ASNR · 2026

Asian–Oceanian Session

Wednesday, May 20, 2026 · 10:25–11:25

Opportunities and Controversies

in the Clinical Application of

AI in Neuroimaging

Rintaro Ito, MD, PhD

Department of Radiology

Innovative Biomedical Visualization (iBMV) · Nagoya University

II.

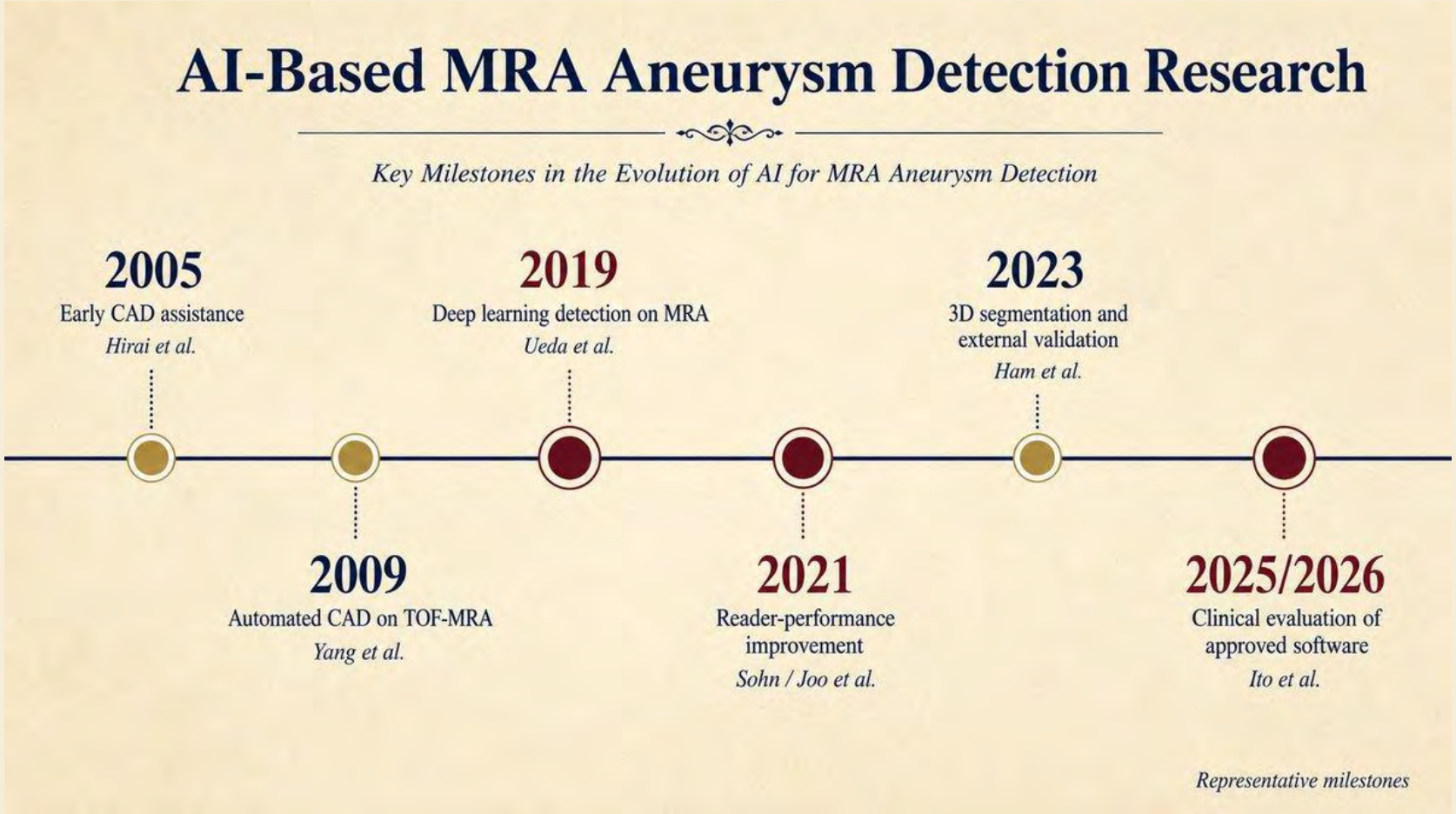
Background

Clinical AI in neuroimaging — opportunity and

In the last ten years, deep learning has moved from research papers to approved medical devices. Brain imaging has led the way.

Many regulators have approved brain-AI products since 2018. Japan — with the highest MRI density in the OECD and a national imaging database (J-MID, 534M images) — is a useful test case.

One key question remains. Do these tools really work in everyday clinical use, for patients and scanners that differ from training data?



References: Fujita, Ito, Naganawa et al. Jpn J Radiol 2024

IV.

Results — primary endpoints

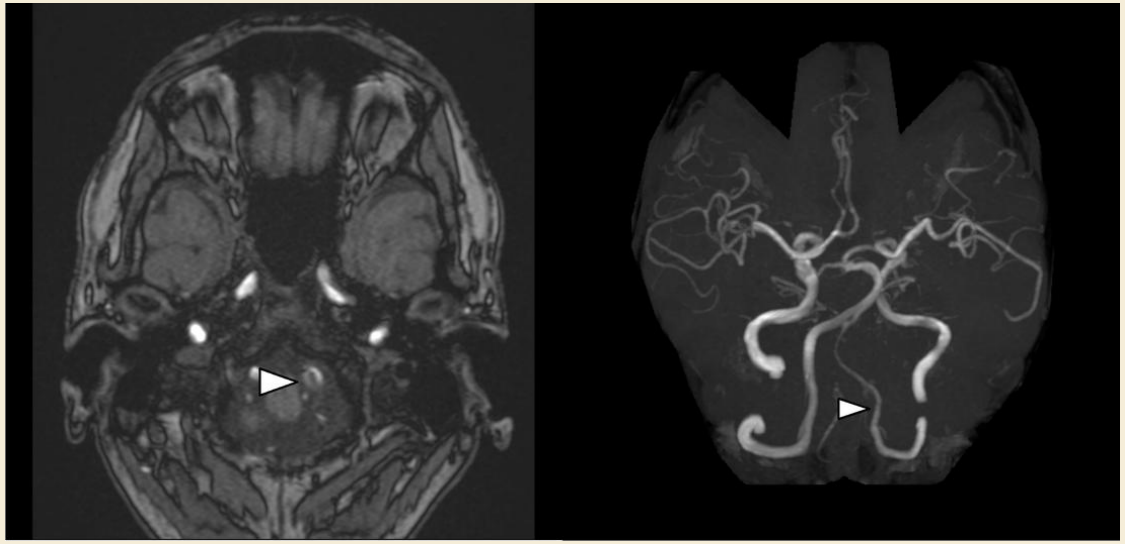
442 cases · 94 reference-standard aneurysms

✦

Table 1. Per-aneurysm performance.

Metric	Value
Sensitivity	**77.7 %**
Positive predictive value (PPV)	**12.3 %**
F1 score	**0.212**
False positives / case	**1.18**

Fig. 4 · One real aneurysm (top) and one false alarm (bottom).



520 false positives vs. 73 true positives → about 7 false alarms for every real aneurysm.

Reference: Ito R, et al. Magn Reson Med Sci 2025. DOI: 10.2463/mrms.mp.2024-0183

IV.

Results — subgroup analyses

Sensitivity by aneurysm size and field strength

Figure 5a · Sensitivity by aneurysm size

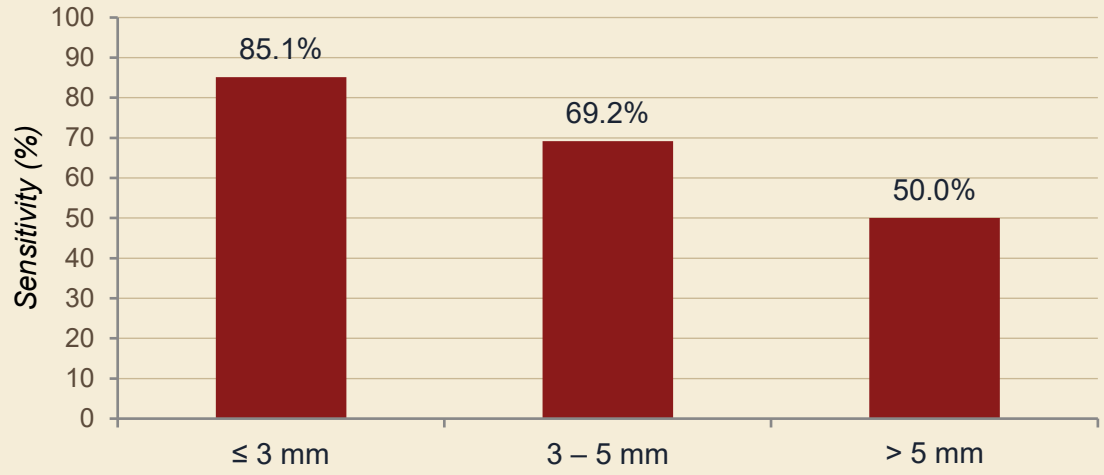
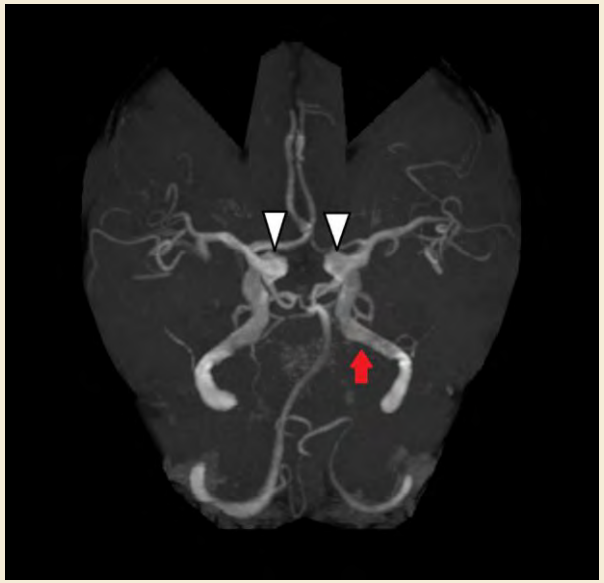


Fig. 5b · A > 5 mm aneurysm that the AI missed



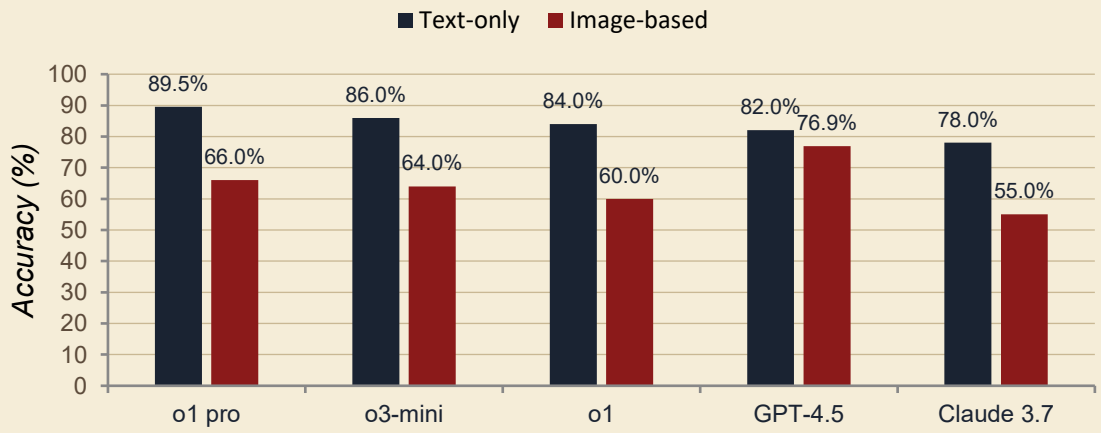
The AI worked well for very small aneurysms (≤ 3 mm: 85% sensitivity), but it missed half of the larger ones (> 5 mm: 50%). These large aneurysms are exactly the ones that may need treatment. The false-positive rate also changed significantly with magnetic-field strength.

IV.

Generative AI on board examinations

How well do large language models do on Japanese board exams?

Figure 7. VLMs on JNMBE (n = 180)



Performance on Japanese boards (2024 – 2026)

- **Nuclear-medicine board (JNMBE) — text 89.5% vs image 66.0%**
- **JRBE 2024 text —** DeepSeek-R1 / o1 → 87.6% (human residents ≈ 67.6%)
- **JRBE 2024 image —** Gemini 2.5 Pro → 76% (human average 72.9%)
- **JDRBE —** AI reaches board level by reading the images directly (Miki 2026)

Figure 8. Sample board question (illustrative)

Q. A 65-year-old patient with a new headache. The image above shows …

(A) Acute infarct (B) Aneurysm (C) Tumor (D) Vasospasm

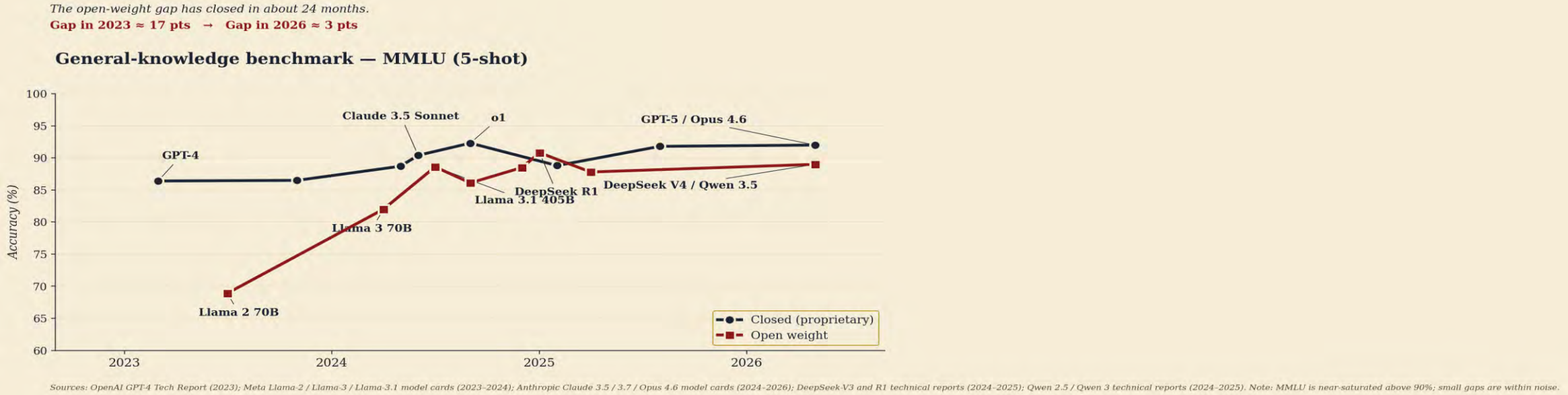
Without the image, AI scores about 89.5%. With the image, only 66.0%.

IV.

AI benchmarks — general knowledge

How well do AI models score on a 57-subject general-knowledge exam (MMLU)?

Closed–open gap: ≈ 17 points in 2023 → ≈ 3 points in May 2026



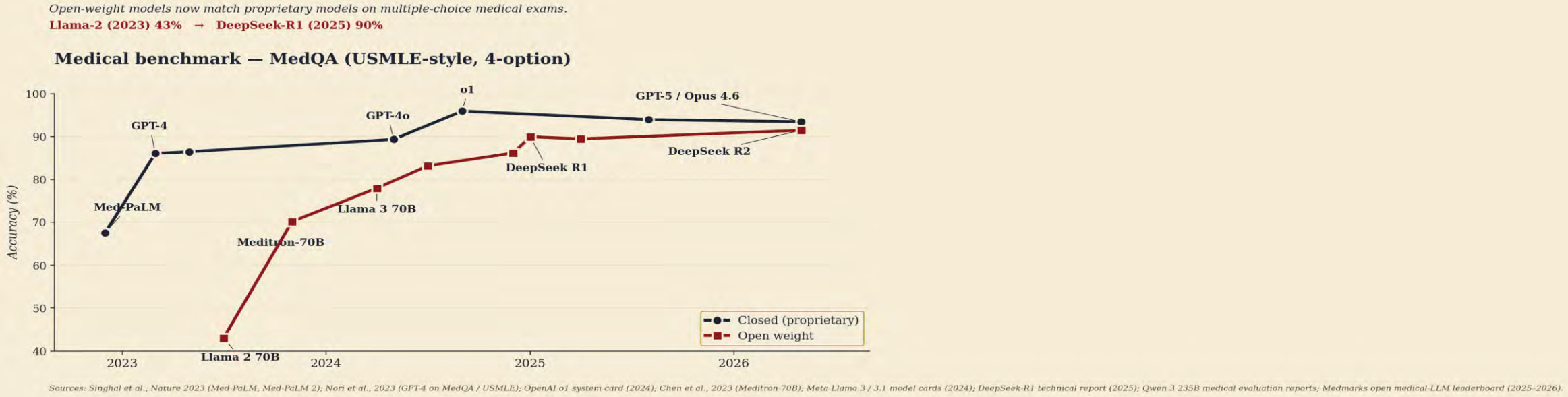
Sources: OpenAI GPT-4 Tech Report 2023; Meta Llama-2 / Llama-3 / Llama-3.1 model cards; Anthropic Claude 3.5 / 3.7 / Opus 4.6 model cards; DeepSeek-V3 and R1 technical reports; Qwen 2.5 / Qwen 3 technical reports. MMLU is near-saturated above ~90%; small differences are within measurement noise.

IV. —

AI benchmarks — medical knowledge

How well do AI models score on a USMLE-style medical exam (MedQA)?

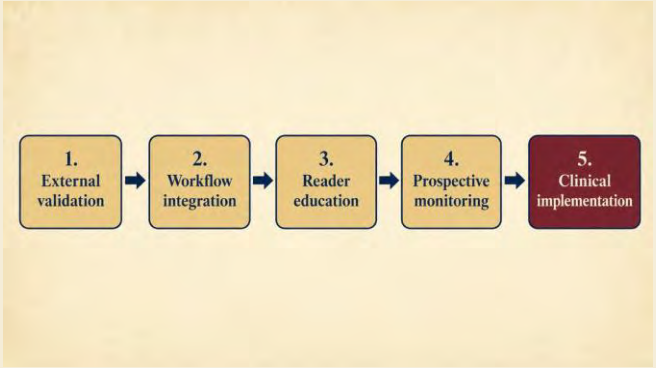
Open-weight models (Llama, DeepSeek, Qwen) now sit within ~3 points of the best proprietary models.



Sources: Singhal et al., Nature 2023 (Med-PaLM, Med-PaLM 2); Nori et al., 2023 (GPT-4 on MedQA / USMLE); OpenAI o1 system card 2024; Chen et al., 2023 (Meditron-70B); Meta Llama 3 / 3.1 model cards; DeepSeek-R1 technical report 2025; Qwen 3 235B medical evaluation; Medmarks open medical-LLM leaderboard 2025–2026.

IV.

Conclusions and future directions



1. **[Controversy] An approved AI is not always a good AI.**
   PPV only 12%. Half of the large aneurysms were missed. Same pattern at many Japanese centers.
2. **[Controversy] AI does not always transfer to new patients and new scanners.**
   Different genes. Different scanners. Different training data.
3. **[Opportunity] AI is improving fast — and open models are catching up.**
   On medical exams, the best open-weight models are now within ~3 points of the best closed models.
4. **[Synthesis] The Japanese lesson is a global lesson.**
   Every country should keep checking its AI — after approval, not only before.