NOTE 005

# Four machines, one question

**Written** 3 November 2026.
**Dataset:** On-Prem Capacity Curve, measurement records of 29 October 2026, with Dataset Console v1. The Zenodo DOI and the GitHub page follow the release.
**Changes:** none.

This note is the story behind the dataset, from a question at a summer school in Seoul to the release I am preparing this week. It is the background to the project and what came after. The numbers live in the dataset and the console.

## Spring 2026

Around March and April, going into the summer break, I was building a retrieval pipeline (RAG) for the lab. The routing gave me trouble because the model kept producing errors, and for the first time I thought I would have to post-train and fine-tune a model of my own. Good material on how was hard to find online. Then a chance came up to attend a medical AI conference, the first I had been to.

## 31 July 2026

The KoSAIM 2026 Summer School ran for two days in Seoul [1], much of it on post-training and fine-tuning local models for health systems. On the first day a lecture covered local language models in clinics, and in the question period someone who ran a private clinic asked about running one in his. The answer from the front was that the limit is the size of the model. A server of about ₩100 million runs a model of about 30B without quantization, perhaps up to 70B, and nothing about that is officially known. I took the question home.

## August

A private clinic does not need a data center the way a university hospital does, and once its records go out through an API wrapper to a server somewhere else, they have left the clinic. As far as I can tell, many private clinics in Korea prefer it that way. What they did not have was the data to weigh a small local setup against its cost: electricity, accuracy, speed, which model to pick and how far to quantize it. A hospital's one server would give one point on that curve, and four machines of different sizes could draw it. Few engineers were on it, and I happened to have the hardware: an M4 Pro Mac mini, an M5 Pro MacBook Pro, an M4 MacBook Air, and a DGX Spark I had my eye on for post-training on CUDA, 24 to 128 GB of unified memory between them.

I wrote the design on 3 August, while I was still deciding whether to add a DGX: sweep model size, quantization and hardware, and record for each cell the accuracy, the steady decode speed and the power at the wall. In the first week of August I went back to school for the semester, picked up the DGX on the way, and by 7 August all four machines were on the bench. It came to two models, a 27B and a 9B, at twelve quantization levels each, on every machine.

The setup was not easy. The smart plugs I wanted first were out of stock or would not ship to the school. The lab's outlets turned out to be 120 V, so the meter had to be a US plug-through type. The one I found sat in customs for over a week, and the import fees came to more than the plugs and the shipping together. Meanwhile the models had to come down over a slow connection, one copy per machine and close to 300 GB for a full set, and the pulls took far longer than I had planned for.

## September

The four plugs finally arrived on 10 September. They turned out to talk only to their maker's cloud, so left alone the power readings would have gone out to the maker's server and come back. I put a small bridge on the DGX to read them inside the lab and cut the plugs off from the internet altogether. Cutting them off was not simple either. It meant writing rules into the lab's router myself, which I had never done, and on 14 September I rewrote them four times in one day. Each time I thought I had it, the plugs found another way out.

Before each machine ran there was a check script, and that alone took fifteen to twenty minutes, while the measurement itself took far longer. The DGX ran on the night of 16 September and the Mac mini the next day, seven and a half hours with nobody touching it. The MacBook Pro I ran overnight on 18 September. I would set it going, do other things, sleep, and check the results in the morning. While a run was going I kept my phone off the lab network too, in case it got in the way of the measurement.

Then the chargers. The first one on the MacBook Pro was a 240 W GaN charger whose label says 200 to 240 V, in a lab where the outlets are 120 V. A first-party 140 W adapter came on 30 September, and this time the Pro would only take 94 W from it. I went with 94 W. Whether the charger changed the numbers I checked later with a small experiment of its own, and the difference was under one percent. Then on 1 October the KoSAIM abstract was accepted, the Pro was needed for the poster, and the measurements stopped where they were.

## October

I ran the Pro's main measurement on the night of 3 October, and in the morning part of the plug's record was empty. There had been a power cut on 24 September, the plug had restarted and changed state, and nobody knew until that night. I set the night aside as it was and ran it again on the morning of 4 October. The Air finished that night, and by 5 October all four machines were done. It was probably the hardest setup I have run. Hardware, electricity, local models and quantization were all on one line, and if any one of them slipped the numbers could not be used.

The rest of October went to the analysis and to Seoul, and the Seoul part is in Note 002 [2]. On 29 October I put the dataset together from the raw records only, 566 files with a checksum each and every redaction written down [3]. The name was settled then too, On-Prem Capacity Curve.

## 3 November 2026

As of today the dataset and the console are done and the first analysis is finished. Now I am building the skeleton of the manuscript. I am proud to put this out, and I hope it is useful to anyone who wants to try the data in a project of their own. It goes up on the lab's channels and the website this week, once one piece of outside paperwork clears.

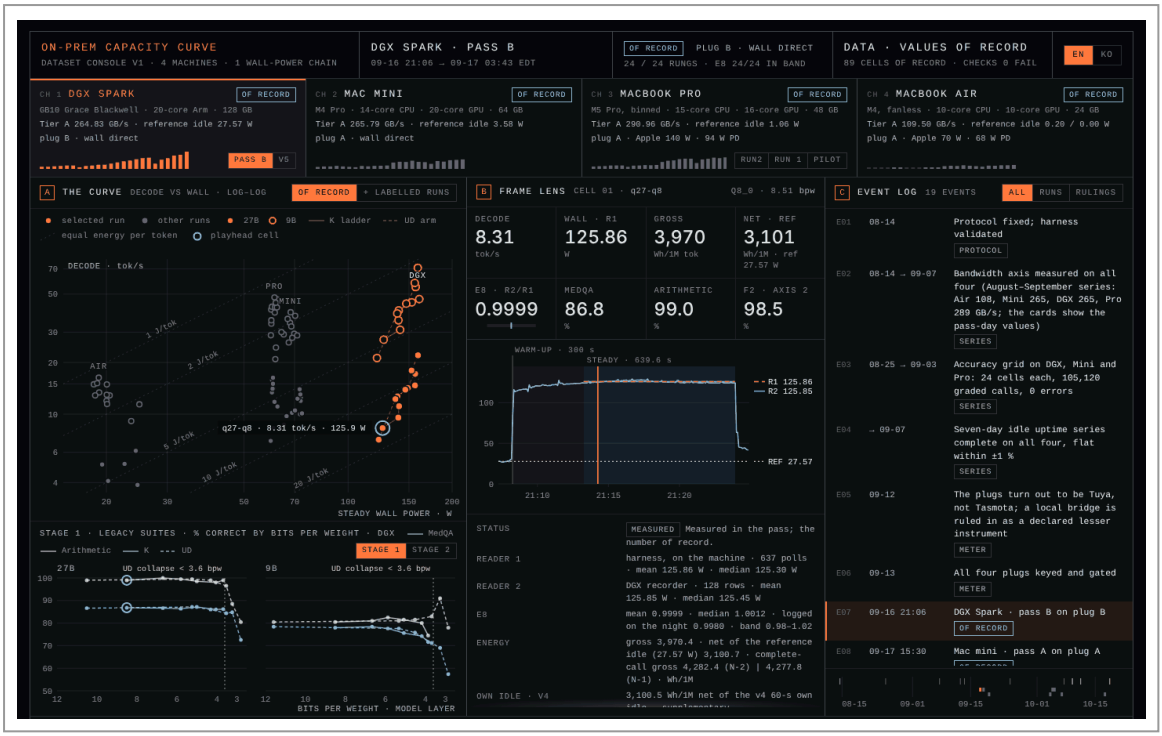



**Figure 1.** The top of Dataset Console v1, the public page that renders the values of record from the deposited files. The four machine channels, then (A) the curve of steady decode throughput against wall power with the stage-1 accuracy below it, (B) the frame lens on one cell and (C) the event log. Below it come the cell stack and the supply A/B, in English and Korean. The data of record are the deposited files.

## Pending

Expert scoring of the 18 meshes for vessel identity and spatial anatomy. Coding of the 27 image-stage responses for the full paper. The second study, pre-registered on OSF as a follow-up to the pilot; sample collection finished overnight between 31 October and 1 November, and coding and anatomical expert scoring remain. Clinician adjudication of a sealed sample of 40 ICU handoff notes. The harness manuscript. The DOI and the public push of the On-Prem Capacity Curve dataset, and then its manuscript. A submission for a project that is not yet public. The formal patent specification.

The list carries over from Note 004 with one item moved along [4]. The next public release is this dataset.

Everything public from this lab carries a DOI or a hash. This note carries dates.

## References

1. Korean Society of Artificial Intelligence in Medicine. KoSAIM 2026 Summer School, Seoul, 31 July to 1 August 2026. Event calendar, accessed 3 November 2026. https://www.kosaim.org/html/?pmode=event
2. Note 002, The Annual Conference of the Korean Society of Artificial Intelligence in Medicine (KoSAIM) 2026, debrief. Written 26 October 2026. https://www.thebioarchitectonic.com/notes/002-kosaim2026-debrief.html
3. Han JP. On-Prem Capacity Curve: measurement records of decode throughput and wall power on four unified-memory machines (dataset 2026-10-29). Zenodo. DOI to be assigned.
4. Note 004, Two timelines, mismatched. Written 1 November 2026. https://www.thebioarchitectonic.com/notes/004-two-timelines.html

---

**Disclosure.** The author is an inventor on a pending patent application related to part of this work.