Nomura Research Institute Group

NEWS RELEASE

Apr. 15, 2025

Nomura Research Institute, Ltd.

NRI Develops Innovative Method for Building Industry- and Task-Specific LLMs

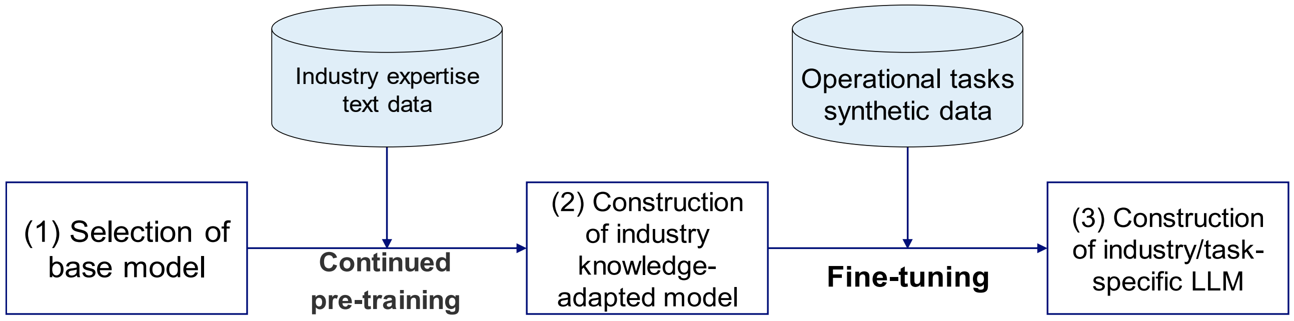
— Small-scale models deliver superior accuracy compared to large commercial alternatives, adaptable across a wide range of industries —

Tokyo, April 15, 2025 – Nomura Research Institute, Ltd. (Headquarters: Tokyo, Japan, Chairman, President and CEO: Kaga Yanagisawa, "NRI") has developed a novel method for constructing large language models (LLMs) tailored to specific industries and tasks ("Industry/Task-Specific LLMs"), leveraging its advanced technologies and deep domain expertise. This method is based on a relatively compact model with 8 billion parameters, yet it has demonstrated task-specific performance[1] that surpasses that of GPT-4o, a leading large-scale general-purpose model. Furthermore, NRI's approach is highly adaptable, allowing the method to be applied across a wide range of industries and use cases.

■ Development Background and Outcome of the Industry/Task-Specific LLM Construction Method

While general-purpose models like GPT-4o are versatile and capable of handling a wide range of tasks, they often struggle with highly specialized use cases—particularly those requiring deep domain expertise, industry-specific terminology, and compliance with complex regulations. These challenges are further exacerbated by the vast number of parameters and high computational costs associated with large-scale models. To address these limitations, NRI has developed a method for building cost-efficient, high-accuracy Industry/Task-Specific LLMs tailored to real-world business operations. This method was established through the following three-stage approach.

Figure: How Industry/Task-Specific LLMs Are Constructed


Industry expertise text data
Operational tasks synthetic data
(1) Selection of base model
Continued pre-training
(2) Construction of industry knowledge-adapted model
Fine-tuning
(3) Construction of industry/task-specific LLM

(1) Selecting a Cost-Efficient, Small-Scale Base Model with Strong Japanese Language Performance

As the foundation, NRI adopted “Llama 3.1 Swallow 8B,” a model with 8 billion parameters jointly developed by the Institute of Science Tokyo and the National Institute of Advanced Industrial Science and Technology (AIST). Known for its exceptional Japanese language processing capabilities, this small-scale model reduces both computational demands and operating costs.

Because NRI’s method utilizes open-weight LLMs, organizations are free to select the most appropriate base model for their specific objectives, rather than being tied to a fixed proprietary model. This flexibility also ensures adaptability to future model updates.

(2) Adapting the Model to Industry Knowledge through Continued Pre-Training

To embed deep industry expertise, NRI curated its own Japanese-language finance corpus, incorporating expert knowledge from banking, insurance, and other financial sectors. Continued pre-training[2] on this specialized corpus enabled the model to retain the broad linguistic and general knowledge of the base model while effectively acquiring domain-specific understanding—creating a versatile foundation for industry applications.

(3) Fine-Tuning for Task Specialization Using Synthetic Data

For use cases like insurance sales compliance checks—where gathering real conversational data is difficult due to privacy and regulatory constraints—NRI generated synthetic data[3] that simulated a wide range of realistic scenarios. Fine-tuning[4] the model with this synthetic data enabled NRI to create a task-specific LLM precisely tailored to the unique requirements of this domain.

As a result of these efforts, during the sales compliance check test for the insurance industry, we achieved an accuracy rate[5] that was 9.6 percentage points higher than that demonstrated by the commercial large-scale model GPT-4o (2024-11-20).

Table: Insurance Industry Sales Compliance Check Performance Evaluation Results

Model	Accuracy rate
GPT-4o (2024-11-20)	76.7%
Base model (Llama 3.1 Swallow 8B Instruct v0.2)	51.7%
NRI’s specialized model (Llama 3.1 Swallow 8B + fine-tuning)	83.1%
NRI’s specialized model (Llama 3.1 Swallow 8B + continued pre-training + fine-tuning)	**86.3%**

In addition, now that this construction method has been standardized, it can be readily adapted to other industries and tasks. By simply modifying the input data or tailoring the generated synthetic data, the scope of application can be significantly expanded.

■ Future Developments and Research Collaborations

Building on these results, NRI will further accelerate its efforts to optimize this technology for a broader range of industries and tasks. By leveraging the key advantages of the Industry/Task-Specific LLM—high accuracy, low cost, and rapid, flexible adaptability enabled by its compact architecture—NRI aims to expand its application to specialized domains that are often challenging for general-purpose models to address.

In FY2025, NRI plans to launch joint research with the Okazaki Laboratory at The Institute of Science Tokyo. This collaboration will involve field testing tailored to real-world industry challenges, with the goal of continuously enhancing the model's performance and promoting the societal integration of generative AI. Additionally, NRI will seek to deepen its partnerships with major tech firms and startups to advance the commercialization and practical deployment of these technologies, including their use in building task-specific AI agents across diverse sectors.

Comments by Professor Naoaki Okazaki at Institute of Science Tokyo’s School of Computing

"The Swallow LLM series was developed and released by the Institute of Science Tokyo and the National Institute of Advanced Industrial Science and Technology to create a general-purpose large language model with strong Japanese language capabilities and commercial potential. I am pleased to see that Llama 3.1 Swallow has been adopted as the base model for an industry/task-specific LLM, and that it has achieved such high accuracy through an innovative approach. Through our joint research, I look forward to further refining this method to develop models tailored to business applications—with the ultimate goal of enabling broad societal implementation."

[1] Tests were conducted on the task of insurance sales compliance checks, which entails reviewing sales conversation records to verify that insurance sales representatives are conducting their activities in compliance with relevant laws and regulations.

[2] Continued Pre-Training: A process that entails taking a pre-trained model that already possesses general knowledge and then training it further with additional data tailored to new fields or tasks, thereby improving its expertise and adaptability.

[3] Synthetic Data: Data artificially created using generative AI, such as an LLM (Large Language Model), specifically for use in training models. In this project, the synthetic dataset was generated using the Llama-3.1-Swallow-70B-Instruct-v0.3 model.

[4] Fine-Tuning: An additional training process in which fine adjustments are applied to the parameters of an already pre-trained model to optimize its performance for a specific task.

[5] The evaluation was conducted using a dataset consisting of 344 conversations—172 conversations involving compliance violations across seven categories, and 172 compliant conversations with no violations. The model's accuracy was assessed based on its correct identification of these violations (including correctly identifying cases with no violations).

Inquiries about this news release:
Kayano Umezawa, Masayoshi Yumino
Corporate Communications Department
Nomura Research Institute, Ltd.
TEL: +81-3-5877-7100
E-mail: kouhou@nri.co.jp

Inquiries about this project:
Tomoyasu Okada, Yuma Okochi
AI Solution Promotion Department
Nomura Research Institute, Ltd.
E-mail: ai-tech-lab@nri.co.jp