Samiksha: What Large-Scale, Community-Driven Evaluation Reveals About Language Models for India

Gayatri Bhat ◇ **Hamna** ♠ **Sourabrata Mukherjee** ♠ **Manan Uppadhyay**♠
Brintha Chandrasekaran ◇ **Monali Shelar** ◇ **Faisal Lalani** ♣ **Evan Heafield** ♣ **Kavitha K** ◇
Vivek Seshadri ◇ **Manu Chopra** ◇ **Divya Siddarth** ♣ **Kalika Bali**♠ **Sunayana Sitaram**♠

♠Microsoft Corporation ♣Collective Intelligence Project ◇Karya
Correspondence: manu@karya.in, divya@cip.org, sunayana.sitaram@microsoft.com

Abstract

Evaluation practices for Large Language Models (LLMs) have largely been shaped by English-centric benchmarks, limiting our understanding of model behavior in multilingual, multicultural contexts. We introduce Samiksha, the first large-scale, community-driven evaluation of Indian language models spanning 11 languages and over 23,000 culturally grounded data points across four high-priority domains. All prompts were developed through a two-step, community-engaged pipeline: topics were first identified in consultation with civil society organizations (CSOs), and native speakers then independently created language-specific prompts to reflect everyday information needs within each context. We conduct a comprehensive mixed-method evaluation combining 150k native-speaker human assessments, expert review, qualitative analysis, and 1.6 million automated LLM-as-judge evaluations. Our results reveal systematic domain and language-specific performance trends, identify model families that generalize consistently across settings, and expose significant divergences between human judgments and automated judges. We further align LLM-based judges with human preferences using evaluation data, improving their reliability while highlighting their limitations in multilingual settings. Beyond constructing leaderboards, our work demonstrates that inclusive, representative, and culturally grounded evaluation at scale is both feasible and necessary for advancing language technologies beyond English. Samiksha provides a blueprint for combining community insight with methodological rigor to better anticipate downstream use and real-world impact.

1 Introduction

Large language models (LLMs) are rapidly being integrated into highly impactful socially relevant domains including agriculture, finance, healthcare, and law (Singh et al., 2024a; Ramjee et al., 2025; Guha et al., 2023). These systems are increasingly being adopted as tools for decision-making, information access, and public service delivery (Sharma et al., 2024). Yet their expanding reach has also drawn attention to broader ethical and social concerns, especially questions of suitability for diverse audiences. Past research shows that LLMs often underperform for non-Western users, reflecting cultural misalignment and underrepresentation that surface as insensitivity, bias, and exclusion (Atari et al.; Kumar and Pratap, 2020). These shortcomings not only undermine user trust but also limit meaningful adoption of AI across global contexts, which is reflected in the uneven diffusion of AI in the Global North and Global South (Mic, 2025).

Evaluation now constitutes the central feedback mechanism through which contemporary AI systems are designed, optimized, deployed, and governed. In the era of large-scale generative models, the most consequential objectives e.g., helpfulness, harmlessness, trust and safety cannot be captured by simple canonical metrics used traditionally for classification and regression tasks. Consequently, evaluation is no longer just "testing", but the primary means by which abstract desiderata are translated into measurable system properties to support development decisions and accountability claims - if you can test a model or system reliably, you can improve it and govern it. However, there is growing consensus within the field that prevailing evaluation practices are inadequate in several important respects, as reflected in the proliferation of dedicated workshops and specialized conference tracks on evaluation across major AI and NLP venues. Unaligned evaluations have also been blamed for critical chal-

lenges such as hallucinations that impede the use of LLMs in high-stakes settings (Kalai et al., 2025).

AI evaluation is even more critical in Global South settings because most AI systems embed Global North assumptions in their data, task formulations, and definitions of success. Standard testing pipelines rarely include Global South benchmarks and release decisions are typically made by a small number of actors in the West. As a result, benchmark gains can obscure systematic failures in the languages and conditions where these systems are deployed, potentially leading to marginalized communities being excluded, stereotyped, or exposed to high-stakes failures. Representative evaluation should also ensure responses feel relatable, are easy to understand, and translate into practical action for community members, aligned with their resources, cultural norms, and day-to-day realities.

Benchmarking has become the dominant paradigm for evaluation, extending to cultural and domain-specific contexts as well as cross-cultural representation (Chang et al., 2024; Watts et al., 2024; Qin et al., 2025). However, today's benchmarks fall short: they often use translated or artificially created data (Nadăş et al., 2025) that overlook real user needs and elevate institutional priorities above those of the communities most affected. In contrast, a culturally grounded benchmark reflects local languages, values, and contexts; ensures representation across diverse identities; and employs community-validated criteria to evaluate models fairly and meaningfully This leads us to our central question: How can we design and implement scalable, culturally grounded, community-centered pipelines to evaluate AI models and systems?

To address this, we propose Samiksha [1], an end-to-end evaluation pipeline that is guided by inputs from Civil-Society Organizations (CSOs) and data workers, who serve as end users. The Samiksha benchmark consists of 23k data points in 11 Indian languages covering four domains (healthcare, legal, education and finance) that are considered to be high-priority sectors for AI in India. The questions within our domains are designed to reflect the everyday information needs and lived realities of community members. They represent the kinds of queries individuals might ask in daily life, rather than specialized or technical questions typically posed by professionals such as doctors, legal experts, or teachers.

Our bottom-up approach combines the domain-specific expertise of CSOs with the lived experiences of Indian users. We conduct domain-specific evaluations of chatbots by eliciting community inputs in three complementary ways. First, we conduct short, focused interviews with CSOs in the relevant domains to capture their perspectives on benchmark creation and evaluation requirements. Second, we use these insights to create task designs for paid data workers, who create the benchmark query dataset and evaluate chatbot responses. These tasks are designed to let data workers draw upon their own lived experiences, interests, and concerns, thereby grounding the benchmark in community realities. Third, we conduct expert workshops with CSOs to ensure that the rubrics we develop are aligned with desired behavior in these domains. This integration of focused insights from CSOs with dispersed contributions from existing or potential AI users ensures that community concerns remain central throughout the process. This approach also channels expertise and engagement from multiple stakeholders, resulting in a benchmark that is both contextually grounded and adaptable to other domains and regions. The novelty of our work lies in this co-creation process and in the systematic integration of community feedback at every stage.

Using this benchmark, we conduct a comprehensive, fine-grained analysis of 17 LLMs using human evaluators, expert evaluators, qualitative methods and automated LLM-as-judge techniques (Zheng et al., 2023) to uncover several key insights that can be used to shape model development for Indian languages and contexts. We evaluate each model not just for language and response quality, but also across the dimensions of cultural relevance and trust, making this the first large-scale evaluation of multilingual models along these dimensions. Evaluators provide judgments grounded in cultural nuance, accounting for local language use, social norms, and context-specific expecta-

[1]The word "Samiksha" is of Sanskrit origin and means collective analysis, review or thorough investigation

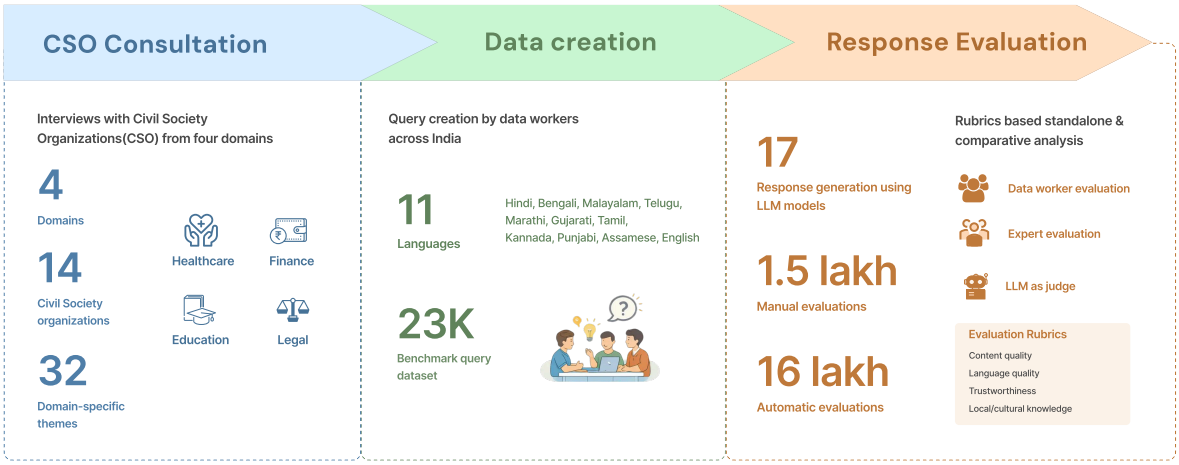

Figure 1: The Samiksha Pipeline illustrates a structured, three-phase evaluation process designed to ground chatbot benchmarking in community realities. *Phase 1 – Civil Society Consultation:* Interviews are conducted with civil society organizations (CSOs) in four domain to inform benchmark and evaluation requirements. *Phase 2 – Query Curation:* Insights from CSOs are transformed into training and task design for paid data workers, who create the query dataset in 11 languages across four domains. *Phase 3 – Response Evaluation:* Chatbot responses generated by 17 large language models are evaluated using mixed-methods, including human and automated assessment.

tions of communication. In parallel, we employ LLM-as-judge methods as a complementary approach to scale evaluation and cross-validate human ratings.

Our work makes the following contributions:

1. We present Samiksha, a co-designed pipeline with CSOs and community members that centers lived realities in what to evaluate, how to build the benchmark, and how to score outputs. We provide a generalizable template that other domains and regions can adapt to build inclusive, context-aware LLM evaluations. We demonstrate that our pipeline enables the integration of deep community insights with large-scale evaluation.

2. Using this pipeline, we develop a culturally grounded, multilingual benchmark spanning 11 Indian languages, comprising over 23k data points. The benchmark covers four high-priority domains for India, with each data point created specifically for its language and context by native speakers.

3. We conduct a mixed-method evaluation comprising of 150k native-speaker human assessments, expert evaluation, qualitative analysis, and 1.3 million automated evaluations using the LLM-as-judge technique. Beyond identifying where models perform well and where they fail, our detailed analysis also surfaces the strengths and limitations of automated evaluation approaches. We use these insights to align LLM-based judges to human preferences using human evaluation data.

4. We construct leaderboards for all domains, languages, and models and identify key trends across domains and languages using the full range of evaluation methods, demonstrating that certain model families perform consistently well across settings.

2 Related Work

2.1 Current Landscape of LLM Evaluation

Large language models (LLMs) have been evaluated in diverse settings, such as multicultural (Sukiennik et al., 2025; Ki et al., 2025), multilingual (Ahuja et al., 2023; Watts et al., 2024), and domain-specific contexts (Budler et al., 2025; Koto, 2025). Traditionally, evaluation has relied on standardized benchmarks (Chang et al., 2024; Chakraborty et al., 2025), synthetic datasets (Nadăş et al., 2025; Ding et al., 2023), and expert-curated tasks (Qin et al.,

2025). While these controlled evaluation methods remain valuable for measuring technical capabilities and enabling comparisons between models, they create a significant gap between how LLMs are tested and how they are actually used in real-world contexts (Adhikary et al., 2025; Ying et al., 2025; Kumar and Pratap, 2020). Their key limitations are threefold. First, benchmarks are often too generic, meaning that models optimized for benchmark performance may struggle in open-ended or culturally specific situations. Second, most LLMs are trained primarily on English-language data, causing them to inherit racial, gender, and cultural biases present in these datasets (Khan et al., 2025; Thakur, 2023; Naous et al., 2023). Third, these approaches lack ecological validity, as they fail to account for local dialects and cultural nuances that are essential for populations that are under-represented in training data and do not primarily speak English (Hussain and Ginige, 2018; Li et al., 2024a).

As LLMs become global tools, evaluation methods must evolve to reflect the cultural and social realities in which they operate. For example, Singh et al. (2025) show that Humanities and Social Science questions dominate culturally sensitive subsets because they rely on contextual and regional knowledge, making them more difficult for LLMs than culturally agnostic STEM questions. Similarly, Chiu et al. (2025) propose CulturalBench, a benchmark of human-written questions covering 45 world regions. Even state-of-the-art models (like GPT-4o) achieve low accuracy on Culturalbench, especially on underrepresented cultures, emphasizing that LLMs often fail on nuanced cultural queries that go beyond textbook knowledge. Several efforts have focused on adapting existing models and creating new language models specifically designed for non-English languages (Kumar and Pratap, 2020; Cul, 2023). Researchers have also worked to build localized benchmarks for different regions, such as India-specific evaluations (Gumma et al., 2024; Kakwani et al., 2020; Gala et al., 2023; Verma et al., 2025), and culture-specific assessment (Chiu et al., 2025; Myung et al., 2024; Li et al., 2024b). However, many multilingual benchmarks are simply translations of English-language benchmarks which can lead to loss of linguistic and cultural context (Watts et al., 2024; Myung et al., 2024). In high-stakes domains such as healthcare, law, and education, model outputs directly influence decisions that affect people's lives. Misinformation or cultural insensitivity in these areas can cause serious harm(Deva et al., 2025; Kumar and Pratap, 2020).

To address these risks, recent work has advocated for domain-specific, culturally-grounded benchmarks. In the healthcare domain, MedMCQA (Pal et al., 2022), MedDialog (Zeng et al., 2020), and MedRedQA (Nguyen et al., 2023), represent prominent efforts to build large-scale benchmarks using expert-authored content, medical board exams, or real-world patient-doctor dialogues. While these datasets offer high topical coverage and clinical relevance, they often lack grounding in community-specific cultural contexts and lived experiences. These efforts highlight the inadequacy of one-size-fits-all evaluation and emphasize the need for specialized, context-aware approaches. AfriMed-QA (Nimo et al., 2025) introduces important African contextual grounding but is constrained to short-answer and MCQ formats and MenstLLaMA (Adhikary et al., 2025) provides meaningful India-specific cultural sensitivity, yet its focus remains narrow, centering solely on menstrual health.

Existing benchmarks rarely reflect the multilayered, conversational, and dilemma-driven nature of real community queries shaped by stigma, family dynamics, myths, power relations, and local resource constraints. Samiksha addresses these limitations by grounding evaluation directly in community-identified priorities and lived experiences, enabling assessment that goes beyond correctness to emphasize community resonance, practical relevance, and cultural alignment, which are essential for the responsible deployment of LLMs in real-world public health ecosystems. Table 1 compares the key features of our benchmark with other benchmarks in the Healthcare domain.

2.2 Community-Centered Approaches to AI Evaluation

There is a growing body of scholarship calling attention for more globally inclusive and community-centered approaches to LLM evaluation (Qadri et al., 2025; Bergman et al., 2024; Shrivastava and Aoyagui, 2025; Hall et al., 2025). One widely used method is crowd-

sourcing, which generally takes two forms. In the top-down model, corporate intermediaries such as Prolific or Amazon Mechanical Turk recruit community members as data workers, define the terms of participation, and dictate what counts as an acceptable submission, often leaving little space for feedback (Miceli and Posada, 2022). In contrast, the bottom-up model directly engages community members in data collection. This approach broadens opportunities for participation and empowers people not only to share cultural knowledge, artifacts, and expertise but also to shape the parameters of their own involvement(Singh et al., 2024b; Birhane et al., 2022; Delgado et al., 2023). STELA (Bergman et al., 2024), for example, introduces a community-centered methodology for eliciting norms to guide AI alignment. STELA conducts deliberative focus groups with underrepresented communities in the U.S. to define rules for chatbot behavior. Similarly, Qadri et al. (2025) argue for "thick evaluations" of cultural representation, showing through workshops with South Asian communities that community-defined metrics offer richer insights than conventional quantitative ones. Our work operationalizes this idea in the context of LLMs, creating a scalable benchmark pipeline that embeds community perspectives into both data generation and evaluation. The importance of co-creation, however, extends beyond method; it also depends on who participates. For example, Hall et al. (2024) found that annotators living outside a region are more likely to view exaggerated or stereotypical depictions of that region as representative, while local annotators can draw on lived experience to provide more accurate assessments. This highlights the value of involving community members directly in AI evaluation. Complementary to this, expert-in-the-loop approaches bring in specific stakeholders whose perspectives strengthen evaluation outcomes(Ramjee et al., 2025). Building on these insights, we position Samiksha as combining community data collection with domain-expert perspectives to co-create evaluation categories and judgments. This hybrid design differs from prior crowd-sourced or expert-only evaluations by centering the lived realities of marginalized user groups while retaining structured input from practitioners familiar with the domain's operational constraints.

3 Benchmark Creation Framework

To assess whether an AI-based solution is truly useful for everyone, it is crucial to consider civil society and community perspectives. However, this can be challenging for several reasons. Firstly, civil-society organizations (CSOs) such as NGOs are stewards of experience and expertise, but resource constraints might prevent them from participating intensively in AI evaluations. Secondly, community members can speak to the usefulness of the tools in their daily lives, but might not have the right opportunities or incentives to participate in an evaluation process. The Samiksha pipeline was developed to gain a holistic and grounded understanding of community needs and preferences in their interactions with AI. The pipeline uses a multi-step process to (1) elicit rich insights via focused CSO consultations (2) expand those insights into a large-scale evaluation via paid data work undertaken by community members or end users. This approach combines big-picture inputs from CSOs, providing knowledge about the general needs and preferences of communities and fine-grained inputs from community members, reflecting their individual interests and concerns, to create a representative benchmark with broad coverage.

Our pipeline follows a structured evaluation approach that encompasses three main phases: query curation via CSO consultations, query generation, and response evaluation, as shown in Figure 1. Each phase was co-designed with community members to foreground their preferences when interacting with LLM based systems.

3.1 Stakeholders

Research Partners: Our research team, made up of three collaborating organizations, is described below along with the role of each group:

1. Karya: Karya is a social impact data platform that delivers ethically collected, high-quality datasets for AI training and evaluation, with a strong focus on linguistic and cultural diversity in India. Their platform empowers rural and economically disadvantaged communities by providing digital

Feature	SAMIK-SHA	AfriMed-QA	MedMCQA	MedDialog	MenstL-LaMA
Multiple stakeholders?	Community members, health CSOs	African medical schools, educators, health orgs	No	No	Medical experts, health educators
Question Types	Queries with long answers	MCQs and short answers	MCQA	Patient–doctor online conversations	QA pairs
Cultural Grounding	Yes	Yes	No	Partial	Yes
Example	*"I cannot speak freely with my gynecologist because she tells my mother everything. How can I go to another doctor without making it awkward?"*	*"A 28-year-old pregnant woman in rural Africa presents with severe abdominal pain. What are the possible causes specific to African healthcare context?"*	*"A 30-year-old man presents with acute onset chest pain radiating to the left arm. What is the most likely diagnosis?"*	*"Patient: I have been experiencing severe headaches. Doctor: How often do they occur and what triggers them?"*	*"I have irregular periods and cramps. What could be the cause and which home remedies are safe in my cultural context?"*

Table 1: Comparison of Healthcare Evaluation Datasets Across Key Features

work opportunities, paying workers substantially above the Indian minimum wage, and building digital skills among participants. Karya also handled the participant recruitment and operationalization of the data work.

2. Collective Intelligence Project(CIP): CIP is dedicated to democratize AI evaluation and governance by involving people from around the world in developing, refining, and setting standards for AI systems. CIP facilitated the ideation and conceptualization of community-centered study design and evaluation by involving different stakeholders.

3. Microsoft Research India: Researchers from Microsoft Research India provided methodological guidance for the pipeline and data creation, conducted automated evaluation, qualitative analysis and analyzed results of human and automated evaluation.

Civil Society Organizations (CSOs): Partnerships with CSOs go far beyond simple data gathering; being community centric and technically equipped, these organizations are able to provide actionable evaluation inputs. Many of them engage directly with marginalized groups such as low-income, rural, or other under-represented communities whose voices are often missing from mainstream datasets and benchmarks. They also bring strong contextual expertise, helping interpret queries with the social, cultural, and linguistic nuances that application developers or model builders might miss. Furthermore, CSOs can help filter and frame user concerns so that public input is more translatable to technical ecosystems. Finally, because the Samiksha pipeline relies on persistent interactions with participants, collaboration with CSOs helps build long-term relationships with communities rather than being one-time data collection exercises. In our work, we engaged with CSOs from four domains i.e. Healthcare, Finance, legal and Education, who all work primarily in India. All CSOs we worked with work closely with communities, and most have experience with chatbot-based interventions, with some having already deployed chatbots, and others in the process of planning deployments. Initial contact with CSOs was made via email, including a brief description of the research goals and a request for a virtual interview. CSOs nominated a representative to participate in the interviews. They were not compensated for

their time.

Communities: In the context of this work, we define communities as a group of people, often but not always geographically proximate (e.g., a village, a specific urban neighborhood), who share a common set of social, economic, cultural, and environmental characteristics. This includes shared local challenges, language nuances, and value systems. Information flow in communities is often governed by local social networks and trusted entities (CSOs, NGOs, local health workers, elders), with reliance on such localized knowledge networks for decision making in critical domains such as healthcare, finance, agriculture and education. The core differentiator between a community-centered AI solution is that the information required for daily life needs to be local, actionable, and culturally situated. For example, a "safe drinking water" question in one community might pertain to specific local well testing, while in another, it might be about the affordability of a municipal connection. In this work, we use two complementary pathways to reach the communities we aim to serve, that is, current and future users of AI-based technologies - via CSOs, that work directly on the ground with communities, and with Karya, which engages native speakers from language communities, ensuring broad and meaningful participation.

3.2 Phase 1: CSO consultation

We conducted a series of semi-structured interviews with 14 CSOs across four domains. This included four CSOs each from the healthcare and legal domains, three from finance, and two from education. Seven of the 14 CSOs had already deployed chatbots, while two others were in the process of development. The medium of communication was primarily English, with some interviews incorporating other languages such as Hindi, Marathi, or Malayalam, depending on participant preference. The interview protocol was designed to be flexible, allowing emerging and unexpected themes to be explored in greater depth. Each interview was conducted by two researchers: one led the discussion, while the other took detailed notes, as the interviews were not recorded. The main aim of these interviews and follow-ups was to capture regionally relevant user needs and concerns, as per each CSO's expertise. Interviewees shared perspectives on domain-specific chatbot use, their own interventions to ensure safe and effective chatbot use in the communities they serve, and broader visions for the future of healthcare chatbots. They described the styles and themes of queries users pose to experts or chatbots, providing examples from on-ground interactions or chatbot deployments. CSOs also articulated the qualities of good and poor LLM responses to domain specific queries, either as general criteria or via specific examples. Conversations with CSOs who had deployed chatbots included lessons learned from development and deployment. Interviewees were encouraged to elaborate on particulars as needed. After each interview, we iteratively refined our questions, continuing until responses reached theoretical saturation. We conducted thematic coding of interview notes, grouping recurring points into multiple themes that captured the query topics identified by CSOs within each domain. Through this process of generalization, eight broader themes per domain emerged, as shown in Table 2. To guide data workers in the subsequent query creation phase, we synthesized topic-specific example queries in Indian English, reflecting diverse styles and perspectives. We also synthesized CSO recommendations to create evaluation rubrics, which are discussed in Section 5.

3.3 Phase 2: Data work

We ran the query creation pipeline in two phases: data creation followed by validation. Each language team included a coordinator who provided data workers with topic lists and examples. Coordinators also supported workers directly by answering questions about topics and providing individualized feedback throughout the process.

3.3.1 Data creation

Karya provided training to data workers for creating questions using the topics and guidelines distilled by us from CSO interviews. The training material was translated into all the languages used in this study by language experts to ensure that data workers understood the task and could raise any questions or concerns with them. First, users were introduced to chatbots, and also informed about how chat-

Healthcare	Finance	Legal	Education
Access to primary/community healthcare	Daily Finance, Savings & Budgeting	Product & Service-Related Consumer Queries	Teaching and Learning Support
Managing injuries & infectious diseases	Loans & EMIs	Family & Marriage Matters	Exam and Job Interview Preparation
Managing chronic illness	National and International Finance	Workplace / Employment Rights & Safety	Career Guidance
Maternal health	Income and Taxes	Safety, Accidents, Theft	Higher Education & Financial Support
Reproductive health	Investments	Financial & Contract Matters	Upskilling and Continuing Education
Senior care	Government Schemes & SHGs	General Knowledge about the Legal System	Educational Policies and Governance
Wellness habits	Insurance	Fraud	Interest-based Learning
Child health	Digital Finance	Property & Inheritance	Student Support and Wellbeing
Other medical questions	Other Financial Questions	Other Legal Questions	Other Education-related Questions

Table 2: Query themes collated from CSO consultation

bots were different from search tools that many of them were familiar with. They were briefed on the types of information users typically seek from chatbots, as well as the common errors chatbots may produce such as factual inaccuracies, irrelevant responses, or unnatural language. Data workers were informed about the goals of the data collection effort - to collect questions that a person would genuinely and naturally ask an expert or a chatbot. We encouraged data workers to think about their own experiences, ask questions that they were genuinely curious about and to create specific and realistic questions. Data workers were asked to understand the topic, ask questions about the concept to the coordinators and to ensure that the questions they asked were not textbook or exam-like. We also asked them not to create multiple questions following the same pattern, or overly generic questions, or questions that contained the topic name, such as "What all comes under [TOPIC]?".

3.3.2 Data validation

Language experts validated data created by data workers. The validation criteria were as follows: the question should be understandable, on-topic, should include relevant details, should not include personally identifiable information, should not directly be based on examples provided and the set of questions created by a data workers should be diverse. Validators were instructed to reject questions that seemed to be textbook-like or from an exam question paper, or if the question was too generic, or if the name of the topic was present in the question.

3.4 Dataset statistics

The final Samiksha benchmark created after the completion of data validation covers the following 11 Indian languages: Indian English, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil and Telugu. These represent the most widely-spoken languages in India and also overlap with the 22 scheduled languages in the Constitution of India (with the exception of Indian English). The benchmark contains 500+ queries per domain and language, with an average of 540 queries per combination of domain and language. It has a total of 23785 queries, covering Healthcare, Legal, Education and Finance domains.

4 Response Generation

We generated responses from several SoTA LLMs for the queries in the Samiksha benchmark. This section elaborates on the models chosen, experimental details for response generation and the final set of models chosen for human evaluation.

4.1 Models

We generated responses to each query using the 17 models listed in Table 4. We used a simple prompt that instructed each model to respond to a question, with the name of the topic provided as context. We selected global models with multilingual support and Indic

Domain	Language	Questions	Translation
Healthcare	Hindi	घर मे हर उम्र के लोग है सबका अलग अलग पसंद है और जब खाने की बात आती है तो मैं हमेशा असमंजस मे रहती हु| क्युकी उसमे बुजुर्गों का अलग से प्रबंधन करना पड़ता है और कोई अच्छा उपाय क्या है जिससे इसको ठीक से मैनेज किया जा सके?	There are people of all ages at home, each with different preferences, and when it comes to food, I always find myself in a dilemma. Because elderly people need to be managed separately, what is a good solution to manage this properly?
Healthcare	Kannada	ನನಗೆ ಸಕ್ಕರೆ ಕಾಯಿಲೆ ಇದೆ ಆದ್ರೆ ಮಾತ್ರೆ ತಗೊಳೊಕೆ ಬೇಜಾರು ನಾನ್ ಮಾತ್ರೆ ತಗೊಳ್ದೆ ಹೇಗೆ ಸಕ್ಕರೆ ಕಾಯಿಲೆಯಿಂದ ದೂರ ಆಗ್ಬಹುದು?	I have diabetes but I don't feel like taking medication. How can I keep diabetes in check without it?
Legal	Malayalam	ഭർത്താവിൽ നിന്നോ കുടുംബത്തിൽ നിന്നോ ഒരു സ്ത്രീക്ക് മാനസികമായോ ശാരീരികമായോ ഉള്ള പീഡനം നേരിട്ടാൽ, നിയമപരമായി അവർക്ക് ലഭിക്കുന്ന സംരക്ഷണവും സഹായവും എന്തൊക്കെയാണ്?	If a woman faces mental or physical abuse from her husband or family, what legal protection and support are available to her?
Legal	Bengali	আমরা যদি আমাদের বাড়িতে চুরির শিকার হই, তাহলে কীভাবে পুলিশকে জানবো? তৎকালীন আমাদের কী করণীয় কাজ করা উচিত হবে ?	If there is a theft at home, how should we inform the police? What immediate steps should we take at that time?
Education	Assamese	কৃষি বিশেষজ্ঞ হ'বলৈ ছাত্ৰ এজনক কি কি অৰ্হতাৰ প্ৰয়োজন?	What qualifications does a student need to become an agricultural expert?
Education	Marati	माझ्या कुटुंबाला वाटत की मी डॉक्टर व्हावं, पण मला संगणक क्षेत्रात रस आहे अशा वेळी योग्य निर्णय कसा घ्यावा?	My family thinks I should become a doctor, but I am interested in the computer field. In such a situation, how should I make the right decision?
Finance	Gujarati	આજે ભારત વિદેશી રોકાણમા કેટલુ આગળ છે? શું આ વિદેશી રોકાણથી ભારતના વિકાસ મા તરક્કી થઈ રહી છે?	How far has India progressed in foreign investment today? Is this foreign investment helping India's development and growth?
Finance	Punjabi	ਜੇ ਕਿਸੇ ਵਿਅਕਤੀ ਨੂੰ ਪੈਨਸ਼ਨ ਦੀ ਰਕਮ ਸਮੇਂ 'ਤੇ ਨਹੀਂ ਮਿਲਦੀ, ਤਾਂ ਉਹ ਕਿਹੜੀ ਸੰਸਥਾ ਨਾਲ ਸੰਪਰਕ ਕਰੇਗਾ?	If a person does not receive their pension amount on time, which organization should they contact?

Table 3: User generated questions

models built specifically for Indian languages, cultures and contexts. Models included both instruction tuned and thinking variants, and varied in size from 2.9B to 405B (known) parameters. We prioritized open-source models to avoid benchmark leakage, with the exception of GPT-5 for which we used the Azure OpenAI service.

4.2 Experimental details

The decoding parameters for response generation, standalone evaluation and comparative evaluations can be found in Tables 5, 6 and 7 respectively.

4.3 Final model choices for evaluation

We selected a subset of the models mentioned in Section 5.1 for human evaluation. Since human evaluation is time and resource-intensive, we did not select models that produced low-quality, inconsistent or unusable outputs to the responses in our benchmark. The models that were not selected for human evaluation are Llama-3-Nanda , gpt-oss-120B and Param-1-2.9B . We observed prompt repetition in gpt-oss-120B , Romanized answers in Llama-3-Nanda and multiple inconsistencies in Hindi outputs of Param-1-2.9B . All models were included in automated (LLM-judge) evaluation. While Aya-expanse-32B was only supported in English and Hindi of the 11 languages, we evaluated its answers on all languages given its strong multilingual performance.

5 Human Evaluation

As the Samiksha benchmark contained questions for which answers were inherently subjective without a single correct answer, we did not rely on ground truth for evaluation. Instead, we developed detailed evaluation rubrics using CSO inputs and provided structured training to data workers to ensure consistent, nuanced, and reliable assessments. We modified the metrics used in our pilot in the healthcare domain (Hamna et al., 2025) to arrive at a new set of metrics. We used two workflows for human evaluation - standalone evaluation, where each answer was scored individually, and comparative evaluation, where two answers were shown to the data worker and the task was to pick the better answer.

5.1 Metrics and rubrics

We used four primary metrics, each comprising multiple sub-metrics, to conduct the standalone evaluations as shown in table 8. We

Model	Model Size	Human Evals	LLM Judge Evals
Aya-expanse-32B	32B	✓	✓
Gemma-3-27B-Instruct	27B	✓	✓
GPT-5	–	✓	✓
gpt-oss-120B	120B		✓
Kimi-K2-Instruct	1T (32B activated)	✓	✓
Krutrim-2-Instruct	12B	✓	✓
Llama-3.1-405B-Instruct	405B	✓	✓
Llama-4-Maverick	400B (17B activated)	✓	✓
Llama-3-Nanda	10B		✓
Llama-4-Scout	109B (17B activated)	✓	✓
Param-1-2.9B	2.9B		✓
Phi-4	14B	✓	✓
Qwen3-235B-Base	235B (22B activated)	✓	✓
Qwen3-235B-Instruct	235B (22B activated)	✓	✓
Qwen3-Next-Instruct	80B (3B activated)	✓	✓
Qwen3-Next-Thinking	80B (3B activated)	✓	✓
Sarvam-M	24B	✓	✓

Table 4: Overview of models evaluated by human evaluators and automated methods.

Parameter	Value
Temperature	0.2
Max Tokens	4096
top_p	1
top_k	-1

Table 5: Hyperparameters used for answer generation.

Parameter	Value
Temperature	0.0
Max Tokens	512
top_p	1
top_k	-1

Table 6: Hyperparameters used for Standalone evaluation

Parameter	Value
Temperature	0.0
Max Tokens	256
top_p	1
top_k	-1

Table 7: Hyperparameters used for Comparative evaluation

combined questions asked in our pilot study related to relevance and completeness/conciseness, as users gave the highest rating in a majority of the cases and it was difficult to establish why lower ratings were given in a small fraction of cases. Our revised framing of a "content quality" score asked users to focus specifically on potential issues in the response, and the final content quality score was computed as a weighted average of the selected options. We kept the Response Quality metric from the pilot, but added sub-rubrics to capture potential issues in the language used in the output. The Response Quality score was also computed as a weighted average of the selected options. In the pilot, data workers who were non-experts in these domains found it challenging to judge the factual correctness of answers. We revised this question to focus on trustworthiness instead. Our pilot study had a question on relevance, which we reformulated to ask users to assess the relevance of the answer to their local and cultural context. In the pilot study, users were asked to provide audio explanations for each of the four metrics, which they found difficult and time consuming. In the revised task flow, we retained an audio explanation component for the question on local and cultural relevance.

5.2 Score calculation

To ensure uniform analysis across diverse metrics, we transformed all scores into a standardized 1–3 scale. To do so, we evaluated responses across the four dimensions using a hybrid scoring approach. Fields that allow multiple selections (multi-select) (*Content Quality*, *Response Quality*) utilize inverse error counting, while fields that ask for a single selected response (single-select) (*Trustworthiness*, *Local/Cultural Knowledge*) use direct categorical mapping.

5.2.1 Multi-Select Scoring

For multi-select metrics, the score S_{err} is derived from the cardinality of the error set E, excluding the "no problem" indicator ($E_\emptyset$):

Evaluation Metrics	Sub-metrics
Content quality	• The answer includes information that has nothing to do with the question • The answer is related to the question, but doesn't fully answer the question • The answer is repetitive or is too long • The answer is missing important details or is too short • The answer does not have any of these problems
Response quality	• Spelling or grammar mistakes • Bad choice of words • The answer doesn't flow smoothly • It is difficult to understand the meaning • The answer does not have any of these problems
Trustworthiness	• I would trust it completely • I would trust it, but only after checking it myself (by searching online, or by asking someone I know) • I would not trust it - I would want an expert to check it
Local/cultural knowledge	• The answer does not show any understanding of my local context • The answer shows only a partial understanding of my local context • The answer shows a complete understanding of my local context • I am unable to judge, as this question would not be asked in my local context

Table 8: Evaluation metrics used for standalone evaluation

$$S_{err}(E) = \begin{cases} 3 & \text{if } |E| = 0 \text{ or } E = \{E_\emptyset\} \\ 2 & \text{if } |E| = 1 \\ 1 & \text{if } |E| \geq 2 \end{cases} \quad (1)$$

5.2.2 Categorical Mapping

Single-select fields are mapped to a numeric scale $\mathcal{M}$ where 3 is optimal. For *Local/Cultural Knowledge*, a value of -1 is reserved for cases where the evaluator is "unable to judge".

- $\mathcal{M}_{trust} = \{\text{Complete} \mapsto 3, \text{Verify} \mapsto 2, \text{No Trust} \mapsto 1\}$
- $\mathcal{M}_{local} = \{\text{Complete} \mapsto 3, \text{Partial} \mapsto 2, \text{None} \mapsto 1, \text{Unable} \mapsto -1\}$

5.2.3 Overall Quality Metric

The final quality score S_{ov} is the arithmetic mean of the four dimensions. This provides a singular ranking metric while preserving granular error data for further analysis:

$$S_{ov} = \frac{1}{4} \sum_{i \in \{cont, resp, trust, local\}} s_i \quad (2)$$

5.3 Results

We ran human evaluation in both standalone and comparative settings for all 14 models. The number of datapoints evaluated varied by model, with five models being evaluated on a larger scale (4000 standalone evaluations) and the rest evaluated on a smaller scale (1000 standalone evaluations). We conducted similar small and large scale evaluations in the comparative setting as well. Sample counts for both settings can be found in Figures 2 and 3.

5.3.1 Standalone evaluation results

Figure 2 illustrates the standalone leaderboard rankings determined by human evaluation. We perform standalone evaluations on a total of 31,898 samples across 14 models. The Qwen3-235B-Instruct model performs best across all models with a mean score of 2.68 across all 4 evaluation metrics. The performance gap between the best performing model (Qwen3-235B-Instruct) and worst performing model (GPT-5) is 0.25 points, indicating substantial variation in model quality. The top three and bottom five models operate similarly, suggesting three tiers of performance. The average score across all 14 models is 2.57, with the best model (Qwen3-235B-Instruct) scoring 0.1 points above this average. The Standard Deviation is 0.06, indicating a relatively small performance spread.

Results by metric: Models rank differently by error type: no single model dominates all four dimensions, suggesting that the error types differ across models. For Content Quality, Qwen3-235B-Instruct performs best, while Aya-expanse-32B ranks lowest with a gap of 0.23 points between them, indicating substantial variation. For Language Quality, Gemma-

Rank	Model	Overall Avg	Content Errors	Language Errors	Trust Rating	Local Relevance	Sample Count
#1	Qwen3-235B-A22B-Instruct-2507	2.684±0.38	2.808±0.39	2.873±0.33	2.484±0.63	2.571±0.85	1010
#2	Qwen3-Next-80B-A3B-Instruct	2.649±0.41	2.783±0.41	2.815±0.39	2.443±0.66	2.554±0.85	1012
#3	Llama-4-Maverick-17B-128E-Instruct	2.638±0.40	2.748±0.43	2.864±0.34	2.430±0.64	2.510±0.84	1025
#4	Gemma3_27B_instruct	2.617±0.43	2.795±0.40	2.808±0.39	2.408±0.65	2.459±1.03	4268
#5	QWEN3_235B_A22B	2.615±0.44	2.766±0.42	2.826±0.38	2.426±0.65	2.442±1.02	4274
#6	Kimi-K2-Instruct-0905	2.611±0.42	2.775±0.42	2.769±0.42	2.418±0.65	2.481±0.90	1026
#7	Krutrim-2-instruct	2.606±0.43	2.696±0.46	2.806±0.40	2.429±0.66	2.492±0.87	1023
#8	SarvamM_24B	2.596±0.44	2.762±0.43	2.793±0.40	2.403±0.65	2.427±1.05	4255
#9	Llama-4-Scout-17B-16E-Instruct	2.586±0.44	2.735±0.44	2.760±0.43	2.390±0.67	2.461±0.92	1018
#10	Llama_3.1_405B_Instruct	2.567±0.46	2.696±0.46	2.795±0.40	2.379±0.66	2.397±1.04	4276
#11	Qwen3-Next-80B-A3B-Thinking	2.536±0.49	2.662±0.47	2.763±0.43	2.337±0.72	2.384±0.97	1015
#12	phi-4	2.526±0.50	2.696±0.46	2.707±0.46	2.354±0.69	2.347±0.99	1026
#13	aya-expanse-32b	2.440±0.53	2.650±0.48	2.601±0.49	2.244±0.74	2.264±1.02	1017
#14	GPT5	2.432±0.49	2.641±0.48	2.552±0.50	2.233±0.66	2.302±1.07	4265

Figure 2: Standalone human evaluation leaderboard across all dimensions. Green indicates high performance (score > 2.40), while blue represents mid-tier performance ($2.20 <$ score < 2.40).

Rank	Source	Score	Matches	Wins	Losses	Draws
1	Gemma3_27B_instruct	**1681.72**	19071	10230	4146	4695
2	GPT5	**1636.17**	19054	6952	9237	2865
3	Qwen3-Next-80B-A3B-Instruct	**1595.22**	1990	795	710	485
4	phi-4	**1558.88**	2000	677	854	469
5	Kimi-K2-Instruct-0905	**1552.59**	1985	1079	537	369
6	Llama-4-Maverick-17B-128E-Instruct	**1493.77**	1980	621	915	444
7	Qwen3-235B-A22B-Instruct-2507	**1480.68**	1985	807	778	400
8	Krutrim-2-instruct	**1452.15**	1980	771	804	405
9	SarvamM_24B	**1429.60**	19067	7965	5916	5186
10	QWEN3_235B_A22B	**1426.09**	17465	6178	6424	4863
11	Qwen3-Next-80B-A3B-Thinking	**1411.96**	1980	570	1013	397
12	Llama_3.1_405B_Instruct	**1411.15**	17466	4266	8990	4210
13	aya-expanse-32b	**1370.02**	1995	485	1072	438

Figure 3: Comparative human evaluation leaderboard ranked by Elo ratings.

3-27B-Instruct performs best, while GPT-5 performs worst with a gap of 0.26 points, showing that language errors are more differentiated across models. For Trustworthiness, there is a smaller gap between the best performing (Qwen3-Next-Instruct) and worst performing (GPT-5) models, suggesting that this dimension is more consistent across models. For Cultural/Local Relevance, Kimi-K2-Instruct performs best while Aya-expanse-32B performs worst with a gap of 0.30 points, the largest gap in all metrics, showing that cultural relevance is

Rank		Overall Avg	Content Errors	Language Errors	Trust Rating	Local Relevance	Sample Count
#1	GPT5	2.7759	2.9656	2.9405	2.2046	2.9930	17,592
#2	Qwen3-235B-A22B-Instruct-2507	2.7705	2.9323	2.9820	2.1930	2.9749	95,140
#3	Kimi-K2-Instruct-0905	2.7684	2.9695	2.9704	2.1432	2.9903	95,140
#4	sarvam-m	2.7377	2.9201	2.9612	2.1315	2.9380	95,140
#5	gemma-3-27b-it	2.7372	2.9598	2.9301	2.1065	2.9524	95,138
#6	Qwen3-235B-A22B	2.7362	2.8884	2.9657	2.1564	2.9345	95,140
#7	Qwen3-Next-80B-A3B-Instruct	2.7251	2.8625	2.9031	2.1703	2.9645	95,140
#8	gpt-oss-120b	2.6705	2.7249	2.8907	2.1028	2.9636	95,138
#9	Qwen3-Next-80B-A3B-Thinking	2.6598	2.7528	2.8399	2.1410	2.9055	95,140
#10	Llama-4-Maverick-17B-128E-Instruct	2.6222	2.6346	2.9012	2.1104	2.8424	95,140
#11	Llama-4-Scout-17B-16E-Instruct	2.5680	2.5971	2.7944	2.0850	2.7954	95,140
#12	Krutrim-2-instruct	2.4629	2.3438	2.7134	2.0465	2.7479	95,138
#13	Llama-3.1-405B-Instruct	2.4626	2.3884	2.7616	2.0613	2.6394	95,140
#14	phi-4	2.3819	2.4312	2.4547	2.0053	2.6364	95,140
#15	Param-1-2.9B-Instruct	2.2149	2.0456	2.5088	1.9633	2.3420	17,640
#16	aya-expanse-32b	2.1892	2.3218	1.9141	1.9623	2.5590	95,140
#17	Llama-3-Nanda-10B-Chat	1.5854	1.5683	1.6941	1.4906	1.5886	95,137

Figure 4: Standalone evaluation leaderboard using LLM-as-a-judge, showing mean quality scores across models. Green indicates high performance (score $>$ 2.70), while blue represents mid-tier performance (2.10 $<$ score $<$ 2.70). Red is indicative of poor performance (score $<$ 2.10)

the biggest differentiator in standalone evaluations. The Qwen3 series (Qwen3-235B-Instruct , Qwen3-Next-Instruct) ranks in the top three across three of the four metrics. GPT-5 and Aya-expanse-32B both rank in the bottom five in all metrics, indicating they struggle uniformly across all dimensions.

Results by language: For Indian English, the average model score is 2.61, with a gap of 0.4 between the best and worst performing models. For Hindi, the average score is slightly higher at 2.65, with the smallest gap between the best and worst performing models, with scores slightly higher than Indian English. For medium resource languages like Gujarati, Punjabi and Marathi the overall average scores are lower and the gap between models widens. Assamese, being the lowest resource language among the languages under consideration has the highest gap of 0.73 between the models. Kannada has the smallest gap (0.17), indicating that model performance is consistent for Kannada despite it being a relatively low-resource language, which warrants further investigation into the dataset. Overall, the gap between models is larger for low-resource languages compared to the medium resource languages considered.

Results by domain: Models score highest overall in the education and finance domain (2.65 average score) and slightly lower on the Legal domain (2.62 average score), with Healthcare having the lowest score (2.58 average). Model performance varies the most in the Education domain (0.26 gap), while Healthcare hast the smallest gap. Qwen3-235B-Instruct performs best across all domains, while Llama-4-Maverick maintains a top-3 ranking across all domains. GPT-5 and Aya-expanse-32B rank in the bottom-3 in all domains.

Rank	Source	Score	Matches	Wins	Losses	Draws
1	GPT5	2543	76,956	75,738	1,218	0
2	Kimi-K2-Instruct-0905	1964	8,796	7,592	1,204	0
3	Gemma3_27B_instruct	1862	76,924	48,841	28,079	4
4	Qwen3-Next-80B-A3B-Instruct	1701	8,800	6,035	2,764	1
5	SarvamM_24B	1583	77,008	38,942	38,060	6
6	Qwen3-Next-80B-A3B-Thinking	1479	8,800	4,066	4,733	1
7	QWEN3_235B_A22B	1439	69,072	23,766	45,300	6
8	phi-4	1289	8,800	2,341	6,449	10
9	Qwen3-235B-A22B-Instruct-2507	1286	8,800	3,457	5,343	0
10	Krutrim-2-instruct	1209	8,800	2,729	6,067	4
11	Llama-4-Maverick-17B-128E-Instruct	1159	8,800	2,230	6,566	4
12	Llama_3.1_405B_Instruct	1072	70,380	3,616	66,759	5
13	aya-expanse-32b	914	8,800	991	7,802	7

Figure 5: Comparative evaluation leaderboard based on LLM-as-a-judge, reporting relative model performance via ELO ratings.Green indicates high performance (ELO score > 1800), while blue represents mid-tier performance ($1400 <$ score < 1800). Red is indicative of poor performance (ELO score < 1400)

Consistency: We check the standard deviation across models and find that domain consistency is highest, indicating that models adapt well across the four domains. Models have similar capabilities across metrics with moderate standard deviation. The largest performance swings are caused due to language, with top performing models like Qwen3-235B-Instruct having 0.15 standard deviation, indicating that multilingual performance is harder than domain expertise. The Qwen3-235B-Instruct model performs best overall, with highest performance in all four domains, and is in the top five in 10 languages, maintaining high consistency across metrics. GPT-5 systematically under-performs in human evaluation, ranking in the bottom-5 for overall performance across error types, domains and languages.

5.3.2 Comparative evaluation results

Figure 3 presents the relative performance rankings via Elo ratings for the comparative analysis. Elo ratings serve as a relative strength index where the difference between two models predicts the probability of one being preferred over the other in a battle. A higher Elo score indicates that a model's responses were consistently judged as superior compared to its peers within the evaluation pool.

Overall, Gemma-3-27B-Instruct performs best in comparative evaluations with a substantial gap with the lowest performing model (Aya-expanse-32B). The second and third best models GPT-5 and Qwen3-Next-Instruct perform similar to the best model, while there is more variance in the performance of the bottom five models. Human evaluators are given four options (Model 1 is better, Model 2 is better, both are good, and both are bad) and a preference for "Model 2 is better" is the most common, comprising of 38.1% of all evaluations with a preference for "Model 1 is better" comprising of 37.6%, indicating the absence of any position bias in human evaluations. In terms of ties (both good or bad), we find that the "both good" option is selected only 0.6% of the time, while "both bad" is selected 23.8% of the time, indicating that model pairs often both struggle rather than excel.

Results by language: Assamese has the highest Elo scores overall, despite being low-resource, with Gujarati having lowest aver-

age Elo. Gemma-3-27B-Instruct maintains a ranking in the top-2 in 9 out of 11 languages showing consistent performance across diverse languages, while GPT-5 ranks first in Kannada and Telugu outperforming Gemma. The Aya-expanse-32B has the widest performance variance across languages, indicating that it performs better on some Indian languages compared to others. In terms of preference patterns, Tamil has the highest number of "both are bad" selections (39.3%), while Assamese evaluations show a slight position bias (39.2% vs. 36.1% for "Model 1 is better" vs. "Model 2 is better").

Results by domain: The Education domain has the highest Elo rating, while Healthcare has the lowest Elo, mirroring results in standalone evaluation. Gemma-3-27B-Instruct performs best across all domains, while GPT-5 ranks second in Education and Finance, but drops to third place in Healthcare and fourth place in Legal domains, showing its performance variability across domains. In terms of preference, the Healthcare domain has the highest number of "both are bad" selections, indicating that this is a challenging domain for all models.

Consistency: Gemma-3-27B-Instruct has the highest variance across languages though it ranks best overall, indicating that it is uneven across languages, while Qwen3-235B-Base shows highest language consistency. Gemma-3-27B-Instruct shows consistency across domains with high performance across all domains, while GPT-5 is less consistent with weaker performance in Healthcare and Legal domains. The best model (Gemma-3-27B-Instruct) has a win rate of 54.5%, almost double of that of the worst performing model (Aya-expanse-32B) which has a win rate of 28.1%.

Human evaluator agreement: We calculate the agreement between human evaluators all data points that are evaluated by two humans. We had 30,510 records for standalone evaluations and 54,009 comparative evaluations that were judged by two experts. The top three models maintain an identical ranking in leaderboards created with both human evaluations, showing high inter-rater consistency at the top. For models in the middle, there is moderate evaluator disagreement.

6 Qualitative Analysis

In order to get an in-depth understanding of the results, we analyzed 4,749 data points across all languages, with 321-533 data points per language taken from a subset of models (GPT-5, Gemma-3-27B-Instruct , Llama-3.1-405B-Instruct , Qwen3-235B-Instruct , Sarvam-M) that were evaluated by two human annotators. We selected data points spanning a spectrum of agreement, ranging from complete agreement to varying degrees of disagreement (e.g., partial disagreement corresponding to differences of one, two, or three rubric points). We assigned scores to each category to cluster similar responses across all 11 languages. We then focused on clusters with medium to low scores to examine the sources of disagreement among data workers and to surface particularly interesting cases. In this section, we present insights from this sample to highlight the general patterns observed in responses that scored highly compared to those that did not.

High-scoring responses: In terms of top-performing languages, Punjabi, Marathi and Hindi were the languages that both human raters marked high. In the following Marathi example from the Education domain,

> "सरकारी शाळा या खाजगी शाळा प्रमाणेच कार्यरत असतात का? आणि जर असतील तर मुलांचा कल खाजगी शाळेकडे जास्त का आहे?"[Gloss: Do government schools function in the same way as private schools? And if they do, why is there a greater preference among children for private schools?]

a high-scoring response compared government vs. private schools across multiple relevant dimensions (curriculum, funding, teacher systems, facilities, accountability, parental preference), showing depth and completeness, and ended with actionable advice to visit schools to assess actual quality. Similarly, in this Marathi example from the Legal domain,

> "मला एक छानसे घर खरेदी करायचे आहे त्यासाठी मला कोणती खबरदारी घेतली पाहिजे आणि त्यासाठी कोणकोणते कायदे आहेत कोणकोणत्या कागदपत्रांची पूर्तता करावी लागेल?"[Gloss: I want to buy a nice house. What precautions should I take for that, which laws apply, and

which documents need to be completed?]

a high scoring answer covered end-to-end home-buying concerns, and reflected real Indian buying risks (fraud, illegal transactions, unclear titles), which annotators valued. The response also explicitly referenced India-specific laws and processes, and was localized to Maharashtra, showing domain and local awareness.

Low-scoring responses: In contrast, only a small number of responses were flagged as having more than two content or language issues, lacking trustworthiness, or demonstrating low local relevance, but these cases yielded valuable insights. While the scores did not always completely align, responses marked as poor by one annotator were generally also rated low by the other. In our sample, Malayalam has the most poor ratings followed by Hindi and Bengali. Some reasons for low scores included generic or vague information in responses and responses without local context that failed to address the user's specific situation, lacked actionable details, or did not adapt to the local setting implied by the question. For example, for this Malayalam question from the Education domain,

> "ഒരു വിദ്യാർത്ഥിയുടെ കഴിവുകളും താല്പര്യങ്ങളും എങ്ങനെ തിരിച്ചറിയാം?അതനുസരിച്ച് അനുയോജ്യമായ കരിയർ തിരഞ്ഞെടുക്കാൻ കരിയർ ഗൈഡൻസ് എങ്ങനെ സഹായിക്കുന്നു?" [Gloss: How can a student's abilities and interests be identified? How does career guidance help in choosing a suitable career accordingly?]

a low-scoring response to was long and detailed, recommending psychometric assessments, job shadowing, mini-internships, and industry visits. It lacked local relevance and practical guidance for Indian students, such as accessible resources, cost-effective options, or region-specific career pathways. Similarly, in this Bengali question in the Education domain,

> "আমি ফার্মাসিতে কাজ করি। নতুন নতুন ওষুধ সম্পর্কে নিজের অভিজ্ঞতাকে আরো বৃদ্ধি করার জন্য কী কোর্স করা উচিত?" [Gloss: I work at a pharmacy. Which courses should I take to further enhance my practical experience and knowledge about new medicines?]

a low-scoring response lists generic courses without concrete career outcomes and misses practical constraints such as cost, duration, and part-time options.

Incorrect geographical context: We also identified instances where questions in Bengali were answered with an apparent focus on the Bangladesh context. Since the models were not provided with any location-specific prompts, it is plausible that they assumed the user was in Bangladesh; however, this is an important consideration for model and application developers to keep in mind. For example, in this Bengali question from the Finance domain,

> "অনলাইনে ইলেকট্রিক বিল জমা করলে কিছুটা টাকা সাশ্রয় হয়। এর কারন কী? অনলাইনে বিল জমা করা নিরাপদ কি?" [Gloss: Paying the electricity bill online saves some money. What is the reason for this? Is it safe to pay the bill online?]

the response used Bangladesh-specific wallets (bKash, Nagad, Rocket) instead of Indian options (UPI, PhonePe). Similarly, for this question from the Legal domain

> "বোন একটি অনলাইনে জামা অর্ডার করেছে, কিন্তু সাইজ ভুল এসেছে এবং রিটার্ন অপশনও বন্ধ দেখাচ্ছে। এই সমস্যায় কাকে যোগাযোগ করা উচিত?"[Gloss: My sister ordered a dress online, but the size is wrong and the return option is showing as unavailable. Who should be contacted for this issue?]

the response referenced jurisdictions from both India and Bangladesh, resulting in confusion. In a question from the Education domain

> "আমার পেশার ক্ষেত্রে উন্নতির জন্য যদি আমি প্রশিক্ষণ নিতে চাই তবে মালিকপক্ষের থেকে কি কোন বিশেষ সুবিধা পাওয়া সম্ভব? আমার প্রশিক্ষণের জন্য ছুটির প্রয়োজন হলে কি আমি সবেতন ছুটি পেতে পারব?" [Gloss: If I want to take training for professional growth, can

I get any special benefits from my employer? If I need leave for training, can I get paid leave?]

the response referred to Bangladesh labor law, which is not applicable in India.

Insufficient local context: For the Content Quality metric, Kannada had a large number of low scores corresponding to "Answer is irrelevant" compared to other languages. The answers addressed the broad topic of the question but fail to understand the user's intent or local context. Responses tended to be generic and high-level, without addressing the user's situational needs, Indian-specific frameworks, or relevant eligibility criteria. As a result, the information provided, while correct, did not effectively resolve the user's query. In this Kannada example from the Healthcare domain

> "ನಮ್ಮ ಹಳ್ಳಿಯಲ್ಲಿ ಒಂದು ಪ್ರಾಥಮಿಕ ಆರೋಗ್ಯ ಕೇಂದ್ರ ಇದೆ ಅಲ್ಲಿ ಯಾವ ಯಾವ ಸೌಲಭ್ಯ ನೀಡುತ್ತಾರೆ? ಮತ್ತು ಅಲ್ಲಿ ಶಸ್ತ್ರ ಚಿಕಿತ್ಸೆ ಮಾಡುತ್ತಾರಾ?"[Gloss: There is a Primary Health Centre (PHC) in our village. What facilities are provided there? And do they perform surgeries?]

the response provided only general information about PHCs without clarifying the user's location or context. The response also incorrectly used phrasing such as "At our village Primary Health Center (PHC)" while listing generic information. Since the models were not provided with location-specific information during our evaluation, they cannot reasonably be expected to answer highly localized queries. In such cases, a stronger response would be to ask an appropriate follow-up question to clarify the user's context. In future rounds, we plan to design separate test conditions to more systematically evaluate how models handle these location-dependent scenarios. Similarly, for this Kannada question from the Finance domain

> "ಮಳೆಯ ಕೊರತೆ, ಬೆಳೆನಾಶ, ಪ್ರಾಣಿಯ ಸಾವು ಇತ್ಯಾದಿ ಸಮಯದಲ್ಲಿ ರೈತರಿಗೆ ವಿಮೆ ಯಾವ ರೀತಿಯ ಸಹಾಯ ಮಾಡುತ್ತದೆ?" [Gloss: During situations such as lack of rainfall, crop loss, or death of livestock, how does insurance help farmers?]

the response just mentions the benefits of crop and livestock insurance without mentioning schemes from the Indian government, eligibility criteria, and how farmers can actually use insurance during crop loss, drought, or livestock death.

Failure patterns by language: Some languages (Kannada, Hindi, Gujarati) had lower Content Quality scores, while others had lower Language Quality scores (Assamese, Gujarati, Tamil). Some reasons for low Content Quality scores included generic responses that were not grounded in Indian realities, or were too high-level without enough detail. In this Gujarati example from the Finance domain

> "વીમાના નોમિની કેવી રીતે નક્કી કરવામાં આવે છે? નોમિનીના અધિકારો કયા છે? નોમિની કેવી રીતે બદલી શકાય છે?" Gloss: [How is a nominee decided in an insurance policy? What rights does a nominee have? How can a nominee be changed?]

the response read like a policy manual with heavy legal phrasing, with long and dense sentence structure and overuse of legal terms, leading to a low Content Quality score. In this Gujarati example from the Education domain

> "નોકરી મળવી અને યોગ્ય નોકરી પસંદ કરવી — બંનેમાં શું અલગ છે? કઈ વાત વધુ સારી છે?" [Gloss: What is the difference between getting a job and choosing the right job? Which one is better?]

the response was overly verbose and read more like an article than like a Q&A response, with sentence construction seeming to be influenced by English making some phrases feel translated. Several phrases were direct translations from English (e.g., "work–life balance, "stepping stone"), which felt unnatural in the Gujarati context. Similarly. for this Assamese question in the Healthcare domain

> "দীৰ্ঘদিনীয়া গোটখা সেৱন কৰা পুৰুষ বা মহিলাই কেনেধৰণৰ বেমাৰৰ সন্মুখীন হ'ব পাৰে? ই স্বামী স্ত্ৰী দুয়োৰে প্ৰজনন ক্ষমতা হ্ৰাস হোৱাৰ অন্যতম এক কাৰক হ'বা পাৰে নেকি?" [What diseases can men or women face due to long-term gutkha consumption? Can it reduce fertility in both husband and wife?]

human evaluators gave low Language Quality scores to the response for grammatical issues, inaccurate word usage, verbosity, and poor flow. In a Tamil example from the Legal domain

> “"நம்ம இந்தியாவுல திருமண சட்டம் என்பது எவ்வாறு உள்ளது? ஒரு சில பேர் வந்து விவாகரத்து வாங்காமலே இன்னொரு திருமணம் எல்லாம் பண்றாங்க. அவங்களுக்கு இந்திய திருமண சட்டத்தின்படி என்ன மாதிரியான தண்டனைகள் வழங்கப்படுது? இந்திய திருமண சட்டத்தின்படி ஆண்களுக்கும், பெண்களுக்கும் திருமண விஷயத்தில் எப்படிப்பட்ட உரிமைகள் இருக்கு? மத அடிப்படையில் அந்த உரிமைகள் ஏன் மாறுபடுது?” [In India, marriage laws are framed in different ways. Some people enter into another marriage without obtaining a divorce from their first spouse. Under Indian marriage law, what kind of punishments are given to such people? Under Indian law, what rights do men and women have in matters of marriage? Why do these rights differ based on religion?]

annotators flagged several language issues, including the mixing of Tamil with English legal terminology, the use of both Tamil and Latin scripts, spelling errors, and the combination of informal language with formal legal phrasing. They also noted problems with sentence clarity and paragraph structure.

Failure patterns by domain: Among the domains we tested, Healthcare showed the lowest scores overall, suggesting that it is the most challenging domain. In this Malayalam example from the Healthcare domain,

> “കൊച്ചു കുഞ്ഞുങ്ങൾക്ക് മുലയൂട്ടുന്ന സ്ത്രീകൾ എന്തെല്ലാം തരത്തിലുള്ള ഭക്ഷണങ്ങൾ കഴിക്കണം? പാൽ ഉണ്ടാകുവാൻ ഏതെങ്കിലും മരുന്നുകൾ കഴിക്കേണ്ടതുണ്ടോ?”[Gloss: What kinds of foods should women who are breastfeeding young babies eat? Is it necessary to take any medicines to increase breast milk?]

the response included medical jargon, and the recommendations toward the end felt overly technical and difficult for the user to relate to. For example, it referenced galactagogue medications such as domperidone and metoclopramide, noted their potential side effects, and emphasized that they should not be taken without a doctor’s prescription. The Legal domain had the lowest trust scores for answers that were otherwise scored high on language and content, suggesting that models struggled to provide credible and actionable responses. The Finance domain showed the highest disagreement between raters, such as in the following Malayalam example

> “ഞാൻ വീട് വയ്ക്കാൻ വേണ്ടി ഹോം ലോൺ എടുക്കാൻ ആഗ്രഹിക്കുന്നു, സുരക്ഷിതവും വിശ്വസനീയമായ വായ്പകൾ വേണ്ടി എവിടെയാണ് അപേക്ഷ നൽകേണ്ടത്?” [Gloss: I want to take a home loan to build a house. Where should I apply to get a safe and reliable loan?]

the response to which was rated as high on trust and local relevance by one rater, with low scores given by the other rater. Overall, errors differed by domain - Legal had more problems around missing content, Healthcare had lower trust scores, while Education and Finance had more balanced error distributions. The sub-topics that models scored lowest on in Healthcare included ”Managing chronic conditions”, in Legal included ”Family and Marriage Matters”, in Finance included ”National and International Finance”, while models fared well on most of the topics in Education.

Failure patterns by model: The Gemma model was the only model that consistently gave disclaimers in responses across domains. For example, it provided the following disclaimer for Legal queries: “I am an AI chatbot and cannot provide legal advice. This information is for general educational purposes only. Consult with a qualified legal professional for advice specific to your situation”. Across domains, GPT-5 responses tended to be verbose yet insufficiently localized. Common weaknesses included overuse of abstract concepts, jargon and abbreviations heavy, lack of actionable guidance tailored to Indian users,

and language-level issues in non-English outputs. These factors collectively contributed to lower human ratings despite the apparent completeness of the answers. For example, in the Indian English example below from the Legal domain,

> **"I joined a private company on a contract basis but they never wrote my salary anywhere. They told me it will depend on my performance but even after working so hard they are paying me very less. Since nothing is written in the agreement I feel stuck and I am not able to quit. What can I do in this kind of situation?"**

the response did not sufficiently reflect the user's constrained position or explain realistic next steps within the Indian legal and employment context. It suggested raising the issue with HR, resigning and requesting a final settlement, and claiming unpaid wages under labor laws.

7 Automated Evaluation

Automated evaluations can be valuable for augmenting, rather than replacing human assessment, particularly when scaling to large datasets or multiple model variants. They enable rapid feedback during the model development cycle, allowing model builders to iterate quickly, identify regressions, and compare system changes efficiently. An automated grader designed to closely approximate human preferences can be used effectively during the model fine-tuning phase. When used alongside human evaluations, automated metrics help balance speed and scale with the depth and contextual nuance provided by human judgment. In this section, we detail experiments on automated evaluation using the Samiksha benchmark and model responses using the LLM-as-judge framework, a popular automated evaluation technique in which LLMs are prompted to provide evaluation scores along with explanations based on a set of rubrics.

7.1 Experimental setup

We use the following LLMs as judges: Qwen3-235B-Instruct , Sarvam-M , Kimi-K2-Instruct , and Gemma-3-27B-Instruct . We evaluated 17 models (listed in Table 4) using all four LLM judges across all languages. Given the significant volume of evaluations required across multiple dimensions and languages, we specifically selected open-weight, non-API-based models. This allowed us to self-host the judges, ensuring computational efficiency and cost-effectiveness while maintaining full control over the inference environment during the large-scale assessment. We then followed a similar setup as in the human evaluation with standalone and comparative evaluations. For standalone evaluations, all 23k queries were evaluated for all 17 models, except Param-1-2.9B which was only evaluated on Hindi and Indian English. Comparative evaluations were done on a smaller subset of 14 models. In all, there were 1.3 million calls made to LLM judges for the evaluation. Hyperparameters for standalone and comparative evaluation can be found in Tables 6 and 7.

The prompt for LLM-judge evaluation followed the rubrics used for human evaluation. The LLM judge is not provided with additional contextual information beyond the question and response, and therefore cannot be expected to meaningfully assess cultural relevance. However, we retain cultural relevance as a rubric to mirror the conditions of the human evaluation. Similarly, while LLM judges do not inherently possess notions of trustworthiness, we include the trustworthiness metric as well, using the same instructions provided to human evaluators. The prompts used for the standalone and comparative evaluation using LLM-judges can be found in Appendix A.2.2 and Appendix A.2.3.

7.2 Results

The leaderboards for automated evaluation can be found in Figures 4 and 5. In the standalone evaluations, GPT-5 has the highest mean score (2.78) followed by Qwen3-235B-Instruct (2.77). The score ranges from 2.78 to 1.58 across most models. For comparative evaluations, GPT-5 has the highest Elo score (2489) followed by Kimi-K2-Instruct (2143), with Aya-expanse-32B being at the bottom of the leaderboard with an Elo of 946.

Language-wise insights: Indian English dominates - the mean score on English is higher

than on Indian languages, showing the gap between these languages. Aya-expanse-32B exhibits good performance in English and Hindi (2.76 and 2.71 respectively), but has low scores for the rest of the Indian languages which it does not explicitly support. Higher resource languages (Indian English and Hindi) get the highest average scores (2.74 and 2.60 respectively) while medium resource languages score between 2.56-2.42 with Assamese, the lowest resource language in our set of languages averaging a score of 2.43. Kimi-K2-Instruct is the most consistent multilingual model with the least variance amongst languages (2.79 (Indian English) to 2.75 (Malayalam)), followed by Qwen 3 235B Instruct (2.83 (Indian English) to 2.73 (Bengali)).

Domain-wise insights: The average score in Education is 2.53, Healthcare is 2.52, Finance is 2.51 and Legal is 2.49. GPT-5 scores highest in all domains. The overall score across domains does not vary a lot, showing that models perform equally well on all domains.

Metric-wise insights: The trustworthiness metric most frequently receives a rating of 2, with LLM judges typically selecting the "trust after checking" option. The metrics are strongly correlated, particularly Language Quality with Trustworthiness, and Content Quality with Cultural Relevance. The mean scores are as follows: Content Quality - 2.58, Language Quality - 2.704, Trustworthiess - 2.04 and Cultural Relevance - 2.722. GPT-5 and Kimi-K2-Instruct get the highest score for Content Quality (2.97) and Cultural Relevance (2.99), Qwen3-235B-Instruct scores the highest in Language Quality (2.98) and GPT-5 scores the highest for Trustworthiness (2.20).

Insights from Comparative evaluations: GPT-5 consistently had the highest win rate across all domains and languages. The model win rate was fairly uniform across all domains, with GPT-5, Kimi-K2-Instruct and Qwen3-Next-Instruct being the top models. GPT-5 was the best performing model across all languages, followed by Kimi-K2-Instruct .

7.3 LLM-judge agreement and bias

We calculate the percentage agreement between judges for the standalone evaluation and find that the average exact match agreement is 45%. So, we modify this slightly to call a difference of more than two points a disagreement. We get low disagreement rates with an average of 6%, with the minimum being 2.9% (Kimi-K2-Instruct -Qwen3-235B-Instruct) and maximum being 9.7% (Sarvam-M -Qwen3-235B-Instruct). We find that the maximum disagreements between LLM-judges occur for content and language quality. LLM-judges based on Sarvam-M and Qwen3-235B-Instruct have the highest disagreement for language quality at 11.9%. Plots for the agreement and distribution across judges can be found at Appendix B.

All LLM-judges showed an agreement of more than 98% for English, where an agreement meant the rubric-averaged score differs less than 0.5. Judges show a significant deviation from their mean score while evaluating English answers, ranging from 0.15 to 0.33. Kimi-K2-Instruct -Sarvam-M and Qwen3-235B-Instruct -Sarvam-M showed slightly lower levels of agreement across Indian languages, with the lowest being 86.1% for Assamese for Qwen3-235B-Instruct -Sarvam-M . We observe high agreement of 92% among LLM-judges for the comparative evaluation task, which is consistent with our prior findings (Watts et al., 2024).

We find some instances of self bias in standalone evaluations, in which all models except Sarvam-M rank second in the leaderboard produced by their own evaluations. In the comparative evaluation setting, we observe position bias in all LLM-judges, with a preference for the first option over the second, with the Sarvam-M LLM-judge having the highest position bias of 4.8%. In terms of length bias, in prior work it has been shown that LLM-judges prefer longer responses, however, we find that shorter responses get higher scores on average.

In summary, GPT-5 consistently ranks first on both the standalone and comparative LLM-judge leaderboards, with Qwen3-235B-Instruct and Kimi-K2-Instruct following closely behind. Most models perform better in English and Hindi; Kimi-K2-Instruct is the most consistent across all 11 Indian languages. The agreement between judges is high, particularly in comparative evaluation.

8 Human-LLM Judge Agreement

We conducted a detailed analysis to compare human evaluation results with those from auto-

mated methods. For this analysis, we consider only overlapping data points evaluated by both humans and LLM-based judges. We use 30,510 standalone evaluations and 54,009 comparative evaluations for this analysis across all domains and languages.

8.1 Human-LLM judge agreement analysis

The Kendall Tau correlation overall between humans and LLM-judges is 0.27, indicating weak-moderate agreement. Both methods identify similar trends but diverge significantly on model rankings, suggesting that the two evaluation techniques are measuring different aspects of model quality. The most stark difference between humans and LLM-judges is on GPT-5, which ranks highest in the LLM-judge evaluation and lowest in standalone human evaluation. All four LLM-judges rate GPT-5 high, which points towards potential bias towards response patterns that GPT-5 produces. LLM-judges rate GPT-5 in the top three in 7 out of 11 languages, while human evaluators never place it in the top five for any language. This is not a marginal difference, but a fundamental disagreement on model quality and suggests that the responses contain linguistic patterns that LLM-judges favor, while humans consider more nuanced aspects. This extreme difference warrants further investigation, and we refer to it as the "GPT-5 paradox".

In terms of domains, both LLM-judges and humans judge models consistently high or low across domains with a few exceptions. GPT-5 is the top performing model in all four domains as rated by LLM-judges, while Qwen3-235B-Base is the top model according to human evaluators in all four domains, with GPT-5 ranked consistently low. There is a massive disagreement between LLM-judges and humans for Hindi evaluation, with LLM-judges rating GPT-5 as the best model and humans ranking it 8th. For Assamese, both humans and LLM-judges identify Aya-expanse-32B to be the lowest performing model, while differing on the best model. Similarly, Kimi-K2-Instruct performs well on some metrics and languages (Gujarati, Kannada) in the LLM-judge evaluation, but human evaluators rate it low.

The Qwen3-235B-Instruct model performs best in human evaluations and second best with LLM-judge evaluations, maintaining a top-5 ranking in both evaluation methods across all 11 languages, making it the most "reliable" model from a cross-evaluation perspective. Humans rate all four Qwen models in the top five, while LLM-judges ratings spread them out more across the leaderboard at 2nd, 7th, 8th and 10th place.

For Trustworthiness, all LLM-based judges assign similar scores across models, with GPT-5 achieving the highest ratings. In contrast, human evaluators make clearer distinctions in perceived trustworthiness and rate the Qwen3-Next-Instruct model as the best. This suggests that LLM-judges may assess confident answers as being trustworthy, while humans use more nuanced judgments.

We calculate the agreement of each LLM-judge with human evaluators to find the LLM-judge that is closest to human evaluation. The LLM-judge based on Qwen3-235B-Instruct performs best, agreeing on the top three rankings with human evaluators in 9 out of 11 languages, suggesting stronger alignment with human preferences.

8.2 Aligning LLM-judges with human preferences

We align LLM-judges to human preferences by providing human evaluations as few-shot examples for the standalone examples during prompting for two of the LLM-judges under consideration, Sarvam-M and Qwen3-235B-Instruct . We select few-shot samples from the same domain and language as the question-answer pair to be evaluated, using the instances where both human evaluators fully agreed, and include one example for each available rubric type. Figure 6 shows the updated LLM leaderboard, where we take the average of the scores provided by the improved Sarvam-M and Qwen3-235B-Instruct LLM-judges. We measure zero-shot vs. few-shot LLM-judge performance by computing the Kendall's Tau, Spearman's rho, Pearson's r correlations with human assessment. The LLM-judge based on Sarvam-M shows dramatic improvement with few-shot prompting. The Sarvam-M LLM-judge goes from non-significant agreement with humans ($p > 0.19$) to very significant agreement ($p = 0.0138$). This 80% improvement in Kendall Tau indicates that few-shot examples substantially

Rank		Overall Avg	Content Errors	Language Errors	Trust Rating	Local Relevance	Sample Count
#1	Qwen3-235B-A22B-Instruct-2507	2.8887	2.9840	2.9664	2.7346	2.8700	2,200
#2	Qwen3-Next-80B-A3B-Instruct	2.8515	2.9464	2.9163	2.6709	2.8723	2,200
#3	Kimi-K2-Instruct-0905	2.8459	2.9682	2.8809	2.6313	2.9032	2,200
#4	GPT5	2.8410	2.9595	2.8338	2.6499	2.9208	8,796
#5	QWEN3_235B_A22B	2.8288	2.9623	2.9436	2.6235	2.7859	8,798
#6	SarvamM_24B	2.8079	2.9346	2.9095	2.5806	2.8067	8,798
#7	Gemma3_27B_instruct	2.8039	2.9627	2.8836	2.5596	2.8097	8,792
#8	Llama-4-Maverick-17B-128E-Instruct	2.7180	2.7754	2.8773	2.5246	2.6946	2,200
#9	Qwen3-Next-80B-A3B-Thinking	2.6785	2.7318	2.6986	2.5377	2.7459	2,200
#10	Llama-4-Scout-17B-16E-Instruct	2.6612	2.7087	2.8018	2.4705	2.6641	2,200
#11	Llama_3.1_405B_Instruct	2.5833	2.5676	2.8179	2.4297	2.5179	8,798
#12	Krutrim-2-instruct	2.5784	2.4909	2.7073	2.4527	2.6627	2,200
#13	phi-4	2.4973	2.5577	2.5482	2.3818	2.5013	2,200
#14	aya-expanse-32b	2.3360	2.4632	2.1650	2.3068	2.4091	2,200

Figure 6: Standalaone LLM-judge leaderboard aligned to human preferences

clarify the evaluation criteria for this judge, bringing its model rankings much closer to human assessment patterns. On the other hand, Qwen3-235B-Instruct is already a relatively strong zero-shot evaluator that aligns reasonably well with human assessment, and few-shot examples provide incremental improvement compared to Sarvam-M .

For the aligned Sarvam-M LLM-judge, we find that models such as Qwen3-Next-Instruct , Qwen3-235B-Base , Llama-4-Maverick , Llama-4-Scout and Llama-3.1-405B-Instruct move up in the rankings, while models such as GPT-5, Kimi-K2-Instruct , Sarvam-M , Gemma-3-27B-Instruct move down. The most significant shift is for GPT-5, which drops five positions from rank 1 to 6.

9 Expert Evaluation

While we had consulted CSOs during benchmark design, we also conducted expert focus group discussions with them to examine how LLM responses align with community needs and contextual realities. We conducted workshops with CSOs from three domains - healthcare, education and finance. The objectives of the workshops were as follows: (1) Trust building - build mutual understanding and trust between the project team and CSOs. (2) Understanding the perspectives of CSO and domain experts - explore how domain experts and CSO representatives working with grassroots communities perceive the use of LLM and chatbots.(3) Evaluation of LLM response - discuss the strengths, gaps and cultural or contextual sensitivities in LLM-generated responses to community-based queries. (4) Evaluation criteria refinement - co-creation of evaluation criteria/ set of principles for community-centered AI evaluation. Each focus group included two to three experts from different CSOs to capture diverse perspectives. Across the three domains, we engaged 11 experts from six CSOs. The participants of these workshops were members of CSOs (both existing, who had already engaged with us during the benchmark creation process and new) and domain experts who work closely with grassroots or marginalized communities and can speak to community needs, sensitivities, and digital inclusion issues.

The research team shared a brief overview of Samiksha, followed by a rating exercise us-

ing a small set of selected user queries and corresponding LLM responses. Experts evaluated each response based on predefined criteria, including relevance and usefulness, accuracy and safety, and cultural sensitivity and appropriateness. A comparative rating exercise was also included to better understand expert preferences. This was followed by an open group discussion where experts reflected on the strengths of current LLM-based chatbots as well as gaps in the responses. The discussions also explored community priorities, cultural and linguistic sensitivities, ethical concerns, and experts' experiences with evaluation metrics and guardrails. The sessions concluded with reflections on what constitutes a good or appropriate chatbot response and on defining effective evaluation metrics. To accommodate participants across locations, sessions were conducted online via Google Meet. Discussions were primarily held in English, with participants switching to Hindi or Marathi as preferred.

The Healthcare workshop highlighted that community evaluation in healthcare chatbots goes far beyond medical accuracy, it must consider practicality, specificity, and emotional sensitivity to ensure that responses are both safe and genuinely useful across diverse user contexts. A key insight was the importance of strict safety boundaries, with experts emphasizing clear "don'ts," especially avoiding prescriptive medical advice, dosage guidance, and long descriptive explanations. One of the interesting discussions centered on urgent health concerns and self-diagnosis user behaviors. In these situations, chatbots should not offer alternative remedies such as yoga or home treatments alongside recommendations to seek medical help, "since Fear of Finding Out (FOFO) may lead users to delay essential care". Experts also noted that bullet-point formatting, empathetic language, and suggestions that acknowledge socio-economic realities help responses feel practical and relatable. They consistently suggested multi-turn conversations so chatbots can clarify background, understand whether the user is a patient, frontline worker, or caregiver, and respond accordingly. Overall, the workshop reinforced that safe and trustworthy chatbots must remain carefully constrained while improving access to reliable health information.

The Education workshop highlighted that chatbot evaluation must account for user intent, language nuance and a learner's context. Getting language nuances are important because these students may reach out for conceptual frameworks. Users can range from high school students to job seekers who aspire to grow academically or professionally, yet what is common across these groups is that they may use imprecise language, combine several questions, or seek direction without knowing how to articulate their goals. In such cases, chatbots should probe to clarify intent and help users reflect, rather than simply provide quick answers. Regional context also matters because it shapes learning pathways and aspirations. Experts noted that models tend to recommend mainstream digital or AI technology-based vocational courses as stable career options. However, a CSO from Nagaland emphasized that this does not reflect their local reality. "The ground reality for a state like Nagaland, is different, our youth aspire to move to bigger cities and pursue beauty and spa courses generally - where such considerations were missed in the responses". Overall, getting a sense of the user's location and language context becomes important in addressing queries in the Education domain.

Discussions with finance experts highlighted a clear misalignment between LLM-generated financial responses and the everyday realities of low-income users, especially women in rural and semi-urban areas. Many LLM responses assume that users understand financial terms and have access to formal banks. In reality, most community members rely on informal financial options and face challenges such as limited mobility, language barriers, documentation requirements, and limited awareness of long-term financial consequences. Experts emphasized that effective financial advice must first understand user intent such as why the money is needed, in what amounts, and under what constraints, rather than immediately offering abstract advice or solutions. They highlighted the value of step by step, conversational responses that clarify assumptions, and use simple numerical examples to support understanding and decision-making. Trust and relevance were seen as closely linked to linguistic sensitivity, alignment with regional prac-

tices (e.g., common borrowing amounts or familiar community norms), and transparent acknowledgment of system limits, including clear disclaimers that an AI chatbot is not a financial expert and should guide users toward local support. Overall, these insights suggest that LLM-based financial systems/chatbots should clearly distinguish between community-level and institutional intents, adapt explanations to users' socio-economic contexts, and move away from neutral global defaults toward culturally and linguistically grounded interactions.

Learnings from expert interviews: We identified several important considerations for designing effective focus group interviews with experts. Selecting participants who are directly involved in or familiar with chatbot development decisions made within CSOs such as community challenges, ongoing experiments, guardrails implemented, and evaluation metrics used, proved particularly valuable. Including experts from different roles within a CSO also helped surface nuanced challenges and diverse perspectives. We found that exploring a smaller set of questions in greater depth led to richer insights than covering many questions superficially. Quantitative ratings can be collected asynchronously, allowing discussions to focus on interpretation and reasoning rather than scoring. Also, to make virtual discussions more engaging, incorporating interactive elements such as live ratings, word clouds, and agree or disagree polls is something we would like to explore in the future. Language emerged as a critical factor during the focus groups. Conducting discussions mainly on English query-responses limited the capture of domain and language specific nuances. Future sessions will accommodate experts' preferred languages and include more non-English LLM responses to better understand how language shapes terminology and expression across domains. Going forward, we plan cross-domain discussions and in-person workshops involving both experts and community members to broaden perspectives.

10 Discussion

We present the first large-scale, community-driven, and methodologically comprehensive evaluation of Indian language models, spanning 11 languages and over 23,000 culturally grounded data points created by native speakers across high-priority domains. Our work combines depth and breadth for non-English languages by integrating native-speaker human evaluation, expert review, qualitative analysis, and LLM-as-judge methods within a unified pipeline. Beyond constructing leaderboards, we surface systematic domain- and language-specific performance trends, reveal divergences between human and automated judgments, and align LLM-based judges with human preferences. Together, this work demonstrates that rigorous, culturally grounded evaluation at scale is both feasible and essential for advancing high-quality language technologies beyond English.

Our overall findings from all evaluation methods suggest that the Qwen family of models performs consistently best across human and automated evaluations. However, human and automated evaluations did not always agree - most notably in the case of GPT-5, where it was ranked very low by human evaluators, while being ranked very high by LLM-judges. We also observed language-specific trends, with many models performing poorly on Assamese, while model performance on Kannada was surprisingly consistent. Another notable finding was that model performance across domains was much more consistent than across languages.

Among the domains we considered, the Healthcare domain proved to be the most challenging, with several pairs of LLM responses being marked by human evaluators as "both bad" in the comparative evaluation setting. Across metrics, cultural relevance emerged as the strongest differentiator among models in the human evaluations. This underscores that while models may produce fluent and reasonably correct responses, they still vary significantly in their cultural knowledge and contextual appropriateness. Our detailed qualitative analysis surfaced several unexpected findings, including incorrect geographical assumptions (such as interpreting Bengali queries as originating from Bangladesh), awkward language use (including literal translations of fixed expressions from English), improper mixing of languages, and impractical or unrealistic suggestions in response to advice-seeking questions.

The low agreement of LLM-judges with human evaluators suggests that LLM-judges cannot be used as a replacement for human evaluation. Further, they may reward patterns that diverge significantly from human preferences, making them risky to use during model training without proper calibration and constant drift monitoring. We see this in the "GPT-5 paradox", where the model gets very high scores in several languages from LLM-judges compared to human evaluators. For comprehensive evaluation at scale, we recommend combining human and automated methods: prioritize human evaluation for overall quality assessment, and use LLM-based judges for identifying specific error types. Because LLM judges emphasize different quality dimensions than humans, they should be calibrated to application-specific requirements and deployed only when their criteria align with those needs; otherwise, human evaluation should take precedence. Comparing and calibrating results from both approaches can improve robustness. In multilingual, multicultural settings, LLM-based judges should be used with particular caution.

To conclude, with Samiksha, we frame model evaluation as a critical step toward understanding how systems may behave in real-world, downstream contexts. Rigorous and contextually grounded evaluation can offer early signals about potential usage patterns, strengths, and risks. However, such evaluation must be situated within a broader, more holistic assessment framework, including deployment studies, user research, and impact evaluation to meaningfully determine real-world outcomes. Our work represents an inclusive, representative, and broad evaluation effort at scale, advancing the field toward more grounded and accountable assessments of real-world impact. Future iterations of Samiksha will include more languages, domains, modalities and models, with a focus on newer models trained specifically for Indian languages, culture and contexts as they are made available.

References

2023. *Intercultural Relations*, page 106–119.

2025. Ai diffusion report: Where ai is most used, developed, and built. Technical report, Microsoft AI Economy Institute. Accessed: February 13, 2026.

Prottay Kumar Adhikary, Isha Motiyani, Gayatri Oke, Maithili Joshi, Kanupriya Pathak, Salam Michael Singh, and Tanmoy Chakraborty. 2025. Menstrual health education using a specialized large language model in india: Development and evaluation study of menstllama. *J Med Internet Res*, 27:e71977.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. Which humans?

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Leona Cilar Budler, Hongyu Chen, Aokun Chen, Maxim Topaz, Wilson Tam, Jiang Bian, and Gregor Stiglic. 2025. A brief review on benchmarking for large language models evaluation in healthcare. *WIREs Data Mining and Knowledge Discovery*, 15(2):e70010. E70010 DMKD-00787.R1.

Sorup Chakraborty, Rajesh Chowdhury, Sourov Shuvo, Rajdeep Chatterjee, and Satyabrata Roy. 2025. A scalable framework for evaluating multiple language models through cross-domain generation and hallucination detection. *Scientific Reports*, 15.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. 15(3).

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and 1 others. 2025. Culturalbench: A robust, diverse and challenging benchmark for measuring lms' cultural knowledge through human-ai red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Roshini Deva, Dhruv Ramani, Tanvi Divate, Suhani Jalota, and Azra Ismail. 2025. "kya family planning after marriage hoti hai?": Integrating cultural sensitivity in an llm chatbot for reproductive health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, pages 1–23. Association for Computing Machinery.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Preprint*, arXiv:2308.11462.

Varun Gumma, Anandhita Raghunath, Mohit Jain, and Sunayana Sitaram. 2024. Healthpariksha: Assessing rag models for health chat-

bots in real-world multilingual settings. *Preprint*, arXiv:2410.13671.

Melissa Hall, Samuel J. Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. 2024. Towards geographic inclusion in the evaluation of text-to-image models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 585–601, New York, NY, USA. Association for Computing Machinery.

Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and Tejumade Afonja. 2025. The human labour of data work: Capturing cultural diversity through world wide dishes. *Preprint*, arXiv:2502.05961.

Hamna, Gayatri Bhat, Sourabrata Mukherjee, Faisal Lalani, Evan Hadfield, Divya Siddarth, Kalika Bali, and Sunayana Sitaram. 2025. Building benchmarks from the ground up: Community-centered evaluation of llms in healthcare chatbot settings. *Preprint*, arXiv:2509.24506.

Shafquat Hussain and Athula Ginige. 2018. Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. pages 698–703.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2151–2165.

Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple llm agents debate for equitable cultural alignment. *Preprint*, arXiv:2505.24671.

Fajri Koto. 2025. Cracking the code: Multi-domain llm evaluation on real-world professional exams in indonesia. *Preprint*, arXiv:2409.08564.

Jyoti Kumar and Surbhi Pratap. 2020. Detriments to cultural sensitivity in hci design processes: Insights from practitioners' experiences in india. In *HCI International 2020 - Late Breaking Papers: User Experience Design and Case Studies*, pages 142–155, Cham. Springer International Publishing.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37:65183–65216.

Milagros Miceli and Julian Posada. 2022. The data-production dispositif. 6(CSCW2).

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

Tarek Naous, Michael Joseph Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648, Nusa Dua, Bali. Association for Computational Linguistics.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu Md, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, and 7 others. 2025. Afrimed-qa: A pan-african, multi-specialty, medical question-answering benchmark dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1948–1973. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Rida Qadri, Mark Diaz, Ding Wang, and Michael Madaio. 2025. The case for "thick evaluations" of cultural representation in ai. *Preprint*, arXiv:2503.19075.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Pragnya Ramjee, Mehak Chhokar, Bhuvan Sachdeva, Mahendra Meena, Hamid Abdullah, Aditya Vashistha, Ruchit Nagar, and Mohit Jain. 2025. Ashabot: An llm-powered chatbot to support the informational needs of community health workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, page 1–22. ACM.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. *Preprint*, arXiv:2310.15461.

Aryan Shrivastava and Paula Akemi Aoyagui. 2025. Dice: A framework for dimensional and contextual evaluation of language models. *Preprint*, arXiv:2504.10359.

Namita Singh, Jacqueline Wang'ombe, Nereah Okanga, Tetyana Zelenska, Jona Repishti, Jayasankar G K, Sanjeev Mishra, Rajsekar Manokaran, Vineet Singh, Mohammed Irfan Rafiq, Rikin Gandhi, and Akshay Nambi. 2024a. Farmer.chat: Scaling ai-powered agricultural services for smallholder farmers. *Preprint*, arXiv:2409.08916.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *Preprint*, arXiv:2504.08863.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *Preprint*, arXiv:2307.09162.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. Milu: A multi-task indic language understanding benchmark. *Preprint*, arXiv:2411.02538.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data. *Preprint*, arXiv:2406.15053.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 LLM as a Judge Leaderboards

A.2 Prompts

A.2.1 Answer Generation prompt

Task: Answer Generation

System Prompt:
(We do not provide a system prompt to mimic chatbots.)

User Prompt:
Please answer the following {domain} question in {language}:
Question: {question}
Answer in {language} only. Do not exceed 400 words.

A.2.2 Standalone Evaluation prompts

Task: Content Quality Check

System Prompt:
You are an evaluator assessing the quality of {language} responses about {topic} in the context of {domain}. You must identify specific content issues based on strict criteria.

User Prompt:
Please check the answer carefully. Does it have any of these problems? Choose all that apply.
Context:
Question: {question}
Answer: {answer}
Evaluation Options:

1. The answer includes information that has nothing to do with the question.
2. The answer is related to the question, but doesn't fully answer the question.
3. The answer is repetitive or is too long.
4. The answer is missing important details or is too short.
5. The answer does not have any of these problems.

Do not include any reasoning or explanation. Return strict JSON only:

```
{
  "includes_irrelevant_info": bool,
  "does_not_fully_answer": bool,
  "repetitive_or_too_long": bool,
  "missing_details_or_too_short": bool,
  "no_issues": bool
}
```

Task: Language Quality Check

System Prompt:
You are a linguistic evaluator assessing the writing quality of {language} responses about {topic} in the context of {domain}.

User Prompt:
Please check how the answer is written. Does it have any of these problems? Choose all that apply.
Context:

Answer: {answer}
Evaluation Options:

1. Spelling or grammar mistakes.
2. Bad choice of words.
3. The answer doesn't flow smoothly.
4. It is difficult to understand the meaning.
5. The answer does not have any of these problems.

Do not include any reasoning or explanation. Return strict JSON only:

```
{
  "spelling_grammar_mistakes": bool,
  "bad_choice_of_words": bool,
  "does_not_flow_smoothly": bool,
  "difficult_to_understand_meaning": bool,
  "no_issues": bool
}
```

Task: Trustworthiness Check

System Prompt:
You are a cautious user evaluating whether a {language} answer about {topic} in the context of {domain} can be trusted.

User Prompt:
Imagine that you had asked this question. Would you trust the answer?
Context:
Question: {question}
Answer: {answer}
Evaluation Options (Select one):

- I would trust it completely.
- I would trust it, but only after checking it myself (by searching online, or by asking someone I know).
- I would not trust it. I would want an expert to check it.

Do not include any reasoning or explanation. Return strict JSON only:

```
{
  "trust_selection": "TRUST_COMPLETELY" |
  "TRUST_AFTER_CHECKING" |
  "WOULD_NOT_TRUST"
}
```

Task: Local / Cultural Relevance Check

System Prompt:
You are a cultural context evaluator judging how well the answer aligns with local expectations for {language} speakers regarding {topic} in the context of {domain}. The cultural

context would be based on the language and the question's content.

User Prompt:
Imagine that you had asked this question. How well does the answer understand the local and cultural context of the question?
Context:
Question: {question}
Answer: {answer}
Evaluation Options (Select one):

- The answer does not show any understanding of the local context.
- The answer shows only a partial understanding of the local context.
- The answer shows a complete understanding of the local context.
- It is hard to judge, as this question would not be asked in the local context.

Do not include any reasoning or explanation. Return strict JSON only:

```
{
  "cultural_relevance_selection": "NO_UNDERSTANDING" |
  "PARTIAL_UNDERSTANDING" |
  "COMPLETE_UNDERSTANDING" |
  "UNABLE_TO_JUDGE"
}
```

A.2.3 Comparative Evaluation prompt

Task: Comparative Evaluation

System Prompt:
You are a highly analytical expert evaluator. Your sole task is to compare two provided answers (A and B) to the same user question, select the superior answer based on a holistic set of criteria, and output the result in a strict JSON format.

User Prompt:

Task: Compare and Select the Better Answer
Goal: Determine which of the two answers (A or B) provides a better, more useful, and safer response to the user's question.

Audience: Internal evaluation system. Base your judgment only on the provided Question, Answer A, Answer B, and general world knowledge. Do not use external browsing or search.

Contextual Variables:

- **Language:** The question and answers are in {language}. Judge within this linguistic context.
- **Topic/Domain:** The content relates to {topic} in the context of {domain}.

Evaluation Criteria (Holistic Priority):

1. **Helpfulness/Relevance:** Does the answer directly and thoroughly address the user's core query?

2. **Factual Accuracy/Plausibility:** Is the information correct and believable? (Do not reward clear hallucinations.)
3. **Clarity/Organization:** Is the information easy to read, understand, and well-structured?
4. **Completeness:** Does it cover all critical aspects of the query without significant, obvious gaps?
5. **Safety:** Does it avoid harmful, misleading, biased, or non-compliant content? (Do not select unsafe guidance.)

Decision Rule:

- **Winner A or B:** Choose the answer that is demonstrably superior overall across the criteria. Even a slight but meaningful advantage warrants selection.
- **Not sure:** Choose this only if the answers are functionally identical (equal merit/flaws) or both are completely unintelligible/empty.

Input:

Question: {question}

Answer A: {answer_a}

Answer B: {answer_b}

Return **only** a single JSON object (no markdown fences, no extra text, no commentary).

```
{
  "winner": "A" | "B" | "Not sure"
}
```

Begin your response with the JSON object immediately.

MODEL	EN	HI	BN	GU	KN	ML	MR	PA	TA	TE	AS
GPT5	2.830	2.801	2.775	2.764	2.766	2.742	2.795	2.780	2.754	2.773	2.756
Qwen3-235B-A22B-Instruct-2507	2.827	2.780	2.734	2.767	2.784	2.753	2.780	2.772	2.765	2.759	2.756
Kimi-K2-Instruct-0905	2.794	2.776	2.760	2.771	2.771	2.748	2.780	2.779	2.757	2.759	2.761
sarvam-m	2.758	2.759	2.702	2.749	2.753	2.727	2.762	2.758	2.702	2.734	2.710
gemma-3-27b-it	2.772	2.754	2.734	2.751	2.732	2.689	2.757	2.758	2.736	2.736	2.697
Qwen3-235B-A22B	2.809	2.735	2.697	2.738	2.756	2.715	2.757	2.736	2.711	2.720	2.727
Qwen3-Next-80B-A3B-Instruct	2.827	2.770	2.722	2.720	2.708	2.664	2.736	2.717	2.702	2.695	2.720
gpt-oss-120b	2.775	2.489	2.678	2.690	2.707	2.643	2.705	2.744	2.647	2.674	2.641
Qwen3-Next-80B-A3B-Thinking	2.826	2.743	2.700	2.633	2.625	2.563	2.700	2.594	2.663	2.574	2.645
Llama-4-Maverick-17B-128E-Instruct	2.728	2.622	2.593	2.653	2.643	2.635	2.682	2.616	2.512	2.552	2.610
Llama-4-Scout-17B-16E-Instruct	2.720	2.588	2.553	2.586	2.584	2.566	2.647	2.602	2.407	2.516	2.482
Krutrim-2-instruct	2.623	2.475	2.408	2.499	2.450	2.413	2.451	2.562	2.354	2.435	2.429
Llama-3.1-405B-Instruct	2.727	2.536	2.464	2.443	2.451	2.414	2.509	2.506	2.260	2.401	2.383
phi-4	2.764	2.678	2.436	2.478	2.352	2.023	2.477	2.515	2.053	2.292	2.164
Param-1-2.9B-Instruct	2.476	1.973	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
aya-expanse-32b	2.765	2.708	2.117	2.083	1.782	2.363	2.099	2.030	2.366	1.707	2.030
Llama-3-Nanda-10B-Chat	2.650	2.313	1.461	1.429	1.300	1.275	1.583	1.761	1.118	1.269	1.287

Figure 7: Language wise Standalone heatmap from evaluations using LLMs as judges.

MODEL	HEALTHCARE	EDUCATION	FINANCE	LEGAL
GPT5	2.793	2.785	2.769	2.757
Qwen3-235B-A22B-Instruct-2507	2.782	2.782	2.769	2.747
Kimi-K2-Instruct-0905	2.776	2.779	2.767	2.751
sarvam-m	2.745	2.758	2.739	2.707
gemma-3-27b-it	2.738	2.754	2.736	2.720
Qwen3-235B-A22B	2.752	2.744	2.733	2.715
Qwen3-Next-80B-A3B-Instruct	2.716	2.756	2.721	2.709
gpt-oss-120b	2.690	2.714	2.600	2.676
Qwen3-Next-80B-A3B-Thinking	2.659	2.650	2.660	2.671
Llama-4-Maverick-17B-128E-Instruct	2.637	2.597	2.636	2.618
Llama-4-Scout-17B-16E-Instruct	2.579	2.558	2.581	2.553
Krutrim-2-instruct	2.452	2.462	2.482	2.457
Llama-3.1-405B-Instruct	2.487	2.428	2.470	2.463
phi-4	2.313	2.470	2.378	2.373
Param-1-2.9B-Instruct	2.334	2.298	2.151	2.057
aya-expanse-32b	2.190	2.252	2.167	2.148
Llama-3-Nanda-10B-Chat	1.605	1.615	1.619	1.501

Figure 8: Domain wise Standalone heatmap from evaluations using LLMs as judges.

MODEL	AS	BN	EN	GU	HI	KN	ML	MR	PA	TA	TE
GPT5	2471	2382	2173	2399	2506	2519	2437	2507	2336	2516	2467
Kimi-K2-Instruct-0905	1853	1858	1837	1897	1907	1889	1906	1895	1886	1800	1900
Gemma3_27B_instruct	1863	1830	1807	1783	1898	1836	1830	1866	1797	1803	1821
Qwen3-Next-80B-A3B-Instruct	1634	1677	1739	1626	1637	1661	1589	1661	1691	1724	1658
SarvamM_24B	1525	1563	1383	1667	1652	1473	1617	1536	1491	1643	1650
Qwen3-Next-80B-A3B-Thinking	1471	1522	1549	1431	1428	1365	1414	1348	1417	1550	1458
QWEN3_235B_A22B	1418	1423	1316	1460	1319	1521	1481	1367	1448	1513	1449
phi-4	1093	1332	1419	1332	1357	1300	1026	1290	1382	1065	1207
Qwen3-235B-A22B-Instruct-2507	1432	1401	1306	1374	1217	1448	1514	1295	1385	1447	1430
Krutrim-2-instruct	1434	1270	1281	1273	1207	1348	1315	1239	1352	1242	1324
Llama-4-Maverick-17B-128E-Instruct	1301	1228	1108	1328	1091	1298	1253	1408	1203	1199	1237
Llama_3.1_405B_Instruct	959	1069	1344	971	1070	931	969	1085	1043	871	994
aya-expanse-32b	1046	945	1238	960	1211	910	1149	1002	1070	1127	906

Figure 9: ELO ratings from the Language wise Comparative heatmap from evaluations using LLMs as judges.

MODEL	EDUCATION	FINANCE	HEALTHCARE	LEGAL
GPT5	2486	2317	2498	2373
Kimi-K2-Instruct-0905	1936	1949	1932	1869
Gemma3_27B_instruct	1755	1786	1815	1749
Qwen3-Next-80B-A3B-Instruct	1636	1703	1727	1744
SarvamM_24B	1816	1778	1535	1532
Qwen3-Next-80B-A3B-Thinking	1257	1485	1413	1568
QWEN3_235B_A22B	1387	1506	1391	1412
phi-4	1352	1296	1194	1170
Qwen3-235B-A22B-Instruct-2507	1311	1296	1462	1349
Krutrim-2-instruct	1386	1217	1329	1329
Llama-4-Maverick-17B-128E-Instruct	1280	1167	1256	1280
Llama_3.1_405B_Instruct	955	1076	1024	1173
aya-expanse-32b	941	924	925	954

Figure 10: ELO ratings for the Domain wise Comparative heatmap from evaluations using LLMs as judges.

MODEL	EDUCATION	FINANCE	HEALTHCARE	LEGAL
Qwen3-235B-A22B-Instruct-2507	2.888	2.889	2.887	2.891
Qwen3-Next-80B-A3B-Instruct	2.853	2.855	2.845	2.854
Kimi-K2-Instruct-0905	2.844	2.831	2.848	2.861
GPT5	2.831	2.826	2.865	2.842
QWEN3_235B_A22B	2.828	2.819	2.852	2.816
SarvamM_24B	2.833	2.812	2.813	2.773
Gemma3_27B_instruct	2.824	2.800	2.803	2.789
Llama-4-Maverick-17B-128E-Instruct	2.706	2.711	2.712	2.743
Qwen3-Next-80B-A3B-Thinking	2.649	2.684	2.665	2.716
Llama-4-Scout-17B-16E-Instruct	2.671	2.658	2.670	2.646
Llama_3.1_405B_Instruct	2.575	2.572	2.611	2.575
Krutrim-2-instruct	2.551	2.575	2.591	2.597
phi-4	2.611	2.493	2.437	2.449
aya-expanse-32b	2.403	2.320	2.329	2.293

Figure 11: Domain wise LLM as a judge heatmap for standalone evaluations after providing few shot examples

MODEL	AS	BN	EN	GU	HI	KN	ML	MR	PA	TA	TE
Qwen3-235B-A22B-Instruct-2507	2.843	2.849	2.916	2.875	2.891	2.905	2.893	2.914	2.945	2.877	2.869
Qwen3-Next-80B-A3B-Instruct	2.800	2.853	2.914	2.843	2.900	2.859	2.811	2.879	2.894	2.798	2.817
Kimi-K2-Instruct-0905	2.800	2.833	2.811	2.827	2.873	2.868	2.859	2.889	2.904	2.816	2.826
GPT5	2.753	2.808	2.870	2.813	2.892	2.848	2.841	2.899	2.904	2.813	2.811
QWEN3_235B_A22B	2.782	2.763	2.912	2.830	2.859	2.850	2.818	2.871	2.870	2.785	2.776
SarvamM_24B	2.735	2.724	2.858	2.828	2.849	2.812	2.809	2.879	2.855	2.765	2.772
Gemma3_27B_instruct	2.702	2.753	2.834	2.836	2.885	2.799	2.769	2.848	2.860	2.758	2.799
Llama-4-Maverick-17B-128E-Instruct	2.708	2.694	2.859	2.730	2.800	2.755	2.690	2.789	2.709	2.590	2.575
Qwen3-Next-80B-A3B-Thinking	2.654	2.732	2.880	2.643	2.834	2.491	2.601	2.796	2.586	2.739	2.507
Llama-4-Scout-17B-16E-Instruct	2.524	2.641	2.837	2.681	2.788	2.645	2.610	2.730	2.669	2.553	2.596
Llama_3.1_405B_Instruct	2.459	2.540	2.874	2.571	2.727	2.499	2.532	2.632	2.636	2.427	2.519
Krutrim-2-instruct	2.506	2.460	2.766	2.556	2.688	2.575	2.593	2.600	2.728	2.438	2.454
phi-4	2.256	2.520	2.867	2.557	2.795	2.486	2.123	2.599	2.676	2.234	2.356
aya-expanse-32b	2.146	2.263	2.856	2.197	2.835	1.956	2.481	2.317	2.250	2.508	1.886

Figure 12: Language wise LLM as a judge heatmap for standalone evaluations after providing few shot examples

B LLM Judge Agreement

13, 14 and 15 show the trends across the 4 LLM evaluators used for the automated evaluation.

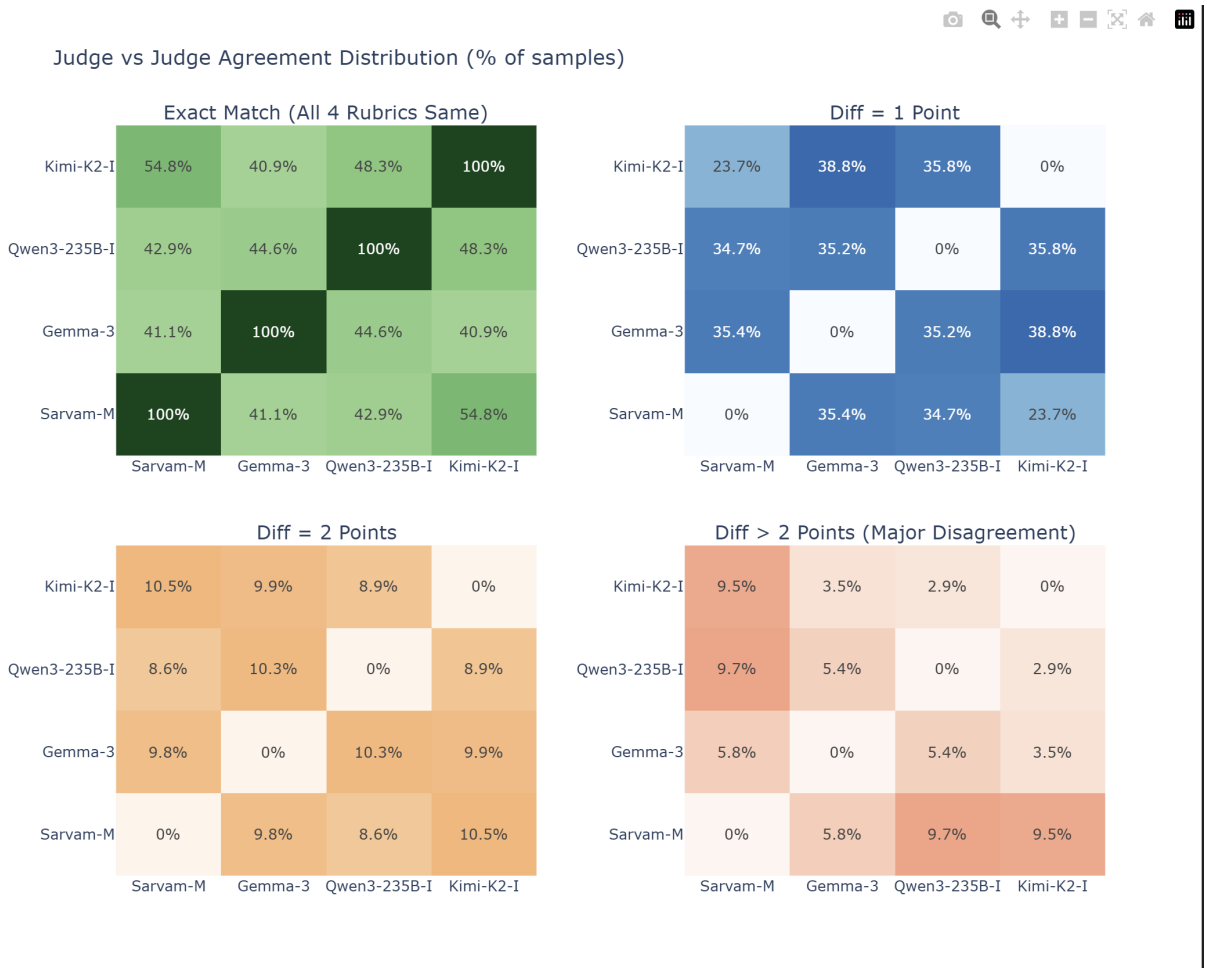

Figure 13: Percentage agreement between LLM judges

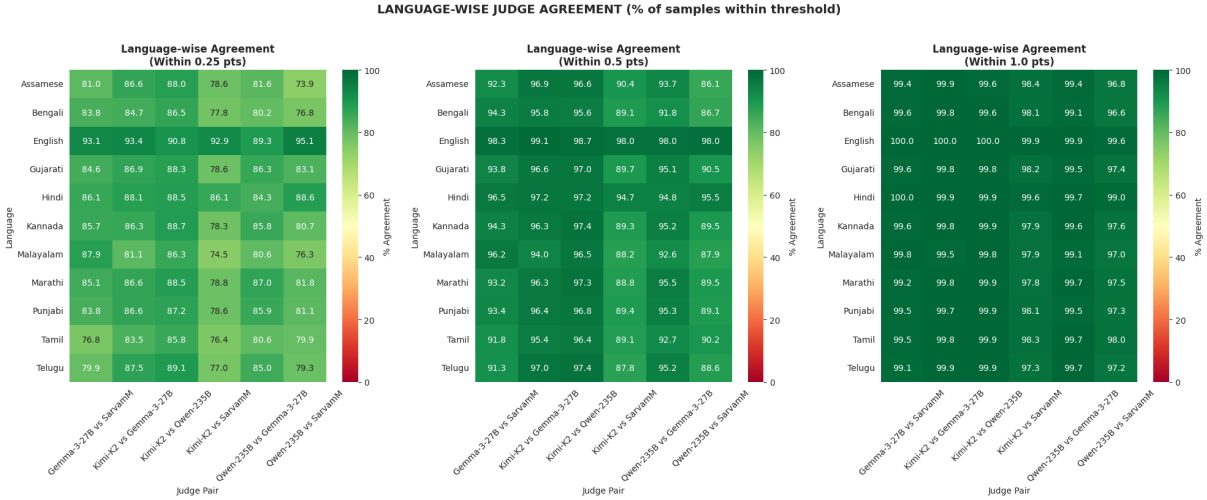

Figure 14: Language wise agreement trends for the LLM judges

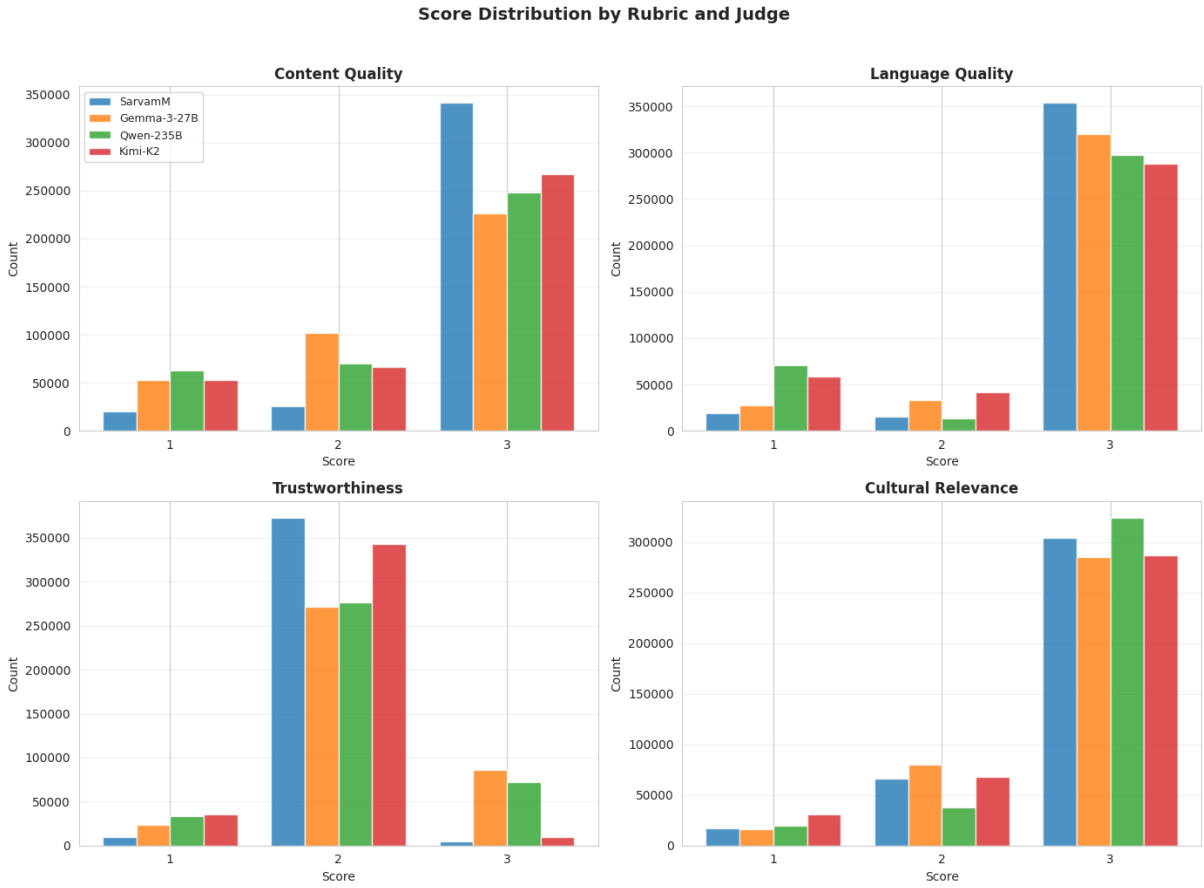

Figure 15: Rubric wise score distribution across judges