# Hardware Performance Analysis for Deploying a Report-Writing Agent in Closed-Network Nuclear Industry Environments

Dae-Young Lee*, Sang-Beom Kang, Seung-Hyeok Yang
*FNC Technology Co., Ltd., 13 Heungdeok 1-ro, 32F, Giheung-gu, Yongin-si, Gyeonggi-do, 16954, Korea*
**Corresponding author: ldy242@fnctech.com*

***Keywords : closed network, local LLM, cost-performance, periodic safety review, report generation**

## 1. Introduction

Organizations in the nuclear industry handle sensitive information on plant operation, design, maintenance, and security, as well as information related to critical national infrastructure, and they strictly control data access, transfer, and external connectivity under national cybersecurity guidelines and their own security regulations. In such environments, feeding internal work documents into external commercial cloud-based large language model (LLM) services is practically impossible. Local deployment of open-weight models has therefore been the only viable option for applying LLMs to document-intensive nuclear work.

The largest obstacle to local LLM deployment, however, has been cost. Datacenter-class GPU servers based on the NVIDIA H100 require a substantial capital investment, which is hard to justify for a department whose document-centric workload is infrequent and involves few concurrent users.

Two recent developments have greatly lowered this cost barrier. First, AI workstations represented by the NVIDIA DGX Spark now provide 128 GB of unified memory in a desktop form factor at a small fraction of the price of a GPU server, although the memory bandwidth of 273 GB/s is far below the roughly 2 TB/s per GPU of the H100 PCIe [1]. Second, small and mid-sized open-weight models that are compact yet capable enough for structured document drafting have been released in rapid succession. Together, these developments open the possibility of department-scale document automation on compact hardware without an expensive GPU server.

In this study, the model family was fixed to Qwen3.6 and four configurations combining architecture (MoE/dense) and precision (FP16/FP8) were served with the same vLLM stack [2] on a DGX Spark and a dual-H100 server, and throughput, latency, and generation quality were measured on a periodic safety review (PSR) safety factor report generation workload. The results are normalized by acquisition cost and energy consumption, providing nuclear organizations with measurement-based evidence for selecting a deployment configuration and platform appropriate to their usage scale.

## 2. Methods

### *2.1 Hardware Platforms*

The specifications of the two platforms are given in Table I, based on vendor documentation and published market prices [1]. The DGX Spark lists at about USD 4,700 (as of 2026), and H100 PCIe 80 GB cards trade at USD 25,000-30,000 each, so the dual-GPU server is estimated at about USD 75,000 including the host. The rated power is 240 W for the entire DGX Spark system and, for the GPU server, the sum of two 350 W PCIe GPUs and the host server.

Table I: Platforms compared in this study

| Item | DGX Spark | GPU server (2x H100) |
|---|---|---|
| Processor | GB10 Grace Blackwell | 2x NVIDIA H100 80 GB |
| Memory | 128 GB unified LPDDR | 160 GB HBM3 (2x 80 GB) |
| Bandwidth | 273 GB/s | ~2.0 TB/s (per GPU) |
| Rated power | ~240 W (system) | ~1.2 kW (system) |
| Cost | ~$4,700 (list price) | ~$75,000 (GPUs + host) |

### *2.2 Models and Serving Configuration*

To compare the effect of deployment configuration rather than differences between models, the model family was fixed to Qwen3.6 and two variants were selected: an MoE variant (35B-A3B) and a dense variant (27B) (Table II). Because the MoE (mixture-of-experts) model activates only 3 B parameters per token [3], its decoding speed is governed by the memory traffic of the active parameters, which is expected to be advantageous on bandwidth-limited hardware. Each model was served with vLLM [2] on both platforms at two precisions, FP16 and FP8 quantization, giving four architecture-precision configurations. FP16 is the baseline configuration with no risk of quality degradation, while FP8 is a deployment option that halves the weight memory traffic to increase throughput.

Table II: Evaluated open-weight models

| Model | Type | Total/active parameters | License |
|---|---|---|---|
| Qwen3.6-35B-A3B [3] | MoE | 35 B / 3 B | Apache 2.0 |
| Qwen3.6-27B [4] | Dense | 27 B / 27 B | Apache 2.0 |

### *2.3 Workload Definition*

The workload of this study is defined as template-based generation of the '8.1 Safety Performance' report, one of the fourteen safety factors of a full PSR (periodic

safety review). The model is given the report template for this safety factor together with input data. The main evaluation items of the 8.1 area are standardized by the PSR guideline framework, and the input data consist of ten years of accumulated records, including unplanned reactor trip history, maintenance, test, and inspection records, and fifteen safety performance indicators evaluated quarterly. From these, the model generates assessment content paragraph by paragraph, following the given table of contents according to user requests; the resulting body of the report runs to several tens of pages. The benchmark scale of 8,000 input tokens and 3,000 output tokens is sufficient to fully generate one representative assessment section (about four pages).

The 8.1 report was chosen as the evaluation target because its format is standardized by regulation and its content consists largely of verifiable figures and operation and maintenance data, making it well suited to quantitative evaluation of the template compliance rate and the factual accuracy rate. In addition, the PSR recurs every ten years and carries a heavy documentation burden, so the practical benefit of automation is evident.

Two execution modes were compared: (i) single-pass generation and (ii) a three-stage multi-agent pipeline (drafting, review, revision) based on the same model, the latter being the configuration intended for actual operation. Tests were also run at concurrency levels of 1, 2, and 4 to emulate a department-scale multi-user environment.

### *2.4 Evaluation Metrics*

Performance is reported as prefill throughput, decode throughput, and end-to-end latency per report. Economic efficiency is reported as decode throughput per USD 1,000 of acquisition cost and as energy per report, obtained by multiplying the rated power by the latency (a conservative upper bound on operating power). Generation quality across configurations and platforms was verified with the following two metrics. The factual accuracy rate is the fraction of responses whose cosine similarity to the corresponding paragraph of the reference report meets a threshold (0.85), and the template compliance rate checks whether the generated report follows the standard template.

## 3. Results and Discussion

Single-stream results for a representative PSR section generation workload are summarized in Table III (S: DGX Spark, H: dual-H100 server; MoE: Qwen3.6-35B-A3B, Dense: Qwen3.6-27B). The DGX Spark measurements are consistent with published benchmarks on the same hardware [5]. At FP16, model architecture was the dominant factor on the compact workstation: the MoE configuration decoded at about 30 tok/s on the DGX Spark, whereas the dense configuration reached only 4.4 tok/s, taking more than 11 minutes per section. Dense configurations of this size are thus impractical on compact hardware, whereas the MoE configuration completed a section in about two minutes. The dual-H100 server was about six times faster in both prefill and decode, and the gap tended to widen with concurrency, so it remains preferable for many users or long-context environments. When acquisition cost is taken into account, however, the comparison reverses: at MoE-FP16, decode throughput per USD 1,000 was about 2.6 times higher on the DGX Spark than on the server. Energy per report for the MoE configurations was similar on both platforms at 5.7-6.9 Wh, showing that the difference in economic efficiency arises from acquisition cost rather than power. With FP8 quantization, decode speeds reached 50.5 (S) / 244.7 (H) tok/s for MoE and 10.4 (S) / 161.9 (H) tok/s for dense, a 1.3-2.4x improvement over FP16, and the per-USD 1,000 throughput advantage of the MoE configuration widened to about 3.3x. For a department generating tens of reports per day at low concurrency, a latency of about two minutes is entirely acceptable, and at that scale the workstation is the more economical choice.

Table III: Single-stream throughput and cost efficiency (concurrency = 1; 8k-token input, 3k-token output)

| Model/Platform | Prefill (tok/s) | Decode (tok/s) | Latency (s) | tok/s / $1k | Energy (Wh) |
|---|---|---|---|---|---|
| MoE-FP16 / S | 3,590 | 29.6 | 104 | 6.3 | 6.9 |
| MoE-FP16 / H | 22,963 | 181.2 | 17 | 2.4 | 5.7 |
| MoE-FP8 / S | 5,147 | 50.5 | 61 | 10.7 | 4.1 |
| MoE-FP8 / H | 34,194 | 244.7 | 12 | 3.2 | 4.2 |
| Dense-FP16 / S | 1,343 | 4.35 | 695 | 0.9 | 46.3 |
| Dense-FP16 / H | 7,787 | 81.9 | 38 | 1.1 | 12.7 |
| Dense-FP8 / S | 2,611 | 10.4 | 292 | 2.2 | 19.4 |
| Dense-FP8 / H | 14,140 | 161.9 | 19 | 2.1 | 6.4 |

Cross-platform quality (factual accuracy and template compliance rates), the latency of the multi-agent pipeline, and throughput scaling at concurrency levels of 2 and 4 were also evaluated on the same workload, and the detailed results will be presented at the meeting. Platform selection guidance as a function of the number of concurrent users and report length will also be discussed.

## 4. Conclusions

In this study, for local LLM deployment in closed-network nuclear industry environments, four configurations combining the architecture (MoE/dense) and precision (FP16/FP8) of a fixed model family (Qwen3.6) were compared on a compact AI workstation and a dual-H100 server. In the MoE configuration, the workstation delivered about 2.6 times the decode throughput per unit of acquisition cost of the server, and its latency was acceptable for department-scale PSR documentation work. Even within the same model family, the architecture and precision configuration determines practicality on compact hardware, and the server retains an advantage only when there are many concurrent users or long contexts. No directly comparable public measurements have been reported for this workload, and these results can serve as a reference for nuclear

organizations planning document automation in closed-network environments when selecting a deployment configuration and platform.

## ACKNOWLEDGMENTS

This work was supported by Korea Hydro & Nuclear Power Co., Ltd. (No. 2025-AI-001-01).

## REFERENCES

[1] NVIDIA, DGX Spark Documentation, 2025; H100 GPU Datasheet, 2023.

[2] W. Kwon et al., Efficient Memory Management for Large Language Model Serving with PagedAttention, Proceedings of SOSP 2023, pp. 611-626, 2023.

[3] Qwen Team, Alibaba Cloud, Qwen3.6-35B-A3B Model Card, https://huggingface.co/Qwen/Qwen3.6-35B-A3B, 2026.

[4] Qwen Team, Alibaba Cloud, Qwen3.6-27B Model Card, https://huggingface.co/Qwen/Qwen3.6-27B, 2026.

[5] Benchmark Results for Qwen3.6-35B-A3B-FP8 on NVIDIA DGX Spark (GB10) Serving via vLLM, https://rikkarth.com/blog, 2026.