DAIH
DEPLOYING AI IN HEALTHCARE

COLM

ACCURACY GETS A MODEL PUBLISHED. SPEED AND COST GET IT DEPLOYED.

Deploying Clinical Language and Vision-Language Models Where the Data Lives

Manpreet Singh Embedded LLM, Singapore manpreet.singh@embeddedllm.com

embedded

SCAN FOR THE PAPER
+ WRITE-UP · CONTACT

"Benchmark performance is a poor proxy for real-world readiness." DAIH 2026 CALL FOR PAPERS

So we stopped benchmarking the model and benchmarked the deployment.

WHY THESE MODELS ARE UNOPTIMISED

Three years of inference work went to LLM serving: **vLLM, SGLang, paged attention, continuous batching**. Almost none of it reaches a pathology encoder, a protein or genomic model, or a bedside time series monitor, because none of them generate tokens. They ship as research code, built to prove accuracy, and stay that way. The main lever a hospital has left is buying more hardware. **The fix belongs upstream, with the lab that ships the model.**

35×
FASTER AT THE BEDSIDE
805 → 23 ms

½
THE ENERGY PER SLIDE
20.2 → 10.0 Wh

24 GB
PEAK MEMORY FITS A MID-RANGE CARD

≤0.001
CHANGE IN AUC
4 tasks × 5 seeds

01 THE PROTOCOL
What the leaderboard leaves out
DAIH asks. We measured.

DAIH ASKS	WE MEASURED → RESULT
✓ EVERY LEADERBOARD · NECESSARY, NOT SUFFICIENT Is it accurate?	Accuracy / AUC **unchanged · |ΔAUC| ≤ 0.001**
03 WORKS Does it still work when it's fast?	Fidelity-gated speedup **35× bedside · 364 slides/shift**
02 EQUITABLE What does one case cost?	Energy per inference **20.2 → 10.0 Wh / slide**
02 EQUITABLE Can it run where the data is?	Deployment locality **24 GB peak · NVIDIA or AMD**
01 SAFE Does the evidence hold at our site?	Cross-vendor reproducibility **not always · §07**

First row is the one every leaderboard reports; the four below are ours. Peak memory fits a 24 GB card, so the fix isn't just buying more hardware.

02 WHERE THE TIME GOES
Why chatbot speedups never reach these models

PATHOLOGY VLM ENCODER tiles per slide	**12,400**
PROTEIN LM letters, on kernels built for 50K-token vocabularies	**20**
GENOMIC MODEL bases at once	**196K**
ICU MODEL of latency is cleaning vitals	**85%**
BINDING MODEL hidden CPU round-trip per sample, §07	**GPU→CPU→GPU**

None of these generate tokens, so vLLM-style serving never touches them.
Same code runs on a cloud GPU or on-prem.

03 HOW WE MEASURED IT
One file, two vendors, one gate

1 source file — OpenAI Triton → **2** GPU vendors — NVIDIA AMD → **0** per-vendor rewrites — compiled unmodified

THE FIDELITY GATE

✓ Output matches reference	reported
✗ Output differs	never reported

Speed is never bought with silent quality loss.
This belongs upstream: the lab that releases a model should ship it fast, the same way it ships it accurate.

04 PATHOLOGY · UNI ENCODER
A shift's worth of slides on one GPU

SLIDES PER 8-HOUR SHIFT

222 → **364**

REFERENCE
OURS · +64%

NVIDIA H100 shown. AMD MI300X: 224 → 348.

ENERGY PER SLIDE

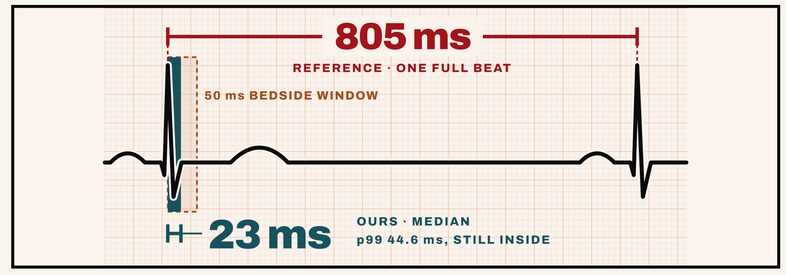
20.2 → **10.0 Wh**

REFERENCE
OURS · HALF

NVIDIA H100 shown. AMD MI300X: 20.4 → 10.6 Wh.
Same diagnosis: |ΔAUC| ≤ 0.001 across four tasks and five seeds.

05 ICU · PHYSIONET 2012
Inside the 50 ms bedside window

The old pipeline took one whole heartbeat.



One R–R interval at 75 bpm ≈ 800 ms; ours is faster than a single QRS complex (≈80–100 ms). Outputs match within 5×10^{-7}, model unchanged.

More than 85% of the delay was cleaning up irregularly sampled vital signs, not running the model.

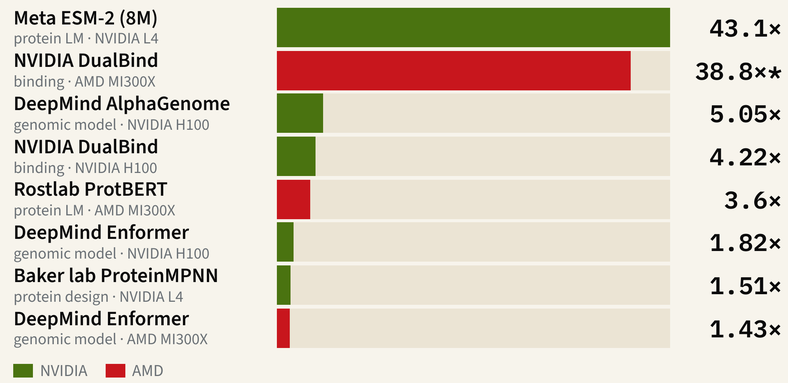
INTERPOLATION STEP
~~2.94 ms~~ **0.05 ms**

ENERGY PER UPDATE
~~52.0 mJ~~ **5.6 mJ**

NVIDIA shown. AMD energy per update: 60.1 → 7.2 mJ.

06 IT GENERALISES
Same source, six models, three GPUs

Models from Meta, DeepMind, NVIDIA and the Baker lab. All accurate. All shipped as research code.



vs reference code as shipped · all pass the fidelity gate (exact or cos ≥ 0.99999) · * inflated by a hidden CPU round-trip, §07

SEPT 2026. Anthropic reported ~4× (≈1.6 to 2× exact-output) across 30+ open bio models, all on H100. Same thesis; we add AMD and clinical deployment metrics.

4.22× ON NVIDIA H100 ≠ SAME CODE 9× APART **38.8×** ON AMD MI300X

07 NEGATIVE RESULT · ONE MODEL, IDENTICAL SOURCE, NOT ONE LINE CHANGED BETWEEN VENDORS
Ask what it does on the card you will actually run it on.

Deployment evidence obtained at one site does not automatically transfer to another.

Cause: the reference code copied results to the CPU for a NumPy score every sample. NVIDIA's runtime hid the cost; AMD exposed it. Corrected, the two vendors land within 1.4× of each other.

METHODS UNI pathology encoder on a single on-site GPU, representative slide ≈ 12,400 tiles; board power sampled at 100 ms across the full encode, idle not subtracted, host CPU and cooling excluded. ICU: PhysioNet 2012, batch 32, model held fixed, retrospective cohort. Speedups are against the reference PyTorch implementation and measure headroom in unoptimised code. Not a clinical validation study.