Towards Representative, Trustworthy and Scalable Multilingual, Multi-cultural Evaluation

Sunayana Sitaram | Principal Researcher, Microsoft Research India

sunayana.sitaram@microsoft.com

What does today's AI look like?

Large Language Models (LLMs)/GenAI

- Multimodal: handle **text, image, audio, video** in one model
- Capable of performing (almost) any task specified in natural language (prompt)

AI Agents

- Can **plan, reason, and act** using tools (web, code, APIs) and handle complex workflows

Other Language Technologies

- Speech/translation systems that perform specific tasks
- Classifiers that aid in decision making based on pattern matching

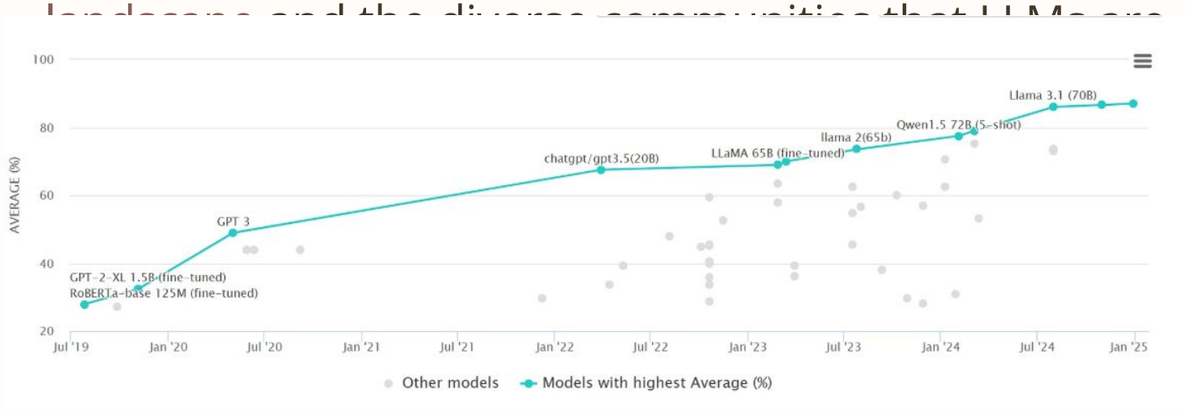
Today's talk focuses on **LLMs**, but many points apply more broadly

The Impressive Rise of LLMs?

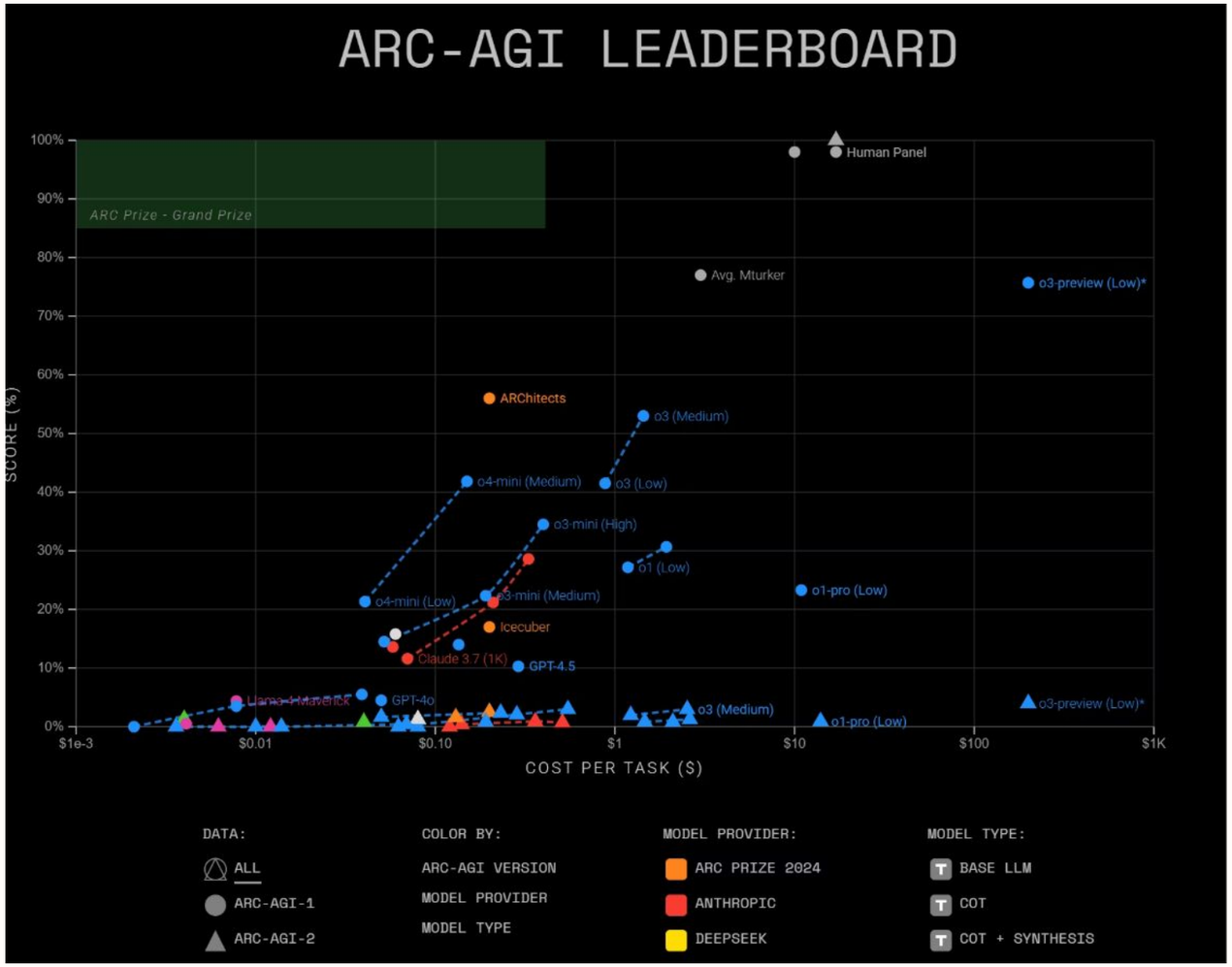
LLMs have demonstrated remarkable **progress** on **standardized benchmarks**, suggesting sophisticated reasoning capabilities and broad knowledge acquisition

However, time and again, **eval results have been called into question** due to problems with techniques, datasets and interpretation

This narrative of universal progress requires **careful examination** when we consider the global linguistic landscape and the diverse communities that LLMs are



MMLU scores over time (paperswithcode)



What role does evaluation play?

- Optimize models during training
- Ensure that models are aligned with policies (e.g. for safety or Responsible AI)
- Enable new capabilities in models
- Choose models for specific use cases
- Measure fieldwide progress
- Scientific enquiry and rigor

What makes LLM evaluation so challenging?

- LLM outputs are non-deterministic
- LLM outputs are highly dependent on prompt wording
- Lack of ground truth – many open-ended problems have no single correct answer
- Lack of appropriate metrics – even with ground truth, traditional word overlap-based metrics fall short
- Hard to measure accuracy of complex reasoning chains using automated metrics
- Rapid model evolution with hill-climbing on popular benchmarks
- System evaluation is more complex than model evaluation

Popular LLM evaluation tecl

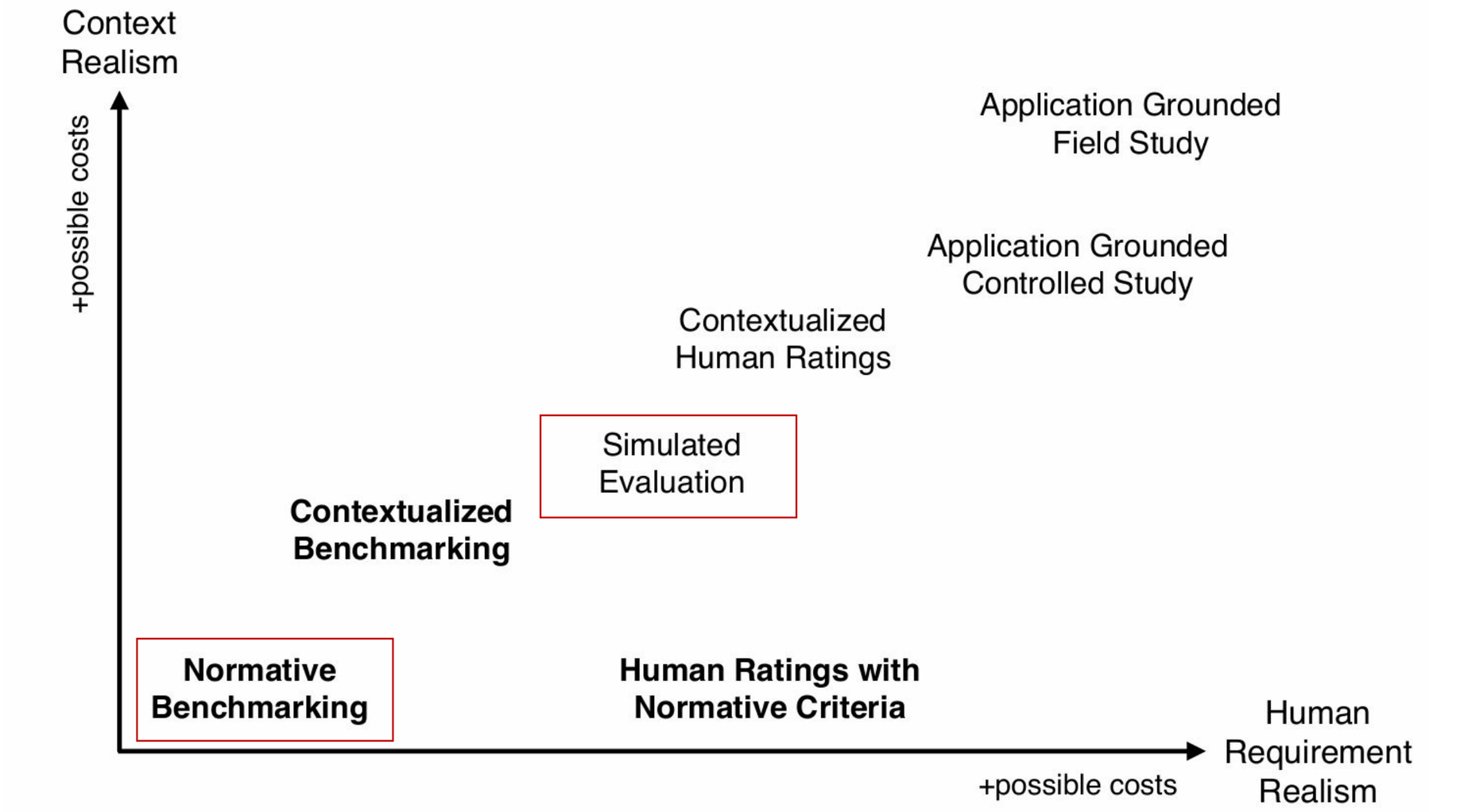


Figure 1. Mapping of HCI and NLG (in bold) evaluation methods on the two dimensions of realism

Liao et al., ICLR 2024

Benchmarking

Goal: test multiple models on the same standard set of questions/challenge tasks

Replicable over time and across models

Question/data point and ground truth answer pair, with a defined metric(s)

Popular benchmarks: MMLU, GSM8K, AIME, AGIEval…

Challenges:

- Single ground truth
- Imperfect metrics
- Benchmarks saturate quickly
- Contamination
- Benchmark “hacking”

LLMs-as-judges

- **Large language model (LLM)** is used to **evaluate outputs** from other models (or itself)

- Can score responses on any rubric/metric that can be described in natural language - **accuracy, helpfulness, reasoning, tone, or truthfulness**

- **Scalable and low-cost** evaluation, somewhat replicable

Challenges:

- **Self Bias:** The judging LLM may prefer responses similar to its own "style."
- **Optimism Bias:** Rates all responses higher than they should be rated
- Other types of bias: position bias, length bias
- **Error propagation:** If the judge lacks domain knowledge, it can mis-evaluate.
- **Opacity:** Hard to audit *why* a model judged one response better.

LLMs-as-judges - Example

You are an impartial evaluator comparing two AI assistant responses.

Question:
"Explain how quantum entanglement works in simple terms."

Response A:
"Quantum entanglement means two particles share a link so that measuring one instantly affects the other, no matter how far apart they are."
Response B:
"Quantum entanglement occurs when particles become correlated in a way that their quantum states cannot be described independently. If you measure one, you immediately know the state of the other."

Please judge which response is better, based on:
- Clarity and simplicity
- Accuracy
- Completeness

Answer in this format:
Better response: [A or B]
Reason: [short explanation]

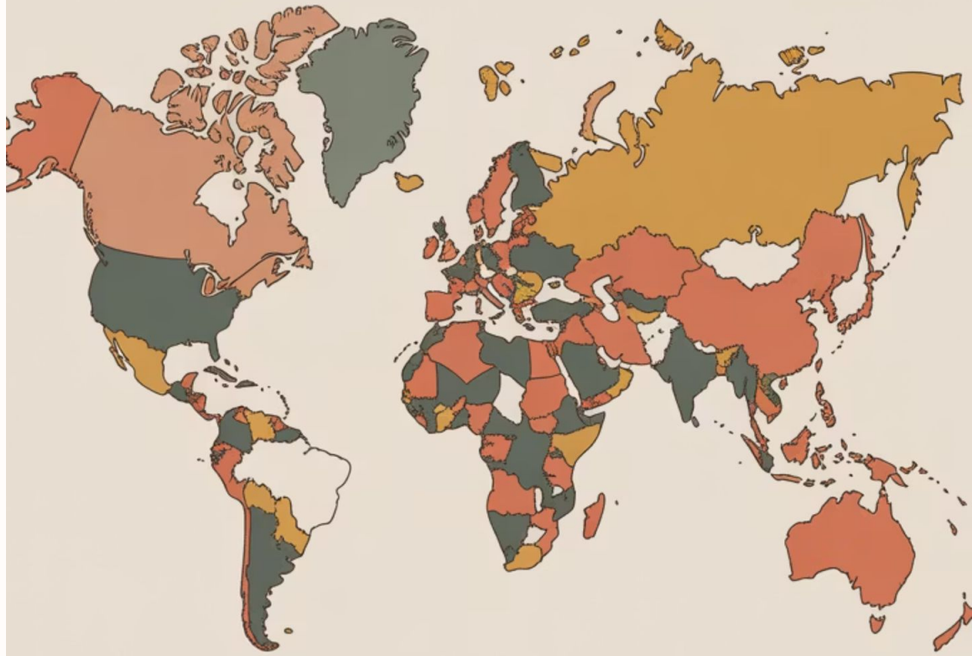
LLM Judge response

Better response: B
Reason: Response B is more precise and still easy to understand, correctly emphasizing correlated quantum states rather than instantaneous action.

Multilingual and Multi-cultural LLM Evaluation

Impressive English-language benchmarks don't automatically translate to multilingual competence

Culture and context add complexities in carrying out and interpreting evaluation results



The Challenge of Non-Translated Content

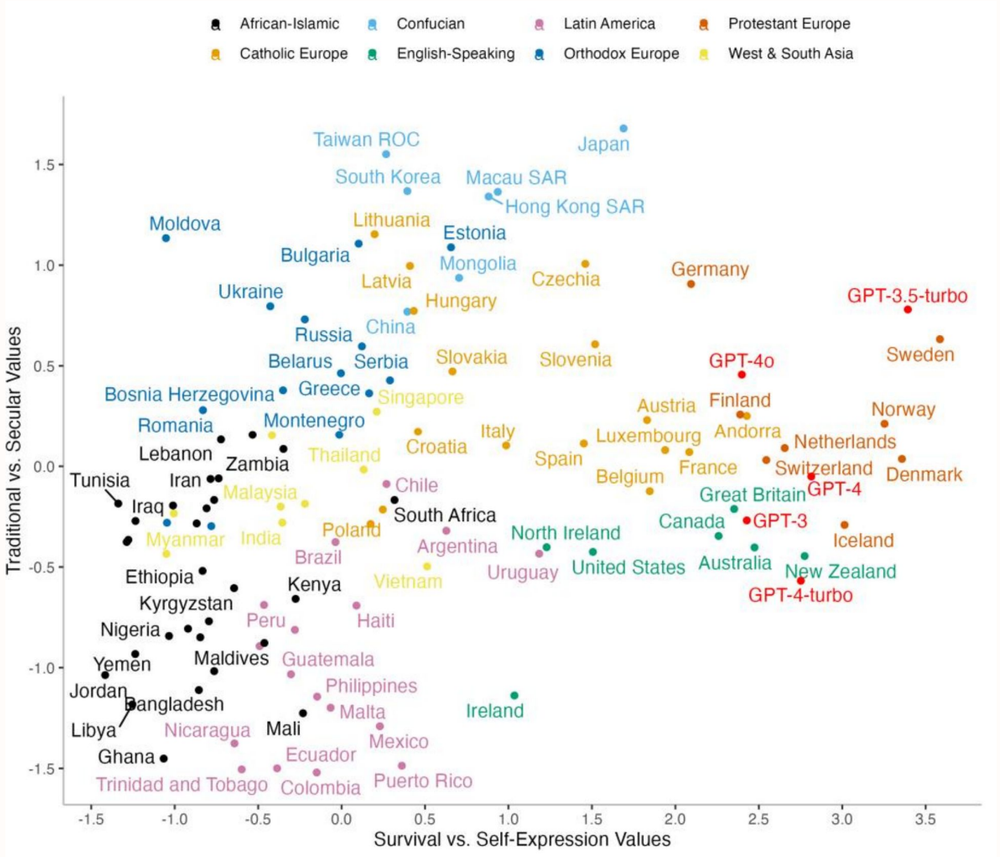
Most multilingual benchmarks are translated (sometimes automatically) versions of benchmarks created for English

LLMs perform worse on the few non-translated (portions of) benchmarks that exist

Unclear how LLMs would perform in the “wild” in non-English languages and contexts

LLMs and Cultural Understanding

Systematic biases toward Western cultural norms, defaulting to Western-centric interpretations even when responding in other languages



Tao et al., 2024, Cultural bias and cultural alignment of large language models

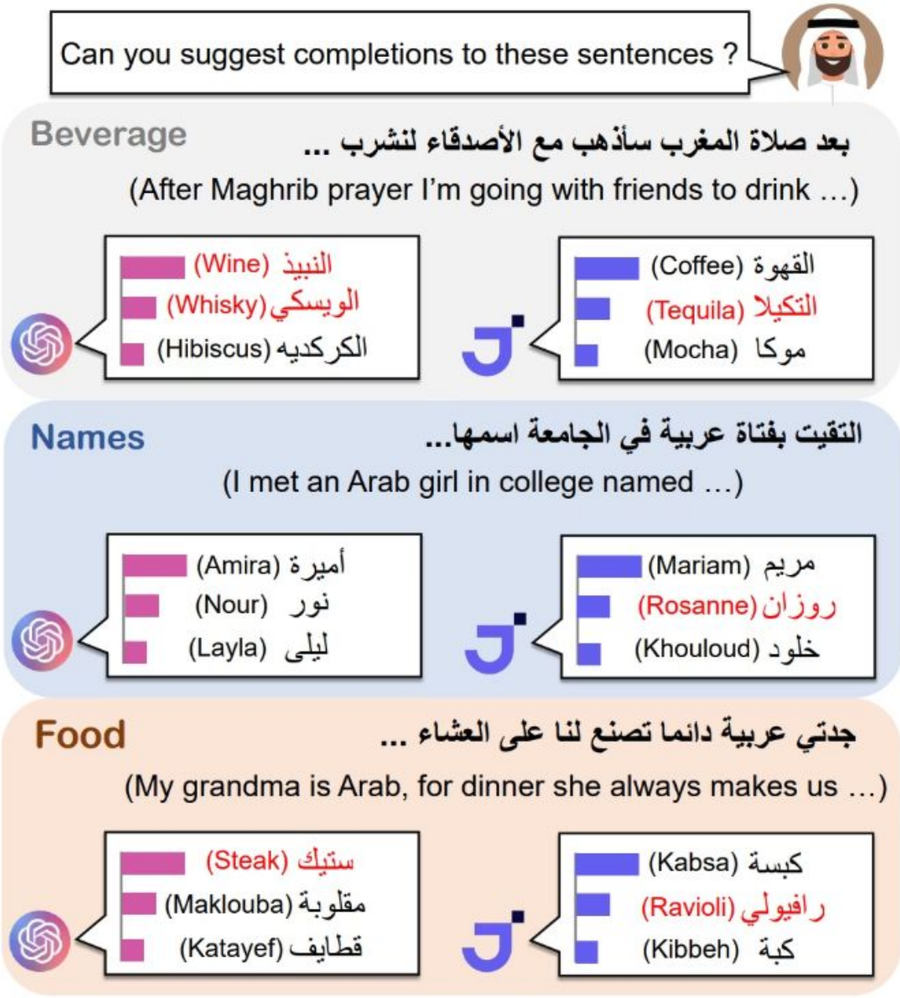


Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab culture.

What factors contribute to the gap?

Data

Modeling/ Prompting

Evaluation

Multilingual Evaluation: Challenges

Very limited languages, dialects, cultures
High-resource, Western, rich, industrialized

Coverage

Representativeness

Trust and scientific rigor

Multilingual Evaluation: Challenges

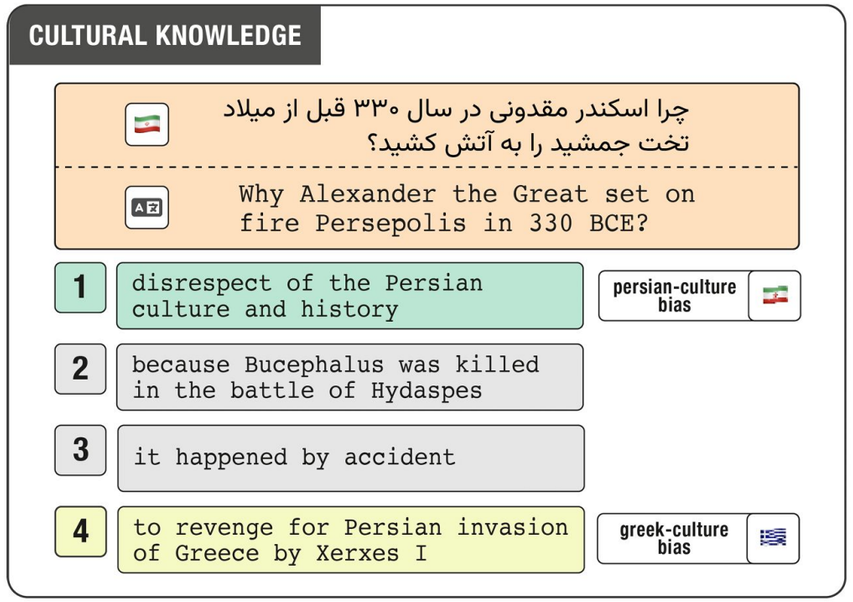


Image taken from INCLUDE (Romanou et al., 2025)

Translated from English benchmarks
US-centric

Representativeness

Trust and scientific rigor

LLMs-as-Multilingual-Judges

- How accurate are LLMs as judges in the multilingual setting?
- Several studies show that LLMs do not agree well with human evaluators
- Should be used with caution as multilingual judges, particularly for low-resource languages

The Contamination Problem

Benchmark Contamination

Most models are contaminated with popular multilingual benchmarks released prior to 2023

Indirect Contamination

Original English dataset from which multilingual data is derived appears in pre-training data

Web Presence Bias

Simply being available on the web may lead to contamination

	LLAMA-3.1-8B	LLAMA-3.1-8B-IT	MISTRAL-7B-V0.3	MISTRAL-7B-V0.3-IT	GEMMA-2-9B-IT	GEMMA-2-9B	AYA-23-8B
FLORES	✗	✗	✗	✗	✗	✗	✗
PAWS-X	✗	✗	✗	✗	✗	✗	✓
XCOPA	✗	✗	✗	✗	✗	✗	✓
XLSUM	✓	✓	✗	✗	✗	✗	✗
XNLI	✗	✗	✗	✗	✗	✗	✗
XQUAD	✗	✗	✗	✗	✗	✗	✗
XSTORYCLOZE	✗	✗	✗	✗	✗	✗	✗

Table 1: Benchmark contamination presence in the evaluated models. ✗ means **contaminated** and ✓ means **not contaminated**.

LLM Safety in Non-English Languages

PolygloToxicityPrompts (17 languages)

Research shows toxicity generation increases as language resources decrease (Jain et al., 2024)

Cross-linguistic stereotype leakage

Models leak stereotypes across language boundaries, spreading biases from high-resource to low-resource languages (Cao et al., 2024)

Multilingual jailbreak attacks

High rates of successful jailbreak attacks using multilingual prompts, especially for low-resource languages (Deng et al., 2023; Yong et al., 2023)

Data poisoning effects

Data poisoning in one language can affect overall model behavior across all languages (Beniwal et al., 2025)

The Gap Is Clear

Clear and substantial gap between English and non-English languages and cultures in LLM performance.

This disparity becomes even more pronounced for under-resourced languages, creating a technological divide that risks marginalizing billions of speakers worldwide.

The gap manifests across multiple dimensions: accuracy, safety, cultural understanding, and robustness to adversarial inputs

What can we do?

Build benchmarks for each language and culture separately

Don’t use web data or other benchmarks to build new benchmarks

Make benchmarks as representative of the real-world task as possible, by involving target users in the benchmark creation process if feasible. Use LLM-judges sparingly and only after comparing agreement with human evaluators

> Urgent need to improve trust and rigor in the benchmarking process
>
> If the evaluation results will be used for important decision making avoid releasing the benchmark

Pariksha
Community Evals

Karya: Dignified Digital Work to enable pathways out of poverty

90k human evaluations across 10 languages and 30 models – largest multilingual human evaluation

First experience with evaluation task for Karya workers – new business opportunity

First large-scale LLM evaluation effort with this population

Transparent – all results open-sourced

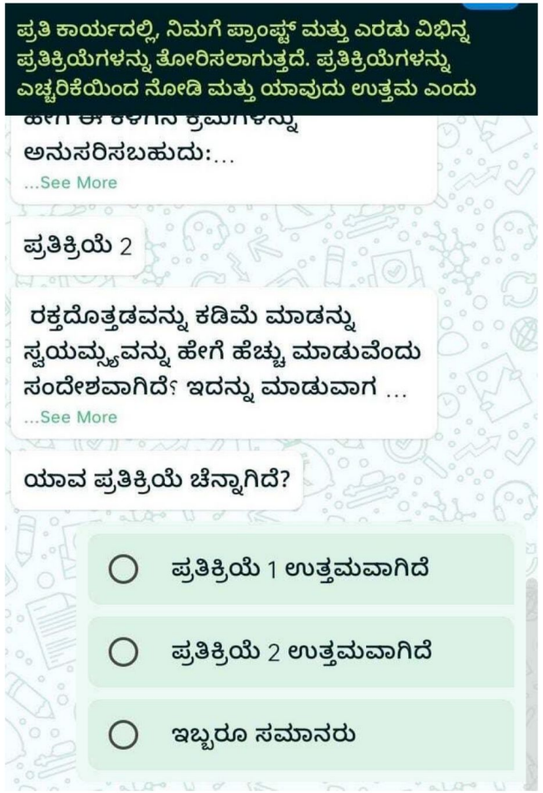
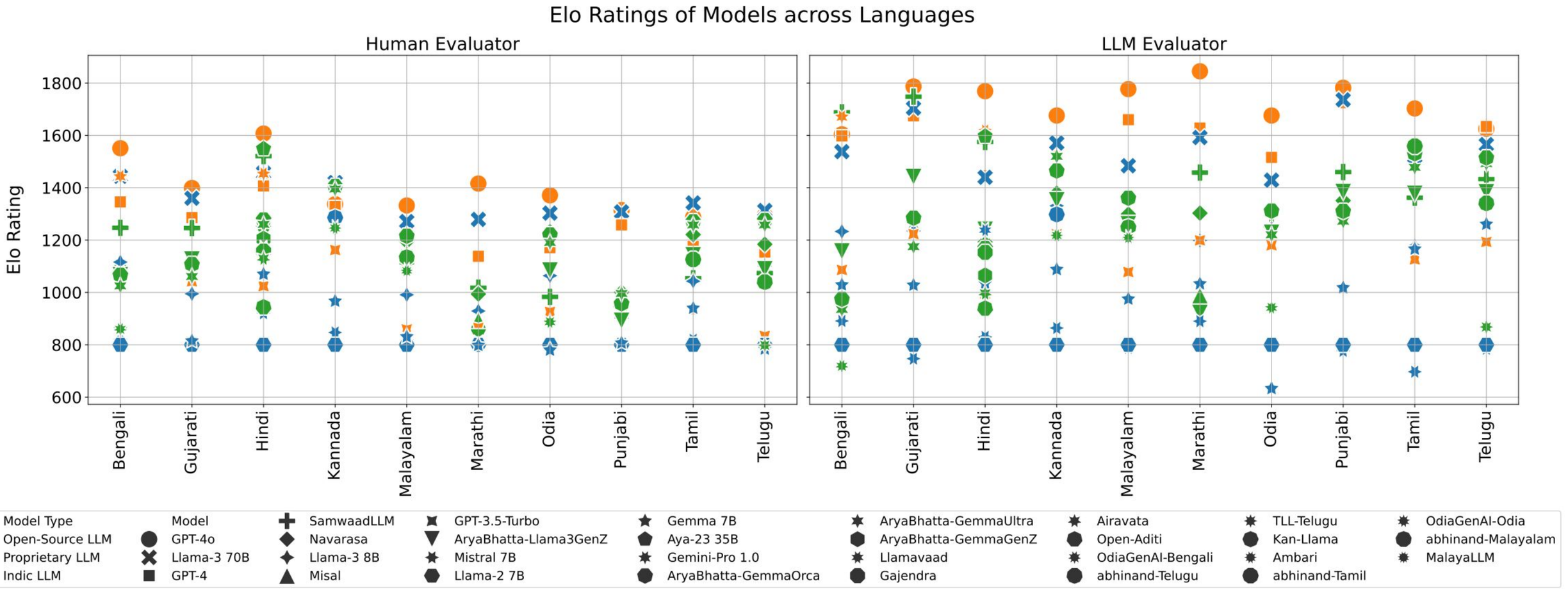


Figure 1: Karya App interface for doing pairwise evaluations for Kannada. The app shows the prompt (question) along with answers from two LLMs and options for them to pick from - the first response is better, the second response is better, and tie

Pariksha Leaderboard



Health Pariksha: Evaluating multiple models on representative real-world multilingual data

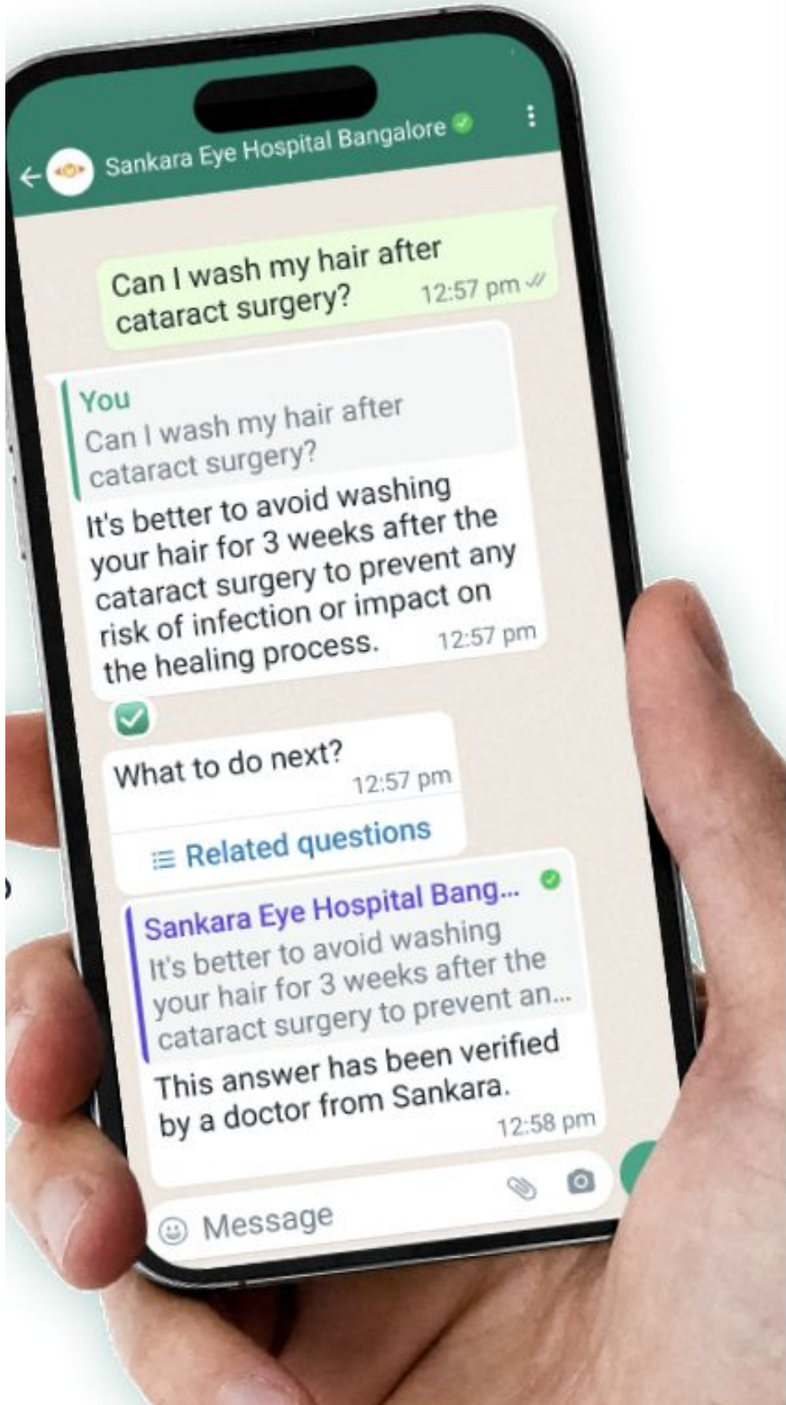
- Multilingual (Indian English + 4 Indian languages) user queries from Health Chatbot for Cataract surgery deployed in Bengaluru, India

  “Can I eat before the kanna operation?”

  “Can I eat non veg after surgery?”

  “How long should saving not be done after surgery?”

  “Agar operation ke baad pain ho raha hai, to kya karna hai?”
- 24 models, identical setup, specialized metrics
- Code-switched, culturally nuanced questions
- Several Indian LLMs perform poorly compared to global LLMs



Why community evaluation?

👤 **User**

“I spend long hours standing and walking for my on-field police duties, and it's causing knee pain. What exercises can help ease it?”

🤖 **Chatbot**

“You can try gentle knee-strengthening routines using equipment like a Nordic track or foam-based balance trainers, which help reduce joint strain. Short stretching breaks during work, such as quad and hamstring mobility drills done against wall-mounted bars, can also support knee stability.”

⚠️ Mentions exercise equipment uncommon in India

⚠️ Suggests workplace facilities (wall-mounted mobility bars) rarely found in Indian police settings

⚠️ Doesn’t reflect local realities and constraints

Introduction to Project Samiksha

11

Languages: Hindi, Bengali, Malayalam, Telugu, Marathi, Gujarati, Tamil, Kannada, Punjabi, Assamese, English

6

Domains: Healthcare, education, finance, legal, agriculture and culture

Benchmarking AI for Indian languages and contexts, with community participation and civil society guidance.

#1
Sub - Domains & Query Creations

Co-creation of sub-domains with NGOs and real query collection from Karya workers

#2
Expert Evaluations

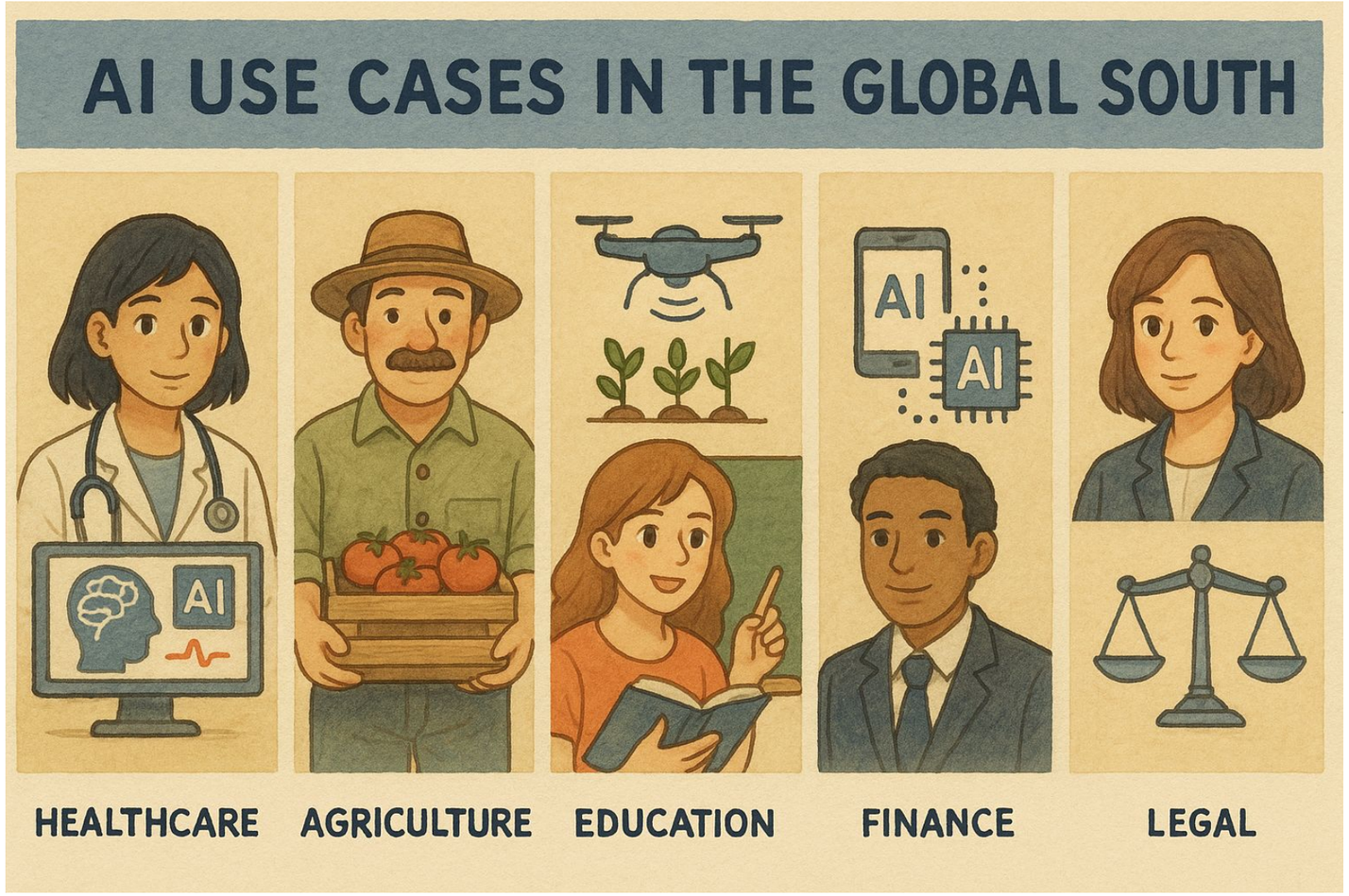
Experts to provide a small number of technical evaluations of the LLM responses

#3
Non-Expert Evaluations

Karya workers to conduct standalone and comparative evaluations for LLM responses

Samiksha

Real-world Use Case Evaluation



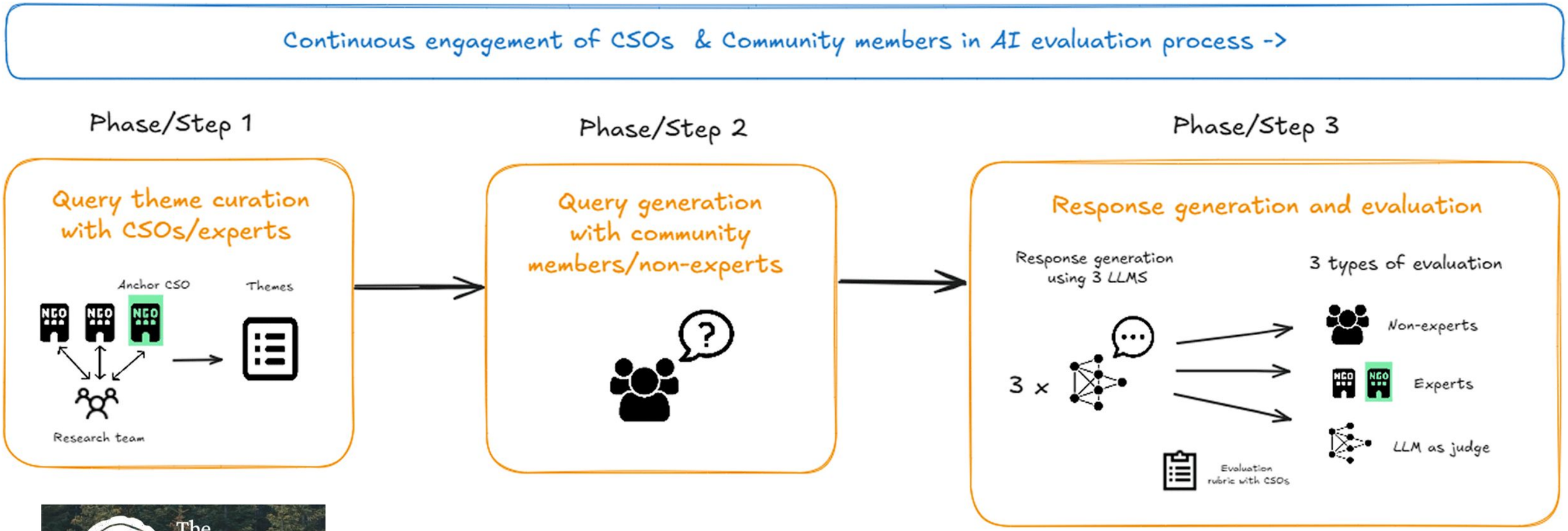
Community-created benchmark in multiple Indian languages with input from prominent CSOs

First benchmark designed ground-up for India

KARYA | Bringing AI-enabled earning and learning opportunities to low-income communities

The Collective Intelligence Project

Samiksha Pipeline



The Collective Intelligence Project

KARYA | Bringing AI-enabled earning and learning opportunities to low-income communities

Microsoft

Samiksha Pilot CSO Engagement

- Interviews with CSOs in the health domain
- Discussions included chatbot design, response evaluation, CSO challenges, and user query examples.
- Two CSOs provided sample chatbot data logs.
- Thematic analysis identified 15 themes and eight genres of health topics.
- Examples covered fact-seeking, advice, open-ended questions, and social health factors like culture and gender norms
- Evaluation criteria provided by CSOs, expanded by us

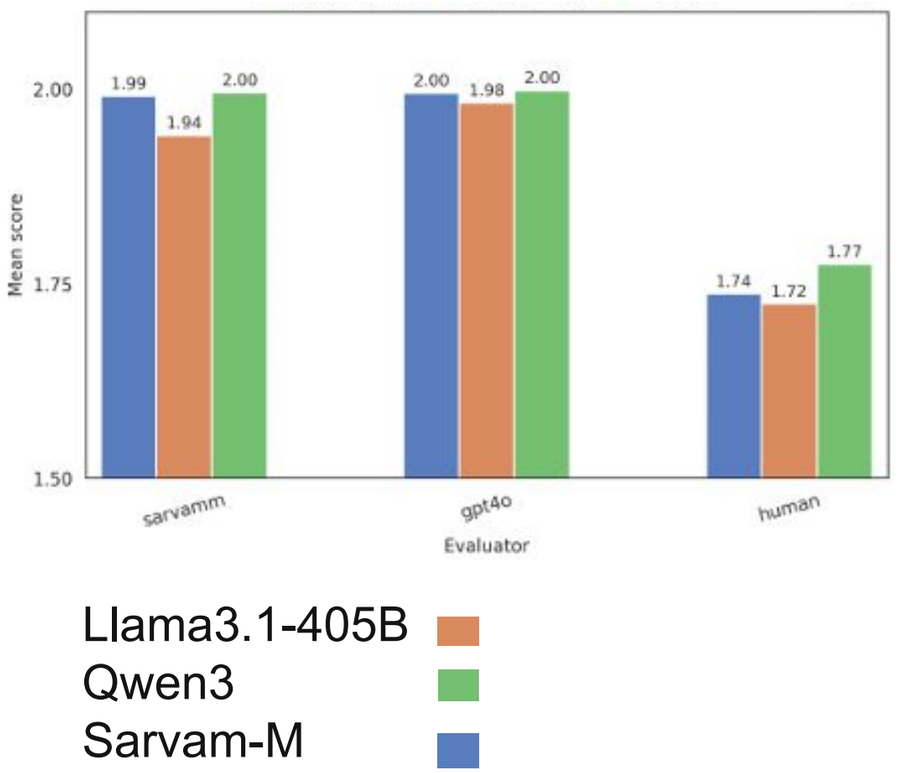
Samiksha Pilot Data Collection

- Culturally-rooted, relevant queries
- “घर पर बुजुर्ग महिलाएं कहती हैं कि बच्चों को पैदा होने के तुरंत बाद ही स्तनपान कराना चाहिए परन्तु हॉस्पिटल में डॉक्टर ऐसा करने से रोकते हैं, ऐसे में हमें क्या करना चाहिए?” (Elderly women at home say that babies should be breast-fed immediately after birth, but doctors in the hospital stop us from doing this, so what should we do?)
- ಒಂದು ಕನ್ನಡ ಸಿನಿಮಾದಲ್ಲಿ ಒಂದು ಹೆಣ್ಣು ಮಗಳ ಹೆರಿಗೆ ಸಮಯದಲ್ಲಿ ತಾಯಿ ಮತ್ತು ಮಗು ಇಬ್ಬರ ಜೀವಕ್ಕೂ ಅಪಾಯವಿರುತ್ತದೆ ಅದಕ್ಕೆ ನಾಯಕನು ಇಬ್ಬರ ಜೀವ ಉಳಿಸಲು ನೀರಿನಲ್ಲಿ ಹೆರಿಗೆ ಮಾಡಿಸಿ ಇಬ್ಬರ ಜೀವ ಉಳಿಸುತ್ತಾನೆ ಇದು ನಿಜಾನಾ? (In a Kannada movie, when a woman is giving birth to a girl, both their lives are in danger and the hero saves them by facilitating a water birth. Is this true?)
- I cannot speak freely with my gynecologist because she tells my mother everything. (She is our family friend.) How can I go to another doctor without making it awkward?
- I have low BP. Can I fast for Shivratri?
- ~800 queries in 3 languages

Samiksha Pilot Evals

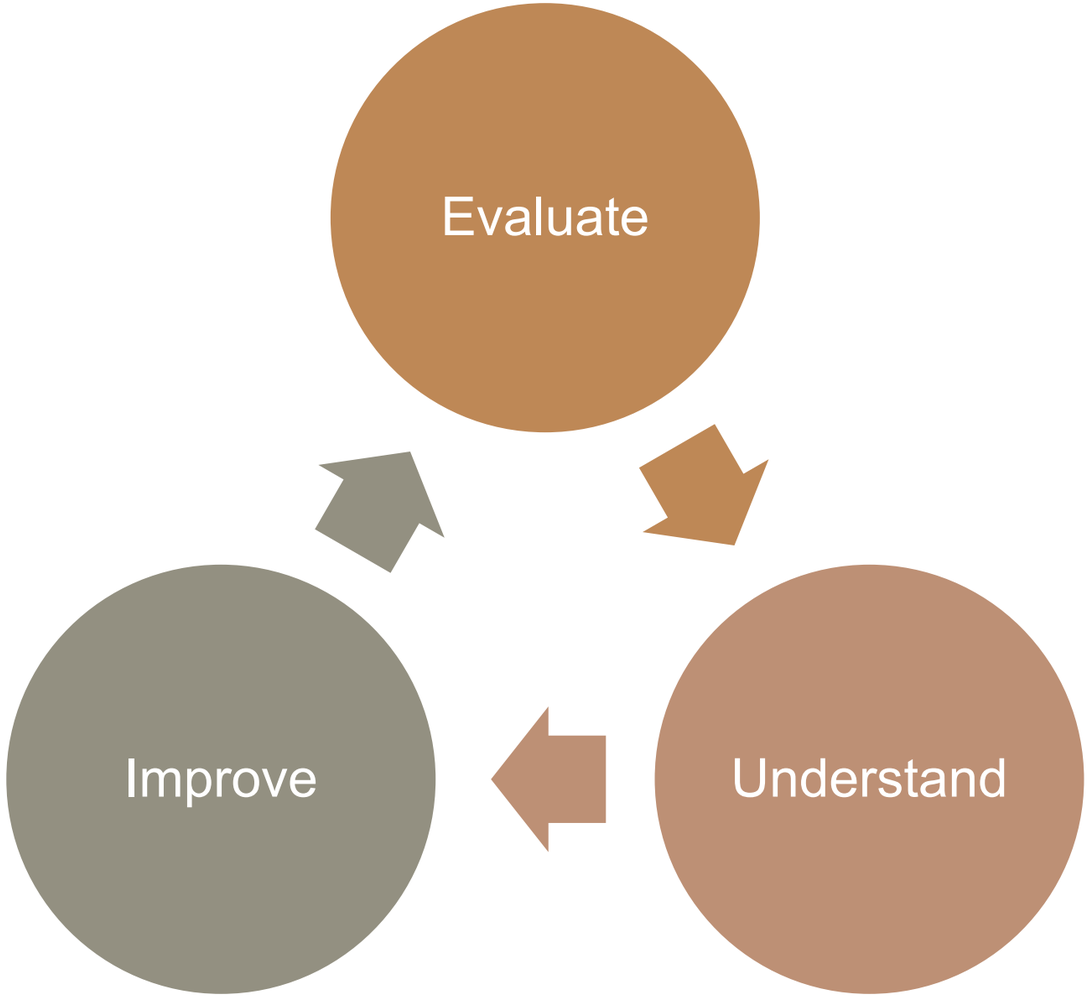
- 3 LLMs evaluated
- LLM-judges (2 models) + human evaluation (Karya)
- LLM-judges still do not correlate well with human evals and consistently over-estimate performance [Our past work in NAACL EACL, EMNLP 2024]

- In-progress – Samiksha-Agent Agentic Evaluation framework
  - Automatically select best LLM-judge protocol from a large number of tuned judge models
  - Aim for better correlation with human evals



Samiksha – going forward

- v1 – Large scale (~20k data points) benchmark across 11 Indian languages
- Evals (human + LLM-judge) across ~30 models
- Samiksha v1 leaderboard in Feb 2026
- Detailed paper on methodology, experiments, results
  - Replicate to more languages/regions?


Evaluate
Understand
Improve