Transactions of the Korean Nuclear Society Autumn Meeting
Gyeongju, Korea, October 29-30, 2026

Comparison of Full Fine-Tuning and LoRA Fine-Tuning for Nuclear Reactor Operator Exam QA Using Gemma Models

Isak Hwang[a,b], Yoonpyo Lee[c*], Hogeon Seo[b, d*]

[a]*Department of Nuclear Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, 04763, Korea*
[b]*Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Yuseong-gu, Daejeon, 34057, Korea*
[c]*The Grainger College of Engineering, Nuclear, Plasma & Radiological Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA*
[d]*University of Science & Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, 34113, Korea*

*Corresponding authors: yoonpyo2@illinois.edu (Y. Lee), hogeony@hogeony.com (H. Seo)

***Keywords : large language models, LoRA, fine-tuning, nuclear reactor operator exam, multimodal model**

1. Introduction

The growth of artificial intelligence (AI) has increased data-center electricity demand and raised the value of stable large-scale power sources such as nuclear plants [1]. As the nuclear sector adopts AI, the security and closed-network requirements of nuclear facilities make cloud-based commercial Large Language Models (LLMs) difficult to deploy [2]. Locally hosted LLMs keep the model and its data inside the facility and are a more practical option for nuclear operations.

A local LLM must be adapted to nuclear-domain tasks under limited on-premises computing. Two adaptation strategies are available. Full fine-tuning (FFT) updates every parameter [3]. Low-Rank Adaptation (LoRA) trains a small set of low-rank adapter weights while the base model stays frozen [4], which enables adaptation of much larger models under the same computational resource constraints. Which strategy better suits a security-constrained nuclear deployment remains unclear.

To assess whether a local model can support nuclear work, we use an operator licensing exam as the benchmark. The exam is written and scored for human candidates, providing a common benchmark for comparing model performance with human examinees. We compare the two strategies on U.S. Nuclear Regulatory Commission (NRC) Reactor Operator (RO) licensing exam questions [5]. The models are Gemma 3 4B-IT and Gemma 4 31B [6, 7], which differ in both generation and parameter scale. The 4B model is fully fine-tuned, and the 31B model is run in base and LoRA-tuned form. We ask whether LoRA on a larger local model outperforms full fine-tuning on a smaller model under the tested conditions.

RO exam items are not text-only. They include plant diagrams, system schematics, and instrument trend plots that cannot be answered from text alone. These items need visual reasoning as well as text understanding, so we use a multimodal (vision-language) model rather than a text-only LLM [8].

This work makes three contributions. First, it compares full fine-tuning of a smaller model with LoRA tuning of a larger one on nuclear operator exam questions under local deployment constraints. Second, it builds a multimodal training set of 3,577 exam items with generated reasoning steps and evaluates the models on 700 held-out questions from both reactor types, with figures processed as images rather than text. Third, it reports accuracy for the four configurations against the 80% score required to pass. LoRA tuning of the 31B model reached 75.5% and outperformed the untuned 31B model at 52.2% and full fine-tuning of the 4B model at 39.4%, but no configuration reached the passing score.

2. Methods

2.1 Dataset and evaluation

We collected questions and answer keys from the U.S. NRC RO licensing exam question bank for both Pressurized Water Reactor (PWR) and Boiling Water Reactor (BWR) exams. The bank includes text-only and image-based items. For training we paired each question with a chain-of-thought (CoT) rationale generated by the gemini-3-flash-preview API [9, 10], giving 3,577 question-rationale-answer examples. The evaluation set kept the questions and answer keys only, with no rationale.

We used the multimodal capability of the Gemma models for the image-based items. Instead of replacing each figure with a text caption, we passed the exam image to the model together with the question text, so the model reads the figure itself as a human examinee would. The same format was used in training and evaluation. The 4B and 31B models were evaluated using an MLX vision-language pipeline.

The evaluation set held 350 questions per reactor type from the March exams of 2015 to 2021, with no overlap with the training set. We read each model's selected option with a simple rule and compared it against the key. Scoring uses the final selected option, as in the RO exam. A few questions gave no readable answer, mostly from

format errors or timeouts, and were excluded from scoring. Fig. 1 outlines this workflow.

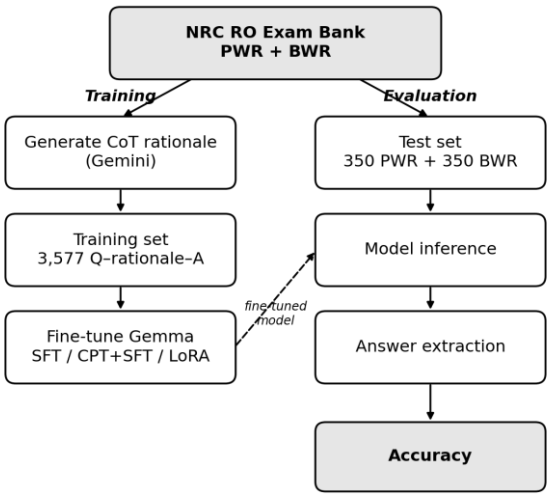


Fig. 1. Data-generation and evaluation workflow.

2.2 Full fine-tuning: Gemma 3 4B-IT

We applied FFT to Gemma 3 4B-IT in two configurations. The first used Supervised Fine-Tuning (SFT) alone for 3 epochs at a learning rate of 2e-5 [11], training only on the answer part of each example. The second added a Continued Pre-Training (CPT) phase before SFT [12], 2 epochs at 1e-5 over the full text, to adapt the model to the exam data before the same SFT step. Table I reports these as "SFT only" and "CPT+SFT".

Both configurations used a batch size of 8, a 2,048-token limit, bf16 precision, a cosine schedule, and gradient checkpointing. Training ran locally on Apple Silicon (M5 Max, 128 GB, MPS), and all 4 billion parameters were updated. CPT loss fell from 2.73 to 0.75 and SFT loss from 1.01 to 0.34.

2.3 Base and LoRA fine-tuning: Gemma 4 31B

The Gemma 4 31B model was run as a base model and a LoRA-tuned model. The base model used pretrained instruction-tuned weights without further domain-specific training and served as the baseline. The LoRA model used the same 3,577 examples as the 4B model to train a low-rank adapter. The 31 billion base parameters were frozen and only the adapter weights were updated. LoRA training used rank 16, alpha 32, and dropout 0.05, targeting the q, k, v, o, gate, up, and down projection layers of the language model while keeping the vision tower and multimodal projector frozen. Training was performed for 3 epochs with an effective batch size of 8, a learning rate of 2×10^{-4}, a cosine learning-rate scheduler, AdamW optimization, bfloat16 precision, and a maximum sequence length of 2,048 tokens. For inference the adapter was merged into the base model to avoid added latency. Both were evaluated with an MLX inference pipeline and no external retrieval. The answer-extraction rules and scoring matched those used for the 4B model.

3. Results

Table I and Fig. 2 report accuracy for the four configurations, computed over valid responses only. A small number of invalid outputs from formatting errors or timeouts were excluded.

Model	Adapt	PWR	BWR	Avg
Gemma 3 4B-IT	SFT only	37.4	38.4	37.9
Gemma 3 4B-IT	CPT+SFT	40.8	37.9	39.4
Gemma 4 31B	None (base)	52.3	52.2	52.2
Gemma 4 31B	LoRA FT	76.5	74.5	75.5

Table I: NRC exam QA accuracy (%) by model and adaptation strategy

FFT of Gemma 3 4B-IT gave limited gains. SFT alone reached 37.9% average accuracy, and adding a CPT phase before SFT raised this to 39.4%. Average accuracy stayed below 40% in both cases. The untuned Gemma 4 31B base model reached 52.2%, 13 points above the best fine-tuned 4B model. However, because the compared models differ in generation and training conditions as well as parameter scale, this performance gap cannot be attributed solely to model size.

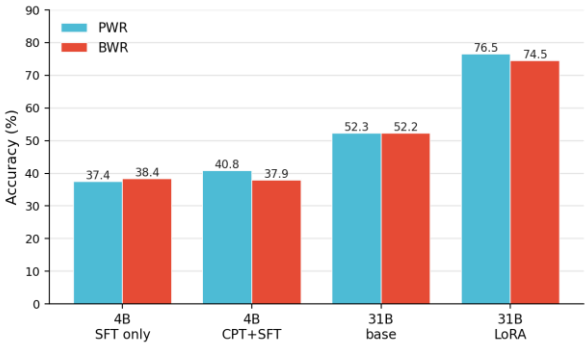


Fig. 2. Accuracy by configuration on the PWR and BWR test sets.

LoRA on the 31B model raised average accuracy to 75.5% while updating only a small fraction of the parameters. This is 23 points above the 31B base model and 36 points above the best 4B model. The 31B LoRA model achieved the highest accuracy on both the PWR and BWR test sets (Fig. 2).

4. Discussion & Conclusion

We expected FFT of the 4B models to approach the untuned 31B model. Instead, both 4B configurations updated every weight but stayed in the high 30s, behind

the untuned 31B model. The stronger performance of the larger base model may reflect more relevant knowledge acquired during pretraining. LoRA adapts it to the exam format while leaving that base intact, whereas FFT on only 3,577 examples can overfit or forget what was learned in pretraining. This may also explain why the extra CPT phase provided only a small improvement.

The comparison therefore reflects two practical adaptation choices under local resource constraints. One option is to fully fine-tune a model small enough to update in its entirety, whereas the other is to freeze a much larger model and train only a small adapter. The second option was the better-performing approach here, since the LoRA-tuned 31B model was the most accurate configuration and needs only a small adapter to be trained and stored, which suits on-site nuclear use with limited computing.

Several limitations remain. Even the best model did not pass the exam. The RO exam requires 80%, and the 31B LoRA model reached 75.5%, so no configuration met the standard. Scoring also used only the final selected option, so we cannot confirm that a correct answer came from correct reasoning. A model can reach the right option through a flawed process, and accuracy does not separate these cases. Still, all configurations performed above the 25% chance level for four-choice items. However, this alone does not establish valid reasoning.

Future work has two aims. One is to reach the passing standard through better training data, retrieval, larger models, or stronger supervision of the reasoning steps. The other is to verify whether the reasoning is valid by analyzing the generated rationale as well as the final selected option. These results should therefore be interpreted as a practical comparison of adaptation strategies under local resource constraints rather than as an isolated measurement of model-scale effects.

ACKNOWLEDGEMENTS

This work was supported by the Institute Strategic Development Project (ISD) funded by the Ministry of Science and ICT (No. 526140-26), the Substantiation Support Program through the Korea Innovation Foundation, funded by the Ministry of Science and ICT, and the Program for Strengthening Technological Capabilities in Nondestructive Testing through the National Research Foundation of Korea (No. RS-2026-25545949), funded by the Ministry of Science and ICT.

REFERENCES

[1] IEA (2025), Energy and AI, IEA, Paris https://www.iea.org/reports/energy-and-ai, Licence: CC BY 4.0.

[2] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, 1(2), 1-124.

[3] Lv, K., Yang, Y., Liu, T., Guo, Q., & Qiu, X. (2024, August). Full parameter fine-tuning for large language models with limited resources. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8187-8198).

[4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *Iclr*, *1*(2), 3.

[5] U.S. Nuclear Regulatory Commission, Generic Fundamentals Examinations for Reactor Operators: PWR and BWR Question Banks, 2021.
https://www.nrc.gov/reactors/operator-licensing/history-rulemaking-activities/generic-fundamentals-examinations/past-exams#pwr

[6] Gemma Team, Gemma 3 Technical Report, Google DeepMind, 2025.

[7] Gemma Team, Gemma 4 Technical Report, Google DeepMind, 2026.

[8] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, *35*, 23716-23736.

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

[10] Google DeepMind, Gemini 3 (gemini-3-flash-preview), 2025. https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

[12] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 8342-8360).