

Designing Culturally Aligned AI Systems For Social Good in Non-Western Contexts

Deepak Varuvel Dennison
Cornell University
Ithaca, New York, USA
dv292@cornell.edu

Tanuja Ganu
Microsoft Research
Bangalore, Karnataka, India
tanuja.ganu@microsoft.com

Mohit Jain
Microsoft Research
Bangalore, Karnataka, India
mohja@microsoft.com

Aditya Vashistha
Cornell University
Ithaca, New York, USA
adityav@cornell.edu

Abstract

AI technologies are increasingly deployed in high-stakes domains such as education, healthcare, law, and agriculture to address complex challenges in non-Western contexts. This paper examines eight real-world deployments spanning seven countries and 18 languages, combining 17 interviews with AI developers and domain experts with secondary research. Our findings identify six cross-cutting factors — Language, Institution, Safety, Task, End-User Demography, and Domain — that structured how systems were designed and deployed. These factors were shaped by Sociocultural (diversity, practices), Institutional (resources, policies), and Technological (capabilities, limits) influences. We find that building effective AI systems required extensive collaboration between AI developers and domain experts, with human resources proving more critical to achieving safe and effective outcomes in high-stakes domains than technological expertise alone. Additionally, we present 12 guidelines synthesizing these dynamics for designing AI for social good systems that are culturally grounded, equitable, and responsive to the needs of non-Western contexts.

CCS Concepts

• **Human-centered computing** → Empirical studies in HCI; HCI theory, concepts and models; • **Social and professional topics** → Cultural characteristics; • **Computing methodologies** → Natural language processing.

Keywords

AI For Social Good, Non-western Contexts, ICTD, AI For High-Stakes Domains, Responsible AI, Culturally-aligned AI

ACM Reference Format:

Deepak Varuvel Dennison, Mohit Jain, Tanuja Ganu, and Aditya Vashistha. 2026. Designing Culturally Aligned AI Systems For Social Good in Non-Western Contexts. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3772318.3791513>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/26/04
<https://doi.org/10.1145/3772318.3791513>

1 Introduction

Artificial intelligence (AI) technologies are spreading quickly across the globe with their largest user growth in non-Western contexts [53, 57]. AI applications now extend beyond productivity or creative tools into education [15, 25], healthcare [58, 76, 77], agriculture [86, 88], and law [9]. In non-Western contexts, this shift carries particular weight as these systems address high-stakes, wicked problems where model outputs can directly affect people’s health, learning, and livelihoods.

Recent research shows that the integration of AI into high-stakes environments have promise. AI can augment information processing [34, 65], strengthen education and training [7], boost productivity [2], and open new creative possibilities [99]. Yet the risks are profound. Prior work shows that they reproduce gender, racial, caste, and ableist biases [4, 35, 72, 84], lack cultural awareness [1, 14, 93], and provide outputs that are more accurate and relevant for Western users than for those in non-Western contexts [12, 63]. Beyond these representational harms, they risk entrenching colonial epistemologies by sidelining diverse knowledge systems [26].

Successfully using AI for social good in non-Western contexts thus requires more than deploying mainstream models; it demands deliberate cultural and linguistic alignment and developing sociotechnical infrastructures that make these systems culturally aligned, contextually relevant, and safe to use. However, given the rapid integration of AI into these settings, and despite the urgency and risks that this integration can pose, little empirical work has examined how AI systems are actually designed and implemented in such settings. We lack a grounded understanding of the tradeoffs developers face, the forms of expertise required, and the human labor that underpins this adaptation. In this paper, we ask three questions:

- RQ1:** How are AI systems adapted and contextualized for high-stakes domains in non-Western contexts?
- RQ2:** What sociocultural, institutional, and technological influences shape design and deployment?
- RQ3:** What lessons can guide the design and implementation of culturally aligned, socially beneficial AI systems in these settings?

To answer these questions, we conducted semi-structured interviews with AI developers and domain experts responsible for systems deployed at scale, reaching from hundreds to millions of users across education, healthcare, agriculture, and law. Our

analysis identifies six cross-cutting factors that shaped how culturally aligned AI systems were built and deployed: **LANGUAGE**, **INSTITUTION**, **SAFETY**, **TASK**, **END-USER DEMOGRAPHY**, and **DOMAIN (LISTED)**. Language required grounding models in local linguistic realities through technical adaptations and community-built data. Institutions determined whether systems could move beyond pilots, since institutional authority and alignment shaped adoption at scale. Safety depended on human oversight and technical safeguards, mediating trust. Task demands introduced practical constraints such as latency, modality, and environmental conditions, often requiring customized AI workflows. End-User Demography shaped system relevance through variation in literacy, age, gender and cultural norms. Domain anchored reliability by defining what counted as correct or trustworthy in each field, which required sustained involvement from domain experts. These factors intersected and reinforced each other, surfacing persistent tradeoffs.

We synthesize these findings into three higher-level influences that shaped the LISTED factors: **SOCIOCULTURAL**, **INSTITUTIONAL**, and **TECHNOLOGICAL**. Together, the LISTED factors and the higher-level influences explain which systems are possible to build, which communities they reach, and how they endure over time in non-Western contexts. Across all cases, one consistent finding emerged: human labor was central to making the systems work effectively. Developers and field workers were responsible for the curation, adaptation, and oversight that made these systems usable, contextual, and safe. This paper makes three contributions:

- An empirical account of how AI-driven applications are designed and implemented in high-stakes, non-Western domains.
- A framework detailing six factors that shape AI system design and three overarching influences that govern how these factors manifest in practice. Together, these components clarify why certain design choices emerge, how systems become usable for diverse populations, and what enables long-term adoption and legitimacy.
- Twelve practical guidelines derived from these findings to support practitioners in building culturally aligned AI systems for high-stakes, non-Western settings.

2 Related Work

HCI and ICTD scholarship has long emphasized that technologies are never value-neutral but are deeply embedded in cultural, political, and economic contexts [38, 39, 69, 82]. Early ICTD work warned that interventions fail when detached from local realities, noting that technology tends to amplify existing intent and capacity rather than substitute for them [94]. The One Laptop per Child (OLPC) project illustrates this point: despite its visionary appeal, OLPC laptops were often underused or repurposed for games because teachers were untrained, curricula were not adapted, and basic infrastructure like electricity and internet access was missing [3]. As Ames argues in *The Charisma Machine*, techno-solutionist approaches risk overlooking structural inequities and local context, leading to disappointing outcomes. Postcolonial computing extended this critique, showing how design practices frequently reinscribe Western epistemologies and marginalize alternative ways of knowing [43, 71]. Together, these insights call for careful cultural

adaptation — technologies must be grounded in local needs and designed with attention to the sociocultural, socioeconomic, and sociopolitical forces that shape their adoption and use. This is especially crucial for technologies meant to address complex social problems, where failures can have outsized and harmful effects on already marginalized communities [74, 97].

With the rise of LLMs, HCI and ICTD scholars have increasingly turned to these technologies to tackle complex social problems in high-stakes domains such as education, healthcare, law, and agriculture — an area often framed as AI for Social Good (AI4SG) [18, 32, 87]. A growing body of work shows that LLMs have potential to improve teaching and training [70], address health information inequities [33], and boost productivity [66]. But these promises come with significant risks. Research also shows that LLMs routinely reproduce identity-based biases [4, 35, 72, 84], lack cultural awareness [1, 14, 93], and can produce outputs that are systematically less accurate and relevant for users in non-Western contexts—creating quality-of-service harms [12, 63]. Anuyah et al. [5] review cultural considerations for AI design in the Global South and show that most models still reflect Western-centric assumptions [73, 100]. Similarly, Karamolegkou et al. [46] identify persistent challenges in NLP for social good, including data scarcity and bias in low-resource languages, misaligned evaluation metrics, safety and ethical risks, limited cross-cultural generalization, and infrastructural gaps. Beyond these representational harms, research shows that LLMs risk entrenching colonial epistemologies by sidelining diverse knowledge systems [26]. These limitations make cultural and contextual alignment essential: without it, AI systems risk not only technical failure but also the erosion of trust and legitimacy in the very communities they aim to serve.

In response to such risks and harms, AI4SG scholars have argued for culturally responsive, community-driven AI solutions that prioritize equity and inclusivity [13, 17, 24, 90, 95]. Yet, despite these calls, little is known about *how* AI developers and domain experts navigate the complex sociotechnical, institutional, and cultural factors that shape these systems in practice and *what* sociocultural, institutional, and technological factors shape their design and implementation.

Our study addresses this gap by analyzing real-world AI deployments at scale across multiple domains — education, healthcare, agriculture, and law — in non-Western contexts. We identify six cross-cutting factors (Language, Institution, Safety, Task, End-User Demography, and Domain) that structure system design and implementation and synthesize them into three higher-order influences that explain adoption, legitimacy, and sustainability of these interventions. In doing so, we provide one of the first empirically grounded accounts of how AI for Social Good systems are built and maintained in high-stakes, non-Western contexts, offering concrete design recommendations for building culturally aligned, socially beneficial AI.

3 Methods

To investigate how AI systems are built and deployed for non-Western populations in high-stakes domains, we examined real-world deployments using key-informant interviews and secondary research drawing on existing reports and documentation across all

of the projects studied. The study protocol was approved by the Institutional Review Board (IRB), and all procedures adhered to established ethical research guidelines.

3.1 AI Project Selection

We used a purposive sampling to select AI for social good projects. The projects were selected based on the following criteria: (i) targeted a problem in a high-stakes domain, (ii) were deployed on the ground with active users in non-Western contexts, (iii) were made available to users at no financial cost and operated by non-profit institutions, and (iv) incorporated at least one AI technology as a core component. In our study, we categorize mainstream language technologies as AI technologies. We focus specifically on Machine Translation (MT), Language Models, and Automatic Speech Recognition (ASR), as they constitute the most widely deployed forms of AI today and therefore provide a useful lens for understanding patterns that extend to other AI domains. We also sampled across domains, geographies, languages, and institutional contexts to ensure diversity.

Table 1 shows the **eight projects** we selected spanning **four domains** (education, healthcare, agriculture, legal). With the exception of two projects (PROMPTS and FarmerChat), all the projects were cross organizational collaborations, where a non-governmental organization (NGO) handled the implementation and a technology company handled the technical aspects. Aforementioned two organizations, however, had sufficient in-house capacity to develop AI technologies given the large scale of their operations.

The projects varied widely in scope – from pilot scale (< 500 users) to large deployments (up to ~350,000 users). The projects were based in **seven countries** (India, Bangladesh, Colombia, Kenya, Ethiopia, Nigeria, Ghana) and supported **18 languages**. Projects leveraged multiple delivery mediums (Web apps, Mobile apps, WhatsApp, SMS) for delivering the solution, and a range of AI technologies (LLMs; speech technologies for Text-to-Speech (TTS)/Speech-to-Text (STT); Neural Machine Translators (NMTs); transformer-based classifiers). Two of the projects – LegalTranslateAI and ReadAI – are pseudonymized in the paper to ensure anonymity as requested by the participants.

3.2 Semi-structured Interviews

After shortlisting the projects, using a combination of cold emails and our personal network, we contacted AI developers and domain experts working on these projects to serve as our primary informants. These two groups were selected because their complementary expertise in technical design and on-the-field implementation provide a comprehensive understanding of the challenges these projects experience when deployed on-the-ground.

AI developers. We define AI developers as technologists directly responsible for building the AI components of the tool. Their work included integrating existing models and pipelines and, when necessary, fine-tuning or training new models. Often in these projects, small development teams took on multiple responsibilities, including both system engineering and AI development.

Domain experts. We define domain experts as practitioners who worked closely with field staff, end users, and AI developers to ensure that the tool meet on-the-ground needs. Their responsibilities included communicating user requirements to the technical teams, coordinating deployment logistics, and monitoring how the system operated in local contexts and impacted on-the-ground workflows.

We conducted interviews between June and July 2025 with **17 participants** (10 AI Developers and 7 Domain Experts) across different projects. Interviews were semi-structured, lasting between 60 and 90 minutes, and conducted either via Zoom or in person (see Appendix B & C for interview protocols). All participants provided informed consent and were assured that they could withdraw at any point. With explicit permission, interview sessions were audio-recorded and subsequently transcribed using the Whisper STT model, followed by manual cleaning to ensure accuracy.

3.3 Secondary Research

We complemented interviews with secondary research [91] conducted beforehand, drawing on published research papers, web articles, and AI model cards related to the chosen projects. This provided background context, guided our interviews, and enabled us to triangulate information across multiple perspectives. To ensure accuracy, we confirmed and cross-checked these details with the key informants during interviews and through follow-up member-checking (see §3.4).

3.4 Analysis

We conducted a qualitative thematic analysis [31] of all study materials. The research team iteratively developed a coding system that combined inductive themes emerging from the data with deductive categories grounded in an AI-adaptation taxonomy, following Liu et al. [55], which distinguishes *in-weight* and *in-context* adaptations for enabling cultural alignment in AI systems.¹

Coding proceeded in iterative cycles. We began with open coding on a subset of transcripts, then consolidated labels into a codebook that was refined through discussion and application to additional data until stable. The final codebook included 63 codes (e.g., *challenges with accent*, *need for good data*, *challenges between modalities*), organized into 10 categories (e.g., *Language Contextualization*, *Social Factors*, *Safety & Validation*), providing both granularity and thematic coherence for full corpus analysis.

Member checking. To strengthen validity, we employed member checking, returning synthesized accounts of each project to participants for review [61]. For each project, we produced a 7–9 page summary document that outlined key findings and situated them within the emerging framework. Participants reviewed these documents, suggesting clarifications or corrections. Following qualitative best practices, we incorporated all factual adjustments, recorded points of interpretive disagreement, and revised our analysis accordingly.

¹**In-weight** adaptations modify model parameters or learnable components (e.g., pretraining or continued pretraining on new corpora, supervised fine-tuning, adapters/LoRA layers, vocabulary or acoustic-model updates), thereby changing the model itself. **In-context** adaptations leave parameters unchanged and instead shape behavior via inputs and orchestration (e.g., prompt and system-instruction design, few-shot exemplars, retrieval-augmented generation, tool use and pipeline routing, translation pivots, and guardrails).

Project	Domain	AI Tech	Usage	Languages	Geography	Year of Launch
VetBot	Agriculture	LLM, STT, TTS	< 500 users	Tamil, English, Kannada, Hindi, Malayalam	India	2025
FarmerChat	Agriculture	LLM, STT, TTS	~15,000 farmers	Swahili, Amharic, Hausa, Hindi, Odiya, Telugu, Kannada, English	Ethiopia, India, Kenya, Nigeria	2023
PROMPTS	Healthcare	LLM, BERT Classifier	~350,000 women; ~1.1M SMS/mo	Swahili, Hausa, Yoruba, Xhosa, Zulu	Ghana, Kenya, Nigeria	2023
CataractBot	Healthcare	LLM, STT, TTS	~2000 patients	English, Kannada, Hindi, Telugu, Tamil, Urdu	India	2024
ASHABot	Healthcare	LLM, STT, TTS	~5,000 health workers	Hindi, English, Telugu, Marathi	India	2024
ReadAI*	Education	STT	~5,000 children	Hindi, Marathi	Colombia, India	2025
Shiksha Copilot	Education	LLM	~1,100 teachers	Kannada, Telugu, Hindi, English	India	2024
LegalTranslateAI*	Legal	NMT	~120,000 Court Judgments	Hindi, Kannada, Tamil, Telugu, Punjabi, Marathi, Gujarati, Malayalam, Bengali, Urdu	Bangladesh, India	2019

*Project name pseudonymized

Table 1: Overview of included projects

Participant ID	Domain	Role	Gender	Experience
P1	Agriculture	Domain Expert	Male	35
P2	Education	AI Developer	Male	3
P3	Agriculture	AI Developer	Female	22
P4	Healthcare	AI Developer	Female	15
P5	Education	Domain Expert	Male	32
P6	Education	AI Developer	Female	6
P7	Healthcare	AI Developer	Male	14
P8	Education	Domain Expert	Female	18
P9	Legal	Domain Expert	Male	28
P10	Agriculture	AI Developer	Male	13
P11	Healthcare	AI Developer	Male	19
P12	Healthcare	Domain Expert	Male	10
P13	Healthcare	Domain Expert	Male	20
P14	Agriculture	Domain Expert	Female	18
P15	Healthcare	AI Developer	Male	2
P16	Agriculture	AI Developer	Male	18
P17	Legal	AI Developer	Male	14

Table 2: Participant Information

With only minor refinements, participants were largely in agreement with our representations.

To protect confidentiality, participants were anonymized. Raw data (audio, transcripts, notes) were stored on encrypted drives

with restricted access, and any potentially identifying details were redacted from materials shared for review.

3.5 Researcher Positionality

We recognize that researcher backgrounds inevitably shape the design and interpretation of this work. All authors are of Indian origin. Two are researchers at a university in the United States, and two are researchers at an industry research lab in India. Two authors were part of the technology development teams for CataractBot, ASHABot, and Shiksha Copilot; however, they did not take part in the study interviews. Their experiences informed the development of the interview protocol, but all insights discussed in this paper originate from interviews conducted with other lead AI developers involved in the projects. Three of the four authors have extensive experience designing and deploying technologies for education, health, and governance in non-Western contexts. This experience gave us both access to participants and deep familiarity with the institutional, cultural, and infrastructural challenges of building AI systems in these settings.

Our methodological choices were informed by this background. We used semi-structured interviews to foreground participants' perspectives and allow them to surface challenges and tradeoffs beyond those anticipated by the research team. To reduce potential bias, all interviews were conducted by the author with limited prior professional connection to participants, helping ensure open and candid conversations. Participants were assured of anonymity, which encouraged them to share both successes and critical reflections on their projects. We see this work as part of a broader commitment to centering local expertise and documenting the opportunities and challenges that developers and domain experts experience in designing culturally aligned AI systems in high-stakes, non-Western contexts.

4 Findings

Our findings show that AI developers and domain experts collaborated extensively, taking into account a range of considerations to ensure that applications worked effectively for their users. We identify six key factors, LANGUAGE, INSTITUTION, SAFETY, TASK, END-USER DEMOGRAPHY, and DOMAIN (LISTED), that shaped both system building and deployment of culturally aligned AI technologies (detailed in §4.1-4.6). To incorporate these LISTED factors, developers employed a mix of in-context and in-weight adaptations to the AI models, summarized in Table 3.

4.1 Language

AI developers and domain experts collaborated to ensure that the AI solutions function effectively in low-resource and non-Western languages. While language also encompasses facets such as domain-, geography-, and culture-specific vocabulary, these are discussed in detail in subsequent sections. Here, we focus on the *basic language capabilities* of the systems — whether they can correctly understand input and produce coherent output in the target low-resource languages. Even such capabilities required dedicated efforts, as these languages are often poorly supported in mainstream LLM models [22, 67, 89].

4.1.1 Language and Technical Considerations. Developers made key decisions about which models best handled target languages, balancing fluency, domain accuracy, and model availability.

Selecting most suitable models. The AI developers began by identifying models best suited to their target language and context. They worked with domain experts to ensure that the linguistic nuances of the target users are accurately captured in the model. Because commercial and open-source models differed widely in language capabilities, the teams tested several configurations before deciding which to deploy. For example, in Shiksha Copilot, GPT-4 generated lesson plan content in English, while the Indic model Sarvam 2B translated it into Kannada. Although the lesson plans were meant for Kannada, they were first created in English because GPT-4 produced pedagogically high-quality content in English but struggled in Kannada. Sarvam 2B, by contrast, generated fluent Kannada but with weaker pedagogical quality. To reconcile these trade-offs, developers adopted a hybrid pipeline, assigning each model to the task it handled best.

Most projects relied on informal, qualitative assessments to evaluate language capabilities, but FarmerChat adopted a more systematic approach. Working with domain experts, they fine-tuned a Llama 3 model with agricultural vocabulary to identify domain-specific terms across languages. They also built gold-standard transcripts for multilingual voice datasets collected from their users. Each time a new ASR model was released, they evaluated it by comparing ASR outputs with these transcripts using the fine-tuned Llama model. Performance was measured with standard metrics such as word error rate and an “agree score” that penalized mistakes on agricultural terms — crop names, pests, diseases, units, numerals — which accounted for about 20% of the vocabulary used by the farmers in FarmerChat. Even a single error (e.g., mistaking *mushroom* for *masoor*(red lentil)) could distort a farmer’s query. This evaluation process revealed substantial variation across ASR models and allowed FarmerChat to select the best available model for each target language.

Geographical disparities in model support. Developers working across multiple regions identified stark geographical disparities in available models. This disparity was especially pronounced for teams like FarmerChat working across nationalities. As P14 explained:

“In India, even though it’s not perfect, at least many of the major languages are supported by the models. But in Africa it’s much harder, because a lot of the languages just aren’t supported even though there are so many users. For example, in Ethiopia we work in two zones that use two different languages, Amharic and Oromo. Amharic is supported to some extent, but Oromo isn’t supported at all in most models.” (P14)

Temporality of model selection. The AI developers emphasized that model selection was not a one-time technical decision but an ongoing process shaped by both language requirements and the evolving availability of technologies at different points in time. For example, LegalTranslateAI’s use of NMT to translate legal documents from English to vernacular languages was shaped by the technologies available to developers at the time. As one LegalTranslateAI developer reflected:







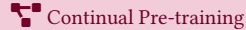
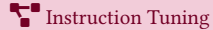

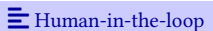


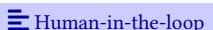


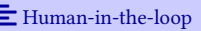



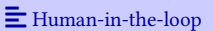


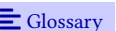
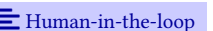
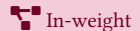

Project	AI Adaptation Techniques
VetBot	 RAG  Prompt Engineering
A web-based veterinary assistant for farmers. It uses RAG over a curated knowledge base and carefully designed prompts to simplify responses in local dialects.	
FarmerChat	 RAG  Prompt Engineering  Glossary  Fine Tuning
A mobile agricultural chatbot for farmers. It combines RAG with curated knowledge bases, glossaries for local terms, and fine-tuned models to improve language and speech recognition in the agriculture domain.	
PROMPTS	 Continual Pre-training  Instruction Tuning  Human-in-the-loop
An SMS-based maternal health service. It uses continually pre-trained and instruction-tuned models for low-resource African languages, with a classifier routing urgent queries to human experts.	
CataractBot	 RAG  Human-in-the-loop  Prompt Engineering
A WhatsApp-based bot supporting patients before and after eye surgery. It uses RAG with vetted clinical and logistical resources, with human doctors reviewing and revising responses before delivery.	
ASHABot	 RAG  Human-in-the-loop  Prompt Engineering  Glossary
A WhatsApp-based bot for community health workers. It combines RAG, glossary adaptations for local terms, and prompts for a friendly tone, with unanswered queries escalated to human experts.	
ReadAI	 Model Training  Human-in-the-loop
A mobile child reading assessment tool. It trains ASR models on children’s speech and integrates human review to correct outputs and improve accuracy over time.	
Shiksha Copilot	 RAG  Instruction Tuning  Glossary  Human-in-the-loop  Prompt Engineering
A web-based lesson-plan generator for teachers. It uses RAG to ground lesson content in official curricula, instruction-tuned models for vernacular translations, and teacher reviews for classroom adaptation.	
LegalTranslateAI	 Fine Tuning  Glossary  Human-in-the-loop
A web-based legal translation system. It relies on fine-tuned NMTs for legal vocabulary, with human translators post-editing and judges verifying outputs.	
	 In-weight  In-context

Table 3: AI adaptation techniques used across projects. The explanations of these techniques are provided in Appendix A.

“LegalTranslateAI was developed in 2019 when using LLM-based approaches was not yet possible. This led us to adopt NMT as the default.” (P17)

These experiences underscore that effective multilingual AI deployment requires treating model selection as a dynamic, iterative process rather than a static technical specification.

4.1.2 Human Infrastructure. Effective language support depended not only on models but on human labor, as domain experts and field staff identified shortcomings, created workarounds, and supplied the resources needed to train or adapt systems.

Role of domain experts. Given the uneven language support in the AI landscape — where certain languages and dialects are better supported than others — the human effort required to contextualize systems varied across both language and modality (text or voice). Domain experts contributed in two main ways: (1) identifying the language shortcomings of the systems and (2) improving their language capabilities.

Language shortcomings were often surfaced through feedback mechanisms from domain experts. All projects conducted regular log analyses and used multiple channels to gather user feedback. For instance, in ASHABot and CataractBot, domain experts received daily logs of interactions by email, enabling them to spot system errors and ask AI developers to apply fixes where possible. In addition, field staff collected feedback through their on-ground programmatic activities. In VetBot, for example, field volunteers not only trained farmers to use the application but also relayed user feedback directly to the developers.

Resource-based adaptation strategies. Acting on this feedback, however, depended on organizational resources. Most teams began with prompt engineering because it was inexpensive. When VetBot’s staff reported that Tamil Nadu farmers used dialectal variants the system failed to recognize — for example, different words for “fever” — developers tried adding these variants to prompts. But as cases multiplied, the approach proved unsustainable, and users were asked to communicate in “standard Tamil.” Better-resourced teams pursued more robust fixes. FarmerChat, for example, supported

Bhojpuri-speaking farmers in Bihar by co-developing a glossary of Bhojpuri words commonly substituted for Hindi. Because models handled Hindi far better than Bhojpuri, the glossary allowed the system to map terms into Hindi before processing — an easier solution than training a Bhojpuri model from scratch.

Human effort for in-weight approaches. When in-context techniques proved ineffective, developers relied on in-weight approaches such as pre-training and fine-tuning, which were far more resource-intensive. ReadAI required STT models for conducting reading assessments in Hindi and Marathi. Existing STT models, designed for adult speech, performed poorly with children, prompting ReadAI's parent organization to build a new model. With support from field staff and partner organizations, they collected ~2,500 hours of voice data from 180,000 children. Annotation was similarly labor-intensive: in addition to trained annotators, 2,900 volunteers contributed by annotating 33,000 samples, spending a total of 475 hours.

Unlike Tamil or Bhojpuri, where prompt engineering or glossaries could mitigate challenges, certain very low-resource African languages such as Hausa and Xhosa lacked even minimal model support. To fill this gap, PROMPTS undertook continual pre-training and instruction tuning of their own LLM — a resource-intensive effort that demanded close collaboration with field partners to annotate training data. Despite the cost and difficulty, the team acted quickly, recognizing the urgency of reaching users otherwise excluded. As P11 reflected:

“If I waited a year for the GPT models to be good enough, that’s hundreds of thousands of moms that may not have had their questions answered.” (P11)

These examples demonstrate that the human resources needed to accommodate various languages and communication formats differed significantly. The approaches ranged from low-effort methods like prompt engineering to highly resource-intensive strategies such as large-scale data collection and model training, depending on both the organization’s available resources and the specific linguistic complexities involved.

4.2 Institution

Institutional factors enabled broader systemic adoption. These involved aligning systems with government regulations and policies, organizational workflows, and institutional priorities.

4.2.1 Aligning with Institutional Requirements and Knowledge. Institutional alignment was essential for adoption, requiring systems to comply with policies, embed institutional knowledge, and safeguard user privacy.

Policy Compliance. For institutional adoption, alignment with institutional requirements was essential: compliance with rules and embedding institutional knowledge such as curricula in education, medical protocols in health, or legal terminology in courts. Domain experts played a critical role in surfacing these requirements, and in working with AI developers to embed them into the system’s outputs. For example, in Shiksha Copilot, lesson plans had to be created in formats specified by government guidelines for teachers. As P8 explained, the format itself was crucial: regardless of

the quality of the content, teachers would not consider using AI-generated lesson plans unless they matched the required structure, since the documents also served as official records for submission to their supervisors. Domain experts were critical in interpreting such requirements, which varied across different states. For example, Telangana mandated a different lesson plan format than Karnataka. Comparable requirements arose in FarmerChat, where institutional requirements shaped how the system had to be adapted. Operating across multiple states and countries meant that its advice had to align with the guidelines of local agriculture departments, which varied by region and were binding on farmers. As P14 explained:

“If you work in Andhra Pradesh, the government there doesn’t want to promote chemical practices at all. They only want regenerative agriculture, natural farming, etc. For that particular location, in the database, it’s only their documents. We would never even call the general knowledge base.” (P14)

Institutional Knowledge. To ensure compliance, AI developers relied on domain experts to curate knowledge base resources that accurately reflected governmental policies, as these experts were closely connected with the relevant government bodies. For instance, in ASHABot, patients often asked health workers about government schemes, and health workers depended on ASHABot for accurate answers. When Rajasthan introduced the LADO scheme, which provides financial incentives to families with a baby girl, the system initially failed to respond because the policy was absent from the knowledge base. Domain experts identified this gap through user feedback and log analysis, then updated the knowledge base so future queries could be answered correctly.

Privacy Compliance. Privacy was another critical concern given the data-intensive nature of these applications. For in-weight adaptations, organizations had to comply with institutional data protection requirements. For example, ReadAI follows India’s Digital Personal Data Protection Act, 2023, which mandates lawful and transparent processing of personal data and requires consent from parents (or guardians) in the case of minors. Accordingly, ReadAI obtains parental consent for the use of children’s voice data. In CataractBot, all personally identifiable information (PII) is removed before queries are sent to the LLM, while non-PII attributes such as age and gender are retained to generate more contextualized responses.

Collectively, these instances show that institutional alignment was as critical as technical performance. From formats and policies to privacy safeguards, systems could gain legitimacy and adoption only when they embedded institutional rules and practices directly into their design.

4.2.2 Institutional Support. Support from institutions, including government agencies, deployment partners, and field staff, played a central role in determining how AI systems moved from development to real-world use. Institutional backing shaped the visibility of these systems, influenced user trust, and ultimately affected patterns of adoption. These forms of support were themselves tied to the institutional capacity available within each country.

Institutional Capacity. Institutional capacity for developing AI systems depends on both organizational resources and the broader

environments in which organizations operate. Some regions benefit from strong government support and established research infrastructures, while others lack them. This produces a geography of uneven technological production in which well-resourced contexts develop new systems and others rely on adapting tools created elsewhere. The trajectories of ReadAI and LegalTranslateAI illustrate this pattern. Both were created by Indian organizations for local users, and their use in Colombia and Bangladesh has centered on adaptation rather than building new systems. These inter-dependencies are reinforced by the global concentration of AI expertise and infrastructure, including widespread reliance on foundation models developed by large US-based companies. As P11, an African developer, noted, many NGOs in resource-constrained settings cannot match the salaries or working conditions of major technology firms, limiting their ability to recruit AI specialists. AI therefore amplifies existing institutional disparities because the capacity to develop and sustain such systems remains unevenly distributed across regions.

On-ground Support. In many cases, such as Shiksha Copilot, the institutional support was mobilized not only through formal structures but also through the broader excitement surrounding AI. P5 explained how the launch of the AI-driven lesson planning tool by the Education Minister significantly boosted its visibility, encouraging more teachers to adopt it. In addition, integrating Shiksha Copilot in the existing structures of the partner organization, such as training and ongoing support from teacher mentors, resulted in more adoption. Similar dynamics played out in health context. In CataractBot, hospital counselors were central to onboarding patients, including guiding them on how to use the application. On-the-ground support also proved indispensable in ASHABot, where early assumptions about WhatsApp’s familiarity proved misplaced. As P12 reflected:

“Initially, we assumed that since health workers already used WhatsApp, no training on the bot was necessary. Over time, we realized training was essential, given the variation in their digital literacy, so in the new district rollout, we now provide it.” (P12)

Institutional Trust. Trust-building was another dimension where institutional partnerships played a decisive role. In PROMPTS, deployments began only after formal MOUs with county governments were signed. The service was branded under the county’s name, rather than as PROMPTS, allowing mothers to engage with it as a trusted government service. In contrast, FarmerChat faced greater challenges in building trust. When the system was advertised on Facebook, many potential users were skeptical, assuming that free services must be scams. As P14 noted, overcoming this skepticism requires time and the gradual building of a critical mass of users, whose participation in turn inspires trust for others to join. Yet, despite these challenges, Facebook ads remained necessary for scale, as in-person onboarding of farmers was too resource- and time-intensive.

Challenges without institutional support. The absence of institutional support was most stark in VetBot. Attempting to scale the AI tool through volunteers, without government or formal institutional involvement, proved to be a major bottleneck. Unlike other

cases, the lack of structured institutional involvement hindered adoption and limited its reach.

These findings show that institutional support was not just helpful but critical. AI systems could scale only at the speed of institutional trust. Without such backing, even well-designed systems struggled to gain traction.

4.3 Safety

Safety factors were central across the design, implementation, and evaluation of these systems, with an emphasis on ensuring outputs were reliable and did not cause harm to end-users. While core concerns such as accuracy and prevention of overtly risky behaviors were systematically addressed, more subtle forms of harm — such as cultural, caste, and identity-based biases — received no attention due to limited time and resources. Safeguards were implemented at multiple points in the workflow, most notably at two stages: before inputs entered the system and during the processing of outputs.

4.3.1 Safety Measures Before Processing Inputs. Safety safeguards began even before inputs reached the main AI pipelines, combining curated knowledge, prompt guardrails, and embedded classifiers to prevent harmful or unlawful outputs.

In-context Techniques. In systems that relied on knowledge bases, as described in §4.6.1, domain experts played a key role in guaranteeing safety by curating knowledge artifacts only from verified and authoritative sources. This process ensured that the foundations of the AI systems were trustworthy before any responses were generated. AI developers also introduced technical safeguards at the input stage. A basic measure was to use prompt guardrails to filter harmful or irrelevant queries. In ASHABot, queries related to a baby’s gender reveal are rejected, since revealing a baby’s gender during pregnancy is prohibited in India.

In-weight Techniques. In PROMPTS, safety mechanisms were embedded directly into the model through continual pre-training and instruction-tuning. The team incorporated the BeaverTails safety dataset [44], which contains adversarial prompts designed to expose unsafe behaviors, making the system more robust. Beyond this, they developed a custom input classifier to filter messages before they reached the LLM. This required substantial human effort: roughly 100,000 messages sent by users to the human helpdesk were manually annotated for training data. A RoBERTa-based model was then trained to classify message intent and assign an emergency-risk score, which determined whether a query was routed to the LLM or escalated to the human helpdesk (§4.6.2). As P11 noted, this classifier was particularly important because incoming messages were highly code-mixed, often blending vernacular languages with English — an area where off-the-shelf intent classifiers consistently failed. The broader challenges of ensuring safety in multilingual settings were underscored by P4:

“Security vulnerabilities are hard to manage in multilingual contexts. Models are trained with guardrails to block dangerous requests, like instructions for weapons or content about self-harm, but these protections can break down in other languages. That creates two risks: malicious users may exploit the gap, and ordinary users

may unintentionally trigger unsafe outputs when discussing sensitive topics the model does not recognize. The problem is especially visible in areas like suicide prevention, where cultural differences and metaphorical language make detection more difficult.”(P4)

These strategies highlight that ensuring safety required not only technical guardrails but also substantial human effort, especially in multilingual settings where risks were harder to detect.

4.3.2 Measures During Workflow Processing. Both machine- and human-driven measures were employed during data processing to ensure safety, with human oversight often taking precedence over automated checks.

Machine Oversight. A common safeguard was the use of LLM validation agents to audit outputs before delivery. This strategy was implemented in applications such as Shiksha Copilot, FarmerChat, and PROMPTS. The validation agents were guided with examples and instructions to review outputs and reject those that did not comply with established guidelines. In PROMPTS, the validation agent checked the accuracy of responses against the mother’s question before releasing them.

Human Oversight. Human oversight remained the cornerstone of safety, with technical safeguards serving as complements rather than substitutes. The balance between automation and human review depended on the stakes of the task and the maturity of the models. In PROMPTS, non-medical queries were handled by LLMs, while medical queries were routed directly to human experts. Oversight was especially strict in new languages: while mature Swahili models could deliver responses directly, Hausa outputs were reviewed by a helpdesk expert until developers gained confidence through feedback. In CataractBot, every message was reviewed by doctors before being sent to patients, with verified responses marked by a badge to reinforce trust. Other domains embedded human oversight into task workflows. In ReadAI, field evaluators could override or correct model outputs. Patterns of errors flagged by evaluators were then fed back to improve the model. In legal contexts, LegalTranslateAI combined AI assistance with rigorous human validation. Expert translators produced drafts with the system, but final documents required judicial approval before release. Even then, judgments carried disclaimers noting that the translations were for general information only and had no legal standing.

These workflows show that reliability depended less on automation alone than on carefully balanced human-machine oversight tailored to the risks of each domain.

4.4 Task

Our participants also had to account for task-specific factors, including the physical environments in which applications were used, model design choices to meet task demands, and latency requirements.

4.4.1 Environmental Requirements. AI developers emphasized the importance of investing effort into understanding the environment within which an AI application would be used. In ReadAI, for example, AI developers and domain experts initially collected children’s voice data using noise-canceling microphones to ensure

high-quality recordings. However, once the model was deployed in real-world settings, they realized that the typically environments where the assessments are done are not noise-free. The model trained on noise-free data performed poorly under those conditions, requiring the team to redo their data collection to reflect these real-world conditions. A similar adjustment shaped FarmerChat’s design. Because farmers primarily interact with the application while working in the fields, voice data was collected in those same farming environments to ensure that the models could handle the background noise typical of such settings.

4.4.2 Task Requirements. Different tasks demanded distinct technical choices, from preserving student errors to producing nuanced translations or adapting tone and style to application needs.

Task dictates design. The goal of a task determined the design of underlying AI systems. For instance, for most ASR systems, the aim is to interpret the intended meaning of speech. But in ReadAI, the purpose was different: to capture students’ mistakes with precision. This required a fundamental shift in design — rather than relying on language models to smooth errors, as standard ASR does, the system emphasized the acoustic signal so that mistakes were preserved instead of corrected.

Typical ASR systems correct user errors by relying heavily on language models. In assessment tasks this is harmful because mistakes must be preserved, so the system should depend more on the acoustic signal than on language modeling. (P6)

Translation requirements. Task requirements also shaped translation strategies. In Shiksha Copilot, the challenge was generating Kannada lesson plans from English. Commercial NMT systems like Azure AI Translate handled short sentences reasonably well but struggled with long lesson plans, where translations needed to be less mechanistic and more contextually nuanced. To address this, the team instruction-tuned an Indic LLM with long academic paragraphs, enabling it to manage long-form academic content translation.

Task-driven stylistic choices. Other applications faced equally task-specific but stylistically distinct demands. In ASHABot, the goal was to sustain an informal, conversational tone that mirrored how health workers spoke with patients — an intent reflected in the WhatsApp bot’s name, *Saheli* (“female friend” in Hindi), which signaled approachability. By contrast, LegalTranslateAI required precise and standardized legal terminology for court translations. Although legal documents are long and complex — making them well suited to LLMs — the team prioritized vocabulary control over flexibility. NMTs offered this control, so despite the advantages of LLMs for handling long-form text, the system continued to rely on NMTs. These examples illustrate how translation and language-processing choices varied across applications depending on task requirements, whether accuracy of speech errors, handling of long-form content, conversational tone, or strict adherence to legal vocabulary.

4.4.3 Latency Needs. Latency — the delay between input and response — was central to usability, pushing developers to weigh quality, cost, and responsiveness.

Critical in realtime chats. Latency was especially important in applications like FarmerChat, where farmers expected real-time conversational exchanges. As P16 explained, fine-tuned LLMs generated more domain-aligned responses but were costly to train and serve. To balance quality and cost, the team experimented with LoRA adapters, which preserved many of the quality gains at lower cost. However, this approach introduced significant inference latency, slowing down responses to the point that it undermined the user experience and proved unsuitable for deployment.

More control with in-weight approaches. Applications that carried out extensive in-weight adaptations had greater control over latency. In ReadAI, reading assessors conducted large numbers of assessments daily and could not afford delays, especially while working with children. To address this, the system was designed to run models offline directly on assessors' devices, substantially reducing latency and ensuring functionality even in areas with poor or no connectivity. Similarly, PROMPTS avoided latency challenges by relying on instruction-tuned models that generated responses directly, without extra steps such as translation or retrieval.

UX-based mitigation. Systems that relied more heavily on in-context adaptations experienced considerable latency delays. In ASHABot and CataractBot, multiple steps — translation before and after RAG workflows, and in some cases human verification — added to latency. Developers adopted expectation management and UX workarounds to mitigate frustration. In ASHABot, health workers were asked to pause at least 30 seconds before sending the next message. In CataractBot, patients were informed upfront that responses would arrive only after a doctor's review. These strategies helped normalize delays but could not eliminate them, leaving responsiveness constrained in time-sensitive interactions.

These cases show that managing latency required both technical adaptations and user-facing adjustments, making latency a design choice as much as a performance constraint.

4.5 End-User Demography

End-User Demographic factors — such as geography, culture, age, gender, and income — were also critical in shaping how end-users engaged with the applications. In particular, aligning systems with cultural symbols and geographic norms was a pervasive concern, as regional variation directly influenced their relevance and adoption.

4.5.1 Geography and Culture. Geographic and cultural diversity demanded careful adaptation, as systems struggled to balance broad regional coverage with hyper-local variation that shaped everyday relevance.

Western-centric defaults. Our participants noted that geographical diversity of the users introduced multiple layers or granularity: across geographic states (different languages), within states (regional variations), and even within regions (dialectal differences). Each layer shaped not just the vocabulary used but also the cultural references and knowledge embedded in applications. Addressing this granularity required deliberate effort, as AI models often defaulted to Western-centric norms and practices. This was especially salient in Shiksha Copilot, where lesson content had to match learners' environments. P2 explained, LLMs often produced Western examples less relevant for Indian learners:

“For example, when teaching a chapter on components of food and trying to show the vitamins or minerals present in dishes that students commonly eat at home, the AI often produces examples like Caesar salad because it is trained on Western data. Indian children, especially in rural areas, are not exposed to such food. This makes it harder for them to understand the context of the lesson, which in turn might reduce the learning outcomes.” (P2)

To mitigate this, Shiksha Copilot relied on prompt engineering to explicitly request culturally relevant Indian examples. In contrast, other projects with domain- and application-specific knowledge bases relied on selective retrieval through RAG to handle regional variation and cultural adaptation. For instance, in CataractBot, the food that the patients consumed after surgery varied based on regions. The system had to retrieve the correct regional knowledge chunks to generate accurate responses. While higher-level adaptations were generally easier to achieve using these approaches, fine-grained contextualization remained nearly impossible at scale, since communities often use hyper-local terms and explanations intelligible only within small geographic pockets. P14 explained:

“I'm sure this is not just in agriculture, it is also in health and other domains. Local communities everywhere have very specific ways of explaining things and use words that are understood only within that small geographic area. These variations change a lot, and this kind of contextualization is almost impossible for an large-scale application like FarmerChat to manage, at least for now.” (P14)

Regional Clustering. This challenge led to a pragmatic compromise: clustering fine-grained localities into broader regional categories that preserved meaningful differences while remaining computationally feasible. In Karnataka, Shiksha Copilot's domain experts identified three regional clusters — North, South, and Coastal — that captured the most educationally relevant variations in dialect, cuisine, climate, and local flora and fauna. In every lesson plan, rather than attempting comprehensive local coverage, the AI system generated multiple regional examples, delegating final selection to teachers' local knowledge.

Data collection efforts for in-weight adaptations also used a similar strategy, prioritizing representativeness over exhaustive coverage. As P6 explained:

“It's not possible to cover the entire state. But of course, within these states, we chose districts such that they covered a wide variety of linguistic diversity. We used the Linguistic Diversity Index, which signifies how many different mother tongues are spoken in a particular region. So, based on that we tried to optimize for the most diversity. For example, in Uttar Pradesh we selected 10 districts that together reflected the maximum possible diversity.” (P6)

Geography-specific Glossaries. Developers also experimented with geography-specific glossaries to address local variation. Just as a Bhojpuri glossary supported farmers in Bihar, FarmerChat has been curating glossaries for different regions, while P9 stressed

the importance of partnering with local experts and government agencies to build such resources.

These approaches reveal a fundamental tension in AI localization: the impossibility of capturing fine-grained local variation within scalable systems. The projects managed this through strategic abstraction — identifying the level of geographic granularity that balanced cultural relevance with operational feasibility. Rather than pursuing exhaustive local coverage, systems created manageable regional clusters, built representative datasets using diversity indices, and relied on local experts to bridge the final gap between regional categories and hyper-local contexts. This division of labor emerged as a pragmatic necessity, with AI handling broader regional variations while human expertise provided the fine-grained cultural translation that no system could anticipate or encode at scale.

4.5.2 Other Demographic Factors. Factors such as age, literacy, income, and gender shaped AI system design across three key areas: interaction modalities, access mediums, and underlying content.

Interaction Modalities. Preferences for text or voice were shaped largely by literacy, age, and accessibility. In FarmerChat, VetBot, and ASHABot, voice was essential because users were far more comfortable speaking than typing. This created technical hurdles, since mainstream ASR systems offered limited or no support for many low-resource languages. Text support in AI models did not imply voice support: for instance, LLMs could generate Kikuyu text for FarmerChat’s users, but no Kikuyu ASR existed. Age and accessibility further reinforced demand for voice. In CataractBot, elderly patients preferred voice notes due to visual impairments and familiarity with spoken communication. Even where voice tools existed, quality and cost shaped implementation. VetBot initially used Chrome’s free STT, but users said its Tamil output sounded like “foreign Tamil spoken on BBC Radio”, leading developers to switch to Google’s paid Speech AI for more natural interaction.

Access Mediums. The medium through which the user accessed the tool was driven by factors such as income and infrastructure considerations. PROMPTS was deliberately delivered as an SMS service so that low-income mothers with only feature phones could access it without mobile data costs. In contrast, many Indian chatbots leveraged WhatsApp, where data was relatively inexpensive and adoption widespread. However, platform choices involved complex tradeoffs beyond user preferences. While WhatsApp offered user-friendly interfaces, using the WhatsApp Business API required lengthy logistical processes to set up accounts with Meta. These constraints forced VetBot to deploy via a web app despite WhatsApp being more accessible to users. Similarly, Shiksha Copilot shifted from a chatbot-based service to a web app primarily because teachers required richer interfaces to edit lesson plans — functionality that simple messaging platforms couldn’t support.

Content Considerations. Demographic factors also shaped the underlying data and content within systems themselves. Gender influenced linguistic patterns: in FarmerChat, women farmers often used non-technical terminology to describe agricultural machinery, while men relied more on technical terms. Age created similar vocabulary differences, with older farmers engaging systems using traditional agricultural terminology and practices, while younger

farmers were more open to exploring modern techniques. These demographic differences required deliberate attention to data collection and model training. In ReadAI, developers made conscious efforts to ensure gender diversity in children’s voice data collection, recognizing that voice models are sensitive to gender. Resource constraints also influenced content design itself. Shiksha Copilot deliberately designed lesson plans containing activities that required minimal or no materials, making them usable in under-resourced classrooms. Similarly, VetBot ensured its recommendations were realistic and actionable for low-income farmers rather than suggesting expensive solutions beyond users’ means.

These examples demonstrate that demographic factors shaped every layer of system design — from modalities and platforms to content and interaction styles. Addressing literacy, age, income, and gender considerations was not an afterthought but fundamental to creating AI systems that could function effectively in diverse, resource-constrained contexts.

4.6 Domain

Domain factors ensured that AI systems could meet the complex demands of specialized tasks by incorporating the necessary domain knowledge and vocabulary needed. While curated glossaries helped with terminology, capturing and applying deeper expertise still required significant involvement from domain experts, as discussed below.

4.6.1 Curated Knowledge Bases over LLM World Knowledge. In high-stakes domains, AI systems relied on curated knowledge bases to ground outputs in verified data, reducing risks from LLM world knowledge while ensuring domain and context specificity.

Knowledge grounding for domain specificity. Developers emphasized the central role of knowledge bases in producing context-sensitive, domain-specific outputs. In high-stakes areas such as healthcare, responses had to be grounded in verified sources rather than an LLM’s world knowledge. Most projects relied on Retrieval-Augmented Generation (RAG) [51], using expert-curated resources as the backbone of the applications. The extent to which developers used an LLM’s world knowledge varied with the stakes. In Shiksha Copilot, lesson plans were anchored in prescribed textbooks, with LLM knowledge used only for supplementary activities. Even so, substantial human oversight was required: teachers had to review roughly 800 lesson plans per grade across six grades in every state of deployment. This review must be repeated whenever curricula change, creating recurring and unpredictable workloads. As P8 noted, this ongoing effort threatens the project’s long-term sustainability, especially as it expands to new regions.

In health applications such as ASHABot, CataractBot, and VetBot, generation was fully grounded in verified knowledge bases maintained by domain experts. Their involvement was essential not only for curation but also for catching retrieval errors. In VetBot, for example, experts found that the system returned treatments for goats when asked about lambs, due to subtle distinctions in Tamil (*aadu* for goat and *semmari aadu* for lamb). Such cases highlight the challenges RAG faces in low-resource languages, where small lexical nuances can lead to serious errors.

RAG systems in English. To mitigate such risks, several projects chose to build their RAG knowledge bases in English, even though the end users interacted with the systems in vernacular languages. As a result, any local-language knowledge artifact had to be translated before it could be added to the RAG system. For example, FarmerChat’s parent organization has disseminated agricultural knowledge through thousands of farmer-created videos since 2008. The transcripts of these videos also inform the knowledge base in the RAG system. Because these transcripts were in local languages, they could not be used directly. Although AI developers implemented machine translation to reduce effort, substantial human oversight was still required to ensure accuracy.

Building upon pre-existing data. A key enabler for applications such as FarmerChat and PROMPTS was the availability of rich pre-existing data, which allowed the use of in-weight adaptation techniques such as fine-tuning and instruction tuning to develop domain-specific models. As P11 emphasized, the main challenge was not the technology but the data, noting that *“having good, local, up-to-date contextual data is essential”* and explaining how they receive thousands of questions daily from mothers to expand and update their databases, ensuring models can be refreshed with the most recent data every six months or annually. Similarly, P16 noted that maintaining fine-tuned models across languages was less a technological challenge than a data one — the real difficulty lay in the human effort required for preparing and managing high-quality datasets:

“Fine-tuning itself is not the difficult part. Around 80% of the effort goes into preparing and collecting the data. Once the architecture, framework, and code are in place, I can adjust hyperparameters and fine-tune another model with relative ease.” (P16)

These cases highlight that while knowledge bases were the backbone of effective AI applications, their success depended less on the sophistication of the underlying models than on the availability, quality, and continual curation of domain- and context-specific data.

4.6.2 Human-in-the-Loop Workflows. Human-in-the-loop (HIL) workflows enabled expert intervention whenever AI systems fell short, ensuring appropriate oversight of domain risks, responsiveness to timely needs, and alignment with local contexts.

Human Fallbacks. Even with curated knowledge bases and fine-tuned models, domain experts remained indispensable during use. Participants noted that chatbots inevitably encountered questions outside their knowledge bases, making HIL workflows essential for filling gaps and ensuring accuracy. In CataractBot, when the knowledge base lacked an answer, queries were routed to the primary doctor. Although responses were sometimes delayed, patients expected their doctors to be busy, and the information still reached them directly. By contrast, the same approach struggled in ASHABot. When health workers posed questions not covered in the knowledge base, the system forwarded them to local knowledge experts via WhatsApp. Multiple responses were then crowdsourced and synthesized by the LLM into a single answer. Crowdsourcing was important because the local knowledge experts were not trained doctors and might have gaps in their knowledge. While this reduced

individual errors, it often failed in practice: health workers usually needed immediate answers to respond to patients, but local experts — occupied with their own responsibilities — frequently replied too late. By then, health workers had either sought information elsewhere or, in some cases, left the patient without guidance. The contrast between CataractBot and ASHABot reveals how timing requirements and expert capacity determine the viability of HIL approaches.

Delegation based on risk. PROMPTS adopted a different HIL model: the AI system handled non-urgent and non-medical queries, while medical and urgent queries were routed to a dedicated help desk staffed by human agents. This hybrid division of labor was shaped by two factors: (1) their use of an instruction-tuned LLM, which generated responses with greater variability and less predictability compared to the more controlled outputs of a RAG-based approach, and (2) the availability of trained staff who could provide mothers with more personalized, context-sensitive support. This division of labor succeeded because it matched system capabilities to query types while ensuring dedicated human capacity for high-stakes interactions.

Accountability considerations. In domains where errors carried legal consequences, HIL workflows shifted toward human-led rather than AI-led processes. In LegalTranslateAI, which was used for translating legal documents in courts, expert court translators worked in a HIL role. They used the AI system to generate initial drafts but performed most of the work themselves, ensuring that legal vocabulary was precise and consistent. Given the official status of these documents, ultimate responsibility for accuracy and validity rested with the human translators rather than the AI.

Layered HIL processes. Educational contexts presented a different challenge: the need for multiple layers of human judgment to address both pedagogical quality and local adaptation. In Shiksha Copilot, lesson plan preparation followed a layered HIL process. First, a group of teachers used the AI system to curate plans — correcting language errors, ensuring pedagogical quality, and aligning content with state-level curricula. These curated plans were then passed to classroom teachers, who conducted a final round of contextualization with AI support. They adapted the plans to their students and local geography, making hyper-local adjustments that could not be anticipated earlier. This step was especially important given the difficulty of capturing fine-grained geographical variations in advance, as discussed in §4.5.1.

The varied approaches across these deployments demonstrate that effective HIL workflows required careful alignment between domain needs, expert capacity, and system limitations. Success depended not on simply adding human oversight, but on designing intervention points that matched domain-specific requirements — whether ensuring medical safety, legal accountability, or educational relevance. Human expertise thus served as the critical bridge between AI capabilities and real-world deployment constraints.

5 Discussion

Our findings identify six cross-cutting factors that shaped how culturally aligned AI systems are built and deployed in high-stakes, non-Western contexts (Figure 1). **Language** required grounding

in local languages through technical adaptations and community-driven resources. **Institutions** acted as mobilizers, with formal rules and informal legitimacy determining scale. **Safety** provided the trust infrastructure through indispensable human oversight complemented by technical safeguards. **Tasks** imposed constraints (latency, text length, environmental conditions) forcing departures from standard practices. **End-User Demography** shaped relevance through literacy, age, gender norms, and access channels. **Domain** considerations anchored reliability by establishing what counted as correct or trustworthy, necessitating sustained domain expert involvement.

The LISTED factors did not operate in isolation but interacted to make AI usable, legitimate, and trustworthy. Language, Domain, and End-User Demography grounded systems in local realities, while Institution, Task, and Safety determined scalability and reliability. Literacy constraints (End-User Demography) intersected with latency requirements (Task) to drive voice-first design; institutional mandates on formats and protocols (Institution) reinforced domain experts' role in safeguarding accuracy (Safety). Building AI systems for social good thus requires balancing multiple, interdependent constraints rather than optimizing individual factors.

While some of these factors apply to any software project, AI's stochastic and data-dependent behavior [10] introduces distinct complexities, as we show in the following sections. Building on prior work in technology design and adoption [23, 83, 96, 98], we synthesize three higher-level influences – Sociocultural, Institutional, and Technological – that shaped the LISTED factors and ultimately determined how culturally aligned AI systems were built, adopted, and sustained. We then present twelve guidelines derived from our findings and situate them within this broader literature.

5.1 Sociocultural Influences

Sociocultural influences shaped how developers addressed Language, End-User Demography, Domain, and Safety factors, determining which communities were served and how trust was established. Prior work has long shown that sociocultural contexts shape technology design [29, 39, 41, 48, 62], but AI intensifies these dynamics. Traditional software has constrained inputs (forms, drop-downs, fixed workflows) and deterministic outputs, making cultural contextualization more bounded and predictable. In contrast, conversational AI systems accept open-ended, unstructured inputs (freeform text, speech, images) embedded with cultural and semantic symbols specific to local communities, and generate probabilistic outputs where the same query may yield different results. This unpredictability makes failures opaque and difficult to debug, and it can erode user trust in systems that behave inconsistently.

G1: Effective systems require AI developers and local domain experts to collaborate as equal partners throughout the full lifecycle of the application. Our findings show that culturally aligned AI systems emerged when AI developers and local domain experts contributed jointly across design, implementation, evaluation, and scaling. Local experts were not simply intermediaries for reaching users; they brought indispensable knowledge of community practices, language use, and on-the-ground workflows (§4.1.2). AI developers contributed complementary technical

expertise that shaped model behavior, data pipelines, and system architecture (§4.1.1). Across projects, the sociocultural complexity of deployments and the long-tailed input and output spaces shaped by community-specific language use required both forms of expertise to be present throughout the lifecycle. Nuances tied to Language and End-User Demography continued to surface as systems were used (§4.5.1), and teams relied on collaborative interpretation and decision-making to refine system behavior over time. Establishing structures that support shared ownership, ongoing dialogue, and continuous feedback enabled both groups to contribute fully and helped ensure that the system remains culturally grounded and responsive.

G2: Culturally aware AI systems require dedicated attention to the linguistic and cultural gaps present in low-resource contexts. Our findings show that teams working in non-Western settings often needed to invest substantial effort simply to achieve basic language functionality. In high-resource languages such as English, models generally perform well out of the box [52], so most work focuses on refining domain-specific behavior. In contrast, many communities in our cases relied on low-resource languages or dialects where baseline model performance was poor, and the underlying NLP inequities were especially visible [12, 45, 63, 78]. Addressing these foundational gaps required curating dialect-sensitive vocabularies (§4.5.1), developing human review processes to check meaning and translation quality (§4.6.2), and instruction-tuning models to better support under-served linguistic and demographic groups (§4.1.2). These activities demanded ongoing human and financial resources rather than one-time fixes, and teams that accounted for these investments early on were better able to build systems that remained usable and culturally grounded across diverse contexts.

G3: Dominant languages often provide workable alternatives to unsupported local languages in contexts where communities are multilingual. Our findings show that even with substantial effort, fully supporting some marginalized languages was not feasible because they were unsupported by existing AI infrastructures [45]. In these situations, teams relied on the multilinguality that was common in many non-Western contexts, where communities often spoke both a local language and a more dominant regional language [16, 19, 20, 47]. Rather than attempting full modeling of severely under-resourced languages, teams adapted the dominant language used by the same population. For example, FarmerChat worked with Hindi but introduced Bhojpuri-relevant glossaries and phrasing adjustments to create more usable interactions for Bhojpuri-speaking farmers (§4.1.2). These cases show that when direct support for a marginalized language is out of reach, building thoughtful adaptations on top of a closely related or widely spoken alternative can improve accessibility while remaining grounded in the community's linguistic reality.

G4: Community trust in AI is a key influence on system design in both mitigating doubt and preventing over-reliance. Our findings revealed that the trust in AI varied widely across communities and domains, echoing prior work that identifies capability, usability, and social influence as antecedents of trust [6, 8, 21, 56]. Across our cases, trust emerged as a sociocultural process shaping

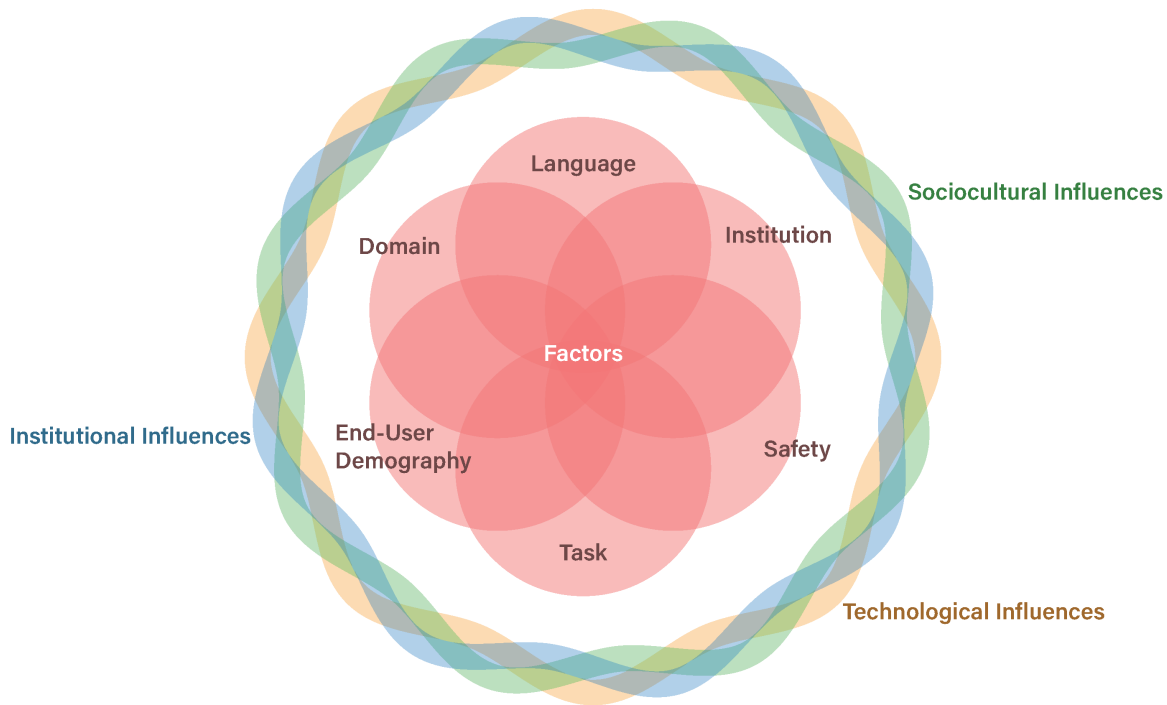


Figure 1: Factors and Influences shaping AI Systems for Social Good in non-Western Contexts

how systems were structured and deployed. In CataractBot, patients expressed confidence in responses because they were reviewed by doctors, aligning system workflows with cultural expectations of medical authority (§4.6.2). By contrast, LegalTranslateAI included disclaimers noting that AI-generated legal translations had no legal standing (§4.3.2) to prevent users from placing unwarranted trust in the outputs. These cases illustrate how system design was shaped by community perceptions of authority, expertise, and risk, requiring attention to both skepticism and over-reliance in context.

5.2 Institutional Influences

Institutional influences shaped how AI systems were adapted, governed, and sustained, often defining the limits of what was possible beyond technical ambition. Prior work shows that AI adoption depends heavily on the institutional environments in which systems operate [68, 85], and our cases reinforce this, particularly in high-stakes domains where institutions determined how responsibilities were shared between humans and technologies. Consistent with research on AI use in business [27, 42, 50], top leadership support from governments and NGOs was pivotal for adoption. Institutions, however, varied widely in technical capacity, staffing, and resources, shaping how deeply models could be adapted, how much human expertise could be sustained, and whether systems could scale over time. These dynamics highlight the need to design for institutional fit: choosing adaptation strategies that match capacity, budgeting for ongoing human labor, and planning for long-term maintenance rather than one-time builds.

G5: Compliance with institutional mandates and workflows shapes whether AI systems achieve practical uptake. Our cases echoed prior research that highlights the importance of aligning new technologies with institutional mandates and workflows [23, 80]. Despite the novelty of AI and the clear productivity gains it offered, institutional alignment remained non-negotiable for successful adoption. In Shiksha Copilot, teachers' initial enthusiasm only translated into use once the system's outputs matched state-prescribed lesson plan formats and curricular expectations, and integrated with the lesson planning workflows of the teachers (§4.2.1). A similar pattern appeared in health deployments. In FarmerChat, the advice provided by the chatbots had to be aligned with the government advisories. These cases show that improvements in service quality were not enough on their own; AI tools gained traction only when they fit smoothly into established institutional routines rather than creating new ones.

G6: AI strengthens existing institutional capacities but cannot substitute for missing foundations. Across our cases, AI systems performed well only when they were embedded within institutions that already had basic structures, workflows, and staffing to support their use. The technology enhanced existing processes but could not compensate for missing or overstretched institutional capacity echoing the argument by Toyama [94]. In CataractBot, the system worked effectively because doctors were already supporting patients through WhatsApp groups, and the AI simply streamlined an established workflow (§4.6.2). By contrast, ASHABot struggled because the chatbot introduced additional demands that supervisors did not have the capacity to meet. Supervisors were expected

Influence	#	Guideline		
Sociocultural	G1	Effective systems require AI developers and local domain experts to collaborate as equal partners throughout the full lifecycle of the application		
	G2	Culturally aware AI systems require dedicated attention to the linguistic and cultural gaps present in low-resource contexts		
	G3	Dominant languages often provide workable alternatives to unsupported local languages in contexts where communities are multilingual		
	G4	Community trust in AI is a key influence on system design in both mitigating doubt and preventing over-reliance		
Institutional	G5	Compliance with institutional mandates and workflows shapes whether AI systems achieve practical uptake		
	G6	AI strengthens existing institutional capacities but cannot substitute for missing foundations		
	G7	Sustainable AI deployment relies on strong institutional capacity		
	G8	Institutional capacity shapes the depth of model adaptation, with in-context methods often preceding in-weight investments		
Technological	G9	Shifting model capabilities call for flexible system designs that can accommodate change without friction		
	G10	Small-scale deployments limit risk early by exposing failures that only emerge in real use		
	G11	Effective AI evaluation accounts for context that conventional benchmarks overlook		
	G12	Effective Safety emerges from combining lightweight models, human-in-the-loop workflows, and frontier models rather than relying on technology alone		

AI Non-Western

While the above guidelines are broadly applicable to software projects across contexts, the icons highlight those that carry heightened relevance for AI projects and for work in non-Western contexts.

Table 4: Design Guidelines for Building Culturally Aligned AI Systems in Non-Western Contexts

to respond to questions the bot could not answer, but their existing workloads and occasional knowledge gaps made this impractical, leading to delayed or absent support. These cases show that AI systems can improve efficiency and reach, but they depend on baseline institutional capacity and cannot act as substitutes for missing infrastructure or human expertise.

G7: Sustainable AI deployment relies on strong institutional capacity. Our findings showed that sustaining AI systems in practice required continuous rather than one-time commitments, unlike many traditional software deployments that demand relatively modest upkeep. Language support needed ongoing updates as user demographics expanded and new gaps surfaced (§4.1.1). Domain

knowledge bases required periodic expert curation across applications (§4.6.1), and Safety oversight, including the validation workflows in LegalTranslateAI, depended on sustained staffing (§4.3.2). Model-related costs also increased as user bases grew, since higher traffic raised inference expenses, infrastructure requirements, and monitoring needs to maintain quality and safety. While prior work has emphasized institutional support and compatibility as enablers of AI adoption [27, 42, 50], our cases show that long-term financial and human capacity strongly shaped whether systems remained viable once deployed and whether they could scale responsibly over time.

G8: Institutional capacity shapes the depth of model adaptation, with in-context methods often preceding in-weight investments. Decisions about whether to rely on in-context methods

or pursue deeper in-weight modifications were shaped primarily by what institutions could realistically sustain. In-weight approaches such as fine-tuning or training new models offered performance gains but required maintaining specialized infrastructure, updating models as technology evolved, and providing ongoing oversight — demands that were especially challenging in low-resource language settings with limited public data [49, 59]. In practice, institutions gravitated toward in-context approaches like prompt engineering and glossaries because they were faster to implement, easier to adjust as conditions changed, and more feasible given existing capacity. Projects such as ReadAI and PROMPTS adopted in-weight modifications only when in-context approaches no longer met their needs (§4.1.2). These patterns show that in-context adaptations often provided an effective, sustainable middle ground for progressing without taking on burdens institutions could not maintain.

5.3 Technological Influences

Technological influences shaped how developers addressed all LISTED factors, determining not only what teams aimed to build but also what was feasible at any moment. Because AI capabilities evolve rapidly and often through experimental progress [92], developers worked within shifting constraints of model performance, evaluation practices, and resource availability. While these technologies offer new opportunities for expanding access to information and services, scholars caution that they can also reinforce global inequities and cultural harms [11]. Our cases show how both the benefits and risks of AI become tangible in non-Western contexts, where uneven model performance across languages, cultures, and tasks [1, 40, 75] directly shapes the design of culturally aligned systems for social good. These technological dynamics required teams to adapt pragmatically through strategies such as anticipating rapid model improvements, developing localized evaluation methods, deploying early to surface real-world gaps, and maintaining strong human fallbacks when automation fell short.

G9: *Shifting model capabilities call for flexible system designs that can accommodate change without friction.* Across our cases, developers operated in a technological landscape evolving faster than traditional development cycles. Model capabilities improved rapidly, making heavy investment in engineering fixes for current limitations inefficient when those limitations were likely to be resolved by broader advances. Teams therefore favored modular designs that allowed new models or components to be integrated without rebuilding entire pipelines. Developers adapted pragmatically to what the technology could support at each moment. In Shiksha Copilot, teams combined multiple models to balance strengths in content generation and translation (§4.1.1). In LegalTranslateAI, early deployments relied on NMT because LLMs were not yet widely available. In PROMPTS, commercial frontier models eventually outperformed the continually pretrained in-house Swahili model. Together, these cases show that system design must anticipate rapid capability shifts and remain flexible enough to incorporate new models with minimal redevelopment.

G10: *Small-scale deployments limit risk early by exposing failures that only emerge in real use.* Early limited deployments played a critical role in managing technological risk because many AI failures surfaced only when systems interacted with the diversity and unpredictability of real users. These failures stemmed not from coding errors but from the probabilistic and context-dependent nature of model behavior, which made pre-deployment testing insufficient. Small rollouts created a safer environment to observe how models handled long-tailed language inputs, community-specific references, and domain nuances absent from development settings. In FarmerChat, for instance, early deployments revealed that the STT model mis-transcribed similar-sounding words, leading to incorrect agricultural advice — problems detectable only when exposed to real accents, pronunciations, and vocabulary (§4.1.1). Across cases, small-scale deployments served as crucial risk-buffering phases, surfacing hidden vulnerabilities before systems reached larger populations.

G11: *Effective AI evaluation accounts for context that conventional benchmarks overlook.* Dominant AI evaluation paradigms often reflected technological assumptions that conflicted with ground realities. Standard benchmarks used for model selection [37] were frequently misaligned with local contexts because of the assumptions embedded in their datasets [28, 60, 79]. Developers treated these benchmarks as starting points but quickly encountered their limits, echoing critiques by Liao and Xiao [54]. Projects therefore created localized evaluation methods. FarmerChat developed gold-standard transcripts to assess agricultural terminology (§4.1.1), and user feedback loops became central to evaluating Safety and cultural relevance across diverse End-User Demography (§4.5.1). These cases show that evaluation practices had to evolve toward participatory, context-sensitive approaches that captured nuances generic benchmarks could not represent.

G12: *Effective Safety emerges from combining lightweight models, human-in-the-loop workflows, and frontier models rather than relying on technology alone.* Across our cases, Safety depends on multiple layers working together rather than on any single technological component. Lightweight models, state-of-the-art LLMs, and human expertise each address different aspects of risk, and systems remain dependable only when these elements are integrated. In PROMPTS, which handles nearly 12,000 questions per day, a lightweight classifier screens each query for sensitive or high-risk content before it reaches the LLM, enabling the AI system to safely handle around 80% of incoming messages. Queries involving critical or high-risk issues are then routed to a dedicated human help desk whose judgments cannot be replaced by automated methods alone. These workflows function effectively because they draw on existing institutional processes, with AI augmenting rather than substituting human capacity. Across deployments, this pattern shows that robust Safety in complex, multilingual, and high-stakes settings arises from combining lightweight triage models, frontier models for general reasoning, and human oversight for critical and high-risk cases.

Together, these 12 guidelines underscore that culturally aware AI systems cannot be achieved through technical fixes alone. Effective system design should address the LISTED factors that are shaped by

Sociocultural, Institutional, Technological influences to ensure that systems remain usable and trustworthy in high-stakes non-Western contexts.

6 Conclusion

We conducted semi-structured interviews with AI developers and domain experts who have built and deployed AI for social good in education, healthcare, agriculture, and law across non-Western contexts. Our findings reveal six critical factors that consistently influence how these systems are built and deployed: Language, Institution, Safety, Task, End-User Demography, and Domain. We document how adaptation strategies varied from lightweight, in-context modifications to more substantial in-weight changes. These technical choices were shaped not only by computational factors but also by Sociocultural, Institutional, and Technological influences. Most significantly, our findings underscore the extensive human labor required for successful deployment. AI developers and domain experts engaged in intensive collaboration throughout the process such as curating datasets, implementing human-in-the-loop systems, and designing safety mechanisms tailored to local contexts. We identify where this human involvement proved most essential, revealing how labor-intensive practices became fundamental to achieving relevance, safety, and trust in real-world applications. Our work concludes with 12 actionable guidelines for practitioners and policymakers aimed at creating AI for social good systems in non-Western contexts that are contextually appropriate, equitable, and responsive to diverse community needs.

Acknowledgments

We would like to thank our participants in the study for their time and thoughtful inputs. We thank Global Cornell and Cornell Global AI initiative for supporting this work. We also thank anonymous reviewers for their constructive and respectful feedback on our work.

References

- [1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21. doi:10.1145/3706598.3713564 arXiv:2409.11360 [cs].
- [2] Humaid Al Naqbi, Zied Bahroun, and Vian Ahmed. 2024. Enhancing Work Productivity through Generative Artificial Intelligence: A Comprehensive Literature Review. *Sustainability* 16, 3 (Jan. 2024), 1166. doi:10.3390/su16031166 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- [3] Morgan G. Ames. 2019. *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. MIT Press. Google-Books-ID: yYy5DwAAQBAJ.
- [4] Urvashi Aneja, Aarushi Gupta, and Aditya Vashistha. 2025. Beyond Semantics: Examining Gender Bias in LLMs Deployed within Low-resource Contexts in India. (2025).
- [5] Oghenemaro Anuyah, Ruyuan Wan, Cornelius Adejoro, Tom Yeh, Ronald Metoyer, and Karla Badillo-Urquiola. 2024. Cultural Considerations in AI Systems for the Global South: A Systematic Review. In *Proceedings of the 4th African Human Computer Interaction Conference (AfriCHI '23)*. Association for Computing Machinery, New York, NY, USA, 125–134. doi:10.1145/3628096.3629046
- [6] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction* 40, 5 (March 2024), 1251–1266. doi:10.1080/10447318.2022.2138826 Publisher: Taylor & Francis eprint: <https://doi.org/10.1080/10447318.2022.2138826>.
- [7] David Baidoo-anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI* 7, 1 (Dec. 2023), 52–62. doi:10.61969/jai.1337500 Number: 1 Publisher: İzmir Academy Association.
- [8] Aaron J. Barnes, Yuanyuan Zhang, and Ana Valenzuela. 2024. AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology* 58 (Aug. 2024), 101838. doi:10.1016/j.copsyc.2024.101838
- [9] Samuel Becher and Benjamin Alarie. 2025. LexOptima: The promise of AI-enabled legal systems. *University of Toronto Law Journal* 75, 1 (Jan. 2025), 73–121. doi:10.3138/utlj-2024-0002 Publisher: University of Toronto Press.
- [10] Emily M. Bender, View Profile, Timmit Gebru, View Profile, Angelina McMillan-Major, View Profile, Shmargaret Shmitchell, and View Profile. 2021. On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623. doi:10.1145/3442188.3445922
- [11] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atulm Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. 2024. Managing extreme AI risks amid rapid progress. *Science* 384, 6698 (May 2024), 842–845. doi:10.1126/science.adn0117 arXiv:2310.17688 [cs].
- [12] Kirti Bhagat, Kinshuk Vasisht, and Danish Pruthi. 2024. Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations. <https://arxiv.org/abs/2411.07320v2>
- [13] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A. Killian. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '21)*. Association for Computing Machinery, New York, NY, USA, 425–436. doi:10.1145/3461702.3462612
- [14] Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' (Lack of) Multicultural Knowledge. doi:10.48550/arXiv.2404.06664 arXiv:2404.06664 [cs].
- [15] Jun Ho Choi, Oliver Garrod, Paul Atherton, Andrew Joyce-Gibbons, Miriam Mason-Sesay, and Daniel Björkregren. 2024. Are LLMs Useful in the Poorest Schools? TheTeacher.AI in Sierra Leone. doi:10.48550/arXiv.2310.02982 arXiv:2310.02982.
- [16] Yunjae J. Choi, Minha Lee, and Sangsu Lee. 2023. Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581445
- [17] Eric Corbett, Remi Denton, and Sheena Erete. 2023. Power and Public Participation in AI. In *Equity and Access in Algorithms Mechanisms and Optimization*. ACM, Boston MA USA, 1–13. doi:10.1145/3617694.3623228
- [18] Josh Cows, Thomas King, Mariarosaria Taddeo, and Luciano Floridi. 2019. Designing AI for Social Good: Seven Essential Factors. *SSRN Electronic Journal* (2019). doi:10.2139/ssrn.3388669
- [19] Diego Moreira Da Rosa, Leandro Soares Guedes, Monica Landoni, and Milene Silveira. 2024. Investigating technology users' behavior and difficulties in two different multilingual contexts. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*. ACM, Brasilia DF Brazil, 1–12. doi:10.1145/3702038.3702071
- [20] Diego Moreira Da Rosa, Maria Luisa Lamb Souto, and Milene Selbach Silveira. 2022. Designing interfaces for multilingual users: a pattern language. In *Proceedings of the 27th European Conference on Pattern Languages of Programs*. ACM, Irsee Germany, 1–7. doi:10.1145/3551902.3551964
- [21] Qinpu Dang and Guiquan Li. 2025. Unveiling trust in AI: the interplay of antecedents, consequences, and cultural dynamics. *AI & SOCIETY* (July 2025). doi:10.1007/s00146-025-02477-6
- [22] Roberts Dargis, Guntis Bärzdiņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating Open-Source LLMs in Low-Resource Languages: Insights from Latvian High School Exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, Mika Hämmäläinen, Emily Öhman, So Miyagawa, Khalid Alnajjar, and Yuri Biz-zoni (Eds.). Association for Computational Linguistics, Miami, USA, 289–293. doi:10.18653/v1/2024.nlp4dh-1.28
- [23] Fred D. Davis and Andrina Granić. 2024. *The Technology Acceptance Model: 30 Years of TAM*. Springer International Publishing, Cham. doi:10.1007/978-3-030-45274-2
- [24] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. doi:10.48550/arXiv.2310.00907 arXiv:2310.00907 [cs].
- [25] Deepak Varuvel Dennison, Bakhtawar Ahtisham, Kavyansh Chourasia, Nirmitt Arora, Rahul Singh, Rene F. Kizilcec, Akshay Nambi, Tanuja Ganu, and Aditya Vashistha. 2025. Teacher-AI Collaboration for Curating and Customizing Lesson Plans in Low-Resource Schools. doi:10.48550/arXiv.2507.00456 arXiv:2507.00456 [cs].

- [26] Lindsey Dewitt Prat, Olivia Nercy Ndlovu Lucas, Christopher Golias, and Mia Lewis. 2024. Decolonizing LLMs: An Ethnographic Framework for AI in African Contexts. *Ethnographic Praxis in Industry Conference Proceedings* 2024, 1 (2024), 46–85. doi:10.1111/epic.12196 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/epic.12196>.
- [27] Anett Erdmann and Luis Toro-Dupouy. 2025. The influence of the institutional environment on AI adoption in universities: identifying value drivers and necessary conditions. *European Journal of Innovation Management* (Feb. 2025). doi:10.1108/EJIM-04-2024-0407
- [28] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation. doi:10.48550/arXiv.2502.06559 arXiv:2502.06559 [cs].
- [29] Vanessa Evers, Agnes Kukulska-Hulme, and Ann Jones. 1999. Cross-Cultural Understanding of Interface Design: A Cross-Cultural Analysis of Icon Recognition. G. V. Prabhu and E. M. del Galdo (Eds.). Rochester, NY, 173–182. http://www.iwips.org/proceedings/1999/papers/cross_cultural_understanding_of_interface_design_a_cross_cultural_analysis_of_icon_recognition/
- [30] Ruchao Fan, Yunzheng Zhu, Jinhan Wang, and Abeer Alwan. 2022. Towards Better Domain Adaptation for Self-supervised Models: A Case Study of Child ASR. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (Oct. 2022), 1242–1252. doi:10.1109/JSTSP.2022.3200910 arXiv:2305.00115 [eess].
- [31] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (March 2006), 80–92. doi:10.1177/160940690600500107 Publisher: SAGE Publications Inc.
- [32] Luciano Floridi, Josh Cowl, Thomas C. King, and Mariarosaria Taddeo. 2020. How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics* 26, 3 (June 2020), 1771–1796. doi:10.1007/s11948-020-00213-5
- [33] Carolina Fuentes, Iyubanit Rodríguez, Gabriela Cajamarca, Laura Cabrera-Quiros, Andrés Lucero, Valeria Herskovic, and Kenton O'Hara. 2024. Opportunities and Challenges of Emerging Human-AI Interactions to Support Healthcare in the Global South. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 724–727. doi:10.1145/3678884.3681834
- [34] Fiona Fui-Hoon Nah, Zheng , Rulin, Cai , Jingyuan, Siau , Keng , and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* 25, 3 (July 2023), 277–304. doi:10.1080/15228053.2023.2233814 Publisher: Routledge _eprint: <https://doi.org/10.1080/15228053.2023.2233814>
- [35] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrah Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524 Place: Cambridge, MA Publisher: MIT Press.
- [36] Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. doi:10.48550/arXiv.2302.07856 arXiv:2302.07856 [cs].
- [37] Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel John Kochenderfer. 2025. More than Marketing? On the Information Value of AI Benchmarks for Practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1032–1047. doi:10.1145/3708359.3712152
- [38] Gillian R. Hayes. 2020. Inclusive and engaged HCI. *interactions* 27, 2 (Feb. 2020), 26–31. doi:10.1145/3378561
- [39] Rüdiger Heimgärtner. 2018. Culturally-Aware HCI Systems. In *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*, Colette Faucher (Ed.). Springer International Publishing, Cham, 11–37. doi:10.1007/978-3-319-67024-9_2
- [40] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. doi:10.48550/arXiv.2203.10020 arXiv:2203.10020 [cs].
- [41] Delene Heukelman and Seraphin Eyono Obono. 2009. Exploring the African Village metaphor for computer user interface icons. In *Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT '09)*. Association for Computing Machinery, New York, NY, USA, 132–140. doi:10.1145/1632149.1632167
- [42] Omar Mohammed Horani, Ahmad Samed Al-Adwan, Husam Yaseen, Hazar Hmoud, Waleed Mugahed Al-Rahmi, and Ali Alkhalifah. 2025. The critical determinants impacting artificial intelligence adoption at the organizational level. *Information Development* 41, 3 (Sept. 2025), 1055–1079. doi:10.1177/0266669231166889 Publisher: SAGE Publications Ltd.
- [43] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 1311–1320. doi:10.1145/1753326.1753522
- [44] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BEAVERTAILS: towards improved safety alignment of Llm via a human-preference dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 24678–24704.
- [45] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. doi:10.48550/arXiv.2004.09095 arXiv:2004.09095 [cs].
- [46] Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Rajarshi Ghosh, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, Sun-Joo Lee, Arushi Mangla, Ishani Mondal, Deniz Nazarova, Poli Nemkova, Dina Pisarevskaya, Naqee Rizwan, Nazanin Sabri, Dominik Stambach, Anna Steinberg, David Tomás, Steven R. Wilson, Bowen Yi, Jessica H. Zhu, Arkaitz Zubiaga, Anders Søgaard, Alexander Fraser, Zhijing Jin, Rada Mihalcea, Joel R. Tetreault, and Daryna Dementieva. 2025. NLP for Social Good: A Survey of Challenges, Opportunities, and Responsible Deployment. doi:10.48550/arXiv.2505.22327 arXiv:2505.22327 [cs] version: 1.
- [47] Naveena Karusala, Aditya Vishwanath, Aditya Vashistha, Sunita Kumar, and Neha Kumar. 2018. "Only if you use English you will get to more things": Using Smartphones to Navigate Multilingualism. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. doi:10.1145/3173574.3174147
- [48] Ji Hye Kim and Kun Pyo Lee. 2005. Cultural difference and mobile phone interface design: icon recognition according to level of abstraction. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services (MobileHCI '05)*. Association for Computing Machinery, New York, NY, USA, 307–310. doi:10.1145/1085777.1085841
- [49] Nir Kshetri. 2024. Linguistic Challenges in Generative Artificial Intelligence: Implications for Low-Resource Languages in the Developing World. *Journal of Global Information Technology Management* 27, 2 (April 2024), 95–99. doi:10.1080/1097198X.2024.2341496 Publisher: Routledge _eprint: <https://doi.org/10.1080/1097198X.2024.2341496>
- [50] Sreejith Kurup and Vivek Gupta. 2022. Factors Influencing the AI Adoption in Organizations. *Metamorphosis* 21, 2 (Dec. 2022), 129–139. doi:10.1177/09726225221124035 Publisher: SAGE Publications India.
- [51] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in neural information processing systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
- [52] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language Ranker: A Metric for Quantifying LLM Performance Across High and Low-Resource Languages. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 27 (April 2025), 28186–28194. doi:10.1609/aaai.v39i27.35038
- [53] Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The Widespread Adoption of Large Language Model-Assisted Writing Across Society. doi:10.48550/arXiv.2502.09747 arXiv:2502.09747 [cs] version: 2.
- [54] Q. Vera Liao and Ziang Xiao. 2025. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap. doi:10.48550/arXiv.2306.03100 arXiv:2306.03100 [cs].
- [55] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics* 13 (July 2025), 652–689. doi:10.1162/tacl_a_00760
- [56] Nicole Tsz Yeung Liu, Samuel N. Kirshner, and Eric T. K. Lim. 2023. Is algorithm aversion WEIRD? A cross-country comparison of individual-differences and algorithm aversion. *Journal of Retailing and Consumer Services* 72 (May 2023), 103259. doi:10.1016/j.jretconser.2023.103259
- [57] Peter John Loewen, Blake Lee-Whiting, Maggie Arai, Thomas Bergeron, Thomas Galipeau, Isaac Gazendam, Hugh Needham Lee Slinger, and Sofiya Yusyopych. [n. d.]. *Global public opinion on artificial intelligence (GPO-AI)*. Technical Report. Tech. Rep., 2024.[Online]. Available: <https://srinstitute.utoronto.ca...>
- [58] Bilal A. Mateen, Vaishnavi Menon, Ambrose Agweyu, Robert Korom, Elizabeth Omoluabi, David McAfee, Natnael Shimelash, Samuel Rutunda, Crystal Rugege, Gwydion Williams, Mira Emmanuel-Fabula, Alastair K. Denniston, Xiaoxuan Liu, and Melissa Miles. 2025. Trials for LLM-supported clinical decisions in African primary healthcare. *Nature Medicine* (July 2025), 1–3. doi:10.1038/s41591-025-03815-3 Publisher: Nature Publishing Group.
- [59] Josh McGiff and Nikola S. Nikolov. 2025. Overcoming Data Scarcity in Generative Language Modelling for Low-Resource Languages: A Systematic Review. doi:10.48550/arXiv.2505.04531 arXiv:2505.04531 [cs].

- [60] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Dan Xu, Paul Watters, and Malka N Halgamuge. 2025. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. *IEEE Transactions on Artificial Intelligence* (2025), 1–18. doi:10.1109/TAL.2025.3569516
- [61] Courtney McKim. 2023. Meaningful Member-Checking: A Structured Approach to Member-Checking. *American Journal of Qualitative Research* 7, 2 (Feb. 2023), 41–52. <https://www.ajqr.org/article/meaningful-member-checking-a-structured-approach-to-member-checking-12973> Publisher: HA Publication.
- [62] Indrani Medhi, Aman Sagar, and Kentaro Toyama. 2006. Text-Free User Interfaces for Illiterate and Semi-Literate Users. In *2006 International Conference on Information and Communication Technologies and Development*. 72–82. doi:10.1109/ICTD.2006.301841
- [63] Shujaat Mirza, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. 2024. Global-Liar: Factuality of LLMs over Time and Geographic Regions. doi:10.48550/arXiv.2401.17839 arXiv:2401.17839 [cs].
- [64] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (April 2023), 3005–3054. doi:10.1007/s10462-022-10246-w
- [65] Khadijeh Moulaei, Atiye Yadegari, Mahdi Baharestani, Shayan Farzanbakhsh, Babak Sabet, and Mohammad Reza Afrash. 2024. Generative artificial intelligence in healthcare: A scoping review on benefits, challenges and applications. *International Journal of Medical Informatics* 188 (Aug. 2024), 105474. doi:10.1016/j.ijmedinf.2024.105474
- [66] Arthur G. O. Mutambara. 2025. *Artificial Intelligence: A Driver of Inclusive Development and Shared Prosperity for the Global South*. CRC Press. Google-Books-ID: UELEQAQAQBAJ.
- [67] Jessica Ojo, Ogunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages?. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 19048–19095. doi:10.18653/v1/2025.findings-acl.976
- [68] Priyadarsini Patnaik and Mahmoud Bakkar. 2024. Exploring determinants influencing artificial intelligence adoption, reference to diffusion of innovation theory. *Technology in Society* 79 (Dec. 2024), 102750. doi:10.1016/j.techsoc.2024.102750
- [69] Roberto Pereira and Maria Cecilia Calani Baranauskas. 2015. A value-oriented and culturally informed approach to the design of interactive systems. *International Journal of Human-Computer Studies* 80 (Aug. 2015), 66–82. doi:10.1016/j.ijhcs.2015.04.001
- [70] Michael A. Peters and Shivali Tukdeo. 2025. Beyond the Utopian/Dyspotic Frames: Towards a Research Agenda for AI in Education and Development (AI4ED) in the Global South. *Contemporary Education Dialogue* 22, 1 (Jan. 2025), 176–184. doi:10.1177/09731849241298118 Publisher: SAGE Publications India.
- [71] Kavita Philip, Lilly Irami, and Paul Dourish. 2012. Postcolonial Computing: A Tactical Survey. *Science, Technology, & Human Values* 37, 1 (2012), 3–29. <https://www.jstor.org/stable/41511154> Publisher: Sage Publications, Inc..
- [72] Mahika Phutane, Ananya Seelam, and Aditya Vashistha. 2025. "Cold, Calculated, and Condescending": How AI Identifies and Explains Ableism Compared to Disabled People. doi:10.1145/3715275.3732128 arXiv:2410.03448 [cs].
- [73] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural Incongruencies in Artificial Intelligence. doi:10.48550/arXiv.2211.13069 arXiv:2211.13069 [cs].
- [74] Sonel Pyram. 2024. Future directions for context in ICT4D: A systematic literature review. *Information Development* (May 2024), 02666669241248149. doi:10.1177/02666669241248149 Publisher: SAGE Publications Ltd.
- [75] 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 (Jan. 2025), 101118. doi:10.1016/j.patter.2024.101118
- [76] 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. doi:10.48550/arXiv.2409.10913 arXiv:2409.10913 [cs].
- [77] Pragnya Ramjee, Bhuvan Sachdeva, Satvik Golechha, Shreyas Kulkarni, Geeta Fulari, Kaushik Murali, and Mohit Jain. 2025. CataractBot: An LLM-Powered Expert-in-the-Loop Chatbot for Cataract Patients. doi:10.48550/arXiv.2402.04620 arXiv:2402.04620 [cs].
- [78] Surangika Ranathunga and Nisansa de Silva. 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. doi:10.48550/arXiv.2210.08523 arXiv:2210.08523 [cs].
- [79] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. *Advances in Neural Information Processing Systems* 37 (Dec. 2024), 21763–21813. doi:10.52202/079017-0685
- [80] EVERETT M. ROGERS, ARVIND SINGHAL, and MARGARET M. QUINLAN. 2008. Diffusion of Innovations. In *An Integrated Approach to Communication Theory and Research* (2 ed.). Routledge. Num Pages: 17.
- [81] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. doi:10.48550/arXiv.2402.07927 arXiv:2402.07927 [cs].
- [82] Luciana Salgado, Roberto Pereira, and Isabela Gasparini. 2015. Cultural Issues in HCI: Challenges and Opportunities. In *Human-Computer Interaction: Design and Evaluation*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 60–70. doi:10.1007/978-3-319-20901-2_6
- [83] Ronny Scherer, Fazilat Siddiq, and Jo Tondeur. 2019. The technology acceptance model (TAM): A meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education. *Computers & Education* 128 (Jan. 2019), 13–35. doi:10.1016/j.compedu.2018.09.009
- [84] Agrima Seth, Monojit Choudhary, Sunayana Sitaram, Kentaro Toyama, Aditya Vashistha, and Kalika Bali. 2018. How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. (2018).
- [85] Haroon Sheikh, Corien Prins, and Erik Schrijvers. 2023. Contextualization. In *Mission AI: The New System Technology*, Haroon Sheikh, Corien Prins, and Erik Schrijvers (Eds.). Springer International Publishing, Cham, 179–209. doi:10.1007/978-3-031-21448-6_6
- [86] Keith Shepherd, [n. d.]. Virtual Agronomist - an AI-assisted chatbot for guiding crop management decisions of smallholder farmers in Africa. ([n. d.]).
- [87] Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial Intelligence for Social Good: A Survey. doi:10.48550/arXiv.2001.01818 arXiv:2001.01818 [cs].
- [88] 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. 2024. Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers. doi:10.48550/arXiv.2409.08916 arXiv:2409.08916 [cs].
- [89] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermiis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 18761–18799. doi:10.18653/v1/2025.acl-long.919
- [90] Katarzyna Stawarz, Dmitri Katz, Amid Ayobi, Paul Marshall, Taku Yamagata, Raul Santos-Rodriguez, Peter Flach, and Aisling Ann O'Kane. 2023. Co-designing opportunities for Human-Centred Machine Learning in supporting Type 1 diabetes decision-making. *International Journal of Human-Computer Studies* 173 (May 2023), 103003. doi:10.1016/j.ijhcs.2023.103003
- [91] David W Stewart and Michael A Kamins. 1993. *Secondary research: Information sources and methods*. Vol. 4. Sage.
- [92] Xuli Tang, Xin Li, Ying Ding, Min Song, and Yi Bu. 2020. The Pace of Artificial Intelligence Innovations: Speed, Talent, and Trial-and-Error. doi:10.48550/arXiv.2009.01812 arXiv:2009.01812 [cs].
- [93] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (Sept. 2024), pgae346. doi:10.1093/pnasnexus/pgae346
- [94] Kentaro Toyama. 2015. *Geek Heresy: Rescuing Social Change from the Cult of Technology*. PublicAffairs. Google-Books-ID: 2w6CBgAAQBAJ.
- [95] Azmine Tushik Wasi. 2025. AI for Social Good and Public Impact in Marginalized Communities: A Cross-Cultural Framework. In *Social Impact of AI: Research, Diversity and Inclusion Frameworks*, Yetunde Folajimi, Leonidas Deligiannidis, Salem Othman, and Shawren Singh (Eds.). Springer Nature Switzerland, Cham, 1–12. doi:10.1007/978-3-031-98949-0_1
- [96] Robin Williams and David Edge. 1996. The social shaping of technology. *Research Policy* 25, 6 (Sept. 1996), 865–899. doi:10.1016/0048-7333(96)00885-2
- [97] Susan Wyche. 2020. Using Cultural Probes in HCI4D/ICTD: A Design Case Study from Bungoma, Kenya. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 63:1–63:23. doi:10.1145/3392873
- [98] Jinying Xu and Weisheng Lu. 2022. Developing a human-organization-technology fit model for information technology adoption in organizations. *Technology in Society* 70 (Aug. 2022), 102010. doi:10.1016/j.techsoc.2022.102010
- [99] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS Nexus* 3, 3 (March 2024), pgae052. doi:10.1093/pnasnexus/pgae052
- [100] Mi Zhou, Vibhanshu Abhishek, Timothy Denderger, Jaymo Kim, and Kannan Srinivasan. 2024. Bias in Generative AI. doi:10.48550/arXiv.2403.02726 arXiv:2403.02726 [econ].

A AI Adaptation Techniques

☰ Prompt Engineering

Prompt Engineering is the crafting and iteratively refining of natural language prompts—sometimes with examples—to guide generative models toward more accurate, relevant, and controlled outputs [81].

☰ RAG

Retrieval-Augmented Generation (RAG) enhances LLM responses by retrieving and grounding them in external knowledge sources (e.g., documents or databases), reducing hallucination and improving factual reliability [51].

☰ Glossary

Glossary Adaptation incorporates domain-specific vocabulary or local terminology—via glossaries—deterministically into the inputs and outputs of the AI system to ensure appropriate and culturally relevant expressions [36].

☰ Human-in-the-loop

Human-in-the-Loop involves routing model outputs—especially high-impact or uncertain ones—to human experts for review, correction, or confirmation before delivering them to end users [64].

🔧 Model Training

Model Training builds models from scratch using domain-relevant data—e.g., training ASR on children’s speech—to create capabilities tailored to specific user needs [30].

🔧 Continual Pre-training

Continual Pre-training continuously updates a pre-trained model with new domain- or language-specific data to enhance its performance in evolving contexts [55].

🔧 Fine Tuning

Fine Tuning further trains a pre-trained model on labeled, domain-specific datasets (e.g., legal text) to better align its behavior with specialized tasks and terminology [55].

🔧 Instruction Tuning

Instruction Tuning adjusts a model using instruction–response pairs so it better understands and follows human-written directives across tasks and styles [55].

B AI Developer Interview Protocol

Introduction (5 mins)

Thank you for taking the time to participate in this interview. We are conducting research on how AI systems are being built for social good applications. We would like to learn more about your experience developing AI solutions that address social challenges. This interview will be approximately 60–75 minutes long.

Do I have your consent for this interview and permission to record this conversation?

Demographics (5 mins)

- What is your current role and organization?
- Age, Gender, Education
- How many years of experience do you have in AI development?
- Can you briefly describe the AI project(s) for social good that you’ve worked on? What specific social problem does each of your AI systems address?

Problem Context & Technology Selection (15 mins)

- What led you to choose AI as the solution approach for this particular problem?
- How was this problem being managed or solved before your AI intervention?
- What unique advantages does AI provide compared to other technological solutions you considered?
- What are the disadvantages of using AI for solving this problem?
- What are the non-negotiable design principles or constraints that guide your technical decisions?
- How does designing AI solutions for social good differ from developing commercial AI applications?

Contextualization & Cultural Adaptation (15 mins)

- What did you do to make your intervention culturally and locally relevant to communities on the ground?
- How do you technically implement this contextualization? What parts of your system require cultural adaptation?
- Who is involved in the contextualization process, and how do you incorporate their input into technical decisions?
- Can you provide specific examples of how cultural context influenced your technical architecture?
- How do you measure whether your contextualization efforts are effective?
- How do you engage with target communities throughout the technical development process?
- What methods have you found most effective for incorporating community feedback into your technical work?
- How is the diversity of the population captured in the contextualization process?
- What technical challenges do you encounter when working with social impact practitioners?
- How is your solution technically delivered to target communities? What infrastructure or accessibility considerations shaped your approach?

AI Harms & Bias Mitigation (15 mins)

- What specific limitations have you identified in the AI models you’re using?
- What biases have you identified in the AI models you’re using?
- What potential harms have you identified in the AI models you’re using?

- What concrete mitigation strategies have you implemented to address these issues?
- How do you test the effectiveness of your bias mitigation approaches?
- How do you test the effectiveness of your harm mitigation approaches?

Development Process (15 mins)

- Can you walk me through your project development lifecycle from start to finish?
- Who is involved at each stage of development, and what are their roles?
- How do you define and measure the success of your technical intervention?
- What are the biggest gaps that you see in the system that is limiting the success of the interventions?

Wrap Up

- What would you do differently if you were starting this project over?

C Domain Expert Interview Protocol

Introduction (5 mins)

Thank you for taking the time to participate in this interview. We are conducting research on how AI systems are being built for social good applications. We would like to learn more about your experience implementing AI solutions that address social challenges. This interview will be approximately 60–75 minutes long.

Do I have your consent for this interview and permission to record this conversation?

Demographics (5 mins)

- What is your current role and organization?
- Age, Gender, Education
- How many years of experience do you have in social impact work?
- Can you briefly describe the AI project for social good that you've been involved with?

Problem Understanding & Solution Approach (15 mins)

- What specific social problem does your AI intervention address?
- How was this problem being managed by your organization before implementing AI?
- What factors led your organization to decide to use AI for this problem?
- What unique value does AI add that other approaches couldn't provide?
- What are the disadvantages of using AI for solving this problem?
- How do the advantages weigh over the disadvantages?
- What are the limitations of the current AI models? What more capabilities are needed to solve the problem more aptly?

- How does this AI intervention fit into your organization's broader work?
- What are the short-term and long-term effects of using the tool?

Contextualization & User Relevance (15 mins)

- How did you ensure that the AI intervention is culturally appropriate to local end-user communities?
- How do you ensure the technology is fully relevant for your end users?
- What do the users primarily value?
- How do you ensure that users' values are reflected in the AI technologies?
- Who are the main stakeholders in the intervention?
- What do you do when values conflict between different stakeholders?
- What processes do you have in place to assess cultural relevance and appropriateness?
- How is the diversity of the population captured in the contextualization process?
- What support mechanisms do you provide to help users better utilize the AI tool?
- What training is provided to support better adoption?

Partnership & Collaboration (15 mins)

- How do you collaborate with the technology team throughout development and implementation?
- What does effective collaboration between social impact practitioners and technologists look like?
- How do you engage with end users throughout the implementation process?
- What challenges do you face when coordinating between technologists, end users, your organization, and the government?
- What are the main tensions you've encountered in these partnerships, and how do you navigate them?
- What opportunities have emerged from working across these different partnerships?

Implementation & Adoption (15 mins)

- What are the main barriers to adoption you encounter with your target users?
- How do you systematically address these adoption challenges?
- Do they find the tool useful? And in what way?
- How do you measure the impact of your AI intervention?
- What potential harms do you see from using AI to solve this particular problem?
- How do you monitor and mitigate these risks?

Wrap-up (5 mins)

- What would you do differently if you were starting this project over?
- What advice would you give to other organizations attempting similar AI for social good work?

- Are there any other aspects of your experience that we haven't discussed that you think are important?
- Do you have suggestions for other practitioners we should speak with for this research?