Development and Deployment of AtomicGPT, a nuclear domain specific LLM for network constrained environments

1st Yeom Seungdon
Department of Artificial Intelligence
University of Science and Technology (UST)
Applied Artificial Intelligence Section (KAERI)
Daejeon, South Korea
tmdehs77@ust.ac.kr

2nd Yu Yonggyun
Department of Artificial Intelligence
University of Science and Technology (UST)
Applied Artificial Intelligence Section (KAERI)
Daejeon, South Korea
ygyu@kaeri.re.kr

***Abstract*—**Deploying Large Language Models (LLMs) in security-critical infrastructures presents unique challenges, particularly in network-constrained environments where cloud connectivity is prohibited. Research institutions like the Korea Atomic Energy Research Institute (KAERI) operate under strict isolation, shifting the computational burden entirely to on-premises resources. This paper introduces AtomicGPT, a domain-specific LLM for nuclear engineering, and proposes a secure on-premises serving architecture optimized for fully closed networks. We evaluate the system using OpenWebUI with two serving frameworks: Ollama and vLLM. Experimental results demonstrate that AtomicGPT outperforms its base models (Gemma2-9B, Qwen2.5-7B) by up to 17 percentage points on a custom nuclear benchmark. Furthermore, a comparative analysis reveals critical system trade-offs between inference latency and GPU resource efficiency. The vLLM-based architecture achieved responses within 2 to 2.5 seconds, compared to about 20 seconds for Ollama—representing up to a 9 times faster speed improvement. However, this latency reduction required higher VRAM consumption (29–34 GB vs. 24 GB), identifying Ollama as a more resource-efficient alternative for hardware-constrained edge nodes. This study validates the feasibility of high-security LLM services and provides architectural guidelines for balancing model specialization, system latency, and hardware resources in isolated network environments.

***Keywords*—** *Domain-specific Large Language Models, Secure deployment, Closed networks, Nuclear Engineering, On-premises service*

I. Introduction

The rapid progress of large language models (LLMs) has enabled intelligent services across a wide range of application domains, including healthcare, finance, and engineering. While general-purpose LLMs demonstrate remarkable capabilities in natural language understanding and reasoning, they often lack the accuracy and reliability required for domain-specific tasks in critical infrastructure. For example, nuclear engineering demands precise terminology and highly specialized knowledge that generic models cannot provide without adaptation. This motivates the development of domain-specific LLMs tailored for safety-critical environments.

At the same time, deploying LLM services in secure research environments introduces unique challenges. Organizations such as nuclear research institutes operate under strict security constraints where external network access is restricted. In such settings, cloud-based APIs and externally hosted models are infeasible, necessitating on-premises deployment solutions. Unlike cloud services where network transmission often dominates latency, in these air-gapped edge environments, the model inference time becomes the critical determinant of Quality of Service (QoS). Furthermore, limited computational resources in closed networks require serving architectures that balance inference performance, efficiency, and ease of integration.

In this paper, we address these challenges through the development and systematic deployment of AtomicGPT, a domain-specialized LLM for nuclear engineering. We first describe the construction of AtomicGPT and demonstrate its effectiveness over base models, achieving up to 17 percentage points improvement on domain-specific benchmarks. We then propose a secure on-premises serving architecture based on OpenWebUI, enabling chatbot-style interaction entirely within a closed network. Finally, we compare two popular serving frameworks, Ollama and vLLM, in terms of inference latency and GPU resource efficiency, showing that vLLM provides lower latency while Ollama offers lightweight deployment advantages.

The contributions of this paper are threefold: (1) The design and evaluation of AtomicGPT, a domain-specific LLM for nuclear engineering, (2) A secure serving architecture for providing LLM services in restricted research networks using OpenWebUI, (3) An empirical comparison of serving frameworks (Ollama vs vLLM) in closed research environments, providing architectural guidelines for optimizing the trade-offs between inference latency, GPU VRAM usage, and ease of management.

II. Related Work

A. Domain-Specific Large Language Models(LLMs)

Recent advances in domain-specific large language models (LLMs) have demonstrated significant performance improvements over general-purpose models. For instance, healthcare-oriented models such as BioBERT achieve

notable gains in biomedical text understanding, while financial models like FinBERT show superior accuracy in financial sentiment analysis [2], [3]. However, existing research has primarily focused on model development and performance evaluation rather than on the practical deployment of such models in secure or resource-constrained environments. In particular, there has been little research on domain-specific LLMs for the nuclear engineering field, where strict security requirements and data sensitivity pose unique challenges.

B. LLM Serving framework

Modern LLM deployment utilizes various serving frameworks optimized for different scenarios. OpenWebUI provides a user-friendly interface based on a containerized architecture, making it suitable for air-gapped deployments. Ollama focuses on simplified local deployment with minimal configuration, emphasizing resource efficiency. vLLM offers high-performance serving with advanced techniques such as PagedAttention and continuous batching, achieving superior throughput and lower latency [4]. However, few studies have systematically compared these frameworks under secure and resource-constrained conditions.

III. Model Development and Deployment

In this study, we first developed a domain-specialized large language model tailored for the nuclear engineering domain. We then deployed this model in the secure, closed network environment of the Korea Atomic Energy Research Institute (KAERI), where external connectivity is strictly restricted. The deployment was realized on-premises using the OpenWebUI framework, enabling interactive chatbot services accessible to researchers without relying on external APIs. This approach ensures both domain reliability and security compliance while providing practical usability within restricted infrastructures.

A. AtomicGPT Development

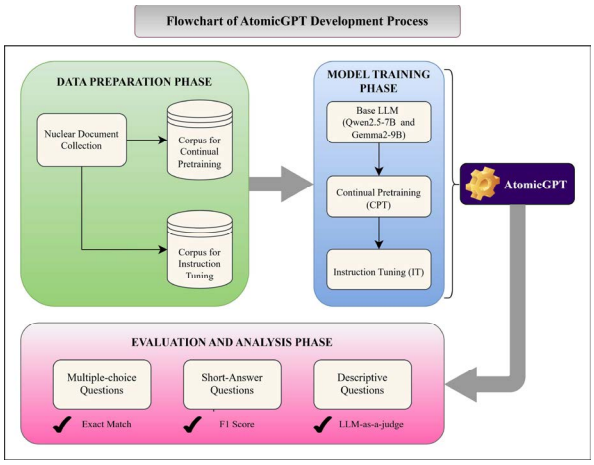

Fig. 1. Overall research workflow for developing AtomicGPT.

Fig. 1 illustrates the overall workflow of this study for developing AtomicGPT. The process is divided into three major stages: data preparation, model training, and evaluation.

In the data preparation phase, we collected nuclear-domain resources from multiple authoritative organizations, including Korea Hydro & Nuclear Power (KHNP) [5], [6], the Nuclear Safety and Security Commission (NSSC) [7], the Korea Atomic Energy Research Institute (KAERI) [8], [9], and the Nuclear Policy Center at Seoul National University [10]. The collected materials consisted of nuclear glossaries, regulatory dictionaries, research papers, and domain-specific knowledge bases. These resources were curated into two types of datasets: a pre-training corpus for continual pre-training (CPT) and an instruction dataset for instruction tuning (IT). In addition, a separate benchmark dataset was constructed for performance evaluation. Data augmentation was applied through bilingual translation (Korean ↔ English) to enrich the corpus, and data quality management steps such as deduplication, filtering, and expert review were performed to ensure reliability. The final dataset comprised approximately 50 million (M) tokens, which served as the foundation for adapting the base models to the nuclear domain.

In the model training phase, we employed Qwen2.5-7B and Gemma2-9B as base models. Two strategies were applied: (1) continual pre-training (CPT) to inject nuclear-domain knowledge, and (2) instruction tuning (IT) to enhance task adaptability.

Finally, in the evaluation phase, we evaluated model performance using the nuclear QA benchmark, which included multiple-choice, short-answer, and descriptive questions [11].

B. Deployment of AtomicGPT

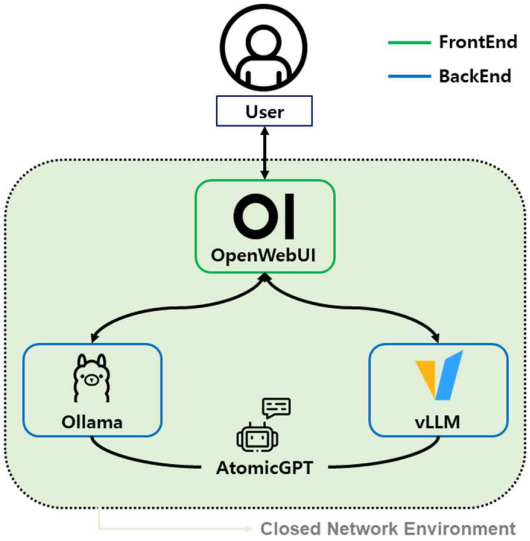

Fig. 2. Framework architecture for AtomicGPT model serving in KAERI internal network-constrained environment.

Fig. 2 presents the deployment architecture of AtomicGPT within the secure research network of KAERI. The primary objective of this design is to provide domain-specific LLM services in a closed environment without any dependency on external connectivity.

At the frontend, researchers interact with the model through OpenWebUI, which offers a web-based chatbot interface. OpenWebUI allows users to submit queries and

receive responses seamlessly, while maintaining usability and accessibility within the restricted intranet.

The backend serving layer is configured with two alternative frameworks: Ollama and vLLM. Ollama is lightweight and easy to configure, making it suitable for restricted systems that prioritize simplicity and portability. In contrast, vLLM is optimized for performance, featuring advanced memory management and scheduling mechanisms that reduce response latency under load. Both frameworks access the same AtomicGPT model, but they provide different trade-offs in terms of deployment efficiency and runtime performance. To further enhance reliability and manageability, the entire service stack is deployed using Docker containers. This containerized approach provides several advantages:

- Environment isolation ensures that all dependencies are encapsulated, minimizing conflicts with the host system.

- Reproducibility allows consistent deployment across different machines in the secure network.

- Simplified management enables rapid updates, scaling, and rollback without interfering with the host environment.

All components, including OpenWebUI, serving engines, and the AtomicGPT model deployment, operate strictly within the closed network environment, as indicated by the dashed boundary in Fig. 2. This ensures that no data or queries leave the internal infrastructure, thereby satisfying stringent security requirements while enabling practical AI services for nuclear research environments.

IV. Experimental Results

In this section, we evaluate AtomicGPT in terms of domain-specific model performance and deployment efficiency. The evaluation consists of two parts: (1) model accuracy on nuclear QA benchmarks compared with baseline LLMs, and (2) response time and GPU usage under different serving frameworks (Ollama and vLLM). The experiments were conducted using two NVIDIA A100 GPUs with 40GB memory each.

A. AtomicGPT model performance

The benchmark consisted of 328 questions, including 100 multiple-choice, 100 short-answer, and 128 descriptive questions. This benchmark design provides a comprehensive assessment of knowledge understanding and problem-solving ability in the nuclear domain.

We compared a total of four models: the base models (Qwen2.5-7B and Gemma2-9B) and their domain-adapted counterparts, AtomicGPT, which were trained with both continual pre-training (CPT) and instruction tuning (IT). For each question type, appropriate metrics were applied: Exact Match (EM) for multiple-choice, F1 Score for short-answer, and LLM-as-a-Judge (using the GPT-4o) for descriptive responses.

TABLE I. Overall Performance by Model and Evaluation Questions Type

Model	*Multiple-Choice (EM, %)*	*Short-Answer (F1, %)*	*Descriptive (LLM-as-a-Judge, %)*
Qwen2.5-7B	28	15.37	36.7
AtomicGPT - Qwen2.5-7B	37	18.08	39.4
Gemma2-9B	23	12.16	36.5
AtomicGPT - Gemma2-9B	**40**	**27.44**	**46.7**

Table I presents the evaluation results. Across all benchmarks, AtomicGPT variants consistently outperformed their base counterparts. In particular, Gemma2-9B AtomicGPT achieved the highest performance, reaching 40% EM on multiple-choice, 27.44% F1 score on short-answer, and 46.7% on descriptive evaluation, representing a maximum improvement of 17 percentage points in EM score over the baseline. Additionally, to ensure the reliability of the GPT-4o-based evaluation, we conducted a human evaluation test on a random sample of 50 responses. The human expert's scoring aligned with the automated judge in over 80% of the cases, verifying the validity of our LLM-as-a-Judge methodology. These results confirm the effectiveness of combining CPT and IT to specialize LLMs for the nuclear engineering domain.

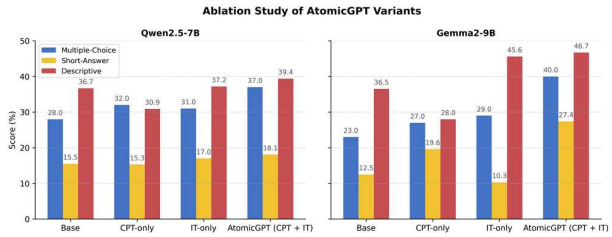

Fig. 3. Ablation study of AtomicGPT variants. The performance comparison across Base, CPT-only, IT-only, and the final AtomicGPT (CPT+IT) models demonstrates the synergistic effect of combining domain knowledge injection (CPT) and task alignment (IT).

To verify the contribution of each training stage, we conducted an ablation study as shown in Fig. 3. The results indicate distinct benefits from each phase: the CPT-only models showed consistent improvement in multiple-choice questions, reflecting enhanced domain knowledge comprehension. Meanwhile, the IT-only models demonstrated superior performance in descriptive questions, highlighting the importance of instruction tuning for generating coherent and detailed explanations. Ultimately, the AtomicGPT (CPT+IT) model achieved the highest scores across all metrics, confirming that combining continual pre-training with instruction tuning is essential for maximizing domain-specific performance.

B. Serving Performance

To evaluate deployment efficiency in secure and network-restricted environments, we compared two popular

serving frameworks, Ollama and vLLM, when running AtomicGPT with OpenWebUI as the frontend. In such closed networks, where external APIs cannot be utilized, selecting an optimal serving framework is critical to ensure both responsiveness and usability for researchers.

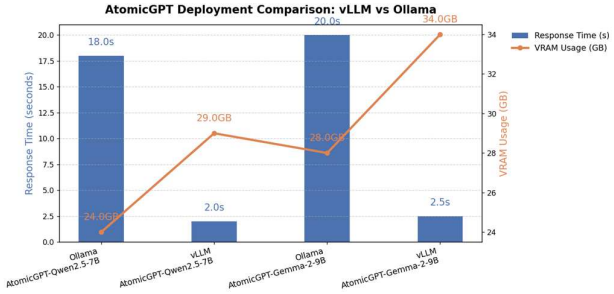

Fig. 4. Comparison of average response time(inference latency) and GPU vram usage for AtomicGPT served with Ollama and vLLM.

Fig. 4 summarizes the comparative results of our deployment performance measurements for AtomicGPT. Across both model variants—AtomicGPT-Qwen2.5-7B and AtomicGPT-Gemma2-9B—the vLLM framework consistently achieved significantly lower response latency than Ollama, confirming its advantage for performance-sensitive deployments even under strict on-premises constraints. For instance, AtomicGPT-Qwen2.5-7B recorded an average latency of 18 seconds when served with Ollama, compared to only 2 seconds with vLLM; similarly, the Gemma2-9B model showed 20 seconds versus 2.5 seconds, respectively. However, this speed improvement required higher GPU memory usage—approximately 29–34 GB with vLLM, compared to 24–28 GB for Ollama—highlighting the trade-off between response speed and resource efficiency.

In addition, Fig. 5 presents the actual deployment of AtomicGPT within the OpenWebUI interface. The system provides a secure, interactive chatbot service entirely within the closed KAERI network, allowing researchers to query nuclear engineering knowledge and institutional data without any external connectivity. This configuration effectively demonstrates the real-world applicability of domain-specific LLMs in air-gapped infrastructures, combining security assurance with user-friendly accessibility. Furthermore, AtomicGPT enables future integration with autonomous agent workflows, where the model can assist researchers by automating document retrieval, report summarization, and data interpretation while maintaining full compliance with internal security policies.

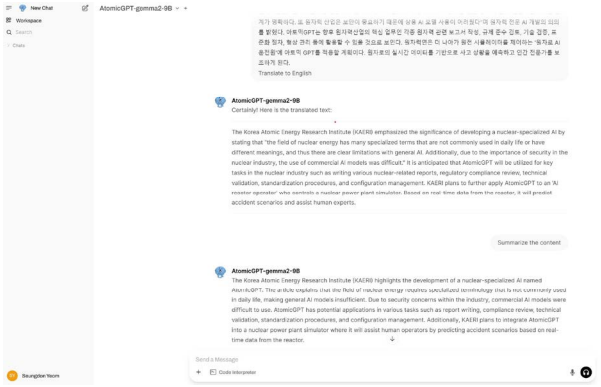

Fig. 5. On-premises deployment of the AtomicGPT-gemma2-9B model as a chatbot interface at the Korea Atomic Energy Research Institute (KAERI).

V. Conclusion

In security-critical research institutions such as the Korea Atomic Energy Research Institute (KAERI), public cloud-based LLM services like ChatGPT or Claude cannot be utilized due to strict network isolation. This constraint highlights the need for domain-specific language models that can be fully operated within air-gapped environments.

This study introduced AtomicGPT, a nuclear domain-specific LLM securely deployed using OpenWebUI, Ollama, and vLLM without external connectivity. Experimental evaluations demonstrated that interactive LLM services can be effectively implemented in closed networks, and that balancing GPU resource utilization and response latency is a key factor in deployment design. The vLLM deployment achieved rapid response times (2–2.5 s) with higher VRAM usage (29–34 GB), while Ollama provided greater resource efficiency (24–28 GB), underscoring the importance of selecting deployment frameworks that align with system constraints.

In this study, we primarily focused on model accuracy and end-to-end response latency to validate the feasibility of on-premises deployment. For future work, we plan to conduct a more granular analysis of inference performance using metrics such as Time To First Token (TTFT) and generation throughput (tokens/s). This will allow for a deeper optimization of the trade-off between user interactivity (latency) and system efficiency (throughput) in resource-constrained network environments. Furthermore, we aim to extend AtomicGPT toward a domain-specific intelligent agent system that integrates nuclear engineering knowledge bases and supports researchers with tasks such as document retrieval, report summarization, and experimental data interpretation. Through sustainable on-premises deployment and operational optimization, AtomicGPT aims to establish a foundation for intelligent research assistants in high-security scientific institutions.

Acknowledgment

This work was supported in part by Korea Atomic Energy Research Institute R&D Program under Grant KAERI-524540-25.

REFERENCES

[1] KAERI-MLP, "AtomicGPT-Gemma2-9B: A domain-specific large language model for nuclear engineering," Hugging Face, 2024. [Online]. Available: https://huggingface.co/KAERI-MLP/AtomicGPT-gemma2-9B.

[2] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[3] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv*:1908.10063, 2019.

[4] W. Kwon et al., "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023, pp. 611–626.

[5] Korea Hydro & Nuclear Power (KHNP), "Glossary of Nuclear Terms," 2025, dataset. [Online]. Available: https://www.data.go.kr/data/15038485/fileData.do?recommendDataYn=Y

[6] Korea Hydro & Nuclear Power (KHNP), "Glossary of Nuclear Law Terms," 2025, dataset. [Online]. Available: https://www.data.go.kr/data/15002295/fileData.do

[7] Nuclear Safety and Security Commission (NSSC), "Dictionary of Nuclear Safety Regulatory Terms," 2025, website. [Online]. Available: https://www.nssc.go.kr/ko/cms/FR_CON/index.do?MENU_ID=2460

[8] Korea Atomic Energy Research Institute (KAERI), "List of Recent Domestic Presentations on Nuclear-Related Trends," 2025, dataset. [Online]. Available: https://www.data.go.kr/data/3077573/fileData.do

[9] Korea Atomic Energy Research Institute (KAERI), "List of Domestic Scholarly Papers on Nuclear-Related Topics," 2025, dataset. [Online]. Available: https://www.data.go.kr/data/3083121/fileData.do

[10] Atomic Wiki, "Nuclear Engineering Encyclopedia (Atomic Wiki)," 2025, website. [Online]. Available: https://atomic.snu.ac.kr/index.php/%EB%8C%80%EB%AC%B8

[11] S. Yeom, "AtomicGPT Evaluation Datasets," 2025, GitHub repository. [Online]. Available: https://github.com/tmdehs77/atomicgpt-eval-datasets