BRIDGE: Benchmarking Large Language Models for Understanding Global Real-world Clinical Practice Text

Jiageng Wu*,[1,2], Bowen Gu*,[1,2], Ren Zhou[3], Kevin Xie[4], Doug Snyder[2,5], Yixing Jiang[6], Valentina Carducci[5], Richard Wyss[1,2], Rishi J Desai[1,2], Emily Alsentzer[6], Leo Anthony Celi[2,4,7], Adam Rodman[7], Sebastian Schneeweiss[1,2], Jonathan H. Chen[6], Santiago Romero-Brufau[2,5], Kueiyu Joshua Lin#,[1,2], and Jie Yang#,[1,2]

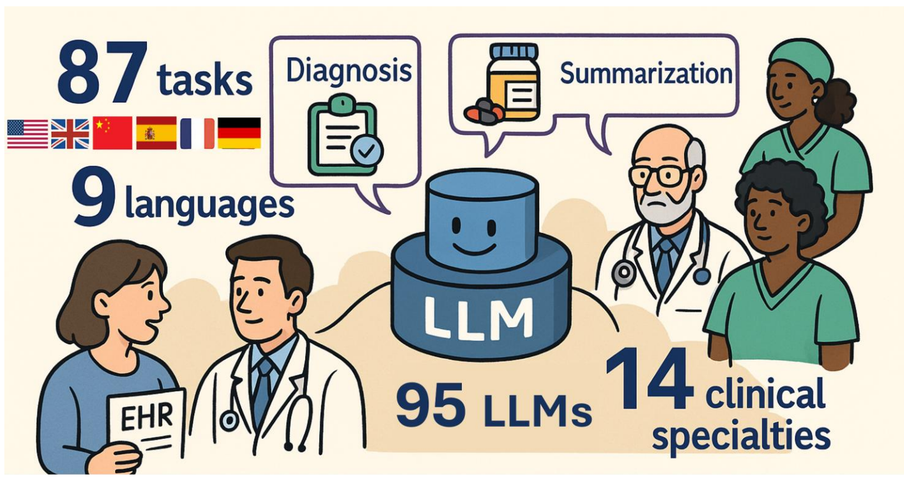
Brigham and Women's Hospital, Founding Member, Mass General Brigham; Harvard Medical School; University of Illinois Urbana-Champaign; Massachusetts Institute of Technology; Mayo Clinic; Beth Israel Lahey Health, Beth Israel Deaconess Medical Center; Stanford

Paper

Leaderboard

Background

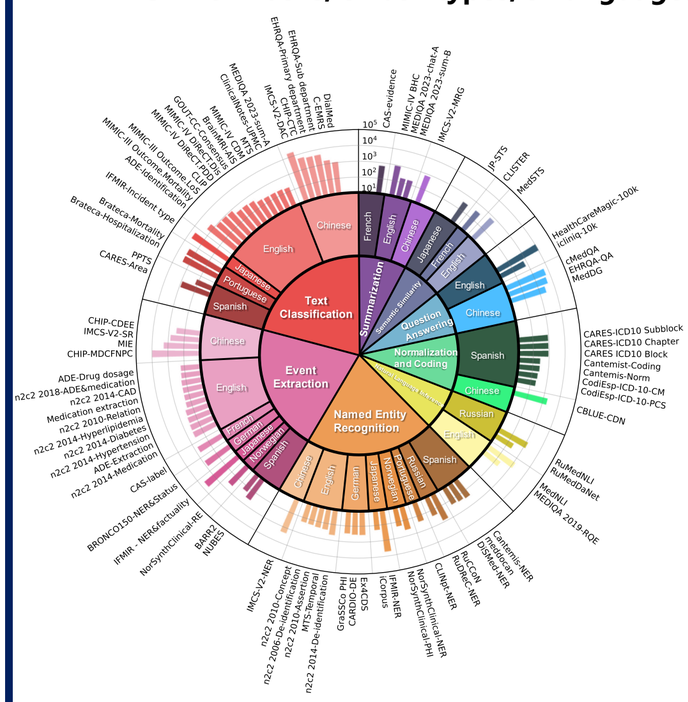
- Existing benchmarks rely on **medical exams** or **biomedical literature** and are hard to reflect LLM capabilities in complex **real-world clinical practice**.
- Isolated evaluations are not comprehensive and are difficult to **generalize**.
- Lack of **multilingual** benchmarks impedes the **global applicability** of LLMs
- Lack of **dynamic** benchmarking for LLMs on diverse real-world clinical tasks as new models continue to emerge.



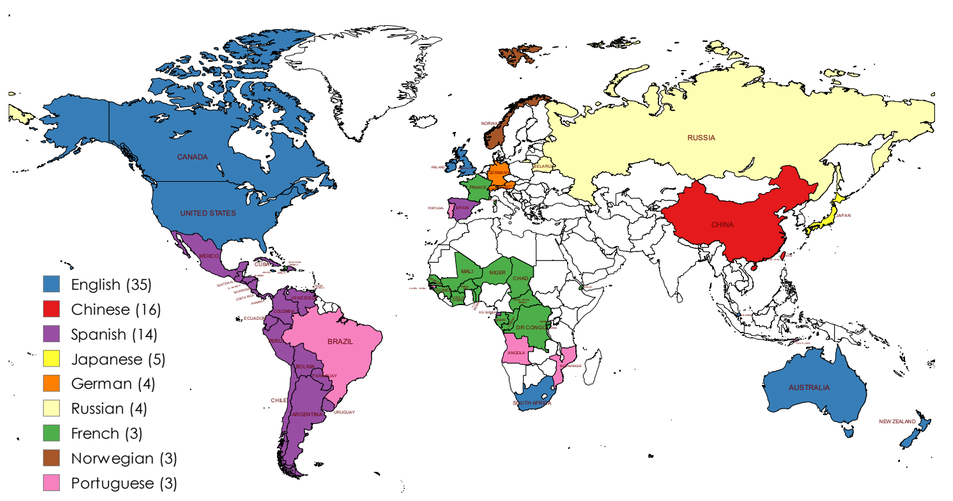
Methods

- **Publicly accessible clinical text datasets**: Diverse datasets collected from literature database (e.g., PubMed, ACL), community challenges (e.g., i2b2), web (Huggingface)
- **Benchmark**: **87 clinical text tasks** were standardized and reformatted into a unified task *instruction – input – output* format, with a systematic task taxonomy.
- **Evaluation**: **95 LLMs:** proprietary (e.g., GPT-4o, Gemini), open-source (e.g., DeepSeek-R1, Qwen3), medical models (e.g., MedGemma, Me-LLaMA), under 3 inference strategies: zero-shot, chain-of-thought, and few-shot.
- **Performance and Leaderboard:** Detailed comparison and analysis, e.g., Open-source vs. Proprietary, across different prompting / languages / task types ….
- **24,795 experiments, 3.4 Billion Outputs**

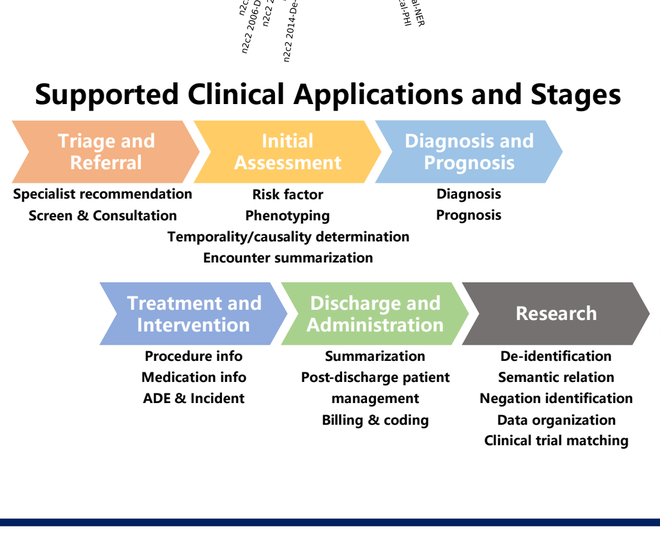
BRIDGE – 87 Tasks, 8 Task types, 9 Languages



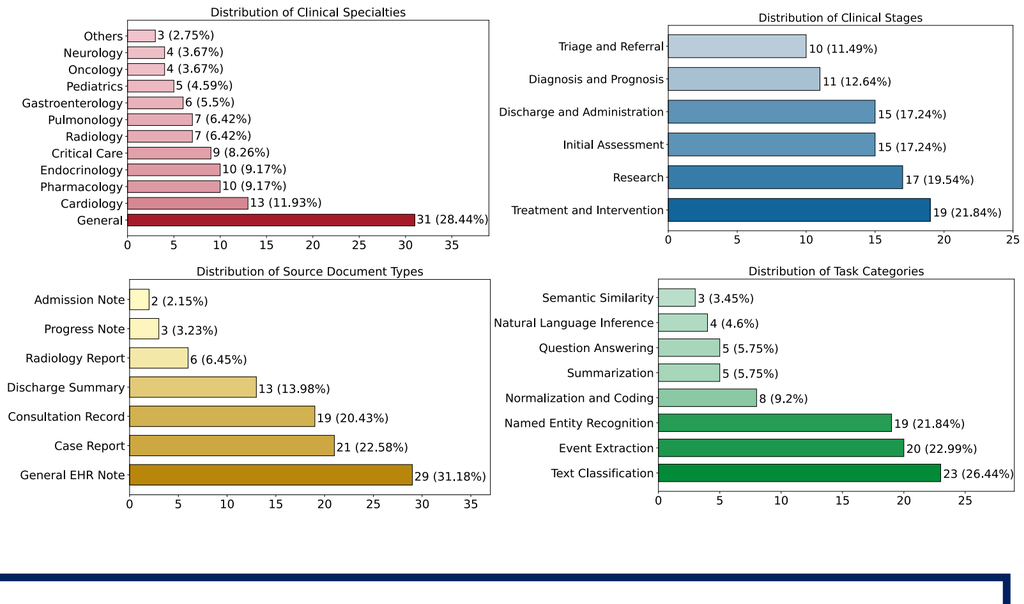
Geographic coverage of BRIDGE-supported languages



Supported Clinical Applications and Stages

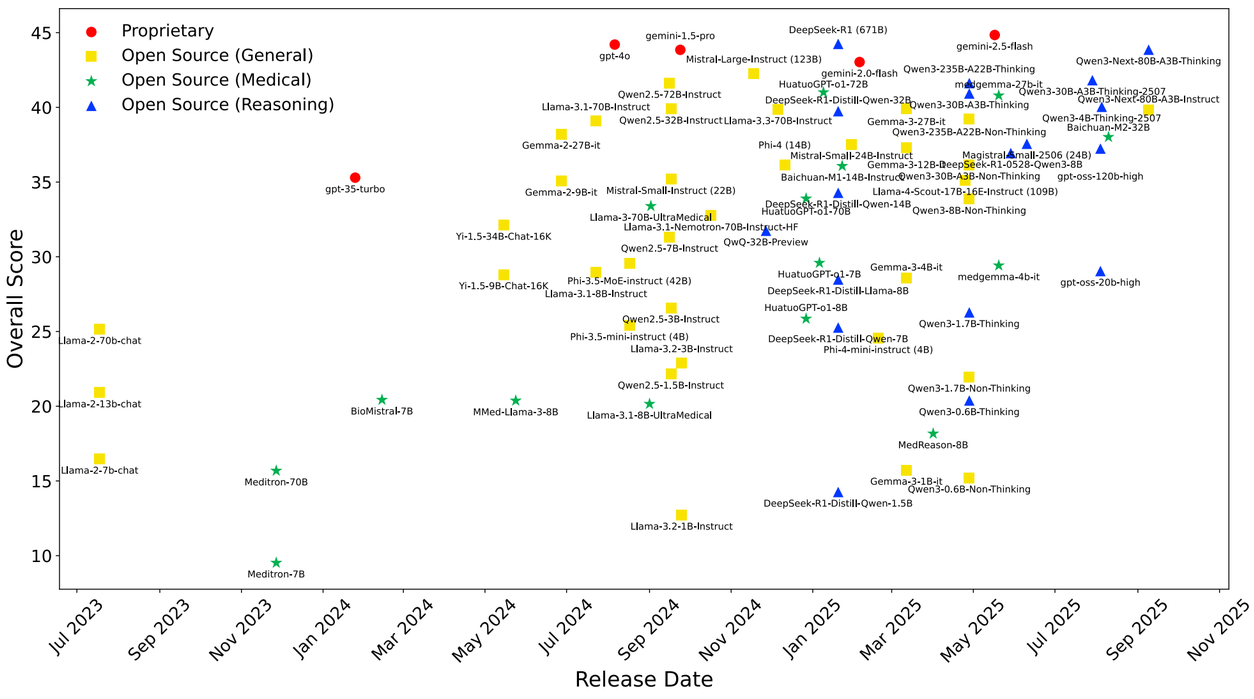


Distribution of Clinical Specialty / Stage / Document type / Task Type

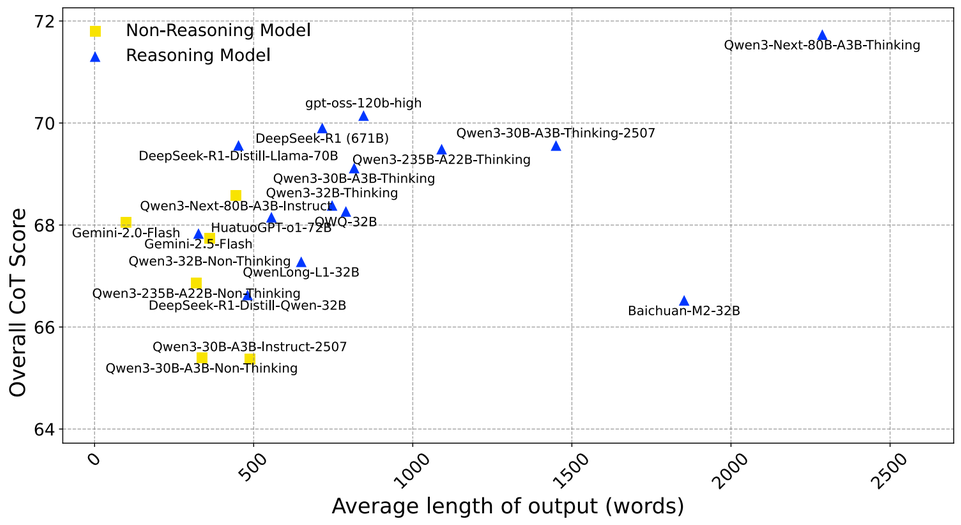


Result

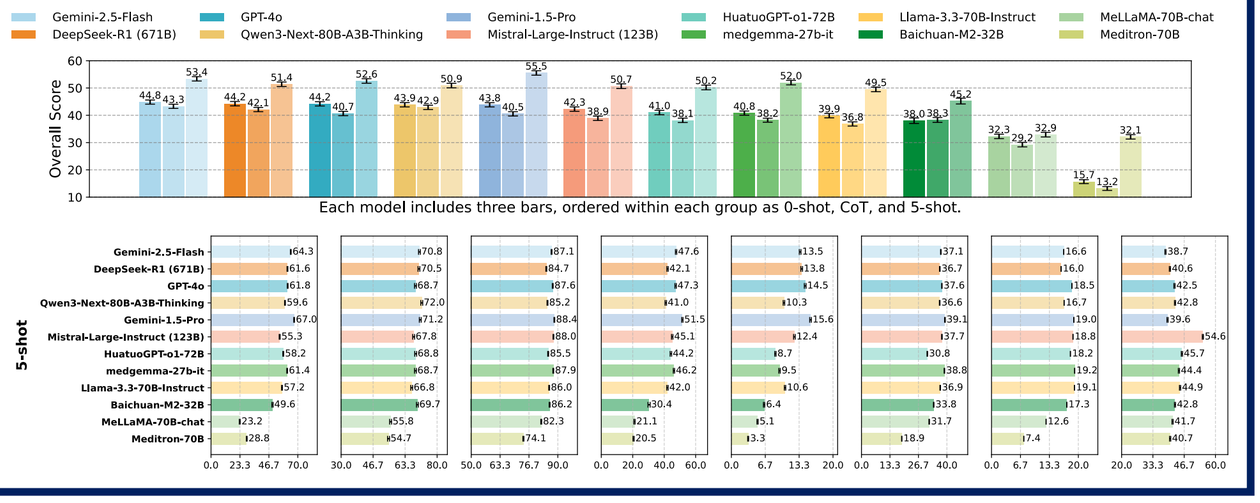
- ✓ **Model performance increases with time**
- ✓ **Open-source LLMs outperform proprietary models**



- ✓ **Performance vs reasoning length (reasoning LLMs)**



- ✓ **Performance variations across task types**
- ✓ **Few-shot consistently boosts performance**



- ✓ **Scaling law still works within clinical text tasks**