



WorldMedQA

Laboratory for Computational Physiology

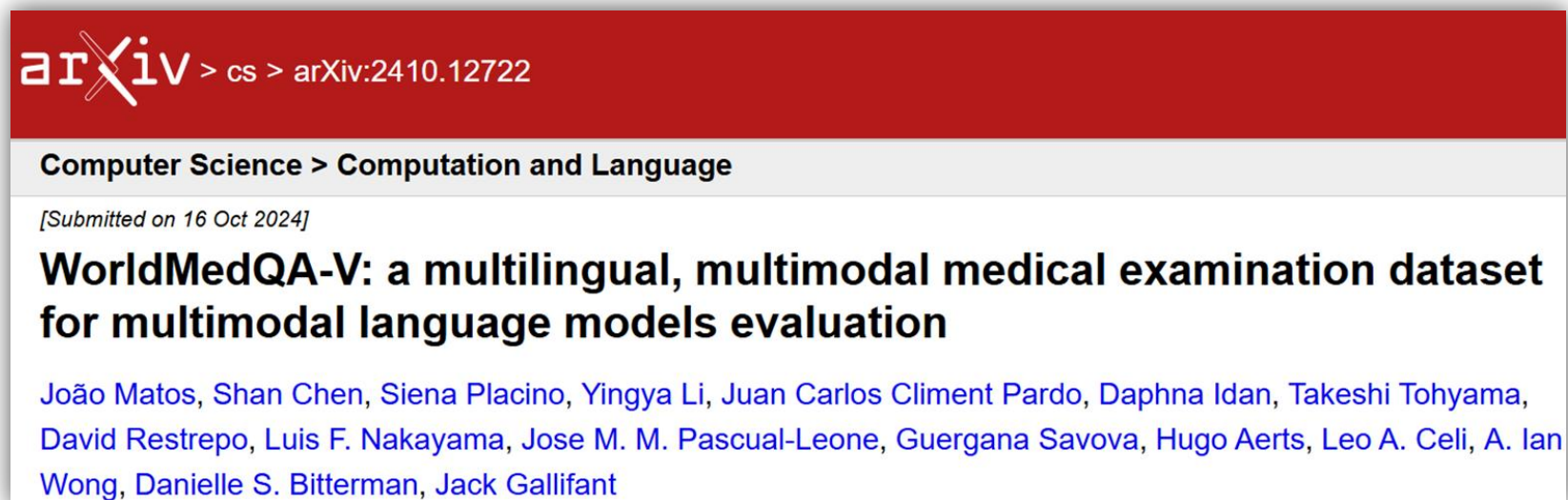
MITHIC Language/AI Colloquium, May 14, 2025



Context

- Generative artificial intelligence (AI) models are increasingly being adopted in healthcare, yet challenges remain:
 - Limited multilingual capabilities
 - Lack of locally contextualized assessment tools & frameworks
- Need for an effective, robust assessment approach
 - Safety, efficacy, fairness, etc.
- Medical exam-based datasets have become essential for evaluating large language models (LLMs)
 - MedQA (USA & China), KorMedMCQA (Republic of Korea), MedQA-SWE (Sweden)
 - Limited representation
- How different are these medical exams across countries ?

1st Version – WorldMedQA–v



To address the **safety**, **efficacy**, and **fairness** of **multimodal** LLMs in global healthcare settings, aspiring to achieve an influence akin to the MIMIC database.

Methods

1. Collect

Collection of medical examinations from different countries

Inclusion criteria:

- Official national medical examination
- Official language version



2. Curate

Selection, cleaning, inspection, and translation of collected QAs

-  Selection of QAs with images
-  Cleaning and data harmonization
-  Clinical Inspection and validation
-  English Translation

3. Evaluate

Evaluation of ten different multimodal language models

10 Models


GPT4o
GPT4o_MINI
GeminiFlash1-5
GeminiPro1-5
llava_next_llama3
llava_next_yi_34b
llava_next_mistral_7b
llava_next_vicuna_7b
Yi_VL_34B
Yi_VL_6B

Experiments

-  With Images
-  Without Images

4. Share

Release the data as a multimodal, multilingual medical benchmark



Data is publicly-available



Code is open-source

1st Version – WorldMedQA–v

Box 1. Example multimodal QA from the Brazilian subset

Original (Portuguese)

Um paciente do sexo masculino, 55 anos de idade, tabagista 60 maços/ano, com tosse crônica há mais de 10 anos, relata que há cerca de três meses observou a presença de sangue na secreção eliminada com a tosse. Refere ainda perda de cerca de 15% do peso habitual nesse mesmo período, anorexia, adinamia e sudorese noturna. A radiografia de tórax realizada por ocasião da consulta é mostrada abaixo. Qual a hipótese diagnóstica mais provável nesse caso?

- A) Aspergilose pulmonar.
- B) Carcinoma pulmonar.**
- C) Tuberculose cavitária.
- D) Bronquiectasia com infecção.
- E) Doença pulmonar obstrutiva crônica.

Image



Translation (English)

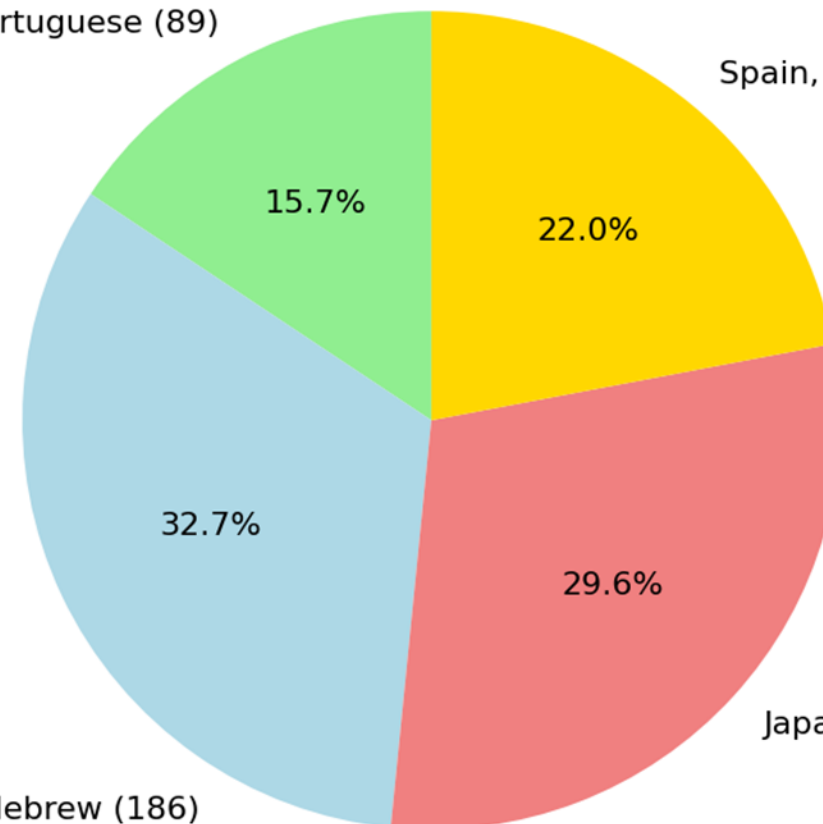
A 55-year-old male patient, with a smoking history of 60 pack-years, has had a chronic cough for over 10 years. He reports that about three months ago, he noticed the presence of blood in the sputum. He also mentions a weight loss of about 15% of his usual weight during the same period, anorexia, weakness, and night sweats. The chest X-ray taken at the time of the consultation is shown below. What is the most likely diagnostic hypothesis in this case?

- A) Pulmonary aspergillosis.
- B) Lung carcinoma.**
- C) Cavitary tuberculosis.
- D) Bronchiectasis with infection.
- E) Chronic obstructive pulmonary disease.

WorldMedQA (N=568)

Brazil, Portuguese (89)

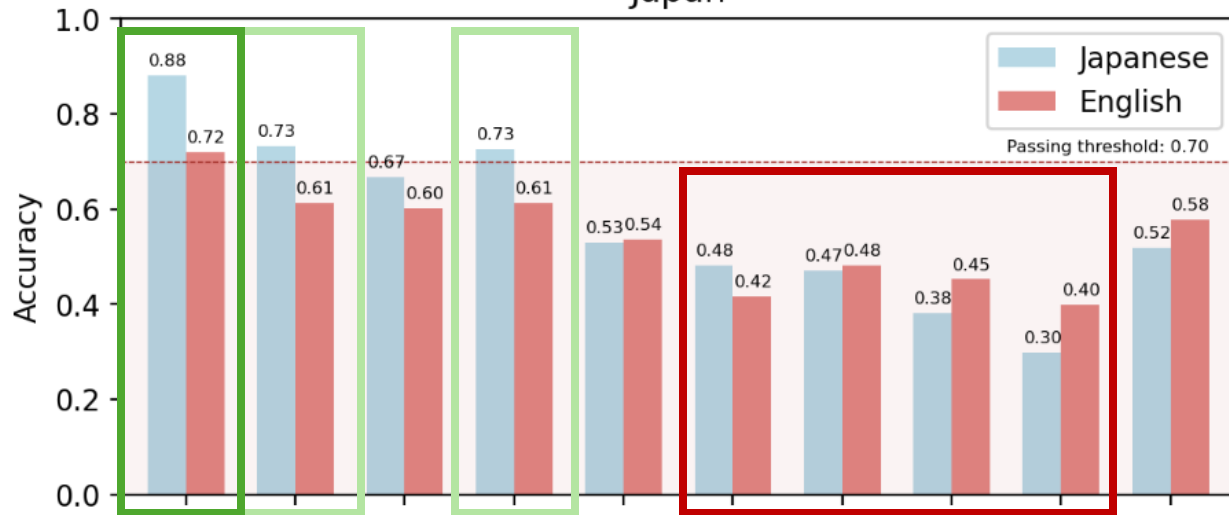
Spain, Spanish (125)



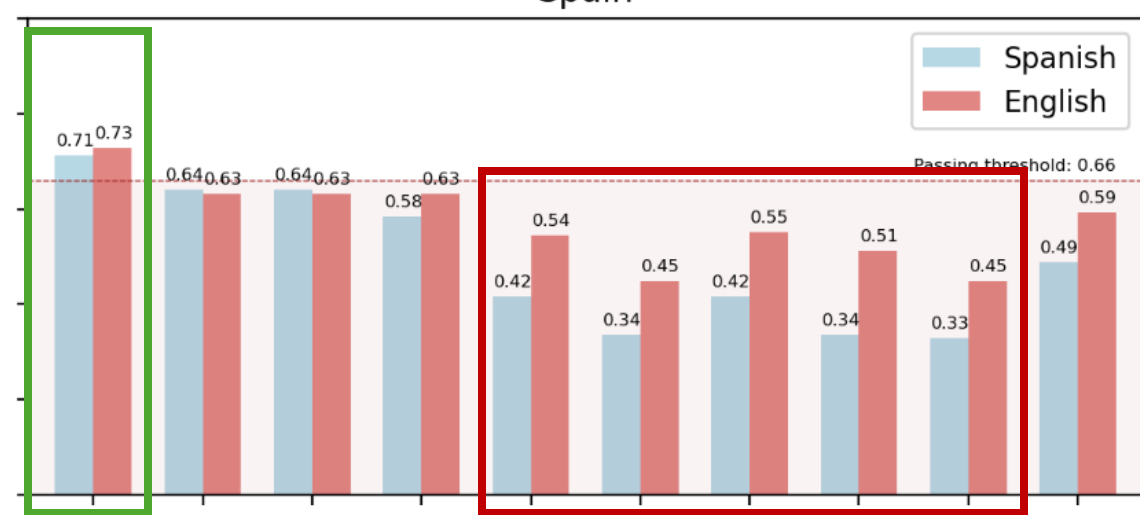
Israel, Hebrew (186)

Japan, Japanese (168)

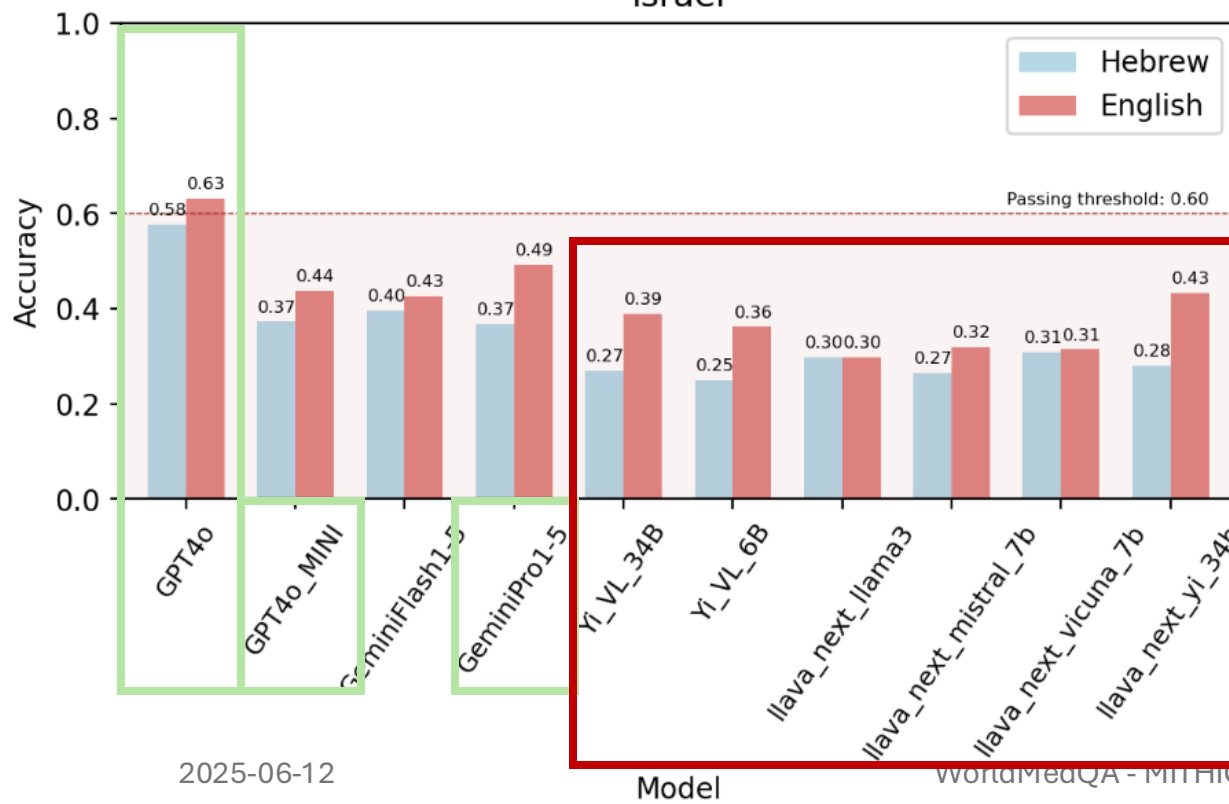
Japan



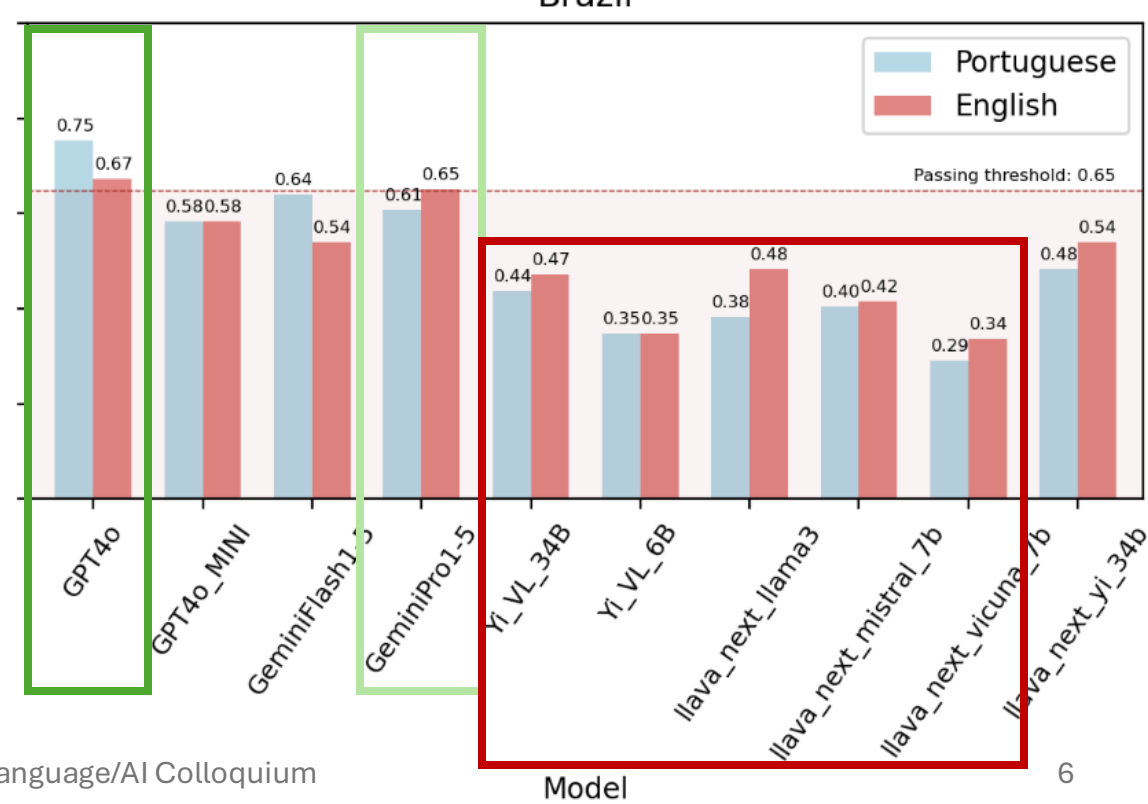
Spain



Israel



Brazil



2nd Version – WorldMedQA-t

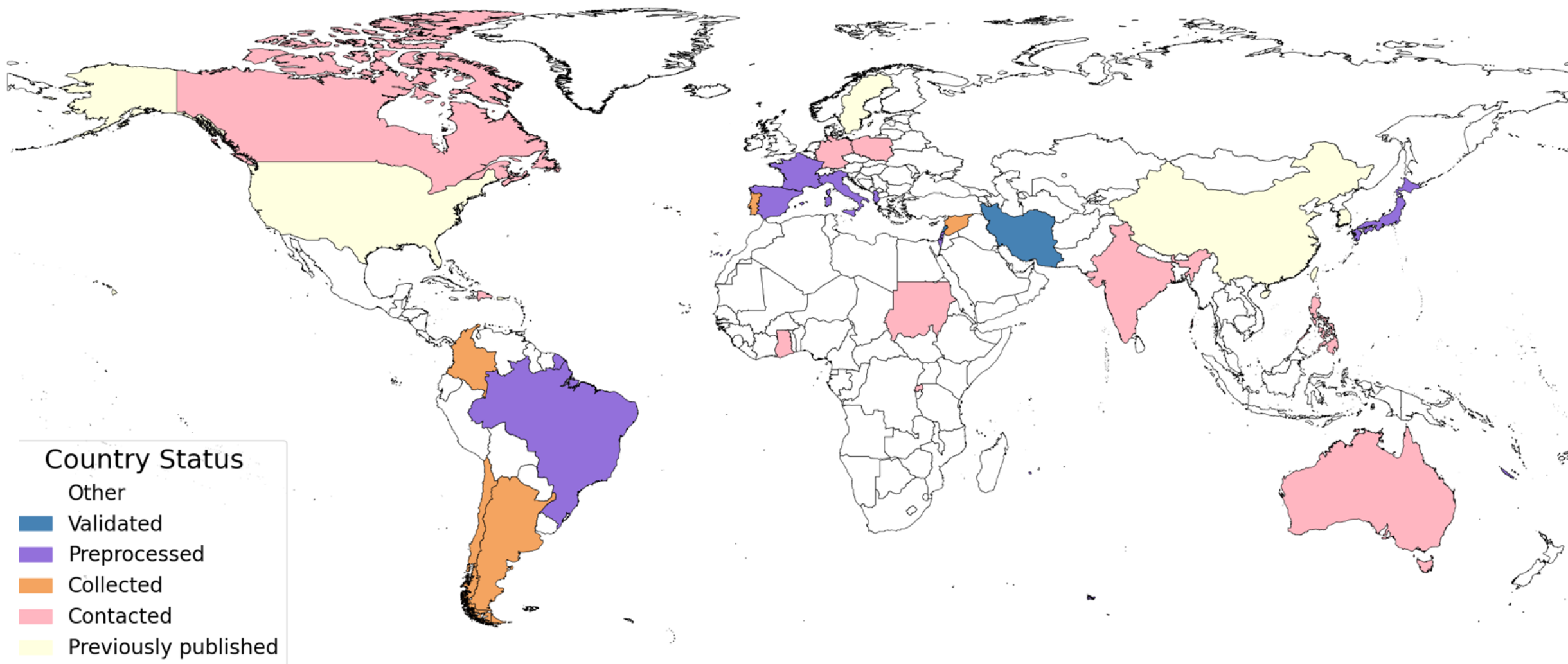
Objectives

- Larger sample size – ~1000 questions/country
- More comprehensive geographical representation - countries spanning six continents
- More diverse language representation
- Up-to-date exam questions: outdated datasets may overlap with LLM training corpora (Zhang et al., 2024; Gallifant et al., 2024)

[1]. Zhang, A. K., Klyman, K., Mai, Y., Levine, Y., Zhang, Y., Bommasani, R., & Liang, P. (2024). Language model developers should report train-test overlap. *arXiv preprint arXiv:2410.08385*.

[2]. Gallifant, J., Chen, S., Moreira, P., Munch, N., Gao, M., Pond, J., ... & Bitterman, D. (2024). Language models are surprisingly fragile to drug names in biomedical benchmarks. *arXiv preprint arXiv:2406.12066*.

WorldMedQA progress Map



Country	Language	Multimodal	Years	No. Q w/ images	No. Q w/o images
Iran	Persian	1	2022-2023	6	394
Lebanon	Arabic, French & English	0	2019-2023	/	1,440
Japan	Japanese	1	2022-2024	168	755
Spain	Spanish	1	2018-2023	158	1,070
Brazil	Portuguese	1	2011-2016 & 2020-2024	89	1,095
Italy	Italian	0	2017-2024	/	11,249
France	French	0	2019-2022	/	1,214
Israel	Hebrew	1	2020-2023	186	(In progress)
Syria	Arabic	0	2004-2006	/	570
Albania	Albanian	0	?	/	3,146
Portugal	Portuguese	1	2019-2023	(In progress)	(In progress)
Argentina	Spanish	0	2019-2024	/	2,400
Chile	Spanish	0	?	/	2,160
Colombia	Spanish	0	2016-2024	/	397
Total	10			607	25,890

Challenges

- Limit data availability
 - Official medical exam sharing policy 
- Inconsistent data quality
 - Off-the-record documents 
- Diverse formats
 - Image, PDFs 
- Labour intensive process 
 - Requires healthcare professionals to validate all Q&A to ensure data quality
 - Need for country-specific teams familiar with language

Next steps

- Covering six continents
 - Australia/New Zealand from Oceania
 - Two ~ three countries from Africa
- Creation of an open-ended Q&A bank
- Benchmarking LLMs against curated dataset
- Sustainability – regular updated project

Now we have the dataset... And?

Where is the bias coming from?

- Modeling
 - Dataset
 - Cultural/Linguistic
-
- To explore cultural and linguistic biases in medical education across countries, such as demographic disparities in pain representation and the prevalence of maternal health issues.

Future steps

- Representation of specific constructs across different cultures
 - Pain
 - Maternal health / women's health / pediatrics
- Sentiment classification on the medical exam questions
- Opportunity to reveal biases in medical knowledge systems globally

New ideas are welcome!



Pain – WorldMedQA

Why pain?

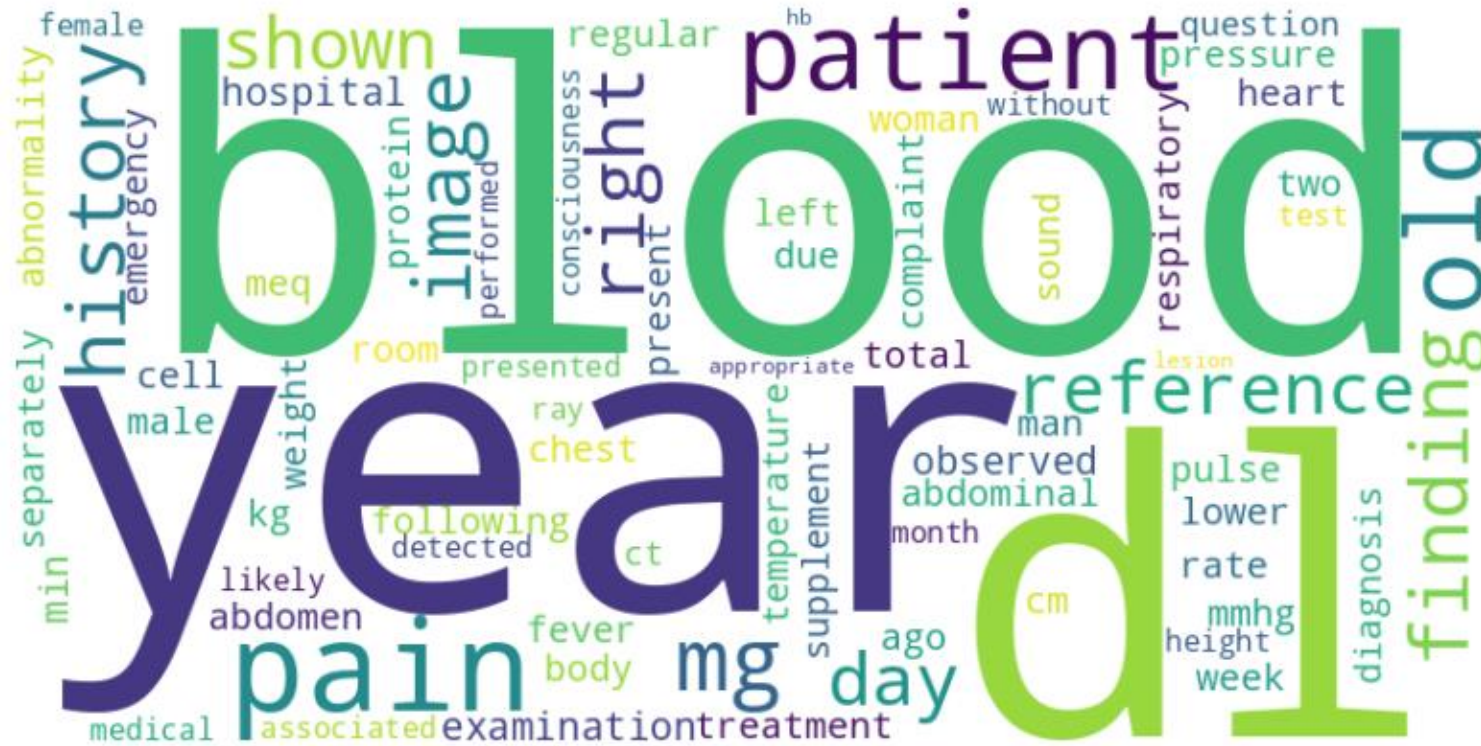
- The complexity of pain experience and the challenging nature of pain measurement (Katz et al., 2025)
- For marginalized and minority patients, the gap between patient-reported pain and physician assessments is even greater
→ significant inequalities in clinical care and patient outcomes (Mathur et al., 2022)

As a universal concept, how is “*pain*” encoded in language through medical exam questions?

[1]. Katz, R. A., Graham, S. S., & Buchman, D. Z. (2025). The need for epistemic humility in AI-assisted pain assessment. *Medicine, Health Care and Philosophy*, 1-11.

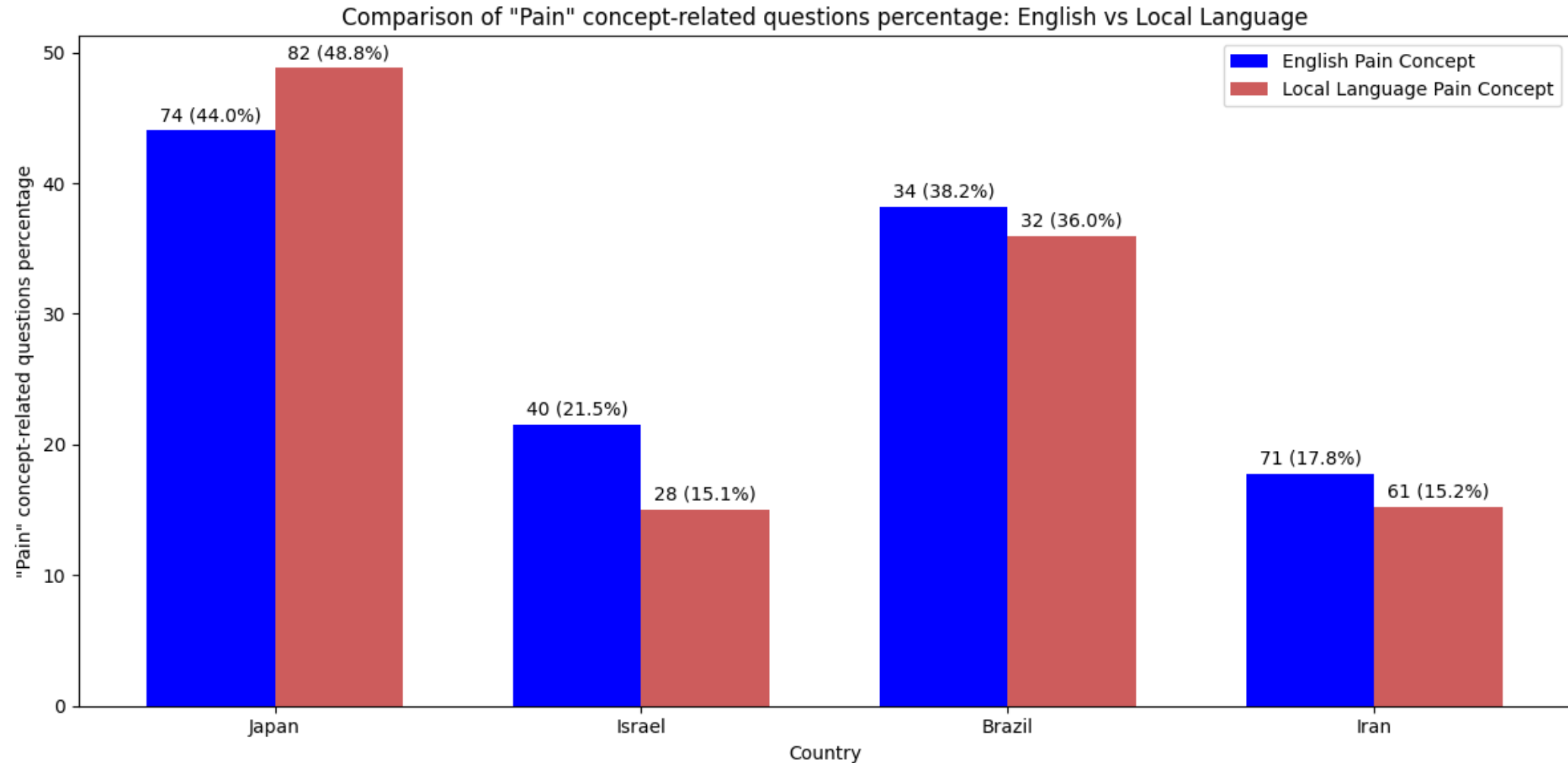
[2]. Mathur, V. A., Trost, Z., Ezenwa, M. O., Sturgeon, J. A., & Hood, A. M. (2022). Mechanisms of injustice: what we (do not) know about racialized disparities in pain. *Pain*, 163(6), 999-1005.

First glimpse – “pain” question (N=267)



- ('abdominal', 'pain'): 45
- ('chest', 'pain'): 39
- ('pain', 'right'): 25
- ('back', 'pain'): 12
- ('eye', 'pain'): 8
- ('pain', 'left'): 7
- ('pain', 'lower'): 6
- ('due', 'pain'): 6
- ('pain', 'swelling'): 6
- ('epigastric', 'pain'): 6

Preliminary analysis (N = 843)



“Pain” concept mentioned in...

Country	Total	In both extractions	Only in English extraction	Only in original language extraction	Information distortion
Japan	88	68 (77.3%)	6 (6.8%)	14 (15.9%)	
Israel	41	27 (65.9%)	13 (31.7%)	1 (2.4%)	
Brazil	35	31(88.6%)	3 (8.6%)	1 (2.9%)	
Iran	76	56 (73.7%)	15 (19.7%)	5 (6.6%)	

Future analysis

As a universal concept, how is “*pain*” encoded in language through medical exam questions?

- Through demographics
 - Sex
 - Age
 - Racial
- Location of pain; Intensity of pain
- Co-occurrence of hurtful words?

New ideas are welcome!

More to come!



WorldMedQA

- Make AI systems more adaptable to different healthcare settings
- Inspire groundbreaking research at the intersection of medicine, linguistics, and anthropology
- Promote fairer, more efficient, and more representative applications.

