JMed48k: A Multi-Profession Japanese Medical Licensing Benchmark for Vision-Language Model Evaluation

Yue Xun[1], Junyu Liu[2], Qian Niu[3]*, Xinyi Wang[1], Zheng Yuan[1], Zirui Li[4], Zequn Zhang[5], Bowen Zhao[6], Shujun Wang[1]
Irene Li[3], Kan Hatakeyama-Sato[3], Yusuke Iwasawa[3], Yutaka Matsuo[3]

[1] The Hong Kong Polytechnic University [2] Kyoto University [3] The University of Tokyo [4] Hohai University [5] University of Science and Technology of China [6] University of Toronto
* qian.niu@weblab.t.u-tokyo.ac.jp

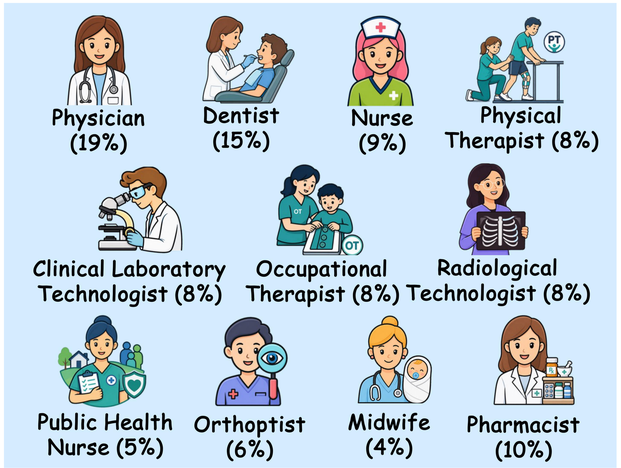
Dataset

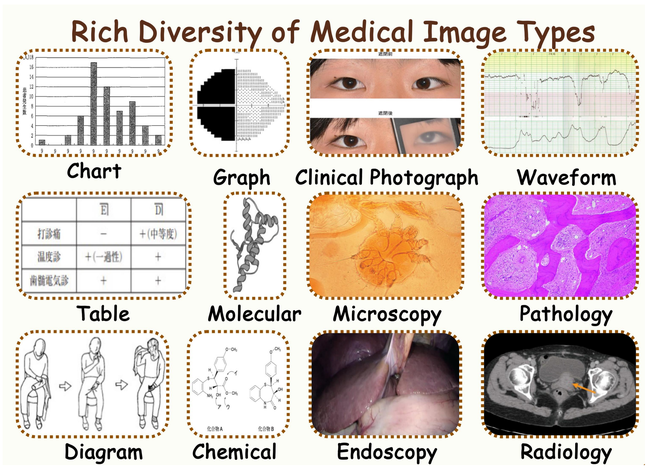
松尾・岩澤研究室 Matsuo-Iwasawa Lab UTokyo

Japanese Licensing Exams

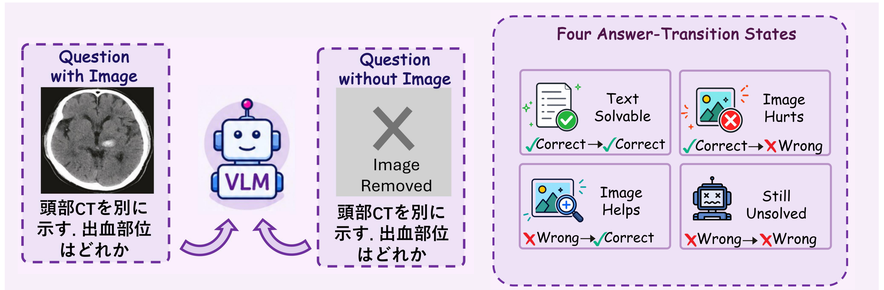
48,862
Official Exam Questions
11 exams, 2005–2025

20,142
Images
8 types, e.g. Chart/Graph/Table





Paired Image-Removal Audit



Medical-Specific Models Gain Little from Images

Medical-specific models: **19.8%** on image-dependent items (random: **20.4%**)

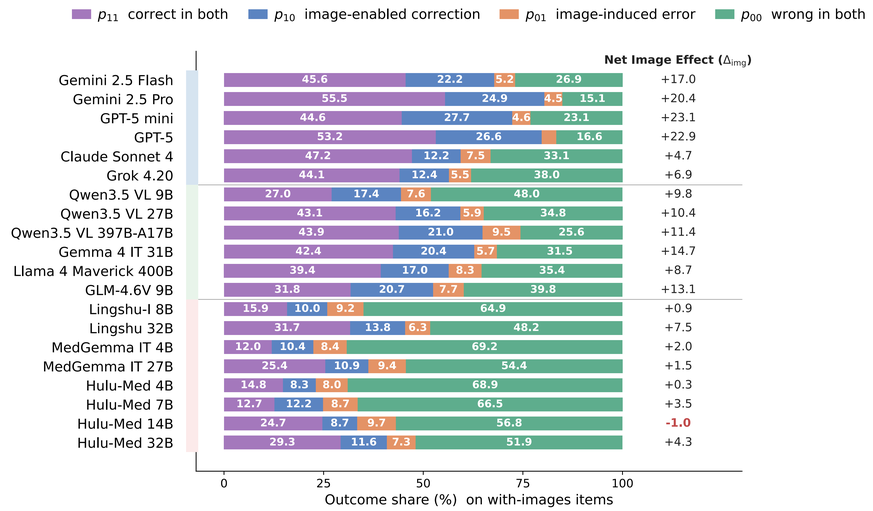


Image Questions: 16–23 pp Lower

21 models, last 5 years

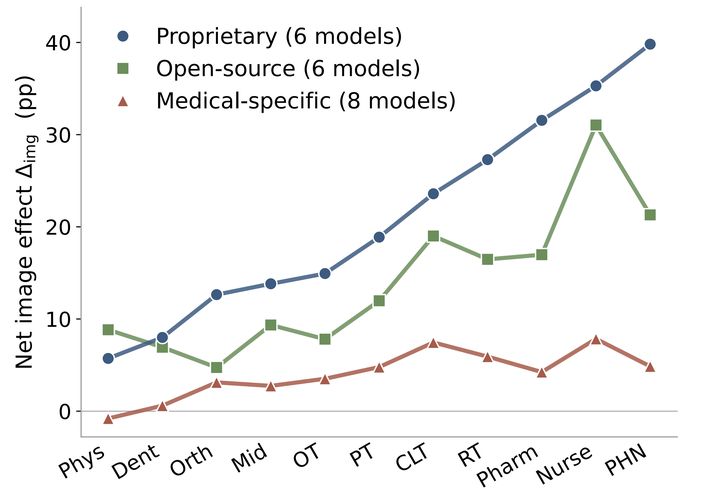
Selected Model	Text-Only (n = 9,905)	With Images (n = 2,579)
Gemini 2.5 Pro	**95.5%**	**79.7%**
GPT-5	95.1%	78.8%
Qwen3.5 VL 397B-A17B	86.5%	63.6%
Lingshu 32B	65.0%	45.0%

Profession-Level Analysis

Proprietary models: image gains vary sevenfold

+5.7 pp
Physician

+39.8 pp
Public Health Nurse

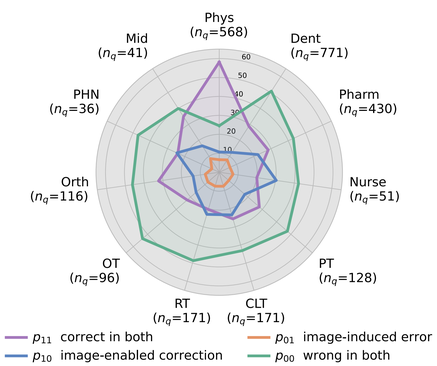


Phys = Physician, Dent = Dentist, Orth = Orthoptist, Mid = Midwife, Pharm = Pharmacist
OT = Occupational Therapist, PT = Physical Therapist, PHN = Public Health Nurse
CLT = Clinical Laboratory Technologist, RT = Radiological Technologist

Answer Transitions by Profession

Pooled Across
20 Multimodal Models

Physician and dentist items are often answerable from the stem alone



Exam images cropped from official licensing exam booklets published by Japan's Ministry of Health, Labour and Welfare. $\Delta img = p_{10} - p_{01}$; text-only and with-image accuracies use different question sets. Paper: arXiv:2605.22080