KAERI-GPT: Extending a Nuclear Domain LLM to Secure Institutional Environments

Seungdon Yeom[a] , Yonggyun Yu[a*]
[a]*Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil , Yuseong-gu , Daejeon, 34057, Korea*
[*]*Corresponding author: ygyu@kaeri.re.kr*

Keywords : Large Language Model (LLM), Domain adaptation, nuclear engineering, on-premise LLM

1. Introduction

Recent advances in large language models (LLMs) have demonstrated strong performance in domain-specific question answering and technical document understanding [1-3]. In the nuclear engineering field, LLMs can support knowledge retrieval, technical analysis, and research assistance. However, direct use of commercial LLM services such as ChatGPT, Gemini, and Claude is prohibited in nuclear research institutes due to strict security policies that forbid transmission of internal information to external cloud-based services.

To address domain adaptation, continual pretraining (CPT) and instruction tuning (IT) have been widely used to specialize LLMs for specific technical fields [4-6]. In our previous work, AtomicGPT was developed as a nuclear domain LLM using publicly available nuclear corpora [7]. While AtomicGPT demonstrated effective domain-level adaptation, it did not incorporate institution-specific internal knowledge required for practical deployment in secured research environments.

In this study, we present KAERI-GPT, an on-premise institutional extension of a nuclear domain LLM designed for secure deployment within a research institute. Building upon the AtomicGPT framework, KAERI-GPT incorporates approximately 4,000 internal technical reports through CPT and instruction tuning, entirely within a secured local infrastructure without reliance on external cloud services. Evaluation on a public nuclear QA benchmark demonstrates that KAERI-GPT outperforms both its base model (Exaone4-32B) and GPT-4 across all question types, suggesting that institutional data integration effectively enhances nuclear domain capability. We also outline a planned ablation study to isolate the contribution of internal data.

2. Methodology

In this section, the methodology for developing and evaluating KAERI-GPT is described, including the model configuration, training data, and evaluation setup. Fig. 1 illustrates the hierarchical knowledge structure and the overall two-stage training framework that extends base LLMs to KAERI-GPT.

2.1 Model Configuration

To establish a nuclear domain baseline, we previously developed AtomicGPT by adapting open-weight base models (Qwen2.5-7B and Gemma2-9B) using publicly available nuclear engineering corpora. The training pipeline consisted of continual pretraining (CPT) on domain-specific text followed by instruction tuning (IT) using nuclear question–answering data.

Building upon this domain adaptation framework, we developed KAERI-GPT as an institutional extension deployed in an on-premise environment. KAERI-GPT is based on Exaone4-32B and trained using (1) publicly available nuclear corpora and (2) internal technical documents from the research institute. The same two-stage training pipeline—continual pretraining followed by instruction tuning—was applied to ensure methodological consistency.

All training and inference processes for KAERI-GPT were conducted within a secured local infrastructure without reliance on external cloud-based APIs. The resulting model has been deployed as an internal chatbot service within KAERI's secured intranet, providing researchers with nuclear domain question-answering and technical document assistance capabilities.

2.2 Training Data

The training data for both AtomicGPT and KAERI-GPT consist of publicly available nuclear engineering corpora and, additionally for KAERI-GPT, internal institutional documents. Table I summarizes the complete training corpus composition.

For AtomicGPT, publicly accessible nuclear-related resources were collected from several sources: terminology dictionaries, regulatory glossaries, academic papers, technical presentations, and a nuclear knowledge encyclopedia [8-12].

For KAERI-GPT, a large-scale institutional corpus of approximately 4,000 internal technical reports (TR) from KAERI was additionally incorporated during CPT. These reports were converted from HWP and DOCX formats into normalized plain text through automated document parsing.

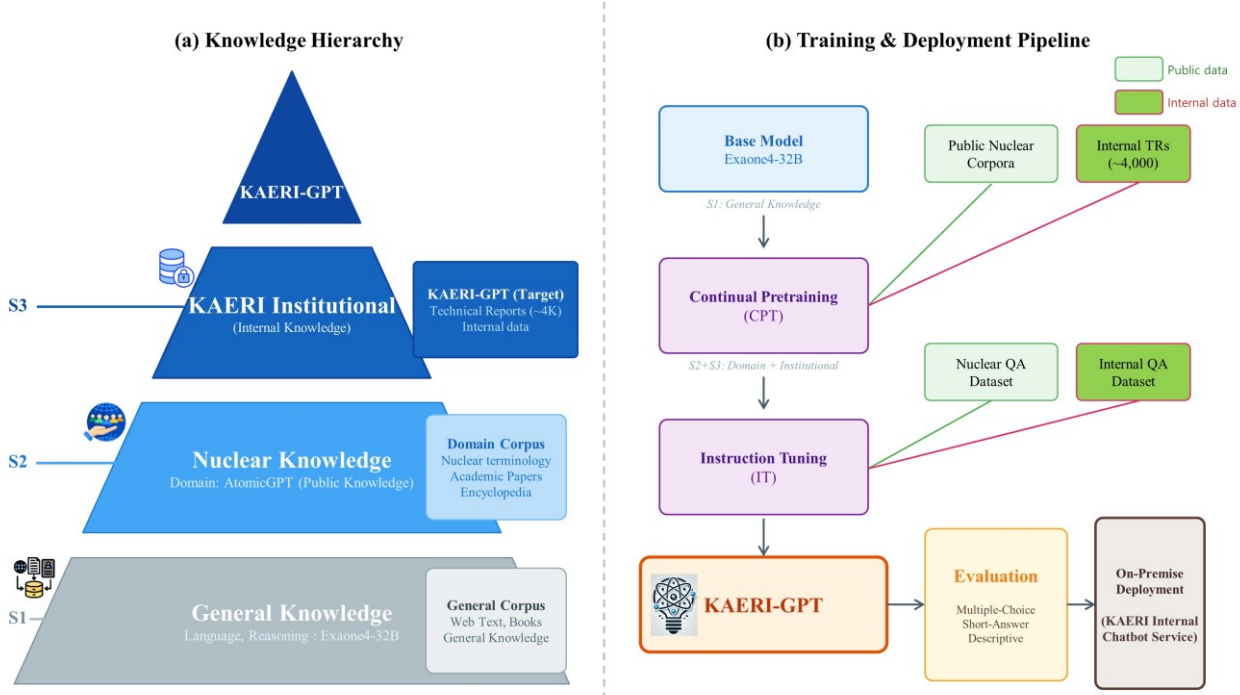

Fig. 1. Knowledge hierarchy and training pipeline of KAERI-GPT: (a) layered knowledge structure from general to institutional, (b) training and deployment workflow

Table I: Training Data Composition

Source	Data Type	Lang	Volume
KHNP	Nuclear terminology	KR	5,886 entries
NSSC	Regulatory glossary	KR	709 entries
KAERI	Academic papers	EN	2,152 papers
KAERI	Technical presentations	EN	335 docs
SNU	Atomic Wiki encyclopedia	KR	697 articles
KAERI (Internal)	Technical reports (TR)	KR/ EN	~4,000 reports

2.3 Evaluation Setup

Model performance was evaluated using a publicly available nuclear QA benchmark. The benchmark consists of 328 questions collected from three sources: (1) Nuclear Engineer License Examination (100 multiple-choice questions), (2) NuclearQA (100 short-answer questions), and (3) Atomic Wiki (128 descriptive questions) [13-15].

To reflect diverse question formats encountered in real-world nuclear knowledge assessment, the evaluation set was categorized into three types: multiple-choice questions, short-answer questions, and descriptive questions.

Type-specific evaluation metrics were applied: Exact Match (EM) for multiple-choice scoring on a 0-100 scale, F1 score(%) for short-answer, and LLM-as-a-Judge (GPT-4o) scoring on a 1–10 scale for descriptive questions. All models were evaluated under consistent inference settings.

3. Evaluation Results

3.1 Performance of AtomicGPT based on Nuclear Domain Benchmark

Table II presents the performance of AtomicGPT on the nuclear QA benchmark. Only the fully adapted models (CPT + IT) are reported.

Table II: AtomicGPT Performance on Nuclear QA Benchmark

Model	Multiple-Choice (EM, 0-100 score)	Short-Answer (F1, %)	Descriptive (LLM-Judge, 1-10 score)
Qwen2.5-7B	28	15.37	3.67
AtomicGPT-Qwen2.5-7B	37	18.1	3.94
Gemma2-9B	23	12.16	3.65
AtomicGPT-Gemma2-9B	40	27.4	4.67

Gemma2-9B-based AtomicGPT achieved higher performance across all question types compared to the Qwen2.5-7B variant. Specifically, the Gemma2-9B-based AtomicGPT achieved an EM score of 40 on multiple-choice, an F1 score of 27.4% on short-answer, and 4.67/10 on the LLM-Judge metric for descriptive questions, outperforming the Qwen2.5-7B variant across all categories.

These results confirm that domain-level continual pretraining and instruction tuning significantly enhance nuclear QA performance compared to base models.

3.2 Institutional Extension: KAERI-GPT

Table III presents the preliminary performance of KAERI-GPT compared to the Exaone4-32B base model and GPT-4. KAERI-GPT demonstrates improved performance across all evaluation categories.

Table III: KAERI-GPT Performance Comparison

Model	Multiple-Choice (EM, 0-100 score)	Short-Answer (F1, %)	Descriptive (LLM-Judge, 1-10 score)
Exaone4-32B (Base)	46	29.83	7.53
GPT-4	48	31.29	7.70
KAERI-GPT	56	34.50	8.21

In multiple-choice questions, KAERI-GPT achieved an EM score of 56, compared to 46 for Exaone4-32B and 48 for GPT-4. For short-answer questions, KAERI-GPT obtained an F1 score of 34.50%, outperforming both Exaone4-32B (29.83%) and GPT-4 (31.29%). In descriptive questions, KAERI-GPT achieved 8.21/10 on the LLM-Judge metric, exceeding both the base model (7.53) and GPT-4 (7.70).

These results indicate that institutional adaptation with internal data enhances nuclear domain performance beyond what model scale alone provides, as KAERI-GPT outperforms both the same-scale base model and GPT-4 across all question types.

4. Discussion

It should be noted that a direct comparison between AtomicGPT (7B/9B) and KAERI-GPT (32B) is confounded by model scale differences. However, the Exaone4-32B base model, despite its 32B scale, underperforms GPT-4 across all question types (e.g., EM 46 vs. 48, F1 29.83% vs. 31.29%, LLM-Judge 7.53 vs. 7.70). Since KAERI-GPT is built on the same Exaone4-32B architecture without any increase in model size, yet consistently outperforms both the base model and GPT-4 across all categories, the observed performance gains are more likely attributable to the integration of institutional data through continual pretraining and instruction tuning rather than to model scale alone.

To further isolate the contribution of internal institutional data, a controlled ablation study is planned: Exaone4-32B will be trained under two configurations—(1) public nuclear corpora only and (2) public + internal institutional corpora—using identical training hyperparameters, pipeline, and evaluation settings. This design ensures that any observed performance difference can be directly attributed to the inclusion of internal data. Extended comparisons with additional general-purpose LLMs (e.g., GPT, Gemini, Claude) are also planned.

5. Conclusions

In this study, we presented KAERI-GPT, an institutional on-premise extension of a nuclear domain large language model. Building upon the AtomicGPT framework, KAERI-GPT integrates internal technical reports and institutional QA data through continual pretraining and instruction tuning within a secured research environment. The model has been deployed as an internal chatbot service on KAERI's secured intranet.

Preliminary results on a public nuclear QA benchmark show that KAERI-GPT outperforms both the Exaone4-32B base model and GPT-4, achieving an EM score of 56, F1 of 34.50%, and LLM-Judge score of 8.21/10. These results demonstrate that institutional adaptation through internal data integration effectively enhances nuclear domain capability.

This work demonstrates the feasibility of extending domain-specialized LLMs to institution-specific on-premise models under security constraints. A controlled ablation study comparing Exaone4-32B trained with public-only data versus public + internal data under identical training configurations will further quantify the contribution of institutional data. Extended comparisons with additional general-purpose LLMs are also planned. The proposed framework provides a practical foundation for deploying secure, domain-adapted LLMs in nuclear research environments.

ACKNOWLEDGMENTS

This study was supported by the Substantiation Support Program through the Korea Innovation Foundation, funded by the Ministry of Science and ICT (No. 76170-26), and the Korea Atomic Energy Research Institute (KAERI) Institutional Program (No. 526140-26).

REFERENCES

[1] Singhal, Karan, et al. "Large language models encode clinical knowledge." Nature 620.7972 (2023): 172-180.
[2] Wu, Shijie, et al. "Bloomberggpt: A large language model for finance." arXiv preprint arXiv:2303.17564 (2023).
[3] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).
[4] Gururangan, Suchin, et al. "Don't stop pretraining: Adapt language models to domains and tasks." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.
[5] Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).
[6] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in neural information processing systems 35 (2022): 27730-27744.
[7] Yeom Seung Don, ChangSu Choi, Lim KyungTae, & Yu Yong Gyun (2024-11-20). Development and Performance Evaluation of a Domain-Specific Language Model for Nuclear: A Comparative Study Using a Custom-Built Dataset. Proceedings of Symposium of the Korean Institute of communications and Information Sciences, Gyeongsangbukdo.
[8] Korea Hydro & Nuclear Power Co., Ltd., *Nuclear Terminology Dictionary* [원자력용어집], 2019.https://www.data.go.kr/data/15038485/fileData.do?recommendDataYn=Y (accessed 2025).
[9] Korea Hydro & Nuclear Power Co., Ltd., *Nuclear-Related Laws and Regulations Glossary* [원자력관련법령 용어집], 2014. https://www.data.go.kr/data/15002295/fileData.do (accessed 2025).
[10] Nuclear Safety and Security Commission (NSSC), *Nuclear Safety Regulatory Terminology Dictionary* [원자력안전규제용어사전].https://www.nssc.go.kr/ko/cms/FR_CON/index.do?MENU_ID=2460 (accessed 2025).
[11] Korea Atomic Energy Research Institute (KAERI), *List of Recent Domestic Nuclear Trend Presentation Materials* [국내원자력관련최신동향발표자료목록],2020.https://www.data.go.kr/data/3077573/fileData.do (accessed 2025).
[12] Atomic Wiki,2023. https://atomic.snu.ac.kr/index.php/%EB%8C%80%EB%AC%B8 (accessed 2025).
[13] ComCBT, *Nuclear Engineer Qualification Exam Past Questions* [원자력기사 기출문제]. https://www.comcbt.com/xe/ccj (accessed 2025).
[14] Pacific Northwest National Laboratory (PNNL), *NuclearQA Benchmark Dataset*. https://github.com/pnnl/EXPERT2 (accessed 2025).
[15] Atomic Wiki, *Nuclear Engineering Q&A Dataset for Descriptive Questions* [원자력, 묻고 답하기]. https://atomic.snu.ac.kr/index.php/%EC%9B%90%EC%9E%90%EB%A0%A5,_%EB%AC%BB%EA%B3%A0_%EB%8B%B5%ED%95%98%EA%B8%B0 (accessed 2025).