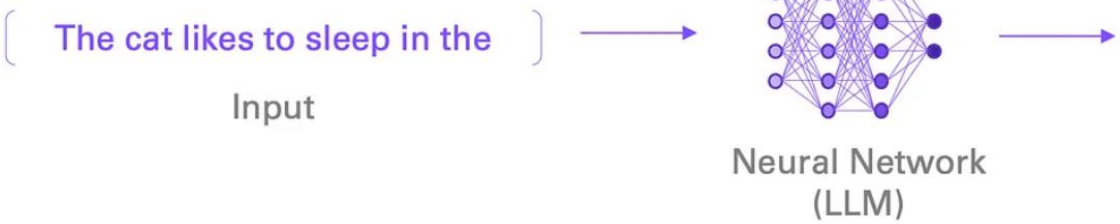
Keeping humans in the loop: aligning open medical LLMs with expert feedback

Giorgia Carra, MEng, PhD

Centre hospitalier universitaire vaudois

What is a Large Language Model?

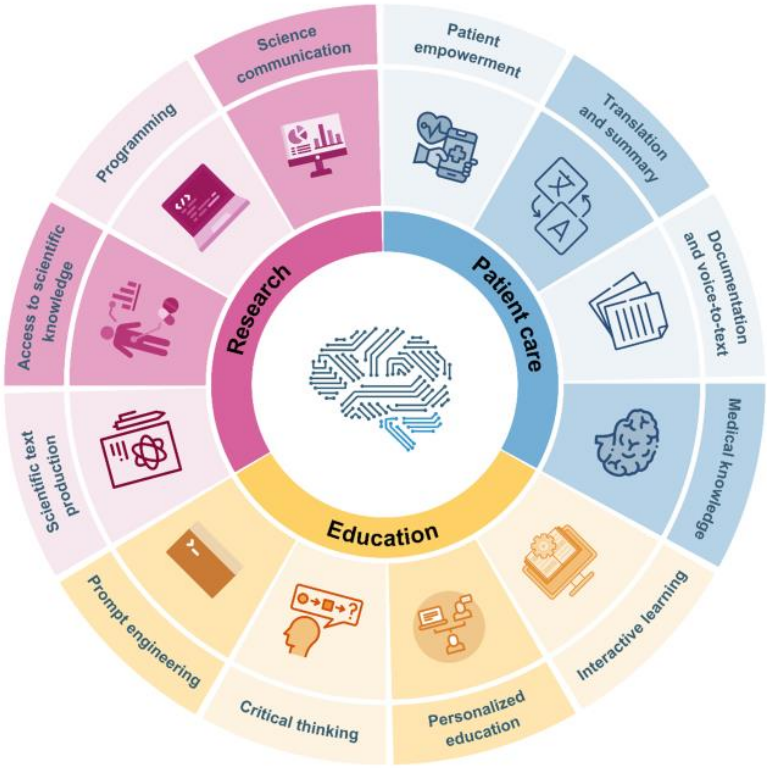


Word	Probability
ability	0.002
bag	0.071
box	**0.085**
...	...
zebra	0.001

Output

LLMs in Medicine

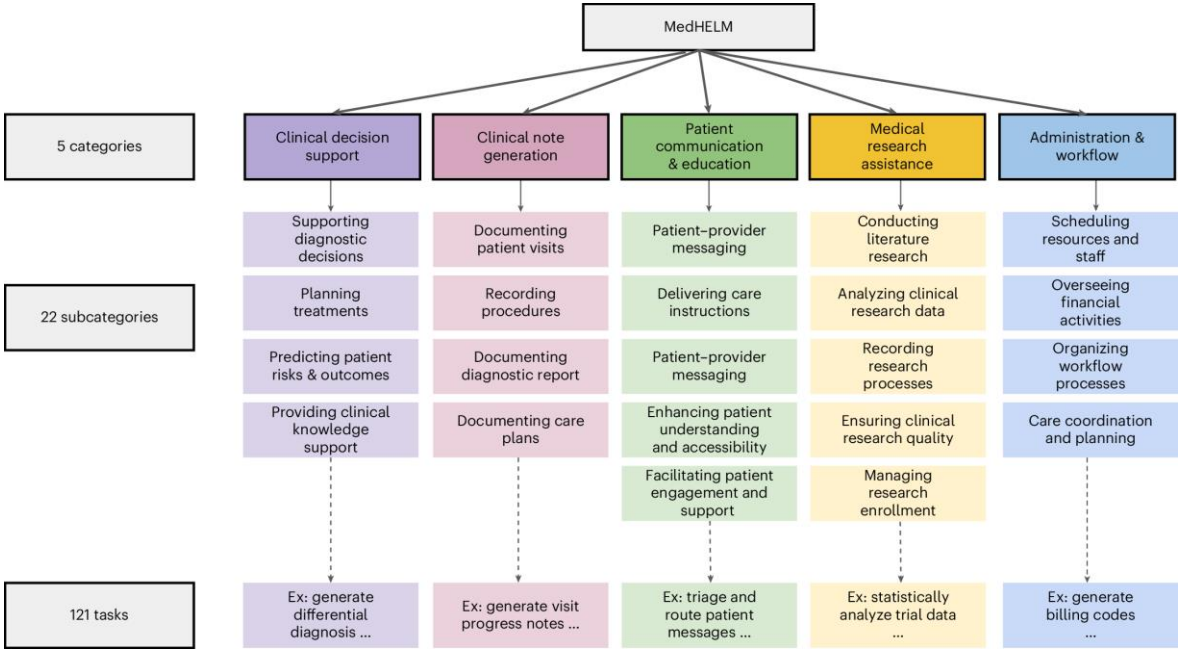
b



LLMs could support medical practice transversally, across many tasks

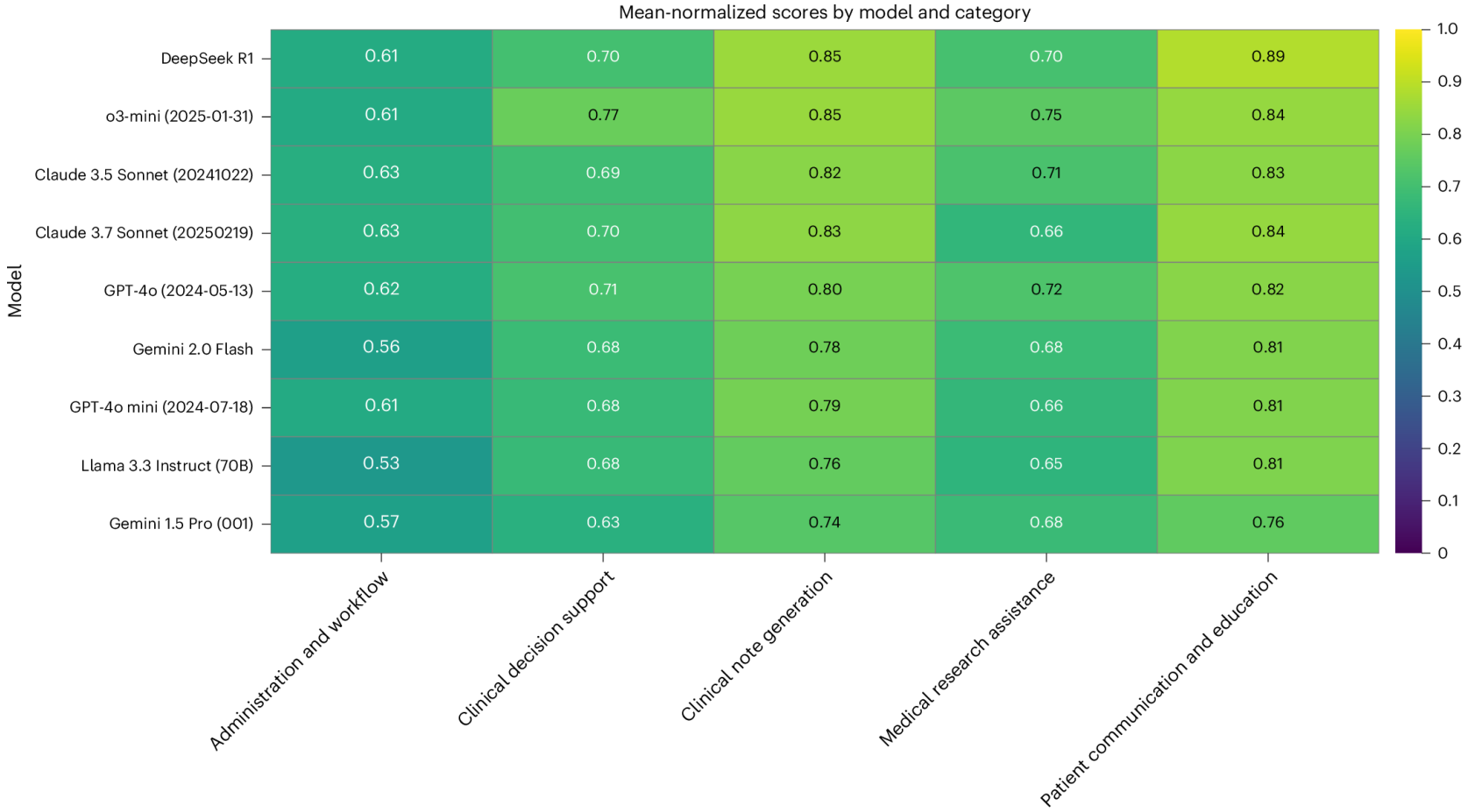
Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023).

LLMs show promising performance on multiple medical benchmarks



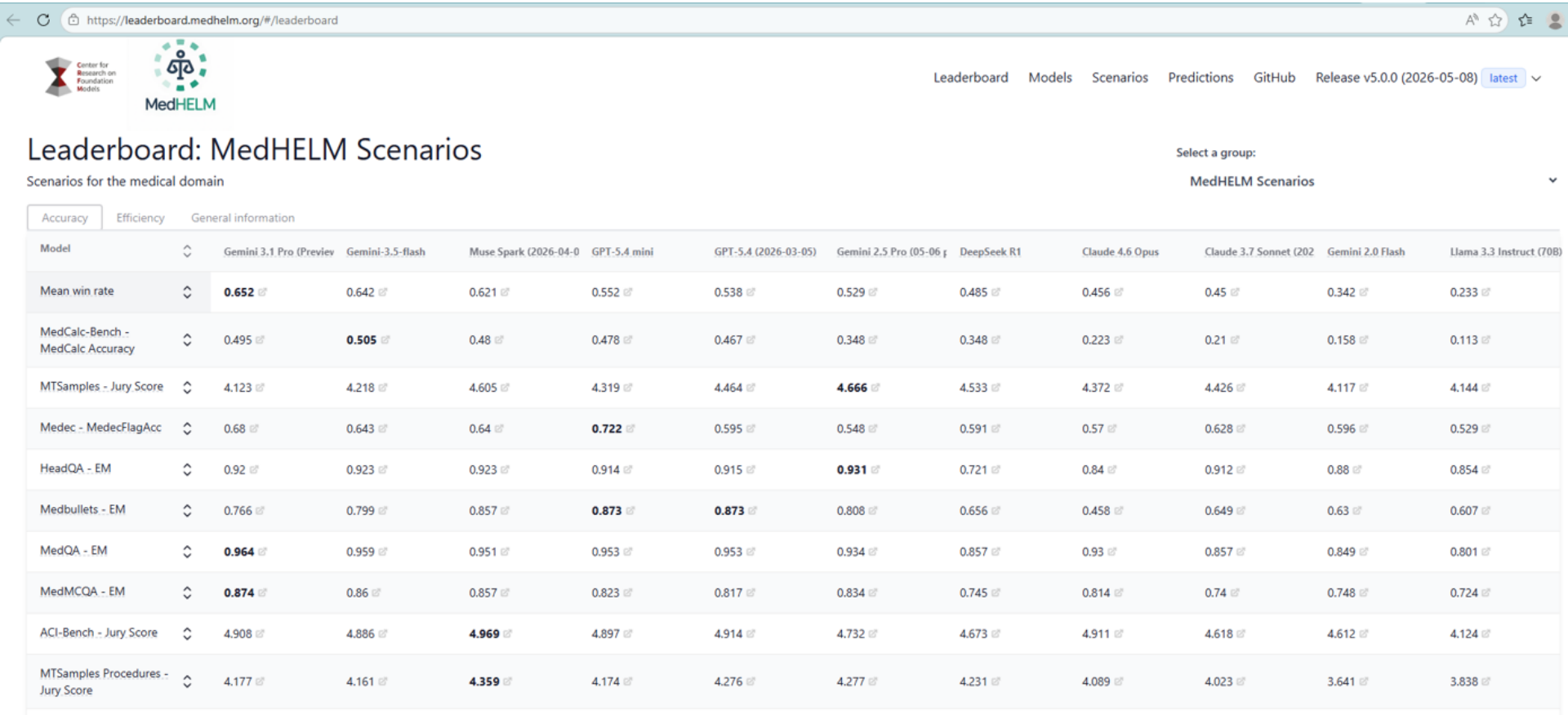
- **9** models
- **37** datasets

Bedi, S., Cui, H., Fuentes, M. *et al.* Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med*



Bedi, S., Cui, H., Fuentes, M. *et al.* Holistic evaluation of large language models for medical tasks with MedHELM. *Nat Med*

Living leaderboard



Model	Gemini 3.1 Pro (Previev	Gemini-3.5-flash	Muse Spark (2026-04-0	GPT-5.4 mini	GPT-5.4 (2026-03-05)	Gemini 2.5 Pro (05-06 p	DeepSeek R1	Claude 4.6 Opus	Claude 3.7 Sonnet (202	Gemini 2.0 Flash	Llama 3.3 Instruct (70B)
Mean win rate	**0.652**	0.642	0.621	0.552	0.538	0.529	0.485	0.456	0.45	0.342	0.233
MedCalc-Bench - MedCalc Accuracy	0.495	**0.505**	0.48	0.478	0.467	0.348	0.348	0.223	0.21	0.158	0.113
MTSamples - Jury Score	4.123	4.218	4.605	4.319	4.464	**4.666**	4.533	4.372	4.426	4.117	4.144
Medec - MedecFlagAcc	0.68	0.643	0.64	**0.722**	0.595	0.548	0.591	0.57	0.628	0.596	0.529
HeadQA - EM	0.92	0.923	0.923	0.914	0.915	**0.931**	0.721	0.84	0.912	0.88	0.854
Medbullets - EM	0.766	0.799	0.857	**0.873**	**0.873**	0.808	0.656	0.458	0.649	0.63	0.607
MedQA - EM	**0.964**	0.959	0.951	0.953	0.953	0.934	0.857	0.93	0.857	0.849	0.801
MedMCQA - EM	**0.874**	0.86	0.857	0.823	0.817	0.834	0.745	0.814	0.74	0.748	0.724
ACI-Bench - Jury Score	4.908	4.886	**4.969**	4.897	4.914	4.732	4.673	4.911	4.618	4.612	4.124
MTSamples Procedures - Jury Score	4.177	4.161	**4.359**	4.174	4.276	4.277	4.231	4.089	4.023	3.641	3.838

Benchmark performance do not necessarily translate in clinical practice

- Benchmark fail to capture the complexity of real practice
- Performance are sensitive to institutional and regional contexts
- Benchmarks do not capture multidimensional dimensions such as bias, harmfulness and alignment to local guidelines

Human-in-the-loop approaches have been proposed as key drivers for LLM adaptation and contextualization

Is it feasible?

useful?

needed?

Why an in-house clinically aligned LLM is needed?

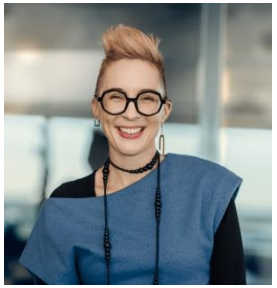
- Patient data privacy
- Context-adaptation
- Control and governance over LLM lifecycle
- Integration with EHR
- No vendor lock-in

Assessing feasibility and effect of human-in-the loop approach at CHUV

Giorgia Carra

Prof. Jean-Louis Raisaro

Prof. Noemie Boillat



Prof. Mary-Anne Hartley

CHUV Centre hospitalier universitaire vaudois

LiGHT Laboratory for Intelligent Global Health & Humanitarian Response Technologies.

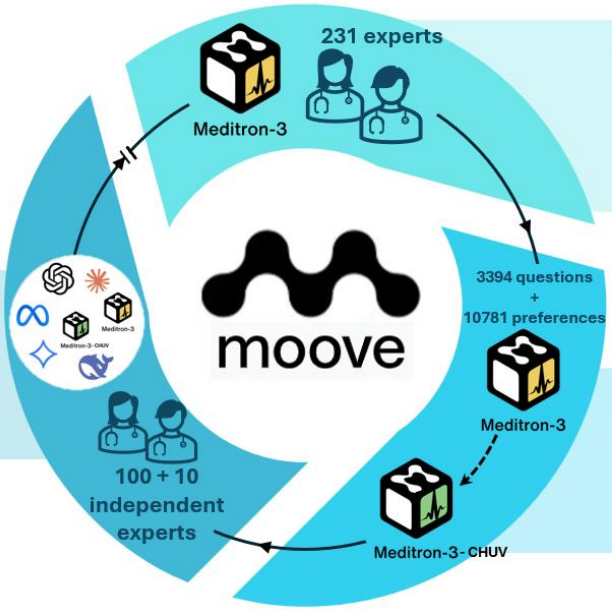
EPFL

	Meditron 3 8B	Meditron 3 70B	Llama 3.1 70B	MedPalm 2 540B	GPT4-Base 1750B
MedQA 4-options	69.9	86.0	79.2	79.7	86.1
MedMCQA	61.1	75.9	72.4	71.3	73.6
PubmedQA	79.2	81.2	79.9	79.2	80.4

Meditron

EPFL

Carra et al. under review



PHASE 1

Data collection

PHASE 2

Model alignment

PHASE 3

Models evaluation

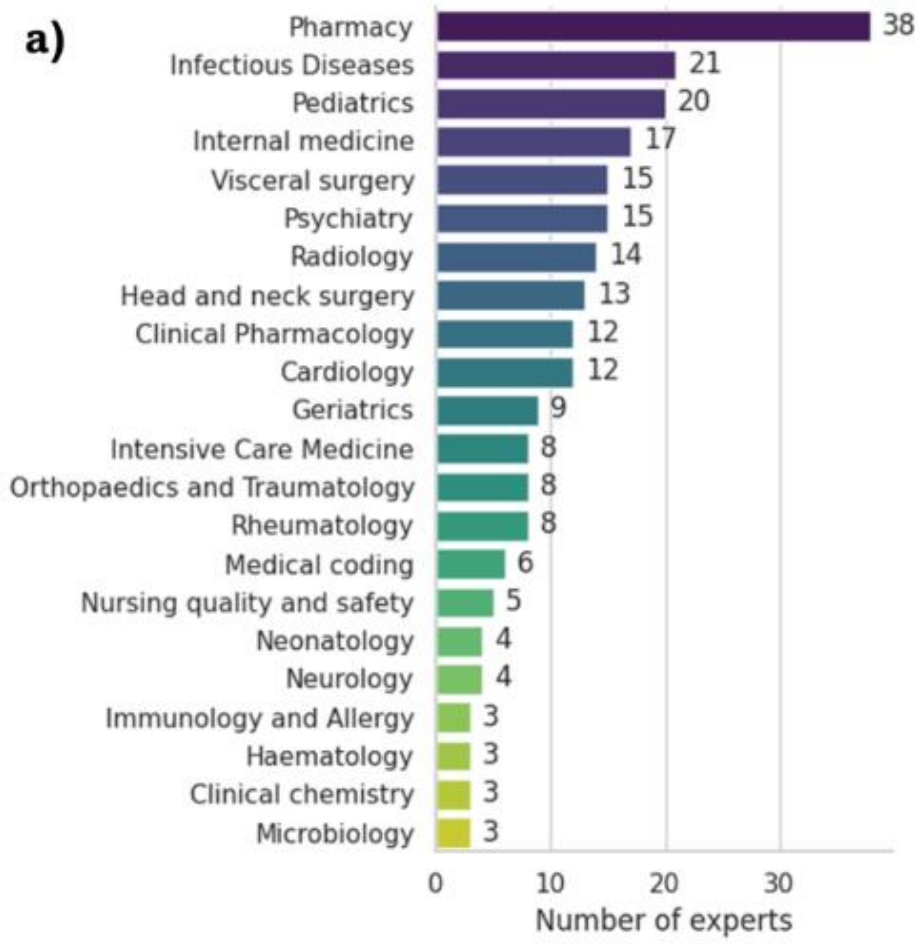
Centre hospitalier universitaire vaudois

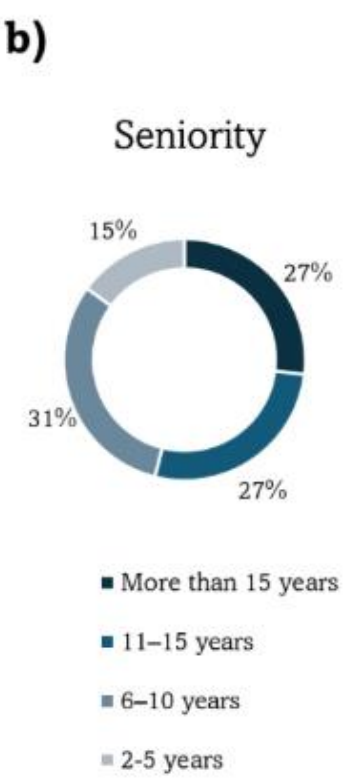
EPFL

LiGHT
Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Carra et al. under review

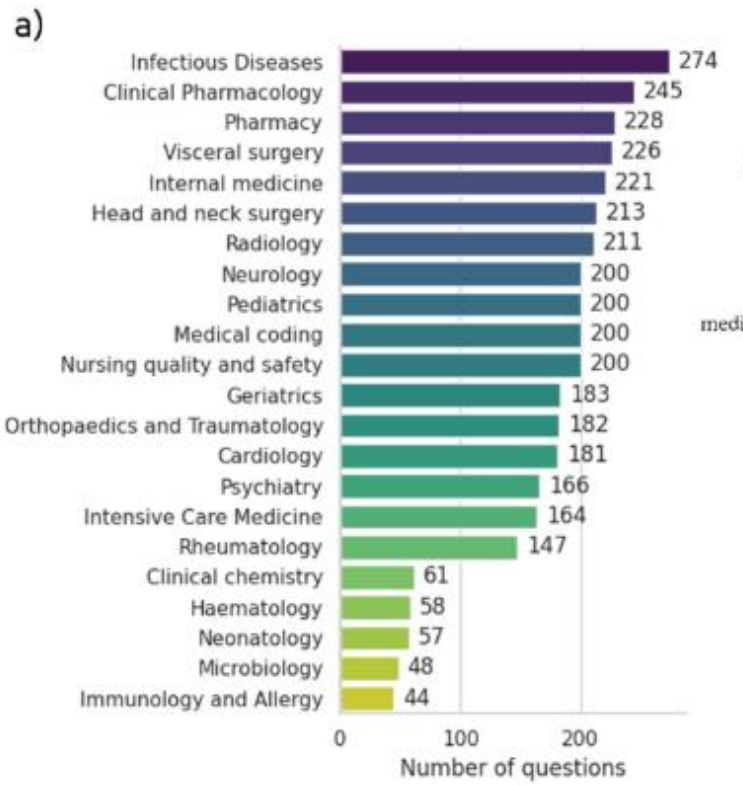
231 experts participated to the study

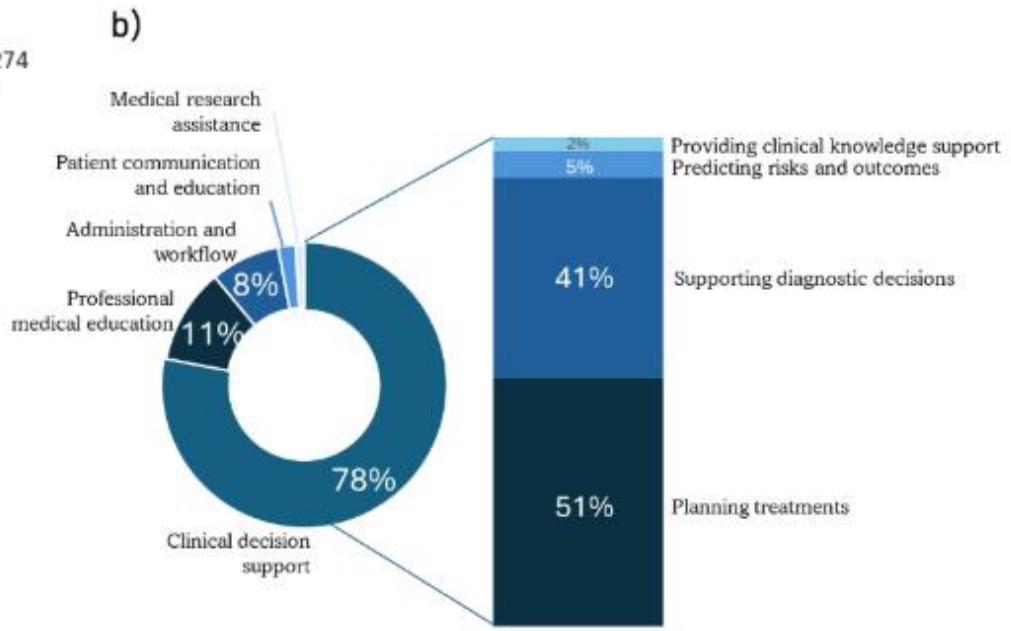




Carra et al. under review

3394 vignettes (10k+ evaluations)





c)

"I am a neurology specialist at the CHUV in Lausanne, who has been asked a question by the emergency room physician. He is taking care of a 42-year-old woman who is 24 weeks pregnant, presenting with a first generalized tonic-clonic seizure. She has no prior epilepsy history but is known for preeclampsia. Blood pressure is 180/110 mmHg, and fundoscopy shows papilledema. What are the differential diagnoses and immediate management priorities in this case?"

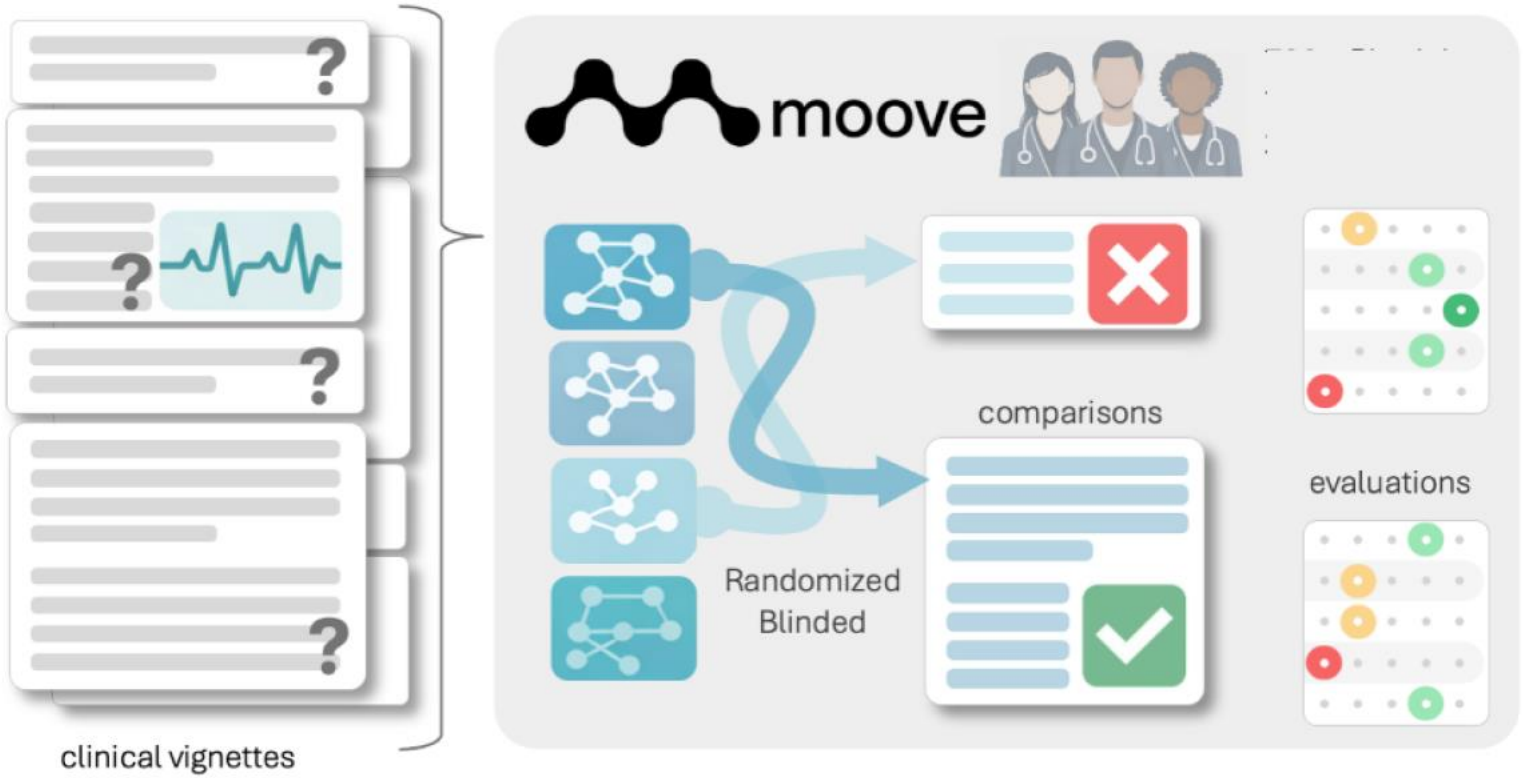
Context

Clinical vignette

Question

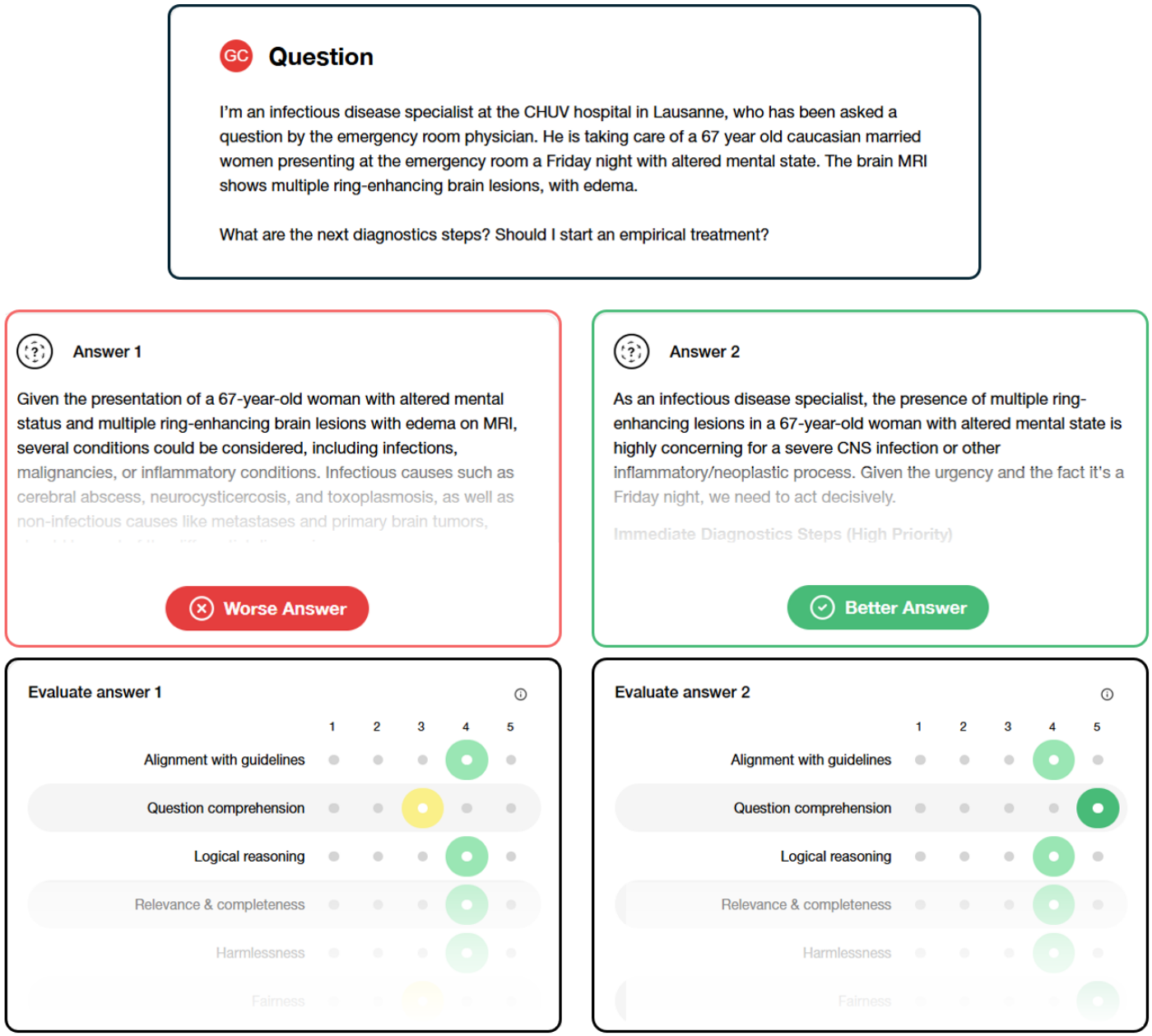
Carra et al. under review

Collecting clinician's preferences with the MOOVE platform

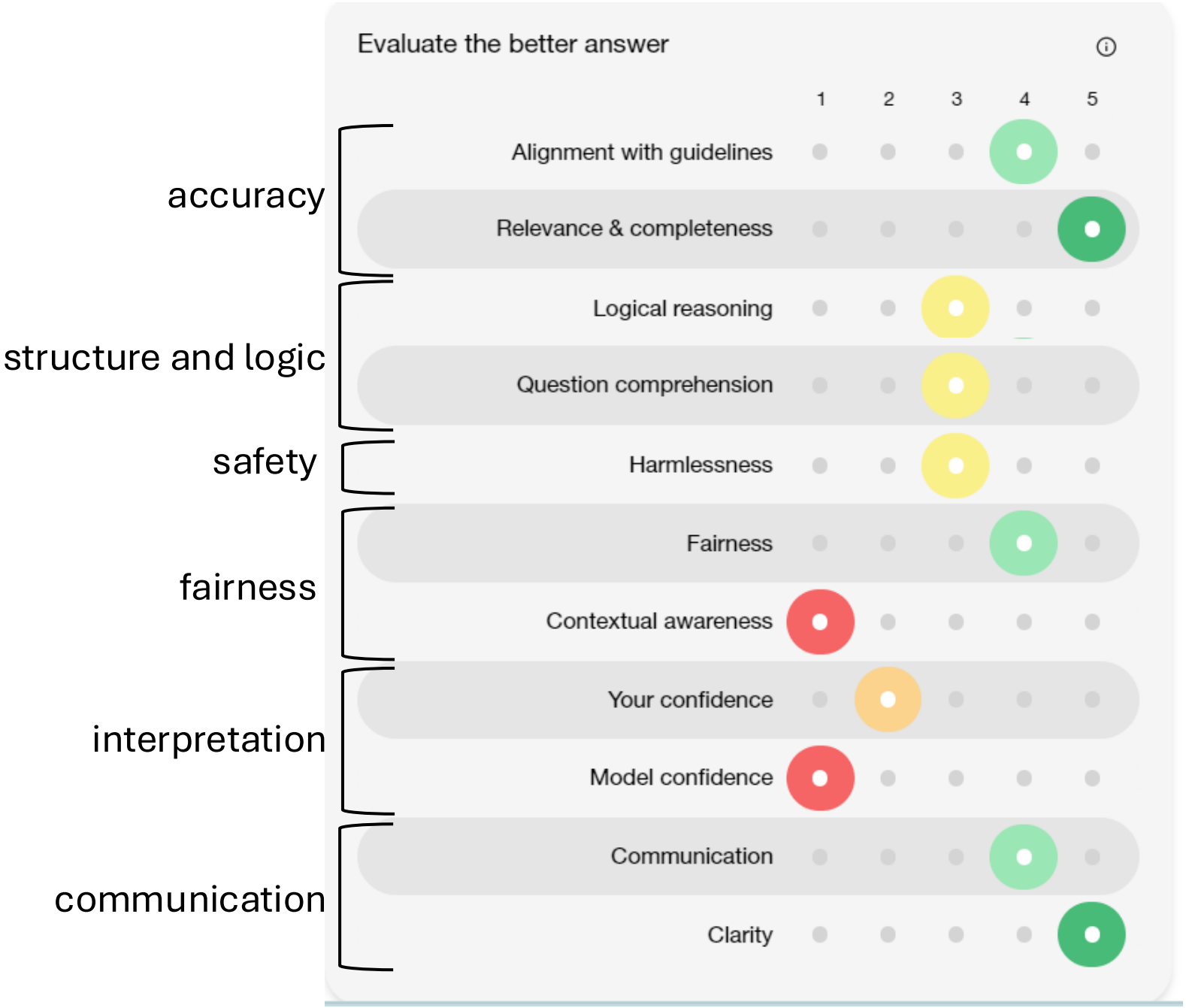


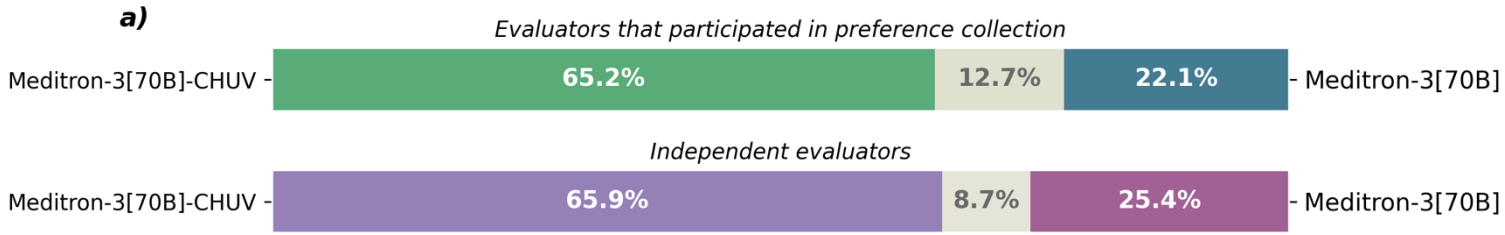
EPFL

LiGHT
Laboratory for intelligent
Global Health & Humanitarian
Response Technologies.

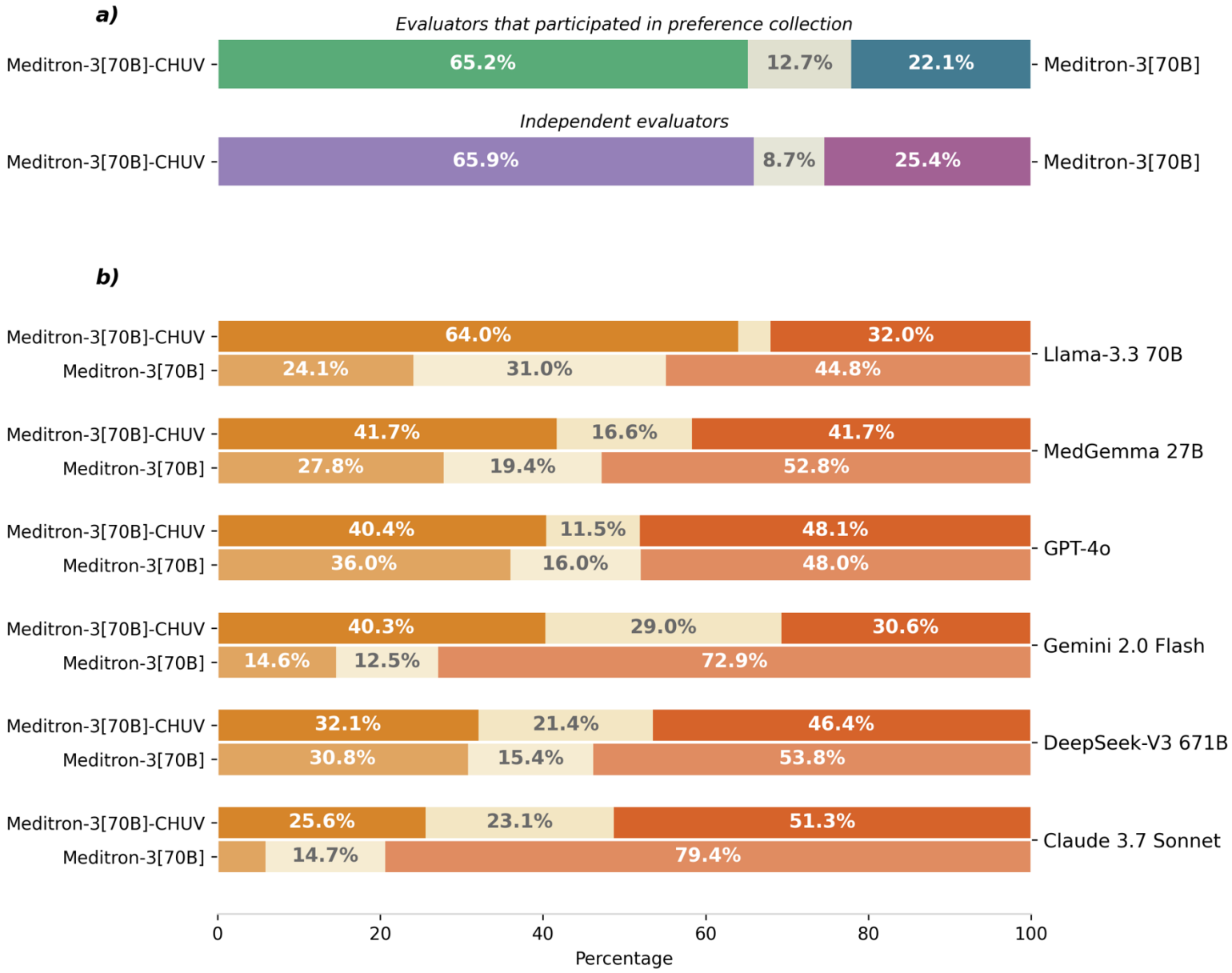


Carra et al. under review

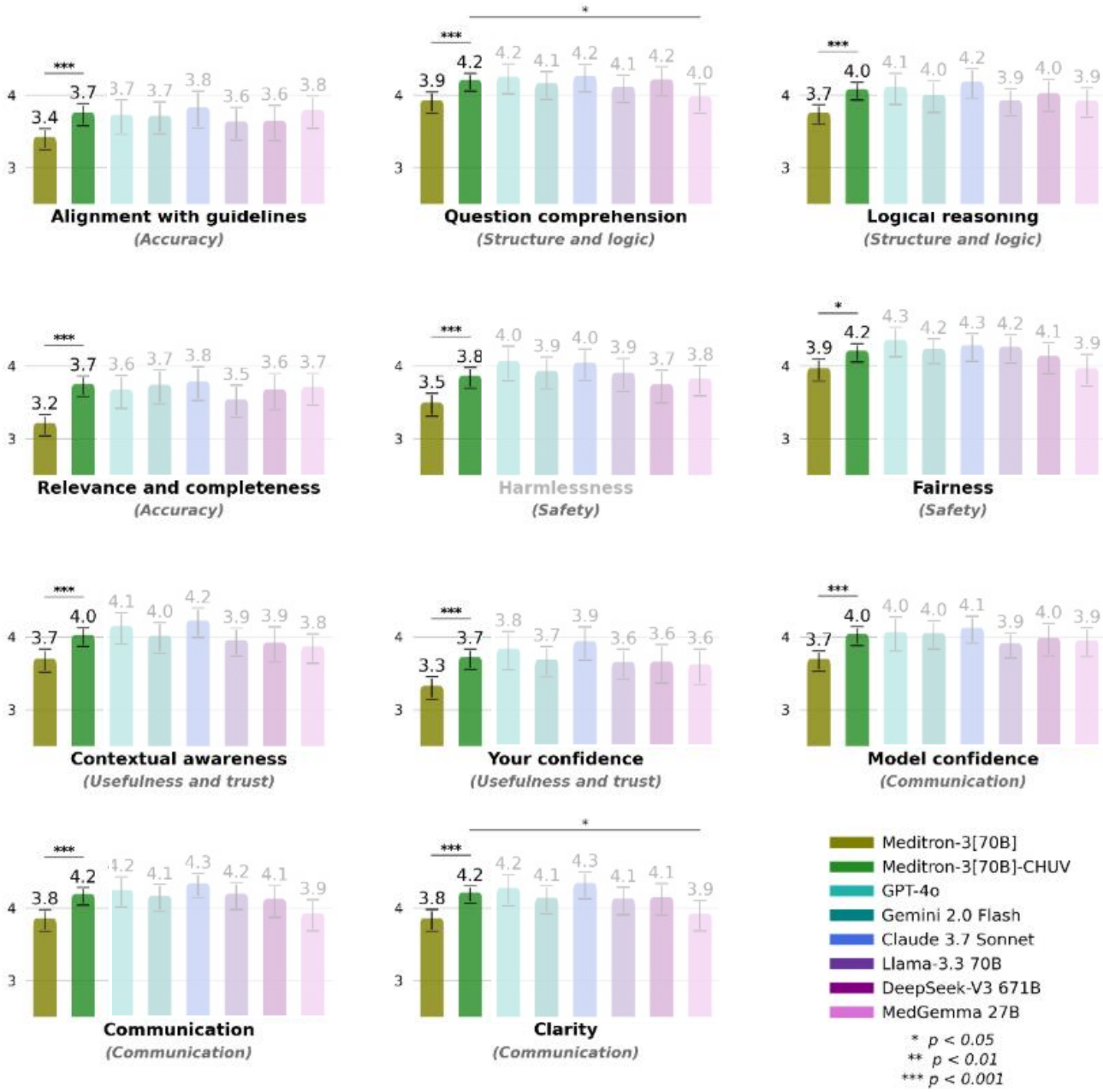

Evaluate the better answer
1
2
3
4
5
Alignment with guidelines
Relevance & completeness
Logical reasoning
Question comprehension
Harmlessness
Fairness
Contextual awareness
Your confidence
Model confidence
Communication
Clarity
accuracy
structure and logic
safety
fairness
interpretation
communication



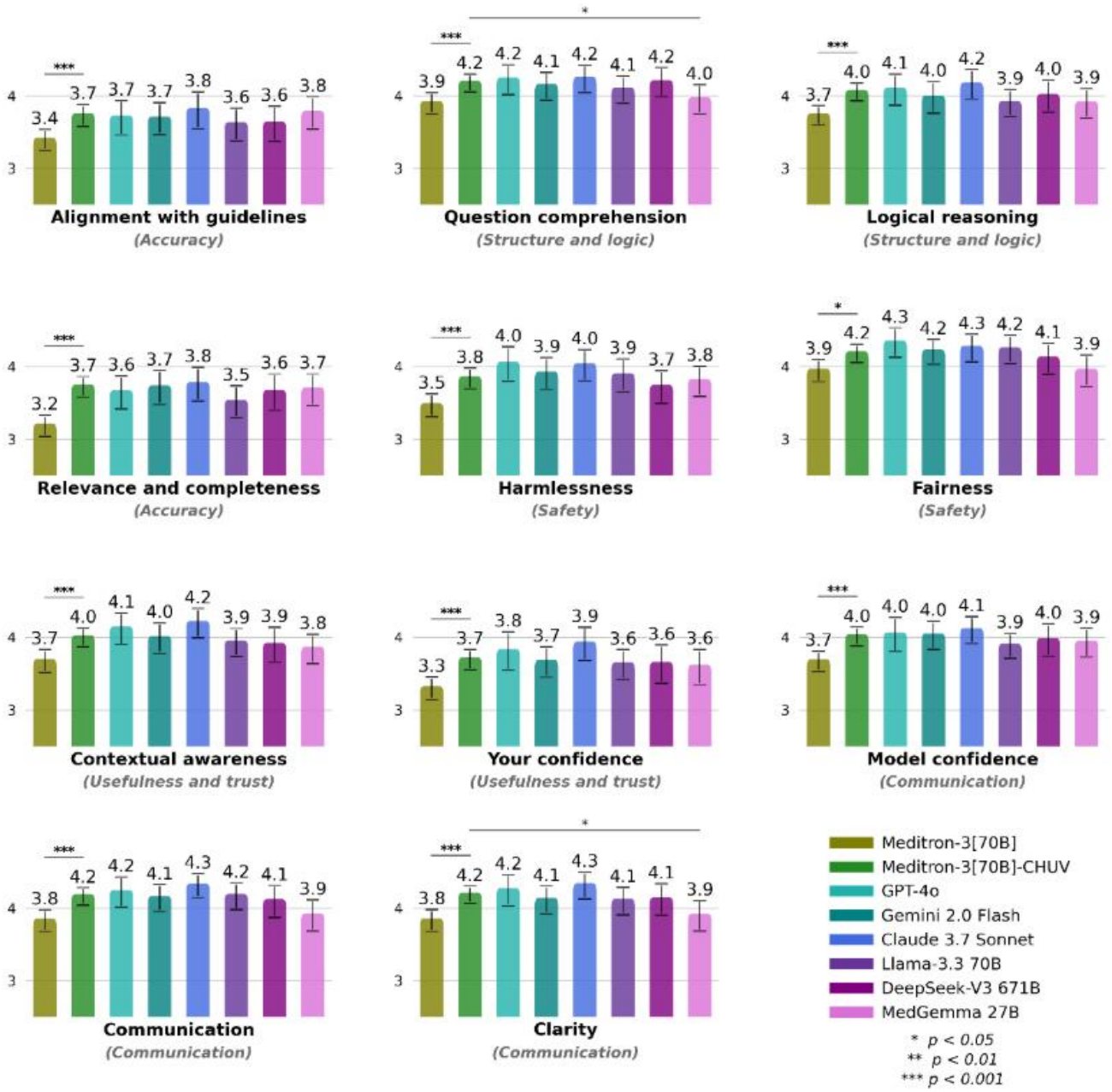
Carra et al. under review



Carra et al. under review

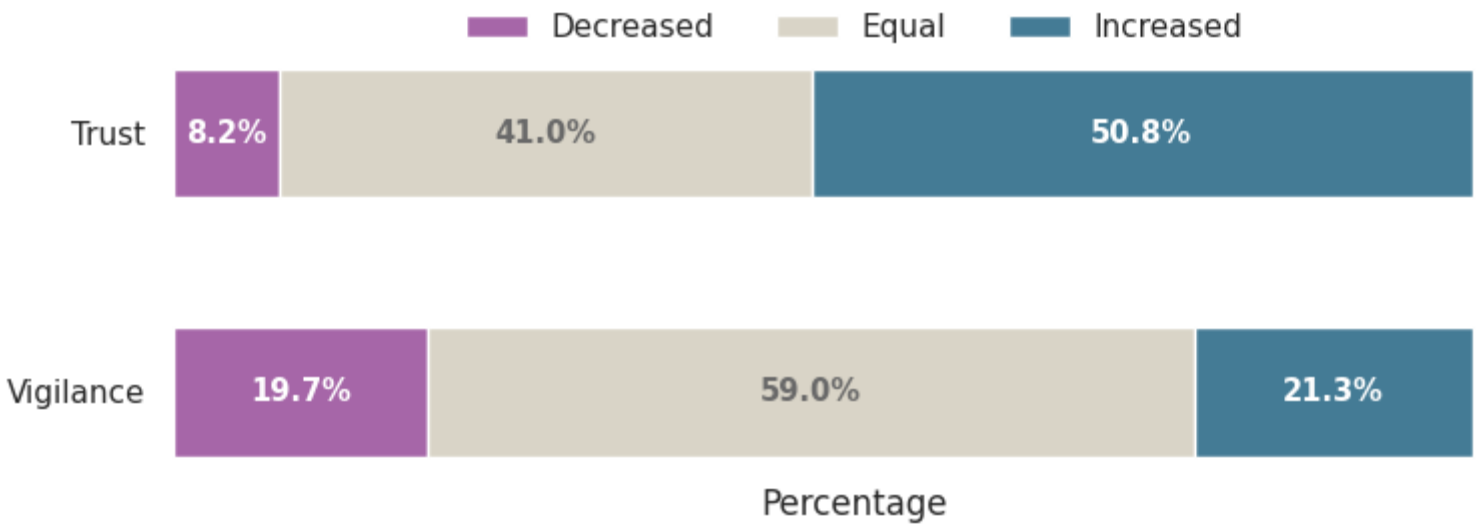


Carra et al. under review



Carra et al. under review

Impact of human-in-the-loop for trust and vigilance



Carra et al. under review

Expert-driven alignment **enhances model's performance** and bridges the gap between **open source and proprietary frontiers models**

Expert-driven alignment **enhances model's performance** and bridges the gap between **open source and proprietary frontiers models**

Experts' engagement in models development and evaluation **builds trust** while **maintaining vigilance**

What could be the role of medical librarians?

- **Provide expert feedback on LLM answers**
  - Responses must be grounded in the best available research; quality of evidence should be summarized accurately.
  - Specialized knowledge (publishing strategy and copyright issues etc.)

- **Support** the creation and curation of **high-quality databases** and institutional repositories that LLM answers draw on.

- **Train healthcare professionals** to reconcile LLM answers with their own critical skills (database searching, critical appraisal of sources etc).

A special thank you to all the experts that participated to the study!