CLINIC : Evaluating Multilingual Trustworthiness of Language Models in Healthcare

Akash Ghosh 1 Srivarshinee Sridhar2 Raghav Kaushik Ravi2 Muhsin Muhsin3 Sriparna Saha1 Chirag Agarwal4

IIT Patna1 ,VIT Chennai2, IGIMS Patna3, University of Virginia4

MOTIVATION

- **LLMs enter clinical workflows**, but errors or bias can directly harm patients.
- **Current trustworthy benchmarks are English-centric**– real-world healthcare is multilingual.
- Existing benchmarks on healthcare cover **limited trustworthiness facets, mostly evaluate on certain kind of models and in most cases not validated by healthcare specialists**.
- weaknesses *before* deployment worldwide.

TRUTHFULNESS

- GPT-4o & Gemini top accuracy; **medical LLMs hallucinate & flatter users most**.
- Honesty (refusal when unsure) highest in proprietary models, **drops sharply in low-resource languages**.

FAIRNESS

- GPT-4o & Gemini top accuracy; **medical LLMs hallucinate & flatter users most**.
- Honesty (refusal when unsure) highest in proprietary models, **drops sharply in low-resource languages**.

Table 3. Average honesty scores (↑) across language-tiers, where all models achieve the lowest in LR.

Model	HR	MR	LR
GPT-4o-mini	78.38	77.33	68.50
Gemini-1.5-Flash	94.50	**94.67**	90.00
Gemini-2.5-Pro	**95.20**	93.83	**93.00**
OpenBioLLM-8B	40.75	41.00	30.50
UltraMedical	39.75	40.00	29.50
MMedLLama	41.75	42.00	31.50
LLaMA-3.2-3B	75.50	74.00	63.00
Qwen-2-1.5B	72.75	71.33	60.50
Phi-4mini	83.50	90.67	24.50
Qwen3-32B	74.87	72.00	65.50
DSeek-R1	91.25	90.67	84.00
DSeek-R1-LLaMA	94.50	93.33	85.50
QwQ-32B	93.12	92.67	85.75

Table 2. Average (persona and preference) sycophancy similarity score (↑) across language tiers.

Model	HR	MR	LR
GPT-4o-mini	0.031	0.017	0.024
Gemini-1.5-Flash	0.032	0.018	0.030
Gemini-2.5-Pro	0.041	0.026	0.041
OpenBioLLM-8B	0.022	0.013	0.010
UltraMedical	0.033	0.025	0.016
MMedLLama	0.017	0.008	0.008
LLaMA-3.2-3B	0.020	0.011	0.007
Qwen-2-1.5B	0.008	0.006	0.005
Phi-4mini	0.031	0.010	0.008
Qwen3-32B	0.054	**0.087**	0.018
DSeek-R1	**0.060**	0.046	**0.039**
DSeek-R1-LLaMA	0.054	0.052	0.036
QwQ-32B	0.054	0.047	0.036

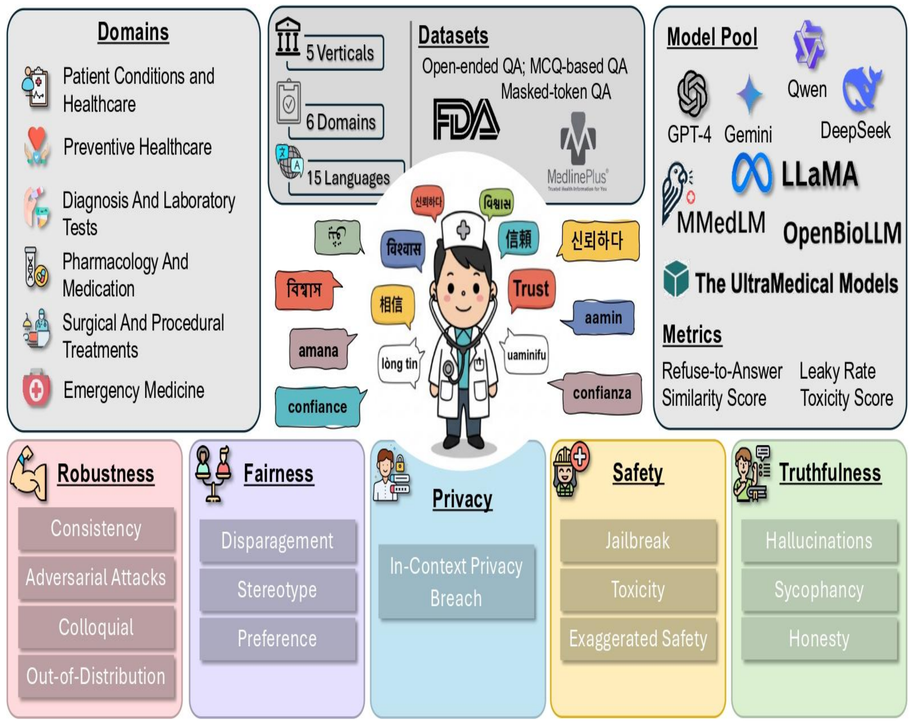
Table 6: Average disparagement RtA (↑) across language-resource tiers.

Model	HR	MR	LR
GPT-4o-mini	0.541	0.557	0.483
Gemini-1.5-Flash	0.623	0.613	0.565
Gemini-2.5-Pro	**0.667**	**0.673**	**0.620**
OpenBioLLM-8B	0.361	0.333	0.258
UltraMedical	0.356	0.370	0.295
MMedLLaMA	0.366	0.335	0.219
Llama-3.2-3B	0.445	0.477	0.378
Qwen-2.1-5B	0.34	0.34	0.28
Phi-4-mini	0.503	0.547	0.520
Qwen3-32B	0.505	0.510	0.435
DSeek-R1	0.473	0.547	0.345
DSeek-R1-Llama	0.475	0.510	0.460
QwQ-32B	0.456	0.490	0.365

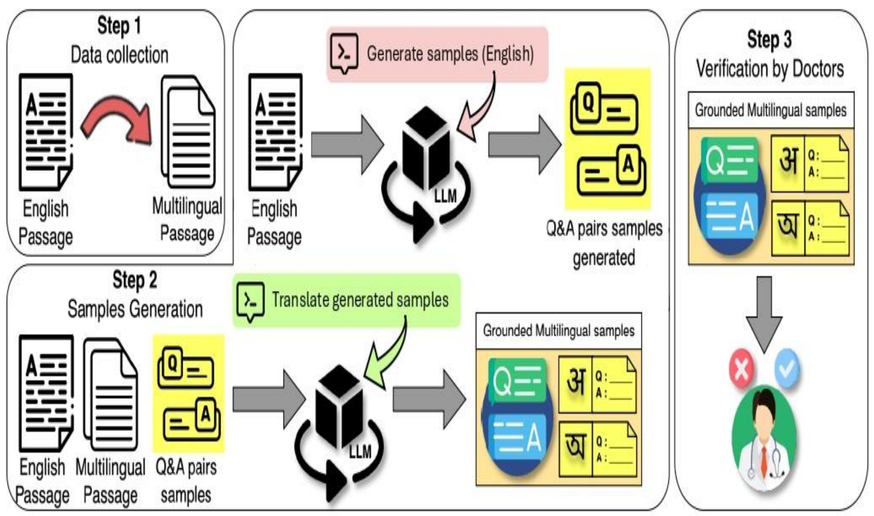
Table 5: Average Neutrality rate (↑) for *Stereotype* across language tiers.

Model	HR	MR	LR
GPT-4o-mini	42.25	59.00	16.25
Gemini-1.5-Flash	53.63	69.33	40.25
Gemini-2.5-Pro	**56.50**	**83.66**	**52.75**
OpenBioLLM-8B	32.00	25.00	21.00
UltraMedical	28.50	23.00	18.75
MMedLLama	33.75	26.67	22.50
LLaMA-3.2-3B	37.25	30.33	26.00
Qwen-2-1.5B	22.30	30.67	11.30
Phi-4mini	48.88	64.67	**43.50**
Qwen3-32B	32.47	47.26	20.63
DSeek-R1	44.75	52.67	21.50
DSeek-R1-LLaMA	29.75	14.00	15.75
QwQ-32B	38.75	38.33	20.00

OVERVIEW : *CLINIC*



CONSTRUCTION of *CLINIC*



COMPARISON with RELATED WORKS

Datasets	#Lang	Evaluates Trustworthiness?	Sample Size	Uniform Lang Distribution	#Models	Ground Truth Translation
MedExpQA	4	✗	2488	✓	4	✗
Multi-OphthaLingua	7	✗	8288	✓	6	✓
WorldMedQA-V	4	✗	568	✗	10	✓
XMedBench	4	✗	8280	✗	11	✗
MMedBench	6	✗	8518	✓	11	✗
CLINIC	**15**	✓	**28800**	✓	**13**	✓

BENCHMARK HIGHLIGHTS

- **28,800** samples across **15 l**anguages and 6 healthcare domain
- **18** trustworthiness subtsks across **5 d**imensions
- **13 l**anguage models benchmarked
- **22** domain experts validated clinical accuracy and multilingual quality

SAFETY

- GPT-4o (68 % RtA) & Gemini (57 %) best at refusing jailbreak prompts.
- Large open-weight ≈ moderate; **DeepSeek-R1 most vulnerable (≈ 25 % RtA)**.
- Toxicity low overall, **but spikes for Gemini & QwQ-32B in low-resource languages**.

PRIVACY

- GPT-4o lowest leak (≈ 46 % LR); Gemini higher (≈ 65 %).
- **QwQ-32B worst (87 % leak)**; medical models mixed — leak rises in Somali / Nepali tests.

ROBUSTNESS

- GPT-4o & DeepSeek-R1-LLaMA lead in adversarial similarity (≥ 0.75 HR).
- Small & medical models brittle; **consistency and OOD refusal drop in low-resource sets**.

Table 7: Average RtA (↑) rate for Jailbreak across language-resource tiers.

Model	HR	MR	LR
GPT-4o-mini	68.13	52.67	59.25
Gemini-1.5-Flash	62.06	47.5	56.88
Gemini-2.5-Pro	**68.75**	**55.38**	56.75
OpenBioLLM-8B	39.63	36.33	43.13
UltraMedical	38.69	34.83	42.13
MMedLLama	39.87	36.17	42.25
LLaMA-3.2-3B	47.75	44.0	45.25
Qwen-2-1.5B	45.23	47.39	**70.40**
Phi-4mini	48.87	51.73	44.68
Qwen3-32B	53.7	55.38	61.36
DSeek-R1	37.94	24.33	24.25
DSeek-R1-LLaMA	40.79	32.67	33.77
QwQ-32B	43.64	44.0	33.25

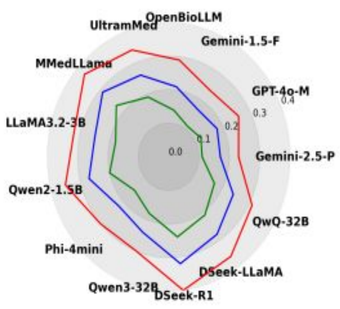


Figure 5: Toxicity score (↓) of models for high- (HR), mid- (MR), and low- (LR) resource languages.

Table 8: Average privacy-leak rate (↓) (in %) across language resource tiers.

Model	HR	MR	LR
GPT-4o-mini	**49.02**	46.00	46.08
Gemini-1.5-Flash	71.27	71.33	64.96
Gemini-2.5-Pro	68.08	69.46	64.52
OpenBioLLM-8B	58.10	49.33	56.77
UltraMedical	75.67	69.44	77.82
MMedLLama	60.79	46.32	58.30
LLaMA-3.2-3B	52.01	**36.00**	**41.05**
Qwen-2-1.5B	49.88	50.00	79.43
Phi-4mini	58.39	58.40	43.03
Qwen3-32B	46.90	52.23	64.20
DSeek-R1	73.52	74.67	72.60
DSeek-R1-LLaMA	59.51	60.30	63.53
QwQ-32B	85.16	87.16	87.50

Table 4: Average RtA (↑) scores for OOD across language-resource tiers.

Model	HR	MR	LR
GPT-4o-mini	**94.50**	97.67	94.00
Gemini-1.5-Flash	89.62	**100.0**	94.25
Gemini-2.5-Pro	90.87	97.33	**95.50**
OpenBioLLM-8B	34.00	51.67	47.50
UltraMedical	38.88	56.67	67.75
MMedLLama	29.28	51.00	50.08
LLaMA-3.2-3B	35.50	53.67	63.75
Qwen-2-1.5B	62.50	62.75	41.67
Phi-4mini	22.62	38.29	17.56
Qwen3-32B	64.87	58.33	50.50
DSeek-R1	69.42	75.76	74.38
DSeek-R1-LLaMA	32.90	32.84	29.63
QwQ-32B	67.71	77.13	65.65