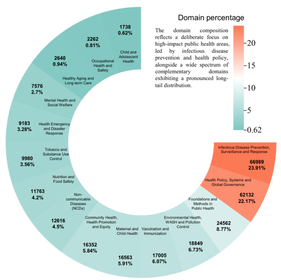
ICML 2026

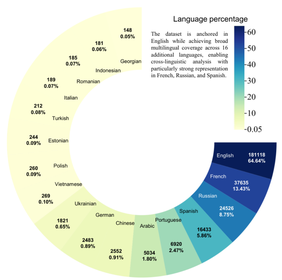
From Knowledge to Inference: Formalizing Specialized Public Health Reasoning on GlobalHealthAtlas

Zhaokun Yan*[1], Shan Xu*[12], Wuzheng Dong*[1], Zhaohan Liu*[1], Lijie Feng[3], Chengxiao Dai[4], Tianqi Chen[1], Binfan Liu[5] ,Yunpu Ma[6], Wenting Wei[7], Yingting Li[1], Yi Zhang[1], Tongning Wu[1]

[1] China Academy of Information and Communications Technology [2] Fudan University [3] CRRC Industrial Academy [4] The University of Sydney [5] BUPT [6] LMU Munich [7] Shanghai AI Laboratory

1. Motivation and Dataset





280,210	15	17	3
instances	domains	languages	difficulty levels

Why this dataset?

Public-health reasoning needs population-level inference, policy awareness, and safety constraints beyond ordinary medical QA.

Existing resources are often narrow, monolingual, and weak in evaluation support.

GlobalHealthAtlas integrates data construction, reasoning traces, and a public-health-specific evaluator.

Core contributions

- Large multilingual benchmark
- LLM-assisted construction pipeline
- Domain-aligned Public-Evaluator
- Benchmarking across 19 models

Data split

Training: 247,599
Test: 27,511
Evaluator set: 5,100
QA: 138,267 | SC: 141,943

Takeaway: this paper contributes not just a dataset, but a reproducible public-health reasoning framework.

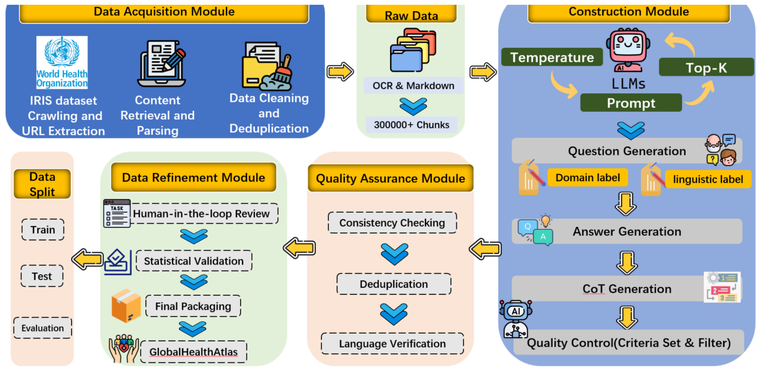
4. Open Resources

Codebase
github.com/Jan8217/GlobalHealthAtlas

Public-Evaluator
huggingface.co/aerovane0/GlobalHealthAtlas_Public_Evaluator

Public-Model
huggingface.co/aerovane0/GlobalHealthAtlas_Public_Model

2. Construction and Evaluation



Acquire & parse

Authoritative public-health documents are collected and converted into structured text.

Construct

LLMs generate questions, answers, domain labels, language labels, and CoT traces.

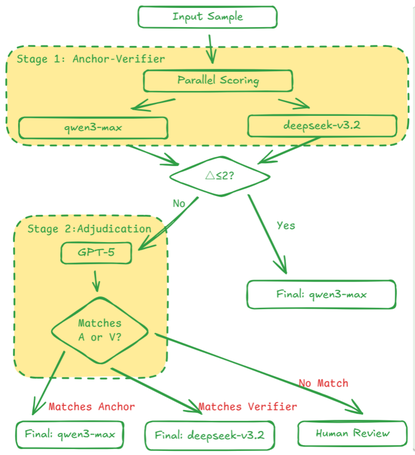
Refine & verify

Consistency filtering, deduplication, language validation, and human review improve quality.

Public-Evaluator

Accuracy · **Reasoning** · **Completeness** · **Consensus Alignment** · **Terminology Norms** · **Insightfulness**

MAE	ICC
1.4259	0.9735
Identical Rate	**StdDev**
0.5533	0.2772



Evaluation design

Anchor–verifier scoring creates high-quality supervision.

GPT-5 adjudicates disagreements between strong evaluators.

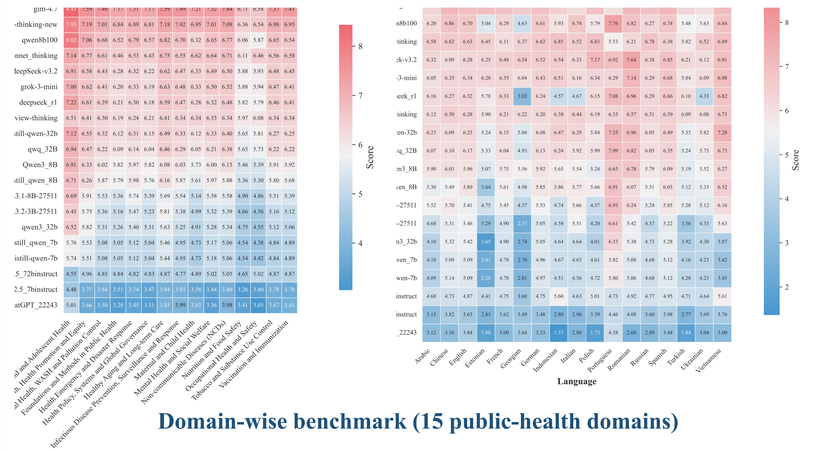
Human review is used when disagreement persists.

The result is a stable evaluator aligned with expert judgment.

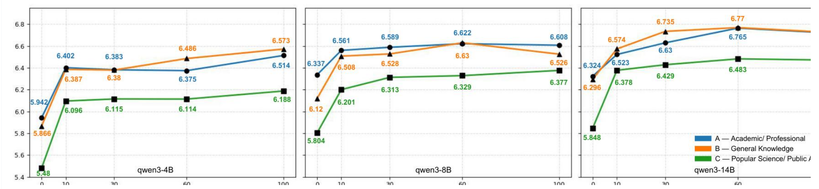
Compared with general evaluators and mainstream LLM judges, Public-Evaluator achieves the best agreement with human experts while maintaining the most stable scoring behavior.

3. Main Results

Model	Score	Domain (Representative)				Language			Difficulty			Type	
		IDP	HPG	V&I	CAH	EN	ZH	ES	A	B	C	SC	QA
Reasoning & Thinking Models													
gemini-3-flash-preview-thinking	6.282	6.344	6.212	6.343	6.506	6.284	6.498	6.388	6.499	6.194	6.328	6.625	5.911
grok-3-mini	6.352	6.478	6.187	6.410	6.997	6.343	6.350	6.683	6.884	6.148	6.413	7.133	5.551
gpt-5-mini	3.634	3.991	3.307	3.810	5.010	3.842	3.159	3.436	3.999	3.443	4.768	4.953	2.612
claude-sonnet-4-5-20250929-thinking	6.534	6.553	6.427	6.585	7.139	6.635	6.622	6.382	6.856	6.416	6.480	7.006	**6.047**
kimi-k2-thinking	**6.930**	**7.023**	**6.806**	**6.945**	7.927	**6.985**	**7.188**	**7.153**	**7.632**	**6.659**	**7.001**	7.912	5.917
qwen Series													
qwen3-8b(Yang et al., 2025)	5.934	6.034	5.822	5.922	6.906	5.956	6.028	6.090	6.648	5.650	6.158	6.987	4.855
Public-Model(based on qwen3-8b)	6.619	6.696	6.566	6.541	**8.022**	6.702	6.857	6.735	7.595	6.230	6.954	**8.063**	5.141
qwen3-32b(Yang et al., 2025)	5.259	5.247	5.309	5.064	6.520	5.415	5.324	5.276	5.979	4.977	5.423	6.357	4.135
qwen2.5-7b-instruct(Qwen et al., 2025)	3.660	3.852	3.469	3.696	4.482	3.634	3.824	3.985	4.127	3.468	3.909	4.418	2.884
qwen2.5-72b-instruct(Qwen et al., 2025)	4.839	4.768	4.829	4.874	4.548	4.869	4.733	4.951	4.578	4.941	4.801	4.578	5.107
qwq-32b(Yang et al., 2025)	6.173	6.286	6.038	6.223	6.937	6.175	6.104	6.348	6.809	5.922	6.340	7.127	5.194
DeepSeek Series													
deepseek-r1(Guo et al., 2025)	6.335	6.470	6.180	6.414	7.217	6.316	6.271	6.661	7.007	6.068	6.554	7.308	5.341
deepseek-v3.2(DeepSeek-AI et al., 2025)	6.365	6.471	6.221	6.448	6.914	6.278	6.089	6.847	6.897	6.159	6.448	7.142	5.569
deepseek-r1-distill-qwen-7b	5.004	4.947	5.039	4.891	5.735	5.091	5.136	5.123	5.570	4.780	5.164	5.899	4.087
deepseek-r1-distill-llama-8b	5.822	5.873	5.765	5.678	6.715	5.892	5.491	6.028	6.473	5.573	5.961	6.785	4.840
deepseek-r1-distill-qwen-32b	6.238	6.326	6.148	6.245	7.122	6.230	6.091	6.487	6.893	5.983	6.373	7.237	5.215
Other Mainstream Open-Source Series													
Llama-3.2-3B-Instruct(Grattafiori et al., 2024)	5.276	5.380	5.233	5.115	6.453	5.465	5.311	5.224	6.183	4.925	5.421	6.589	3.933
Llama-3.1-8B-Instruct	5.464	5.539	5.394	5.386	6.689	5.406	5.696	5.854	6.261	5.149	5.666	6.722	4.174
glm-4.7	**7.326**	**7.490**	**7.175**	**7.433**	**8.432**	**7.314**	**7.351**	**7.545**	**8.108**	**7.019**	**7.525**	**8.484**	**6.151**



Domain-wise benchmark (15 public-health domains)



Language-wise benchmark (17 languages)

Key findings

Reasoning-oriented frontier models achieve the highest overall scores, especially on academic/professional tasks.

The domain-specific Public-Model substantially improves over the vanilla qwen3-8b baseline.

Performance differences across languages and domains reveal persistent cross-lingual and cross-domain reasoning gaps.