series data) and with a marine terminology dictionary. First, from S-101 (point) and S-102 (gridded) bathymetry Excel files, we extracted latitude, longitude, and depth information, and then generated instruction–response pairs by constructing Korean queries of the form "What is the water depth at a specific latitude–longitude location (or within a given sea area)?" and concise numerical answers in the format "value + unit (m)". For meteorological and current data, comma-separated value (CSV) files containing wind speed, wind direction, current speed, current direction, and maximum instantaneous wind speed were organized by location, variable type, and date–time. For each day (00:00–23:00), we grouped hourly observations together with the daily maximum, mean, and minimum values into a single response, and generated training samples by pairing them with queries of the form "On YYYY-MM-DD, please provide the data for a given variable at a specific location." In addition, terms and definitions were extracted

from the IHO Dictionary S-32 CSV file to construct English QA data of the form "Instruction: What is the definition of 'Term'?" and "Response: Definition". All data were converted into user/assistant conversational text using the Llama 3 chat template. For the construction of the RAG knowledge base, we used the SentenceTransformer model all-MiniLM-L6-v2 for embeddings and a Facebook AI similarity search (FAISS)-based L2 distance index. From the bathymetry QA data and S-32 terminology definition data, only the "question–answer" text was cleaned, combined, and used to form the document set of the knowledge base. Each document was embedded into a fixed-length vector and stored in the FAISS index. When a query is given, the same embedding model is used to vectorize the query, and a retriever function is defined that searches the index for the top-k most similar documents and returns the original "Question/Instruction–Answer/Response" text as context.

For the meteorological and current time-series training samples, this RAG context was combined to generate RAG-inspired augmented training samples with a "Context + Instruction + Response" structure. Specifically, for each meteorological/current sample, we parsed the Llama 3 template to separate the original query (Instruction) and answer (Response), fed the query into the retriever, and searched the bathymetry and terminology definition knowledge base for the most relevant document. The retrieved document was formatted as a bullet-list Context string. Finally, the user message was reconstructed as "Context:\n[retrieved sentences]\nInstruction:\n[original query]", and the assistant message retained the original answer; this pair was then wrapped back into the template to produce the retrieval-augmented data construction sample. Finally, the RAG-augmented meteorological/current data, the bathymetry QA data, and the S-32 terminology definition data were merged into a single integrated training corpus. For the S-32 data, the question and answer segments were separated from the existing "Instruction: … Response: …" strings using regular expressions and converted into the same Llama 3 chat template format as the other data. This integrated text corpus was then converted into the Hugging Face Datasets format, and batch tokenization was performed with a maximum sequence length of 512 tokens. Using the SFT Trainer in the TRL library, we performed SFT on the Llama 3.1-based model with the applied quantization and LoRA settings, enabled resumption of training via intermediate checkpoints, and finally saved the fine-tuned model and tokenizer to a separate directory. Through this process, we constructed a specialized Llama 3.1 model capable of generating more accurate and informative responses by leveraging ocean and meteorological domain knowledge together with RAG-based context.

4. Result

In this study, we proposed an end-to-end fine-tuning pipeline with an integrated RAG framework to specialize a Llama 3.1-based language model for the ocean and meteorological domain, built on an integrated environment connecting Google Colab, Google Drive, and Hugging Face. Specifically, bathymetry data (S-101/102), time-series observation data such as wind speed, wind direction, current speed, current direction, and maximum instantaneous wind speed, and IHO S-32 terminology definitions were normalized into an instruction–response format based on the Llama 3 chat template. By performing SFT on a lightweight model with 4-bit quantization and LoRA, we demonstrated that a specialized language model reflecting

domain knowledge can be implemented even under limited computational resources. In addition, we constructed a RAG knowledge base centered on bathymetry and terminology definition data using SentenceTransformer and FAISS, and created a dataset with a "Context + Instruction + Response" structure by combining this context with meteorological and current time-series training samples. In this way, the model was trained to acquire response patterns that not only perform simple question answering but also reference and utilize external knowledge. This provides a foundation for handling frequently requested numerical, terminological, definitional, and time-series pattern information in the ocean and meteorological fields in a consistent manner, while generating richer and more explanatory responses to new queries.

5. Conclusions

The proposed pipeline has practical and technical significance from three perspectives: a structured training data generation procedure that reflects the domain characteristics of ocean and meteorological data, an efficient domain fine-tuning strategy that combines 4-bit quantization with LoRA, and retrieval-augmented data construction that exploits a knowledge base of bathymetry and terminology definitions. Through this, we present the potential of an ocean- and meteorology-specialized language model capable of providing more reliable and consistent answers to expert queries related to port and route design, maritime traffic safety, weather monitoring, and marine disaster response. Future work includes expanding the data to incorporate additional ocean variables such as waves, tides, and tidal currents, temperature, and salinity; supporting international chart and specification documents through multilingual (e.g., parallel Korean–English) training; and conducting field-based quantitative evaluation by integrating the model with operational navigation and forecasting systems. Furthermore, introducing online or continual learning to reflect the latest observations and regulations, and optimizing the RAG components (embedding model, index structure, and retrieval strategy) to simultaneously improve response quality and inference speed remain important tasks.

Acknowledgments

This work was supported by the 2025 Research Fund of the University of Seoul.

References

1 T. W. Kang: J. Korea Acad.-Ind. Coop. Soc. **25** (2024) 177. https://doi.org/10.5762/KAIS.2024.25.11.177
2 J. Tak, K. Choi, H. Na, and M. Kim: Korean J. Artif. Intell. **13** (2025) 17. https://doi.org/10.24225/kjai.2025.13.2.17
3 K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan: Proc. 2024 IEEE/CVF Conf. Computer Vision and Pattern Recognition (2024) 27831–27840.
4 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen: arXiv:2106.09685 (2021). https://arxiv.org/abs/2106.09685
5 Y. Zhang, J. Li, Z. Wang, Z. He, Q. Guan, J. Lin, and W. Yu: Int. J. Appl. Earth Obs. Geoinf. **136** (2025) 104312. https://doi.org/10.1016/j.jag.2024.104312
6 C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy: arXiv:2305.11206 (2023). https://arxiv.org/abs/2305.11206
7 H. Yang, X.-Y. Liu, and C. D. Wang: arXiv:2306.06031 (2023). https://arxiv.org/abs/2306.06031

8 J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu: AI Open **5** (2024) 181. https://doi.org/10.1016/j.aiopen.2024.09.002
9 C.-S. Jeong: J. Intell. Inf. Syst. **29** (2023) 129. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11674564
10 J. I. Park, M. J. Lim, and P. K. Kim: Smart Media J. **13** (2024) 44. https://doi.org/10.30693/SMJ.2024.13.6.44
11 S.-J. Jo and S.-S. Park: J. Korean Inst. Intell. Syst. **34** (2024) 373. https://doi.org/10.5391/JKIIS.2024.34.5.373
12 S.-G. Lee, J.-H. Yoo, and I.-G. Wang: Saf. Cult. Res. **45** (2025) 19. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART003235941
13 J.-W. Lee: Land (2025) 43. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE12424114
14 S.-Y. Lee, D.-W. Lee, H.-B. Gan, and H.-H. Jeong: Mech. J. **66** (2026) 35. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE12597715
15 G.-H. Kim and D.-G. Kim: Smart Media J. **14** (2025) 50. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART003217000
16 Y.-J. Kim, H.-Y. Kim, J.-W. Kim, S. Jang, and D.-H. Seo: Proc. Korean Inf. Sci. Soc. Conf. (2025). https://discos.sogang.ac.kr/file/2025/dome_paper/KSC_2025_D_Seo.pdf
17 S.-I. Lee, Y.-S. Nam, S.-H. Oh, S.-Y. Han, and Y.-M. Jeong: Proc. Korean Inf. Sci. Soc. Conf. (2025). https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART003148081

About the Authors

Je-hyung Tak received his master's degree from the University of Seoul in March 2024 and is currently pursuing his Ph.D. degree in the Department of Geoinformatics at the same university. His research interests include hydrographic data models, artificial intelligence, and large language model (LLM) development. He is also interested in marine GIS research related to the S-100 standard. (wpgud23@uos.ac.kr)

Yun-Soo Choi received his Ph.D. degree from the Department of Civil Engineering, Sungkyunkwan University, in 1992. From 1991 to 2001, he was an associate professor at Hankyong University. Since 2001, he has been a professor at the University of Seoul, and since 2018, he has been the president of the Hydrography Society Korea. His current research interests include geographic information and S-100. (choiys@uos.ac.kr)

Min-Sung Kim received his B.S. degree from Kyungpook National University in February 2024. Since March 2024, he has been pursuing his M.S. degree at the University of Seoul, with research interests in S-100, maritime geographic information systems, and the application of international standards for hydrographic data.
(mskim9804@uos.ac.kr)

Chan-woo Lee is a graduate student in the Department of Geoinformatics at the University of Seoul. His research interests include the S-100 standard, marine geographic information systems, autonomous navigation, and artificial intelligence. (cwlee97@uos.ac.kr)