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

Jie Yang, PhD
Assistant Professor

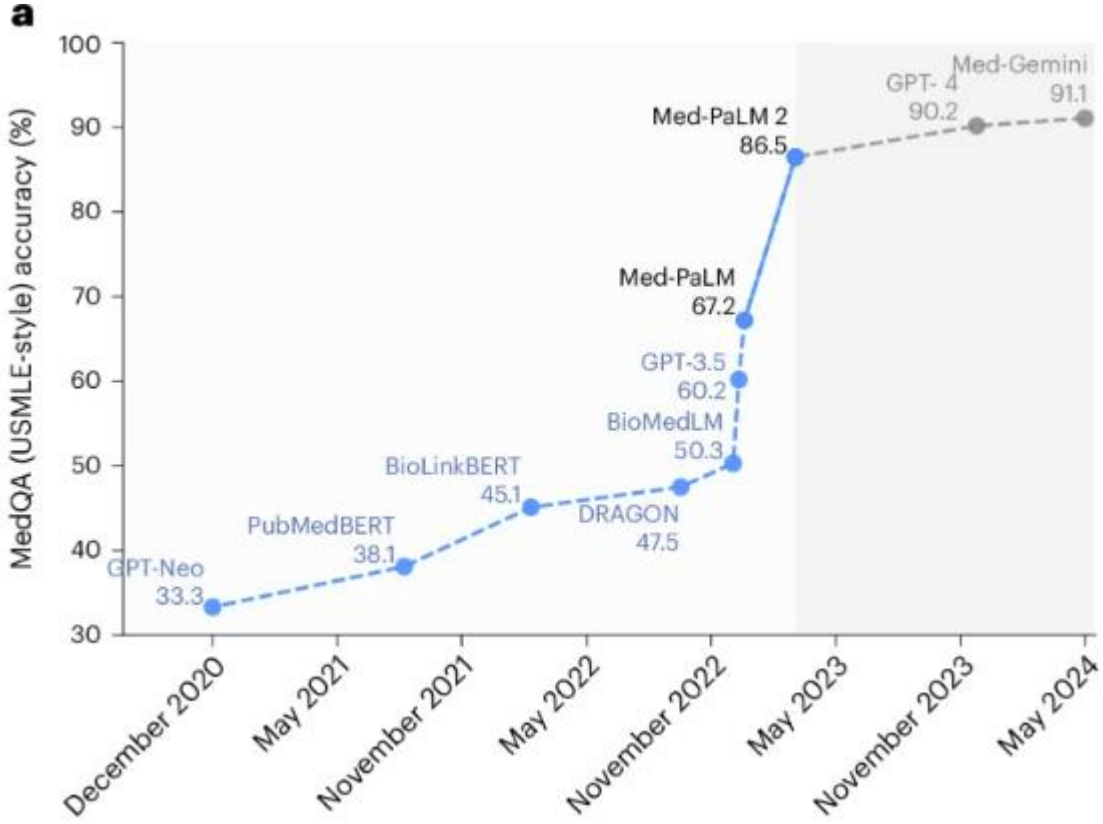
Jiageng Wu, PhD
Postdoc Fellow

Brigham and Women's Hospital, Harvard Medical School

@MITHIC-2025
Aug 2025

- **Background**
- **BRIDGE Overview**
- **Dataset Introduction**
- **Potential Topics**

➢ Large language models have demonstrated strong capabilities in healthcare.



GPT-4 and Med-Gemini reach > 90 in USMLE

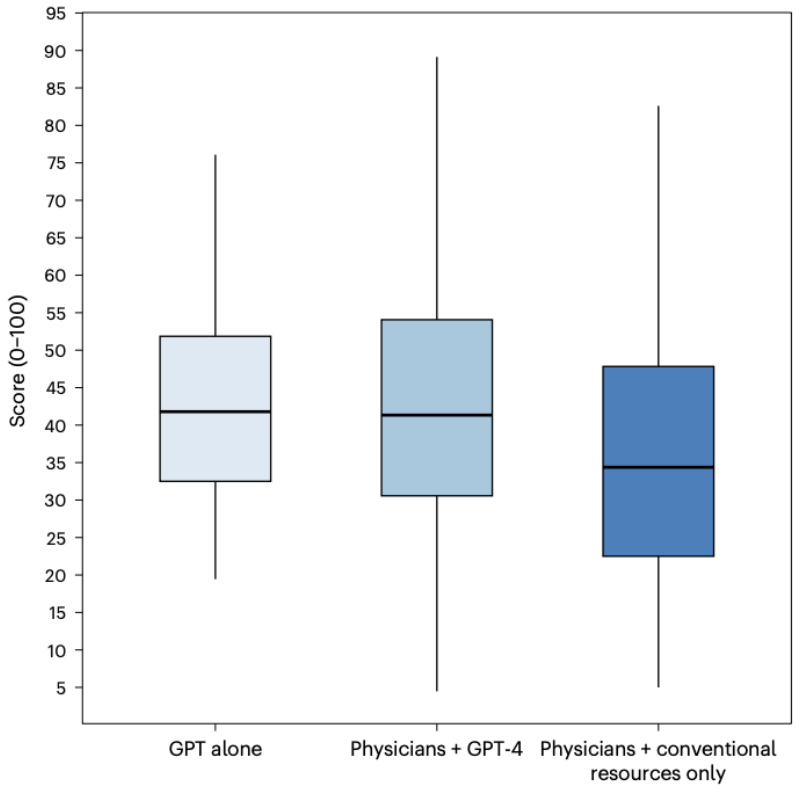


Fig. 3 | Comparison of the primary outcome according to GPT alone versus physician with GPT-4 and with conventional resources only (total score standardized to 0–100). The GPT-alone arm represents the model being

GPT-4 > GPT-4+Physicians > Physicians

1. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H. and Neal, D., 2025. Toward expert-level medical question answering with large language models. Nature Medicine, pp.1-8.
2. Goh, E., Gallo, R.J., Strong, E., Weng, Y., Kerman, H., Freed, J.A., Cool, J.A., Kanjee, Z., Lane, K.P., Parsons, A.S. and Ahuja, N., 2025. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nature Medicine, pp.1-6.

Table 1. ICD-9-CM, ICD-10-CM, and CPT Code Generation Overall Performance Metrics, Full Code Set.*

Coding System	Metric	GPT-3.5 Turbo (March)†	GPT-3.5 Turbo (June)†	GPT-3.5 Turbo (Nov)†	GPT-4 (March)†	GPT-4 (June)†	GPT-4 (Nov)†	Gemini Pro†	Llama2-70b Chat†
ICD-9-CM (n=7697)	Exact match, % (95% CI)	26.6% (25.6%–27.6%)	26.7% (25.7%–27.7%)	28.9% (27.9%–29.9%)	42.3% (41.2%–43.4%)	44.1% (43.0%–45.2%)	45.9% (44.8%–47.0%)	10.7% (10.0%–11.4%)	1.2% (1.0%–1.5%)
	cui2vec cosine similarity, mean (95% CI)	0.747 (0.741–0.753)	0.750 (0.744–0.756)	0.765 (0.760–0.771)	0.833 (0.828–0.838)	0.837 (0.832–0.842)	0.843 (0.838–0.848)	0.641 (0.635–0.648)	0.418 (0.409–0.428)
	METEOR score, mean (95% CI)	0.415 (0.406–0.424)	0.414 (0.405–0.422)	0.437 (0.428–0.445)	0.564 (0.555–0.573)	0.579 (0.569–0.588)	0.593 (0.585–0.602)	0.255 (0.247–0.262)	0.100 (0.094–0.106)
	BERTScore, mean (95% CI)	0.857 (0.855–0.860)	0.856 (0.854–0.859)	0.863 (0.861–0.866)	0.899 (0.896–0.901)	0.903 (0.901–0.906)	0.907 (0.904–0.909)	0.812 (0.809–0.814)	0.749 (0.747–0.751)
ICD-10-CM (n=15,950)	Exact match, % (95% CI)	17.1% (16.5%–17.7%)	17.8% (17.2%–18.4%)	18.2% (17.6%–18.8%)	27.5% (26.8%–28.1%)	28.4% (27.7%–29.1%)	33.9% (33.2%–34.6%)	4.8% (4.5%–5.1%)	1.5% (1.4%–1.7%)
	cui2vec cosine similarity, mean (95% CI)	0.571 (0.564–0.577)	0.576 (0.570–0.583)	0.566 (0.559–0.572)	0.669 (0.663–0.675)	0.680 (0.673–0.685)	0.733 (0.728–0.739)	0.414 (0.406–0.421)	0.287 (0.280–0.294)
	METEOR score, mean (95% CI)	0.399 (0.393–0.405)	0.405 (0.399–0.410)	0.400 (0.394–0.406)	0.510 (0.504–0.516)	0.522 (0.516–0.528)	0.581 (0.575–0.587)	0.250 (0.245–0.254)	0.129 (0.125–0.132)
	BERTScore, mean (95% CI)	0.866 (0.864–0.868)	0.870 (0.868–0.871)	0.866 (0.864–0.868)	0.899 (0.897–0.900)	0.902 (0.901–0.904)	0.918 (0.917–0.920)	0.824 (0.822–0.826)	0.774 (0.773–0.776)
CPT (n=3673)	Exact match, % (95% CI)	28.4% (27.0%–29.9%)	26.2% (24.7%–27.6%)	31.9% (30.4%–33.4%)	44.0% (42.4%–45.6%)	42.6% (41.0%–44.2%)	49.8% (48.2%–51.5%)	11.4% (10.3%–12.4%)	2.6% (2.1%–3.1%)
	METEOR score, mean (95% CI)	0.461 (0.448–0.474)	0.433 (0.421–0.446)	0.495 (0.482–0.507)	0.596 (0.583–0.609)	0.586 (0.573–0.599)	0.655 (0.642–0.667)	0.295 (0.284–0.306)	0.182 (0.172–0.192)
	BERTScore, mean (95% CI)	0.868 (0.864–0.871)	0.859 (0.855–0.863)	0.878 (0.874–0.882)	0.904 (0.901–0.908)	0.901 (0.897–0.904)	0.921 (0.918–0.925)	0.816 (0.813–0.820)	0.770 (0.766–0.773)

* CI indicates confidence interval; CPT, Current Procedural Terminology; ICD-9-CM, International Classification of Diseases, 9th edition, Clinical Modification; ICD-10-CM, International Classification of Diseases, 10th edition, Clinical Modification; and Nov, November.

† The application programming interface was accessed between December 26 and 27, 2023.

GPT-4 and Gemini models are pool medical coders.

Soroush, A., Glicksberg, B.S., Zimlichman, E., Barash, Y., Freeman, R., Charney, A.W., Nadkarni, G.N. and Klang, E., 2024. Large language models are poor medical coders—benchmarking of medical code querying. NEJM AI, 1(5), p.AIdbp2300040.

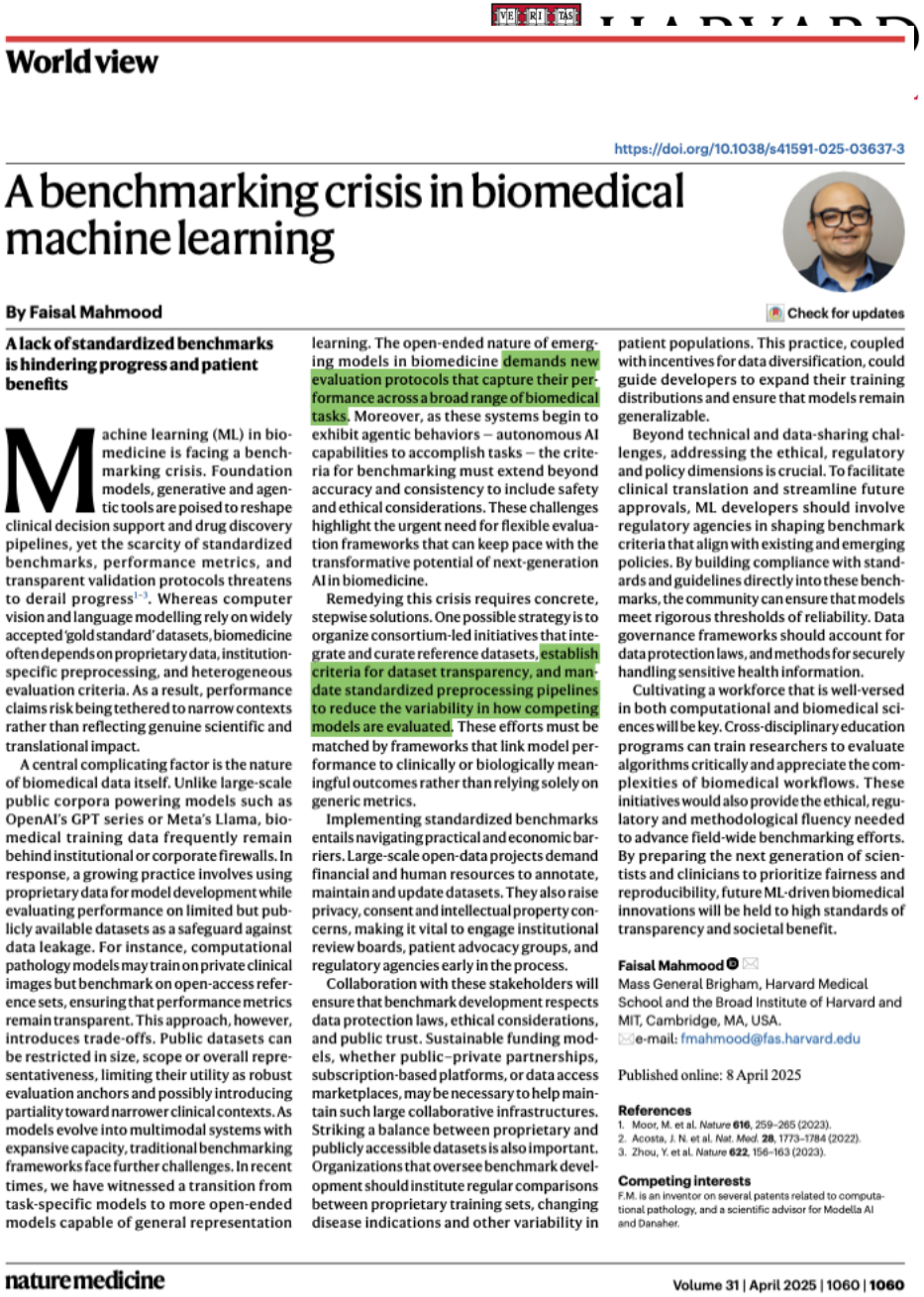
Table. Representative Cases and Diagnostic Outcome for a Large Language Model (LLM)

Representative case	LLM diagnosis	Physician diagnosis	Artificial intelligence diagnosis outcome	Outcome frequency (N = 100)
15-y-old girl with unexplained intracranial hypertension	Adrenal insufficiency (Addison disease)	Primary adrenal insufficient (Addison disease)	Correct	17
Rash and arthralgias in a teenager with autism	Immune thrombocytopenic purpura	Scurvy	Incorrect	72
Draining papule on the lateral neck of an infant	Branchial cleft cyst	Branchio-oto-renal syndrome	Did not fully capture diagnosis	11

GPT-4 has an error rate of 83% for pediatric case studies.

Barile, J., Margolis, A., Cason, G., Kim, R., Kalash, S., Tchaconas, A. and Milanaik, R., 2024. Diagnostic accuracy of a large language model in pediatric case studies. JAMA pediatrics, 178(3), pp.313-315.

- When and which LLM performs best on my medical applications, and how good or bad the performance is?
- Existing studies have shown inconsistent LLM performance across different medical applications
- LLMs evolve rapidly, with new models released every week
- It is important to build benchmarks to assess advanced LLMs in medicine and track newly released models.



World view

https://doi.org/10.1038/s41591-025-03637-3

A benchmarking crisis in biomedical machine learning

By Faisal Mahmood

Check for updates

A lack of standardized benchmarks is hindering progress and patient benefits

Machine learning (ML) in biomedicine is facing a benchmarking crisis. Foundation models, generative and agentic tools are poised to reshape clinical decision support and drug discovery pipelines, yet the scarcity of standardized benchmarks, performance metrics, and transparent validation protocols threatens to derail progress[1–3]. Whereas computer vision and language modelling rely on widely accepted 'gold standard' datasets, biomedicine often depends on proprietary data, institution-specific preprocessing, and heterogeneous evaluation criteria. As a result, performance claims risk being tethered to narrow contexts rather than reflecting genuine scientific and translational impact.

A central complicating factor is the nature of biomedical data itself. Unlike large-scale public corpora powering models such as OpenAI's GPT series or Meta's Llama, biomedical training data frequently remain behind institutional or corporate firewalls. In response, a growing practice involves using proprietary data for model development while evaluating performance on limited but publicly available datasets as a safeguard against data leakage. For instance, computational pathology models may train on private clinical images but benchmark on open-access reference sets, ensuring that performance metrics remain transparent. This approach, however, introduces trade-offs. Public datasets can be restricted in size, scope or overall representativeness, limiting their utility as robust evaluation anchors and possibly introducing partiality toward narrower clinical contexts. As models evolve into multimodal systems with expansive capacity, traditional benchmarking frameworks face further challenges. In recent times, we have witnessed a transition from task-specific models to more open-ended models capable of general representation learning. The open-ended nature of emerging models in biomedicine demands new evaluation protocols that capture their performance across a broad range of biomedical tasks. Moreover, as these systems begin to exhibit agentic behaviors – autonomous AI capabilities to accomplish tasks – the criteria for benchmarking must extend beyond accuracy and consistency to include safety and ethical considerations. These challenges highlight the urgent need for flexible evaluation frameworks that can keep pace with the transformative potential of next-generation AI in biomedicine.

Remedying this crisis requires concrete, stepwise solutions. One possible strategy is to organize consortium-led initiatives that integrate and curate reference datasets, establish criteria for dataset transparency, and mandate standardized preprocessing pipelines to reduce the variability in how competing models are evaluated. These efforts must be matched by frameworks that link model performance to clinically or biologically meaningful outcomes rather than relying solely on generic metrics.

Implementing standardized benchmarks entails navigating practical and economic barriers. Large-scale open-data projects demand financial and human resources to annotate, maintain and update datasets. They also raise privacy, consent and intellectual property concerns, making it vital to engage institutional review boards, patient advocacy groups, and regulatory agencies early in the process.

Collaboration with these stakeholders will ensure that benchmark development respects data protection laws, ethical considerations, and public trust. Sustainable funding models, whether public–private partnerships, subscription-based platforms, or data access marketplaces, may be necessary to help maintain such large collaborative infrastructures. Striking a balance between proprietary and publicly accessible datasets is also important. Organizations that oversee benchmark development should institute regular comparisons between proprietary training sets, changing disease indications and other variability in patient populations. This practice, coupled with incentives for data diversification, could guide developers to expand their training distributions and ensure that models remain generalizable.

Beyond technical and data-sharing challenges, addressing the ethical, regulatory and policy dimensions is crucial. To facilitate clinical translation and streamline future approvals, ML developers should involve regulatory agencies in shaping benchmark criteria that align with existing and emerging policies. By building compliance with standards and guidelines directly into these benchmarks, the community can ensure that models meet rigorous thresholds of reliability. Data governance frameworks should account for data protection laws, and methods for securely handling sensitive health information.

Cultivating a workforce that is well-versed in both computational and biomedical sciences will be key. Cross-disciplinary education programs can train researchers to evaluate algorithms critically and appreciate the complexities of biomedical workflows. These initiatives would also provide the ethical, regulatory and methodological fluency needed to advance field-wide benchmarking efforts. By preparing the next generation of scientists and clinicians to prioritize fairness and reproducibility, future ML-driven biomedical innovations will be held to high standards of transparency and societal benefit.

Faisal Mahmood
Mass General Brigham, Harvard Medical School and the Broad Institute of Harvard and MIT, Cambridge, MA, USA.
e-mail: fmahmood@fas.harvard.edu

Published online: 8 April 2025

References
1. Moor, M. et al. *Nature* **616**, 259–265 (2023).
2. Acosta, J. N. et al. *Nat. Med.* **28**, 1773–1784 (2022).
3. Zhou, Y. et al. *Nature* **622**, 156–163 (2023).

Competing interests
F.M. is an inventor on several patents related to computational pathology, and a scientific advisor for Modella AI and Danaher.

Mahmood, F., 2025. A benchmarking crisis in biomedical machine learning. Nature Medicine, 31(4), pp.1060-1060.

➢ Existing benchmarks of LLMs on medicine:

MedQA (USMLE)	PubMedQA (PubMed abstract)	MedMCQA (medical exams)	HealthBench (Expert created scenarios)
MMLU medical related subsets (Books, exams)	Benchmark for biomedicine (PubMed+)[1]	Isolated evaluations on specific clinical scenarios	

1. Chen, Q., Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M.B., Ai, X., Lai, P.T., Wang, Z. and Keloth, V.K., 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, *16*(1), p.3280.

- Existing medical exam benchmarks **overly simplify** the task and do not reflect LLM performance in real-world clinical understanding.
- Isolated clinical evaluations are not comprehensive and are difficult to **generalize**
- There is a lack of **dynamic** benchmarking for LLMs on diverse real-world clinical tasks as new models continue to emerge.

NEJM AI

NEJM AI 2025;2(2)
DOI: 10.1056/AIe2401235

EDITORIAL

It's Time to Bench the Medical Exam Benchmark

Inioluwa Deborah Raji, B.A.S.,[1] Roxana Daneshjou, M.D., Ph.D.,[2,3] and Emily Alsentzer, Ph.D.[2]

Received: December 18, 2024; Accepted: December 19, 2024; Published: January 23, 2025

Abstract

Medical licensing examinations, such as the United States Medical Licensing Examination, have become the default benchmarks for evaluating large language models (LLMs) in health care. Performance on these benchmarks is frequently cited as evidence of progress and used to justify the deployment of LLMs into clinical settings. However, we argue that these benchmarks are fundamentally limited as signals for assessing true clinical utility.

Raji, I.D., Daneshjou, R. and Alsentzer, E., 2025. It's Time to Bench the Medical Exam Benchmark. NEJM AI, 2(2), p.AIe2401235.

- There is a need for benchmarks to evaluate SOTA LLMs that:
  - Specifically focus on Real-world clinical (e.g. EHR) text understanding.
    - Clinical status, history, and procedures are documented in EHRs.
    - Clinical decisions are based on EHR data in practice
    - Real-world clinical text is messy in format and presents many more challenges.
  - Include multiple language
  - Include large-scale and diverse tasks
  - Include diverse LLMs to support diverse use cases :
    - open-sourced vs commercial
    - medical vs general
    - small vs large
  - With Open leaderboard, dynamically updated with new models

- **Background**
- **BRIDGE Overview**
- **Dataset Introduction**
- **Potential Topics**

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

Jiageng Wu, Bowen Gu, Ren Zhou, Kevin Xie, Doug Snyder, Yixing Jiang, Valentina Carducci, Richard Wyss, Rishi J Desai, Emily Alsentzer, Leo Anthony Celi, Adam Rodman, Sebastian Schneeweiss, Jonathan H Chen, Santiago Romero-Brufau, Kueiyu Joshua Lin, Jie Yang

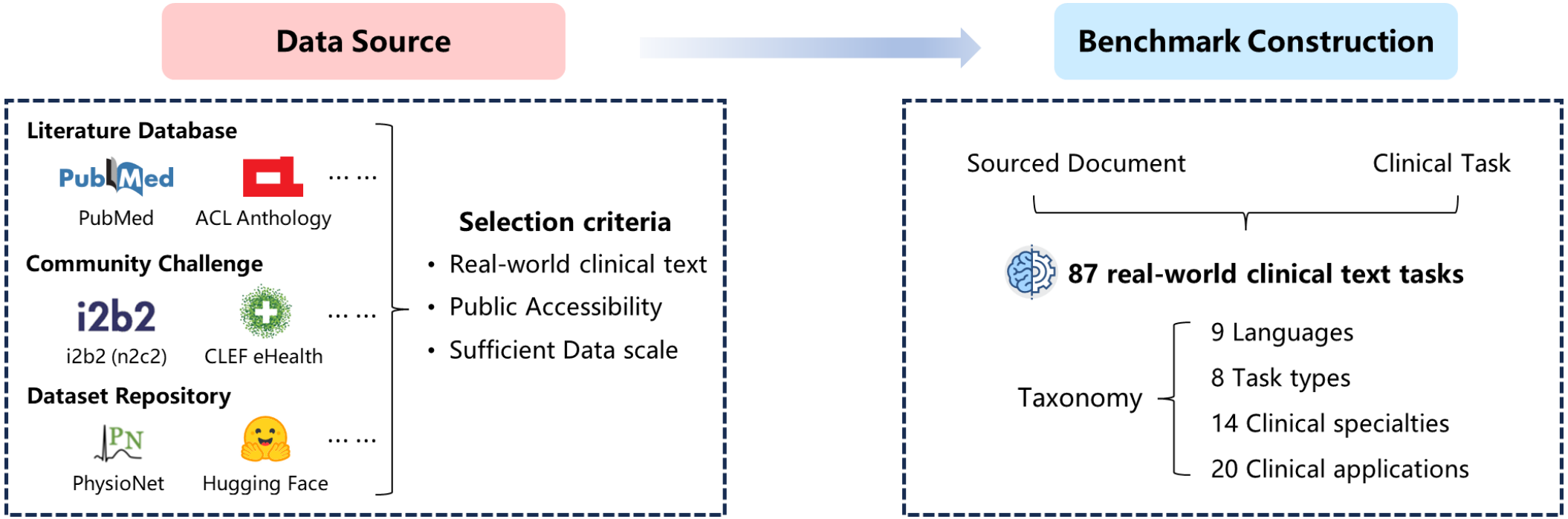
Paper:

Leaderboard:

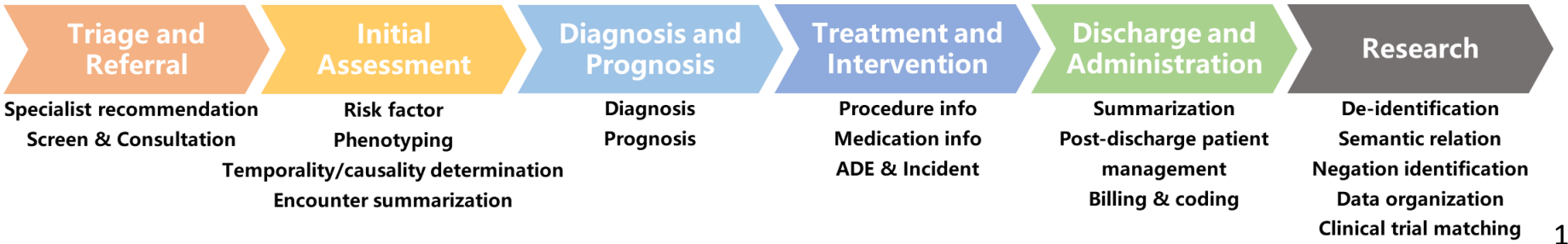
Dataset:

Wu, J., Gu, B., Zhou, R., Xie, K., Snyder, D., Jiang, Y., Carducci, V., Wyss, R., Desai, R.J., Alsentzer, E. Celi, L.A., …, Yang, J. 2025. BRIDGE: Benchmarking Large Language Models for Understanding Real-world Clinical Practice Text. arXiv preprint arXiv:2504.19467.

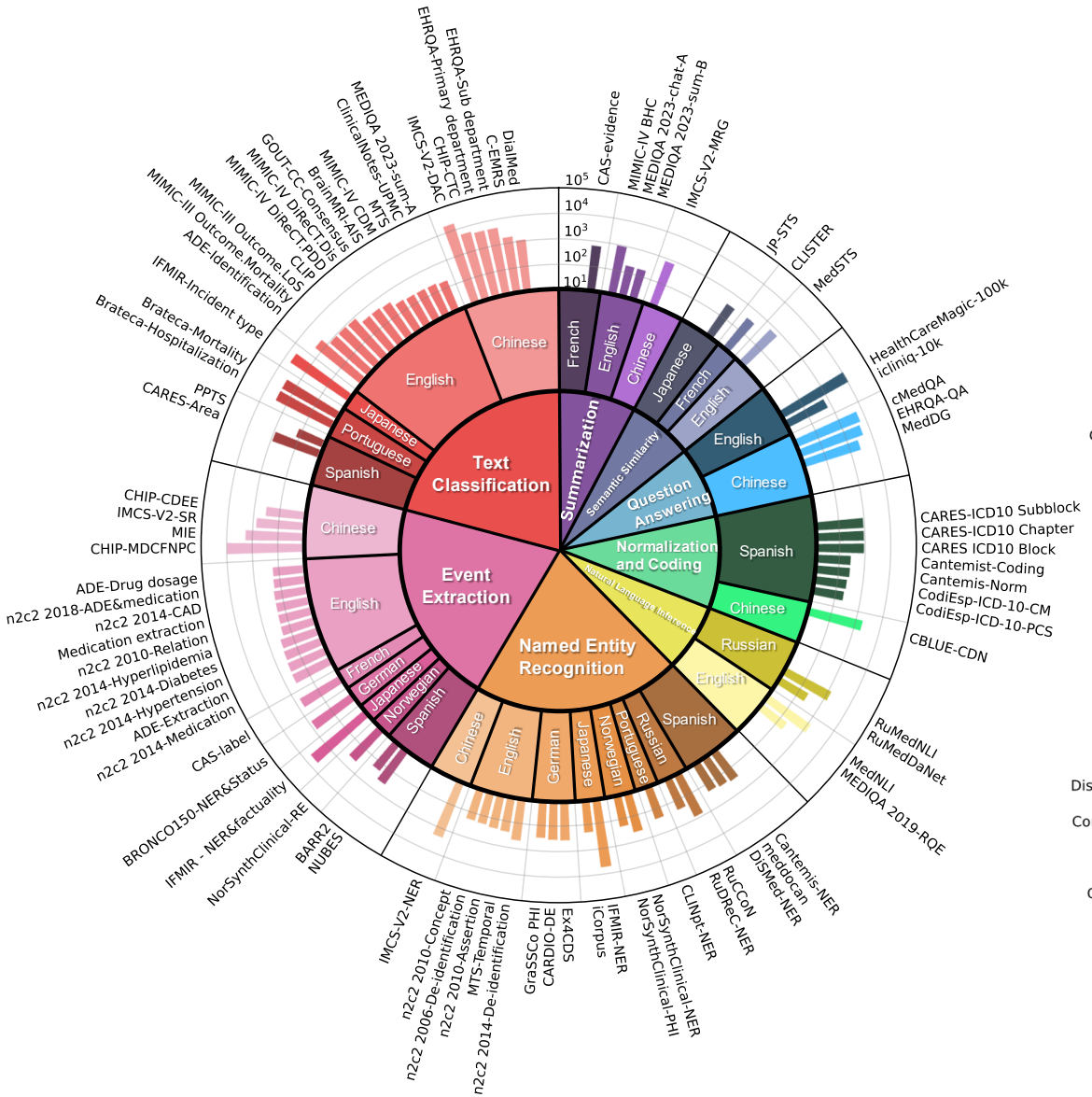
➢ **BRIDGE:** Benchmarking Large Language Models for Understanding Real-world Clinical Practice Text



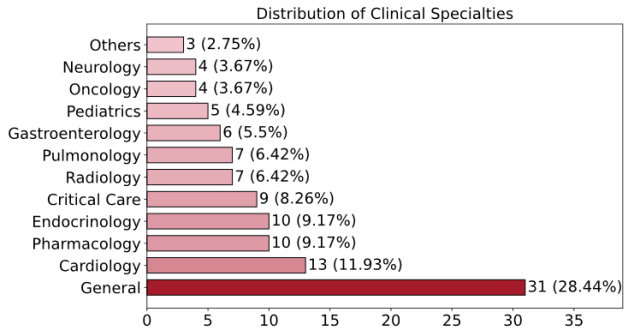
Support clinical scenarios

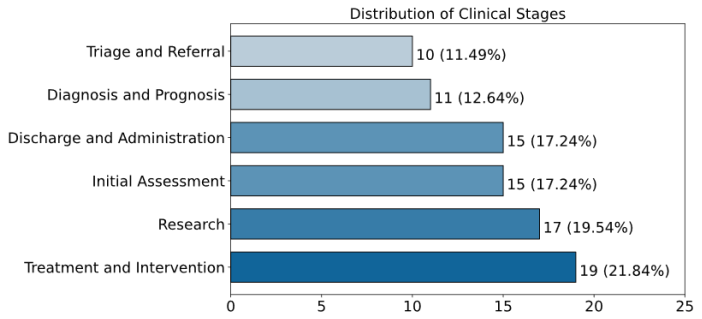


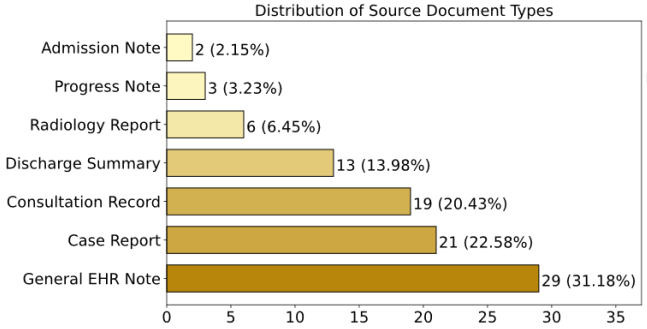
➢ Task type and include languages

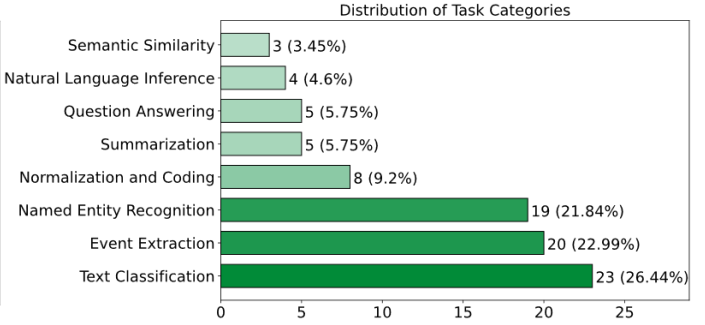


➢ Distribution of Clinical Specialty / Clinical Stage / Document type / Task Type

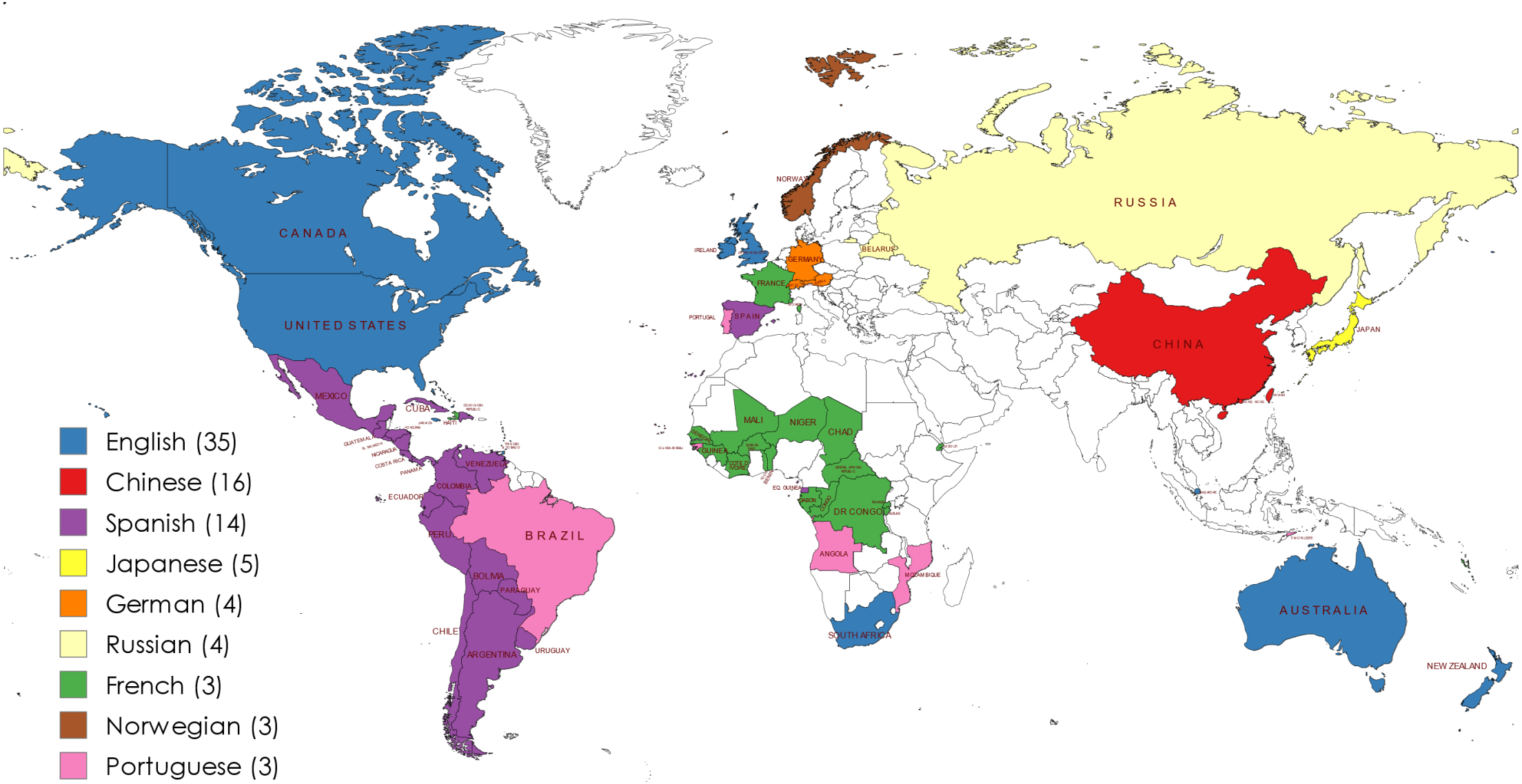








➢ Geographic distribution of countries where BRIDGE covers the official languages.



- **Background**
- **BRIDGE Overview**
- **Dataset Introduction**
- **Potential Topics**

➢ Dataset Information

○ Appendix Section 5: BRIDGE Dataset and Task Information

Paper link

Dataset information

5.7 BrainMRI-AIS

BrainMRI-AIS[9] dataset comprises 3,024 brain MRI radiology reports, focusing on the identification of acute ischemic stroke (AIS). It consists of free-text radiology reports collected between January 2015 and December 2016 from Hallym University Sacred Heart Hospital in Korea. Reports were manually annotated with AIS or non-AIS labels by medical experts to ensure high-quality labeling.

- **Language:** English
- **Clinical Stage:** Diagnosis and Prognosis
- **Sourced Clinical Document Type:** Radiology Report
- **Clinical Specialty:** Neurology, Radiology
- **Application Method:** Link of BrainMRI-AIS Dataset

Task information

Task type: *Text Classification*
Instruction: *Given a brain Magnetic Resonance Imaging (MRI) radiology report, determine whether the patient has acute ischemic stroke (AIS).*
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
AIS: label
The optional list for "label" is ["Yes", "No"].
Input: *[A brain magnetic resonance imaging (MRI) radiology report of a patient]*
Output: *AIS: [Yes / No]*

- **BRIDGE-Open**
  - 55 totally publicly accessible datasets
  - The others were regulated

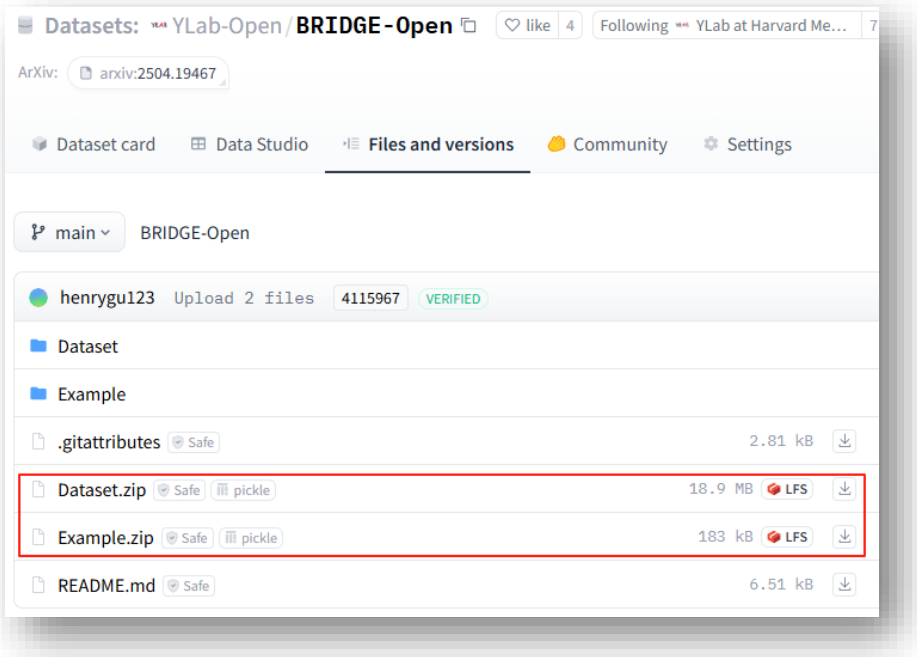
- **Dataset link:**

  HuggingFace Dataset: BRIDGE-Open
  (https://huggingface.co/datasets/YLab-Open/BRIDGE-Open)

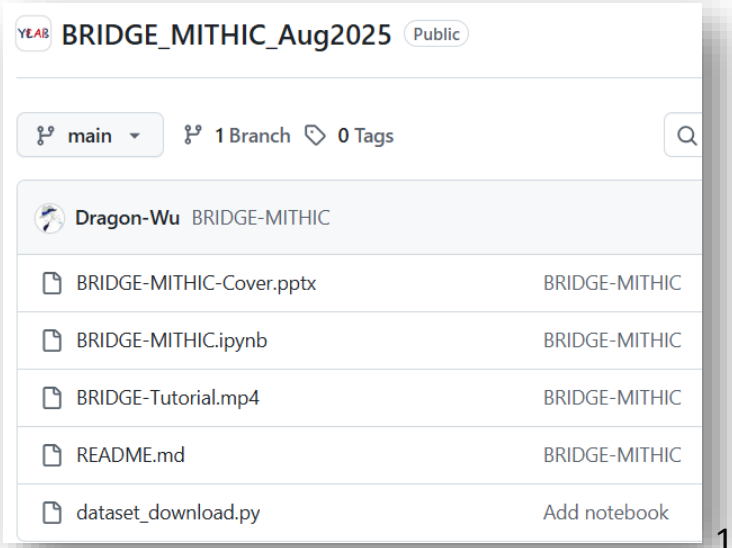
- **Demo code for dataset:**

  Github repo: BRIDGE_MITHIC_Aug2025
  (https://github.com/YLab-Open/BRIDGE_MITHIC_Aug2025)

Dataset page



Demo code



```
dataset_raw/
├── task_1.SFT.json
├── task_1.SFT.json
└── example/
    ├── task_1.example.json
    ├── task_1.example.json
```

➢ File format:

- Each task contain: testing data, examples
- Data: metadata, instruction, input (raw clinical text), output (only example provided)

task: BrainMRI-AIS
language: en
task type: Text Classification
id: 1661
split: train
instruction: Given a brain Magnetic Resonance Imaging (MRI) radiology report, determine whether the patient has acute ischemic stroke (AIS).
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
AIS: label
The optional list for "label" is ["Yes", "No"].
input: Unremarkable finding of brain parenchyma and cerebrospinal fluid space. magnetic resonance angiography: No gross abnormal finding.
output: AIS: No

➢ Task type

1. **Text classification**: Determine or predict categorical labels (e.g., diagnosis, risk stratification)
2. **Semantic similarity**: Assessing the similarity of two sentences or clinical notes.
3. **Natural Language Inference (NLI)**: Evaluating the logical relationships between paired texts.
4. **Normalization and coding**: Map the whole clinical note or the extracted entities to standardized clinical code systems (e.g., ICD, SNOMED)
5. **Named Entity Recognition (NER)**: Identify the medical entities and label them with types.
6. **Event extraction**: Identify the medical entities and capture additional attributes or relations beyond simple entity types (e.g., temporal status, severity).
7. **Question-Answering (QA)**: Generating accurate responses to healthcare inquiries.
8. **Summarization**: Condense clinical notes into concise summaries by extraction or generation.

➢ Task Example – Text classification (diagnosis / prognosis /…)

BrainMRI-AIS (English)

Task: To determine whether the patient has an acute ischemic stroke.

Input: A brain magnetic resonance imaging (MRI) radiology report of a patient

Output: AIS: [Yes / No]

> **Instruction:** *Given a brain Magnetic Resonance Imaging (MRI) radiology report, determine whether the patient has acute ischemic stroke (AIS).*
> *Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:*
> *AIS: label*
> *The optional list for "label" is ["Yes", "No"].*
> **Input:** *1. Left parietal depressed skull fracture. 2. Diffuse left cerebral subdural hygroma. 3. Multiple enhanced nodules and both cerebellar hemispheres, cerebellar vermis, both thalamus and both cerebral corticomedullaryy junctions. - Associated with surrounding edema, internal hemorrhage. - Multiple hemorrhagic metastasis. 4. Chronic both otomastoiditis. 5. Scant left parietal subdural hemorrhage. 6. Multiple old small infarcts in both cerebral deep white matter. 7. Diffuse hydrocephalus. 8. No diffusion restriction. 9. Venous malformation in left parietal subcortical white matter. 10. Focal stenosis of left proximal cervical internal carotid artery and right vertebral ostium. 11. Dolichosis of vertebro-basilar artery. 12. Steno-occlusive disease of left cavernous internal carotid artery.*
> **Output:** *AIS: [Yes / No]*

➢ Task Example – Natural Language Inference (logical relationships between texts)

MEDIQA 2019-RQE (English)

Task: To identify entailment between two questions in the context of QA.

Input: Question A: [content of Question A]

Question B: [content of Question B]

Output: Answer: [true / false]

Instruction: *Given the following two clinical questions labeled as "Question A" and "Question B", determine if the answer to "Question B" is also the answer to "Question A", either exactly or partially.*
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
answer: label
The optional list for "label" is ["true", "false"].
Input: *Question A: "What is the incidence of dyspareunia after a vaginal wall cyst? This was removed about a year ago by a gynecologist and for the past year she has had this problem."*
Question B: "What is the incidence of dyspareunia after a vaginal wall cyst removal?"
Output: *answer: true*

➢ Task Example – Semantic similarity

CLISTER (French)

Task: To capture semantic textual similarity between French sentence pairs sourced from clinical cases.

Input:

Clinical sentence pair in French

Output:

similarity score: [0/1/2/3/4/5]

Instruction: *Given the following two clinical sentences that are labeled as "Sentence A" and "Sentence B" in French, decide the similarity of the two sentences. Specifically, analyze the potential similarity, including: Surface similarity: concerns the structural similarity. This similarity is based on grammatical words or words that are not related to the domain. Two sentences that have a surface similarity can be syntactically close but semantically distant. Semantic similarity: concerns medical concepts. The closer the concepts are to one another, the higher the similarity. These concepts can refer to medications, diseases, procedures, and others. Clinical compatibility: going further into the semantics, clinical compatibility is an assessment of whether sentences in a pair can refer to the same clinical case.*
Then, assign a similarity score to the sentence pair based on the following scale:
- "0": For sentence pairs with only surface similarity, such as words non-specific to the medical domain or stop-words.
- "1": For sentence pairs with only surface similarity, concerning at most one medical entity.
- "2": For sentence pairs containing medical concepts with low semantic similarity, but no clinical compatibility. Typically, sentences in a pair can concern a disease, a procedure, or a drug.
- "3": For sentence pairs with semantic similarity on several medical concepts making them partially clinically compatible.
- "4": For sentence pairs with high semantic similarity and clinical compatibility. One sentence may contain more information than the other may, and vice-versa.
- "5": For sentence pairs with high semantic similarity and full clinical compatibility. The sentences have globally the same meaning, while one may be more specific than the other.
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
similarity score: score
The optional list for "score" is ["0", "1", "2", "3", "4", "5"].
Input: *Sentence A: L'examen cytobactériologique des urines a permis d'isoler un colibacille.*
Sentence B: L'examen cytobactériologique des urines était négatif.
Output: *similarity score: 3*

➢ Task Example – Named entity recognition (medical entity and type: drug / PHI / …)

GraSSCo PHI (German)

Task: To extract all instances of Protected Health Information (PHI) from the patient discharge summaries and identify the corresponding PHI type.

Input:

A discharge summary of a patient in German

Output:

entity: xxx, type: xxx; …

Input: *Werte Frau Kollegin, werter geehrter Herr Kollege!*
Wir berichten über lhre Patientin Beate Albers (4.4.1997), die sich vom 19.3. bis zum 7.5.2029 in unserer stat. Behandlung befand.*
Vorgeschichte Befund
- *Verbrennung 1. – 3. Grades, Kopf I Hals,5% v KOF*
- *Handamputation LI*
- *Akute Psychose aus dem schizophrenen Formenkreis*

Output: *entity: Beate Albers, type: NAMEPATIENT;*
entity : 4.4.1997, type : DATE;
entity : 7.5.2029, type : DATE;
entity : 19.3., type : DATE;

Instruction: *Given the clinical summaries of a patient in German, extract the following types of entities from the clinical text:*
- *"LOCATION_COUNTRY": Country name or reference*
- *"NAME_DOCTOR": Name of a medical professional*
- *"AGE": Numeric representation of a person's age*
- *"CONTACT_FAX": Fax number for communication*
- *"LOCATION_ZIP": Postal or ZIP code*
- *"LOCATION_ORGANIZATION": Name of an organization or institution*
- *"CONTACT_PHONE": Phone number for communication*
- *"DATE": Calendar date*
- *"LOCATION_CITY": Name of a city*
- *"CONTACT_EMAIL": Email address*
- *"NAME_PATIENT": Name of a patient*
- *"LOCATION_HOSPITAL": Name of a hospital or medical facility*
- *"PROFESSION": Job title or occupation*
- *"NAME_TITLE": Honorific or title before a name*
- *"NAME_USERNAME": Username for online identification*
- *"ID": Identification number or code*
- *"NAME_RELATIVE": Name of a family member*
- *"NAME_EXT": Name suffix, extension, or additional identifier*
- *"LOCATION_STREET": Street address or name*

Return your answer in the following format. Do not output entities whose types do not exist in the clinical text. DO NOT GIVE ANY EXPLANATION:
entity: ..., type: ...;
...
entity: ..., type: ...;
The optional list for "type" is ["LOCATION_COUNTRY", "NAME_DOCTOR", "AGE", "CONTACT_FAX", "LOCATION_ZIP", "LOCATION_ORGANIZATION", "CONTACT_PHONE", "DATE", "LOCATION_CITY", "CONTACT_EMAIL", "NAME_PATIENT", "LOCATION_HOSPITAL", "PROFESSION", "NAME_TITLE", "NAME_USERNAME", "ID", "NAME_RELATIVE", "NAME_EXT", "LOCATION_STREET"]

➢ Task Example – Event extraction (drug dosage / temporal info / phenotyping / …)

NUBES (Spanish)

Task: To extract the negation and uncertainty cues and scope.

Input:

Clinical text of a patient

Output:

entity: [clinical entity],

type: [NegSynMarker / NegLexMarker / NegMorMarker / UncertSynMarker / UncertLexMarker],

scope: [scope of the negation cue];

Instruction: *Given the clinical text of a patient in Spanish, extract the following types of entities from the clinical text:*
- "NegSynMarker": Syntactic negation cue.
- "NegLexMarker": Lexical negation cue.
- "NegMorMarker": Morphological negation cue.
- "UncertSynMarker": Syntactic uncertainty cue.
- "UncertLexMarker": Lexical uncertainty cue.
Then, for each extracted entity, extract the scope that the entity affects.
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
entity: ..., type: ..., scope: ...;
...
entity: ..., type: ..., scope: ...;
The optional list for "type" is ["NegSynMarker", "NegLexMarker", "NegMorMarker", "UncertSynMarker", "UncertLexMarker"].
Input: *Durante el ingreso no vuelve a presentar sangrado .*
Oi : normal
Se difiere intervención por alergia al Latex .
Edema a nivel periorbitario izquierdo. visión y movilidad ocular conservada Se retira taponamiento FNI , no se objetiva sangrado activo .
EXPLORACIÓN 2/3/20 orl Trago OD : + oTOSCOPIA : CAE eritematoso y leve edema con exudado , permeable , visualizo timpano integro .
Se decide traslado a Hospital Dolors Aleu por persistencia de sangrado .
Se añade antibiotico y corticoide , desapareciendo el edema periorbital .
Se programara segun protocolo .
Se retira el taponamiento anteroposterior y se mantiene en observación durante 6 horas sin sangrar , por lo que es dado de alta .
Se habla con el S. de ORL y UCI .
Output: *entity: no, type: NegSynMarker, scope: vuelve a presentar sangrado;*
entity: no, type: NegSynMarker, scope: se objetiva sangrado activo;
entity: Se retira, type: NegLexMarker, scope: taponamiento FNI;
entity: desapareciendo, type: NegLexMarker, scope: el edema periorbital;
entity: Se retira, type: NegLexMarker, scope: el taponamiento anteroposterior;
entity: sin, type: NegSynMarker, scope: sangrar;

➢ Task Example – Normalization and Coding (term normalization / ICD code /…)

CARES (Spanish)

Input: To identify the ICD-10 code that corresponds to the condition mentioned in the radiology report.

Input:

anonymized radiology reports

Output:

ICD-10 sub-block: detailed code

Instruction: *Given a radiology report in Spanish, determine the appropriate ICD-10 sub-block codes corresponding to the conditions mentioned in the report. Specifically, the sub-block code is the third level of the ICD-10 classification and represents several related diseases. Each sub-block code is identified by a code containing a character, two digits, and a decimal, which indicates its chapter, block, and detailed sub-block. This report may contain multiple conditions and is related to multiple sub-block codes. Assuming the number of appropriate sub-blocks is N, return the codes for N appropriate sub-blocks in the output. Notably, the required sub-block code is a combination of the chapter, the block, and the sub-block, such as "I00.0", "I01.0", rather than the coarse range of the sub-block, such as "I00.0-I99.9".*
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
ICD-10 sub-block: code 1, code_2, ..., code N.
Input: *Edad: 21 años y 9 meses. EXPLORACIONES: RMI de silla turca e hipófisis sin y con contraste. INFORMACION CLINICA. Hiperprolactinemia. COMPARADO CON: No existen estudios previos de esta región anatómica. HALLAZGOS. Morfología globulosa de la glándula hipofisaria (diámetro craneocaudal de 12 mm) con ligero aumento de tamaño de la silla turca. En la porción más craneal de la adenohipófisis, próximo al tallo hipofisario se identifica imagen de 4 mm que muestra captación de contraste de forma precoz con respecto al resto de la glándula, que podría corresponder a microadenoma. Se distingue la neurohipófisis. El tallo está situado en línea media. La cisterna supraselar y quiasma óptico son normales. No alteraciones en los senos cavernosos ni en el seno esfenoidal. CONCLUSION: Morfología globulosa de la glándula hipofisaria con ligero aumento de tamaño de la silla turca, y probable microadenoma en la porción más craneal de la adenohipófisis.*
Output: *ICD-10 sub block: D35.2, D35.4*

➢ Task Example – Question answering (generating responses to healthcare inquiries)

MedDG (Chinese)

Task: To generate the doctor's response based on the provided dialogue history of a medical consultation.

Input: Dialogue history of medical consultation between a doctor and patient

Output: Doctor: [generated response from doctor's perspective]

Instruction: *Given the medical consultation in Chinese, generate the next response of the doctor based on the dialogue context.*
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
医生: ...
Input: 患者：我昨天下午到今天肚子疼，而且是一会疼一会不疼，昨晚还拉肚子了，这是怎么回事啊？（女，15岁）
Output: 医生: 肚子具体哪个部位疼？拉了几次了？

Patient: I have been experiencing stomach pain from yesterday afternoon to today, and it's sometimes painful and sometimes not. I also had diarrhea last night. What's going on? (Female, 15 years old)

Doctor: Which specific part of the stomach hurts? How many times of diarrhea?

➢ Task Example – Summarization (summarize clinical notes by generation or extraction)

IMCS-V2-MRG (Chinese)

Task: To generate a structured summarization based on the patient's chief complaint and the full doctor–patient dialogue

Input: A whole medical consultation

Output:

A structure summarization include chief complaint, present illness history, auxiliary examination, past history, diagnosis and suggestion

Output: 主诉: 发热一天
现病史: 患儿今下午出现发热症状，退热栓治疗，目前症状无明显改善。精神食欲尚可，二便正常。
辅助检查:
既往史:
诊断: 小儿发热
建议: 完善血常规

Instruction: *Given the medical consultation in Chinese, generate the brief report based on the dialogue between the patient and doctor. The report should include the following sections:*
1. 主诉*(Chief complaint):* 病人自诉（*Self-report*）的总结，包括主要症状或体征;
2. 现病史*(Present illness history):* 对话中病人涉及到的现病史的总结，如主要症状的描述（发病情况，发病时间）;
3. 辅助检查*(Auxiliary examination):* 对话中病人涉及过的医疗检查的总结，如病人已有的检查项目、检查结果、会诊记录等;
4. 既往史*(Past history):* 对话中医生对病人的过去病史的总结，如既往的健康状况、过去曾经患过的疾病等;
5. 诊断*(Diagnosis):* 对话中医生对病人的诊断结果的总结，如对疾病的诊断;
6. 建议*(Suggestion):* 对话中医生对病人的建议的总结，如检查建议、药物治疗、注意事项。
Return your answer in the following format. DO NOT GIVE ANY EXPLANATION:
主诉: ...
现病史: ...
辅助检查: ...
既往史: ...
诊断: ...
建议: ...
Input: 患者：宝宝一岁半，去了医院大夫说嗓子红，没有别的症状，建议验血我没验，今天下午突然发烧，*38*现在放了退热拴一个小时现在体温是*38*还可以用美林么?
医生：你好
患者：你好
医生：请问退热栓的成分是什么?
患者：就是清热吧
医生：有说明书吗?
患者：在医院开的就是放肛门的
患者：有的
...

- **Background**
- **BRIDGE Overview**
- **Dataset Introduction**
- **Potential Topics**

Research Topics

- Is there a topic correlation between **clinical notes vs medical exam questions**?
- Are patients of **different genders, races, or socioeconomic backgrounds** documented differently in clinical notes?
- How do **discharge summary contents vary across countries with different health systems** (e.g., single-payer vs. privatized)?
- Do clinicians in **different countries document differently when it comes to family involvement, religious beliefs, end-of-life care, or alternative medicine**?
- How have **certain conditions (e.g., HIV, COVID-19, opioid use disorder) been described and framed over time** in discharge summaries?
- Are there notable **differences in how the same clinical concept is expressed across languages (e.g., pain)?**

MITHIC-BRIDGE Workshop

- 2:40 – 2:55 BRIDGE Tutorial
- 2:55 – 4:20 Group discussion
- 4:20 – 4:40 Presentation

Link of this slide