

**WORKING PAPER NO: 739**

**Next-Billion AI Index: The compass for AI utility and adoption in the global majority**

**Ambrish Rawat**

*IBM Research*

[ambrish.rawat@ie.ibm.com](mailto:ambrish.rawat@ie.ibm.com)

**Jessica He**

*IBM Research*

[jessicahe@ibm.com](mailto:jessicahe@ibm.com)

**Subhabrata Majumdar**

*Assistant Professor, Decision Sciences  
Indian Institute of Management Bangalore*

[smajumdar@iimb.ac.in](mailto:smajumdar@iimb.ac.in)

**Claudio Pinhanez**

*University of Sao Paulo*

[pinhanez@gmail.com](mailto:pinhanez@gmail.com)

**Yann Le Beux**

*YUX Design*

[yann@yux.design](mailto:yann@yux.design)

**Satyapriya Krishna**

*Harvard University*

[krishna.satyapriya123@gmail.com](mailto:krishna.satyapriya123@gmail.com)

**Rahul Gupta**

*Amazon*

[gupra@amazon.com](mailto:gupra@amazon.com)

**Rumman Chowdhury**

*Humane Intelligence*

[rchowdhury@gmail.com](mailto:rchowdhury@gmail.com)

**Kush R. Varshney**

*IBM Research*

[krvarshn@us.ibm.com](mailto:krvarshn@us.ibm.com)

# Next-Billion AI Index: The compass for AI utility and adoption in the global majority

Ambrish Rawat<sup>1</sup>, Jessica He<sup>1</sup>, Subhabrata Majumdar<sup>2</sup>, Claudio Pinhanez<sup>3</sup>, Yann Le Beux<sup>4</sup>, Satyapriya Krishna<sup>5</sup>, Rahul Gupta<sup>6</sup>, Rumman Chowdhury<sup>7</sup>, Kush R. Varshney<sup>1</sup>

<sup>1</sup>IBM Research <sup>2</sup>Indian Institute of Management Bangalore <sup>3</sup>University of São Paulo <sup>4</sup>YUX Design  
<sup>5</sup>Harvard University <sup>6</sup>Amazon <sup>7</sup>Humane Intelligence

## Abstract

Generative AI assessments remain dominated by frontier capability benchmarks that often fail to capture whether systems can be sustainably deployed, adapted, and trusted in locally grounded and infrastructure-constrained settings. This paper introduces the Next Billion AI Index (*nexbax*), which we believe is the first diagnostic framework to treat economic viability, operational deployability, and governance alignment as co-equal determinants of AI utility in next-billion-user contexts. Rather than treating usefulness as a single outcome, *nexbax* operationalizes the preconditions for useful AI through 10 dimensions organized under three themes: Effective Efficiency, Operational Practicality, and Societal Integrity. These dimensions assess whether systems are economically viable, deployable under infrastructure and workflow constraints, and aligned with local needs, user expectations, and collaborative development practices. We pair the framework with rubrics for weak, moderate, and strong performance, and conduct a formative expert evaluation through eleven semi-structured interviews with founders, developers, product leaders, and technical practitioners building AI systems for next-billion markets. Participants found the index useful for reasoning about adoption trade-offs and effective at capturing factors shaping real-world AI uptake—particularly cost, usability, reliability, and trust. They also identified the need for contextual explanations, domain-specific evidence, and broader stakeholder validation. *Nexbax* is therefore proposed not as a universal score of social value, but as a diagnostic for artificial *useful* intelligence: a way to make visible the technical, economic, and governance properties that make inclusive AI deployment more viable.

## 1 Introduction

What determines whether AI meaningfully benefits the next billion users? While AI has shown significant promise to multiply user productivity in western, technological contexts, it is yet to replicate that promise for users in the global majority. In next-billion settings characterized by people using digital technologies for the first time in emerging markets, artificial intelligence (AI) adoption is limited by fundamental constraints: meaningful connectivity and affordability, mobile-first device realities, optimized compute and energy reliability, fragmented and informal economic activity, and heterogeneous language and literacy contexts (Oluwatuyi et al. 2026; Varuvel Dennison et al. 2026). Under these conditions, the primary determinant of

real-world AI impact is not peak benchmark performance, but whether systems can be deployed and sustained under local constraints (UNDP 2025). This stands in sharp contrast to high-income markets, where AI diffusion is increasingly shaped by enterprise software integration, abundant cloud compute, mature data pipelines, and relatively affordable connectivity, allowing “model capability” to translate more quickly into productivity gains. In line with this mismatch, interviews with practitioners deploying AI in the Global South found that dominant AI evaluation benchmarks were insufficient for local contexts, and instead, “evaluation practices had to evolve toward participatory, context-sensitive approaches that captured nuances generic benchmarks could not represent” (Varuvel Dennison et al. 2026). Cross-country evidence also suggests that the character of adoption differs across these contexts. In University of Melbourne’s and KPMG’s 2025 “Trust, Attitudes and Use of Artificial Intelligence” global study (KPMG 2025), which surveyed over 48,000 respondents in 47 countries, people in emerging economies such as India, Egypt, and Nigeria reported markedly higher trust in AI systems than those in high-income economies, with emerging economies showing ~57% trust versus ~39%. Likewise, Anthropic’s Economic Index (Anthropic 2026) and Microsoft’s 2025 Future of Work report (Microsoft 2025) suggest that, whereas AI use in high-income economies is concentrated in white-collar knowledge work and enterprise workflow automation, in emerging markets it more often functions as a general-purpose problem-solver for informal work, small businesses, education, and frontline services, where flexibility, cost efficiency, and usability matter more than frontier model performance (Belcak et al. 2025). Taken together, these patterns underscore the need for evaluative frameworks that assess AI not by abstract model capability or aspirations toward artificial general intelligence, but by what might be termed artificial *useful* intelligence: economic viability and contextual utility under real-world constraints.

India and markets across Africa represent large and diverse next-billion-user contexts that make these dynamics concrete.

In India, the “next billion AI market” is not a monolithic consumer segment but a digitally amplified, institutionally mediated adoption environment in which AI value is expected to scale through public platforms, regulated sectors,

and micro, small and medium enterprise participation. The government of India’s policy think tank positions AI as a macroeconomic lever, but it also makes clear that readiness depends on more than model capability: it depends on shared digital rails, compute access, interoperable datasets, sector-grade governance, and domain-linked skilling (NITI Aayog 2025). As a result, product-market fit in India privileges affordability, robustness, multilingual access, and compliance-by-design, while foregrounding actors such as digitally powered infrastructure (DPI) custodians, regulators, and sector-specific developers as key market makers shaping what can be deployed, at what cost, and for whom.

Across Africa, the “next billion AI market” is better understood as a people-centered, development-oriented adoption environment where value is defined less by frontier capability than by practical utility under binding constraints. Participatory research with members of a marginalized community in South Africa found a strong desire to leverage AI for socioeconomic advancement (Oluwatuyi et al. 2026). Similarly, the African Union’s Continental AI Strategy (African Union 2024) frames AI as an engine for inclusive growth and treats governments as anchor adopters that can catalyze deployment through DPI, open data, procurement, and partnerships. At the same time, the strategy highlights that uptake depends on closing foundational gaps in electricity, broadband, data infrastructure, compute access, and locally relevant datasets, while treating governance and risk management as first-order adoption conditions. This vision is operationalized through targeted national missions, such as Nigeria’s National Artificial Intelligence Strategy (2025–2029), which explicitly links AI infrastructure to local talent development and social inclusion (FMCIDE 2025) — and Kenya’s National AI Strategy 2025–2030 which underscores the use of community-based ‘Digital Innovation Hubs’ to ensure AI technologies remain accessible to underserved groups and local SMEs (Government of Kenya 2025). In this context, systems are likely to be rewarded for being cost- and energy-efficient, multilingual and locally grounded, deployable through public and SME channels, and governed in ways that sustain trust.

Together, India and African markets illustrate a shared structural reality: AI adoption in next-billion contexts is determined less by frontier performance and more by the interplay of affordability, infrastructural variance, linguistic diversity, institutional capacity, and the centrality of public and small-enterprise channels. A related theme is “sovereign AI”: the ability of states, firms, and communities to retain legitimate control and agency across the AI stack, rather than becoming dependent on monopolistic providers at critical layers of data, compute, models, or deployment infrastructure (Roberts 2024). These determinants sit largely outside prevailing, capability-centric benchmarks (METR 2025, 2026; Liang et al. 2023), which implicitly assume mature infrastructure and enterprise capacity. In such settings, “AI progress” cannot be equated with benchmark supremacy; it must be defined by economic viability, operational resilience, and contextual utility under real-world constraints.

The Next Billion AI Index (*nexbax*) addresses this mea-

surement blind spot. It evaluates generative AI systems—including models and agentic architectures—not by abstract performance metrics alone, but by whether they exhibit system-level properties that can support economically sustainable, inclusive, and developmentally relevant deployment across constrained environments. Inspired by the precedent set by Stanford’s Foundation Model Transparency Index (FMTI) (Wan et al. 2023), which spurred greater openness across the AI industry, *nexbax* extends that momentum toward inclusivity, accessibility, and innovation for the next billion users, asking not only how powerful systems are, but whether they create the technical, economic, and governance conditions under which meaningful use becomes more likely. Partly informed by earlier frameworks for inclusive innovation, including C. K. Prahalad’s work on technology innovation for the next billion (Prahalad 2006, 2012), *nexbax* reinterprets these concerns for the generative AI era through **10 measurable dimensions** organized under **3 themes: Effective Efficiency, Operational Practicality, and Societal Integrity**. These capture how well AI solutions balance affordability, adaptability, and social responsibility. In this version, we pair the dimensions with broad rubrics and a formative expert exercise to explore how such a diagnostic can support comparison across AI technologies. In doing so, it aims to **incentivize innovation that is not only powerful but also socioeconomically viable**, while recognizing that actual usefulness and cultural appropriateness must be established through situated evaluation with local stakeholders and affected communities.

*Nexbax* is not a universal measure of usefulness for the global majority. It is a first-stage diagnostic for identifying whether AI systems possess the pre-conditions—technical, economic, and governance properties—likely to support useful deployment for the next-billion. Determining actual usefulness, cultural appropriateness, and social value requires situated evaluation with local stakeholders and affected communities.

The remainder of the paper situates *nexbax* in related work on inclusive innovation and AI evaluation (Section 2), defines the index dimensions and rubrics (Section 3), reports findings from a formative expert evaluation (Section 4), discusses limitations and future directions for participatory, context-specific validation (Section 5), and concludes by reflecting on the broader implications of evaluating AI through the lens of useful and inclusive adoption (Section 6).

## 2 Related Work

### 2.1 Influence of AI Indices

Practically, indices play a crucial role in aligning innovation incentives by translating complex, multidimensional goals into comparable and trackable signals. In the AI ecosystem, FMTI (Wan et al. 2023) demonstrated how clearly defined indicators can shift industry behavior by elevating transparency from a normative aspiration to a measurable performance dimension. *Nexbax* builds on this logic by extending evaluative attention beyond transparency toward the conditions that enable broad-based adoption and value creation.

By benchmarking AI systems across dimensions that matter for low-resource and underrepresented contexts, the index helps redirect optimization efforts toward efficiency, interoperability, and user empowerment areas that are often underincentivized despite their outsized impact on real world uptake (Rismani et al. 2025).

Crucially, *nexbax* functions as a diagnostic instrument for equitable development rather than a prescriptive ranking of technological superiority. By making gaps in affordability, resilience, localization, and openness empirically visible, the index enables developers, investors, and policymakers to identify where targeted innovation can unlock disproportionate social and economic value. Over time, tracking through a standardized index allows stakeholders to distinguish between nominal advances in AI capability and substantive progress in inclusive utility. In this way, *nexbax* supports a development-aligned innovation trajectory, one in which success is measured not only by what AI systems can do in ideal settings, but by how effectively they serve diverse populations at scale.

## 2.2 Related Indices and Frameworks

This subsection reviews adjacent efforts that inform *nexbax* while also clarifying the gap it is intended to fill. The most relevant literature spans AI-wide indices and comparative frameworks; national missions and public-sector assessment efforts; accountability and transparency frameworks; domain-specific application literatures; cross-cutting capability verticals such as multilinguality, efficiency, and robustness; and broader traditions of inclusive innovation. Taken together, these streams show that AI progress is already being measured from multiple angles, but they rarely integrate deployability, inclusivity, governance alignment, and developmental relevance into a single comparative framework for next-billion-user contexts.

Conceptually, *nexbax* is informed by value-sensitive design (Borning and Muller 2012), which emphasizes embedding human values such as accessibility, fairness, and accountability directly into technical systems rather than treating them as downstream constraints. By operationalizing these values as measurable system-level properties, the index moves beyond aspirational principles toward actionable signals that can shape design and optimization choices.

**AI indices and benchmarks.** A growing body of work has introduced indices and benchmarking frameworks to measure different aspects of AI development, diffusion, and governance. Early and influential efforts such as the Stanford AI Index (Shoham et al. 2017) provide ecosystem-level trend tracking by aggregating indicators across research output, model capabilities, investment flows, and deployment patterns, offering a longitudinal view of the “state of AI”. Similar macro-level indices and readiness assessments—developed by institutions such as the OECD.AI index (OECD 2026) and Oxford Insights’ Government AI Readiness index (Oxford Insights 2025)—evaluate national or regional capacity for AI adoption, emphasizing infrastructure, talent pipelines, regulatory frameworks, and public-sector readiness. Complementing these, the Global

Index on Responsible AI (GIRAI) by the Global Center on AI Governance based in South Africa provides a comparative global benchmark specifically focused on the implementation of responsible AI commitments across diverse national contexts (Adams et al. 2024). Capability-focused efforts such as METR’s autonomous-agent evaluations and time-horizon studies, along with HELM’s holistic language-model evaluation framework, provide more system-level evidence about model performance and emerging autonomous capabilities (METR 2025, 2026; Liang et al. 2023). While these indices and benchmarks are valuable for comparability and agenda-setting, they are primarily descriptive or capability-centric, and provide limited insight into how individual AI systems perform across heterogeneous, culturally diverse, and resource-constrained contexts, or how their downstream societal impacts should be evaluated. Emerging initiatives are beginning to address these gaps by foregrounding cultural context, participatory design, and societal impact in AI evaluation and deployment. For example, Microsoft Research Africa’s Atlas: A Playbook for Cross-Cultural AI emphasizes culturally grounded approaches to AI system design and deployment across diverse sociotechnical settings (Microsoft Research Africa 2025), while the ImpactBench initiative seeks to develop benchmarks for evaluating the societal impacts of AI systems beyond conventional capability metrics (MIT Media Lab 2026).

**National and multilateral AI missions.** An adjacent but distinct category of measurement has emerged through large-scale national AI missions and public-sector programmes, particularly in countries with large populations and diverse socio-economic conditions. Initiatives such as the IndiaAI Mission (Government of India 2026b), as well as continental and regional strategies articulated through bodies such as the African Union (African Union 2024), increasingly incorporate scorecards, dashboards, or progress indicators to track investments, skilling efforts, startup ecosystems, and public-sector AI deployment. While these efforts are critical for coordinating state capacity and signaling national priorities, their evaluative focus is typically top-down and policy-centric, centered on programme execution rather than on the technical and operational characteristics of AI systems themselves. In contrast, *nexbax* is intentionally bottom-up and system-oriented: it does not assess national readiness or mission success, but instead evaluates whether the AI technologies emerging within and often deployed across such ecosystems are practically usable, affordable, and adaptable for large, diverse user bases.

**Accountability, transparency, governance.** A second line of work focuses on accountability, transparency, and responsible AI practices. The Stanford’s FMTI (Wan et al. 2023) demonstrates how structured rubrics can operationalize transparency and incentivize improved disclosure by AI developers across data, evaluation, compute, and governance dimensions. Related assessment frameworks, including the AI Policy Observatory of the OECD (OECD 2026), the Readiness Assessment Methodology for AI proposed by UNESCO (UNESCO 2023), and ISO/IEC 42001’s AI management-system standard (ISO and IEC 2023), simi-

larly emphasize institutional practices, documentation quality, and ethical alignment. These approaches have proven effective as soft-governance mechanisms; however, they largely evaluate organizational processes and reporting practices rather than downstream properties such as affordability, deployability, usability, or robustness of AI systems in low-infrastructure or culturally diverse settings (Kim et al. 2025).

**Application verticals.** Another important stream of related work comes from domain-specific AI and digital system deployments in sectors such as healthcare, education, environment, agriculture, finance, and crisis response (Varuvel Dennison et al. 2026). This literature is typically solution-oriented, emphasizing task-specificity and pragmatic adaptations such as offline-first architectures, low-bandwidth interfaces, voice and SMS-based interaction, support for low-literacy users, and compatibility with low-end devices to enable operation under constraints of infrastructure, capital, and institutional capacity. Collectively, this literature shows that meaningful functionality in resource-constrained settings depends less on maximal technical sophistication than on fit with local realities.

At the same time, this body of work remains highly contextual. Benchmarks or evaluations developed for one vertical often privilege domain-specific outcome measures, data conditions, and workflow assumptions, making them difficult to compare across sectors or to generalize into a common framework for AI system design. *Nexbax* builds on the practical lessons surfaced in these sectors, but abstracts them into cross-cutting dimensions that can be applied across model families, developer tools, and deployment settings.

**Capability verticals.** Parallel to application-specific benchmarking, a growing set of studies evaluates cross-cutting AI capabilities that are especially salient in next-billion-user contexts, including multilingual performance, low-resource language support, efficiency under compute and bandwidth constraints, robustness to noisy inputs, and interoperability across tools and platforms (Bailey et al. 2026; Vibhasha Team 2025; Muchai et al. 2026; Microsoft Research Africa 2026; Oluwatuyi et al. 2026; Varuvel Dennison et al. 2026). These efforts are valuable because they expose failure modes that broad benchmark suites often miss, particularly where users, developers, and institutions operate outside high-resource English-centric settings.

However, capability-vertical assessments usually isolate one attribute at a time. A multilingual benchmark may reveal language inequities without speaking to affordability; an efficiency benchmark may highlight compute trade-offs without addressing transparency, governance, or user empowerment (Miranda et al. 2026). For *nexbax*, these verticals are therefore best understood as necessary but partial inputs into a broader evaluative frame that asks how these properties interact in real deployment environments.

Taken together, the literatures above provide many of the ingredients that *nexbax* seeks to integrate: macro-level AI tracking, mission-driven public adoption goals, governance-oriented disclosure frameworks, lessons from sectoral de-

ployments, and focused assessments of specific capabilities. What remains less developed is a unifying framework for comparing AI systems according to whether they are economically viable, operationally practical, and socially aligned for large, diverse, and resource-optimized settings. To motivate that synthesis, we next turn to broader frameworks for inclusive innovation on which *nexbax* builds.

**Inclusive innovation.** Prahalad’s framework provides an early and influential articulation of how technologies must be rethought to serve large, underserved populations (Prahalad 2012, 2006). Many of its core principles—such as radical price-performance improvements, robustness to hostile environments, scalability across cultures and languages, and redesign from first principles—remain relevant for AI systems intended for next-billion-user contexts. In particular, requirements around affordability, resilience, adaptability of user interfaces, and functionality under infrastructural volatility closely parallel challenges faced by contemporary AI deployments in low-resource settings.

More recent work on inclusive and globally situated AI has extended these concerns by emphasizing structural inequality, local participation, representational harms, and the need to account for diverse social and cultural contexts in AI design and governance (Mohamed, Png, and Isaac 2020; Sambasivan et al. 2021; Prabhakaran et al. 2022; Clancy, Zhu, and Majumdar 2025). This literature highlights that inclusion cannot be reduced to access alone, but also depends on whose values, languages, institutional assumptions, and lived experiences are reflected in AI systems.

However, the broader inclusive innovation tradition has also been extensively critiqued for overemphasizing provider perspectives, market opportunity, and the framing of low-income populations primarily as consumers. Karnani, for example, argues that poverty alleviation requires attention to people as producers and workers, not only as buyers of low-cost products (Karnani 2007). More broadly, the experimental development economics tradition associated with Banerjee, Duflo, and Kremer emphasizes that claims about poverty alleviation should be grounded in specific, empirically evaluated interventions among the people most affected, rather than inferred from broad market logics (Banerjee and Duflo 2009, 2026). These critiques are directly relevant to AI evaluation: a system can be affordable, technically robust, or attractive to providers while still failing to be useful, appropriate, or empowering in a particular community. *Nexbax* therefore treats Prahalad’s principles as one historical input rather than as a sufficient foundation.

### 3 *Nexbax*: Next Billion AI Index

Building on these indices and frameworks, we developed the initial version of *nexbax* using an iterative process, then validated and refined it through a formative user study. We describe the iterative process below and the study in Section 4.

#### 3.1 Development of *nexbax*

We began by reviewing existing literature on digital inclusion, identified through searches for “inclusive innovation,” “digital inclusion,” “open source AI” and “AI adop-

tion” (Belcak et al. 2025; Kim et al. 2025; Kirk et al. 2024; Melville et al. 2013; nat 2025; Odame-Darkwah and Tsado 2025; Prahalad 2006; Varshney 2023; Seger et al. 2023; Casper et al. 2025; Eiras et al. 2024). We mapped and reinterpreted the principles outlined in this literature in the context of AI systems using documentation of two leading closed-source and open-sourced models as guidance: GPT (Achiam et al. 2023), Claude (Anthropic 2024), Mistral (Jiang et al. 2023, 2024), Llama (Touvron et al. 2023), and Nemotron (Chandiramani et al. 2026; Adler et al. 2024). Within their documentation, we identified instances of existing principles and adjacent concepts, then examined how their meanings and implications changed in the context of AI development and deployment. For example, Prahalad (2006)’s principle of designing for “hostile environments” manifested as mechanisms for “robustness” in constrained environments, and “deskilling” was reinterpreted as “usability” of installation and deployment processes for AI systems.

Furthermore, to ensure relevance to diverse stakeholders — including end users, developers, business decision-makers, and policy-makers — one author held informal conversations based around the empathy map framework (Gibbons 2018) with 5 people from these stakeholder groups. Their feedback and priorities further informed the development of AI-specific considerations for global adoption.

Once an initial set of dimensions had been drafted, we referenced industry reports, such as the India AI Impact Summit goals (Government of India 2026a) and the “Trust, Attitudes and Use of Artificial Intelligence” global study (KPMG 2025), to prioritize and refine components of our framework based on considerations grounded in real-world practice. For example, we removed sustainability as it primarily reflected infrastructural concerns rather than adoption priorities, and validated “adaptability” and “multiculturalism” due to their prevalence in such reports.

Finally, the leading author clustered the dimensions into three themes to reduce fragmentation and promote conceptual clarity. The resulting index is comprised of ten dimensions under three overarching themes (Figure 1) with concrete evaluation criteria for each dimension (Table 2).

### 3.2 Themes and Dimensions

The *nexbax* dimensions are organized into three themes that reflect where inclusivity is negotiated in modern AI systems: economic feasibility, operational deployability, and societal alignment. This thematic grouping is intentional. Rather than treating affordability, usability, and responsibility as independent attributes, *nexbax* recognizes them as interacting layers of system design that jointly determine whether generative AI can be meaningfully adopted at scale in next-billion-user contexts. The three themes also align with distinct decision horizons such as cost and resource choices made early, deployment and adaptation challenges faced during rollout, and longer-term social and institutional consequences, allowing the index to surface trade-offs that are often obscured in capability-centric benchmarks.

**Theme A: Effective Efficiency** *Why economic efficiency must be foundational, not downstream.*

Effective Efficiency captures whether an AI system can deliver acceptable utility under binding economic and infrastructural constraints. In global-majority contexts, affordability is not a secondary optimization objective but a gating condition: systems that are performant yet economically misaligned are functionally inaccessible. *Nexbax* therefore treats cost-effectiveness and resource efficiency as first-order evaluative dimensions, reflecting both monetary cost (pricing models, total cost of ownership) and non-monetary resource intensity (compute, energy, bandwidth).

This theme goes beyond nominal “cheapness.” It evaluates whether efficiency is structurally encoded, for example, through model size choices, edge or offline capability, and pricing flexibility, rather than achieved through temporary subsidies or scale advantages. Table 1 offers indicative signals for assessing these dimensions across AI systems.

**Theme B: Operational Practicality** *Why deployability, not capability, determines real-world impact.*

Operational Practicality assesses whether AI systems can be adapted, maintained, and evolved in heterogeneous and often fragile environments. While many AI benchmarks implicitly assume stable connectivity, skilled operators, and homogeneous users, *nexbax* foregrounds the frictions that dominate real deployments: intermittent infrastructure, linguistic diversity, limited developer bandwidth, and evolving use cases.

The dimensions grouped here—adaptability, interoperability, robustness, usability/automation, and education/empowerment—collectively capture the translation layer between raw AI capability and sustained use. Importantly, several of these dimensions are developer-facing rather than end-user-facing, reflecting the reality that inclusivity is often decided upstream through tooling, abstractions, and integration pathways. Table 1 summarizes possible signals for these properties without privileging any single deployment paradigm.

**Theme C: Societal Integrity** *Why inclusive AI requires shared stewardship, not just safeguards.*

Societal Integrity captures whether AI systems are aligned with the values, institutions, and pluralism of the populations they serve. Rather than treating ethics, inclusivity, and openness as compliance checklists, *nexbax* frames them as enablers of legitimacy and long-term adoption. Systems that are opaque, culturally narrow, or closed to local participation and ownership may achieve short-term deployment but struggle to earn trust or evolve responsibly.

The dimensions in this theme assess whether AI systems support multicultural representation, transparent governance, and collaborative development models that allow diverse stakeholders—developers, public institutions, and communities—to meaningfully shape outcomes. As suggested in Table 1, these dimensions emphasize observable design and governance choices over aspirational commitments.

### 3.3 Measurement Signals

To move from conceptual dimensions to evaluable properties, *nexbax* treats each dimension as a composite construct that can be examined through observable signals. In this

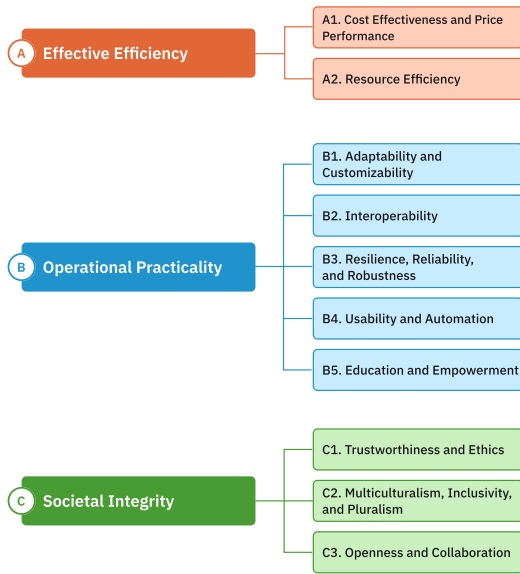


Figure 1: Overview of the Next Billion AI Index dimensions.

first version, these signals are intended as a practical starting point rather than a fixed scoring protocol. They combine standardized benchmarks where available with system-level and deployment-level evidence such as documentation, deployment requirements, pricing structures, language support, and integration pathways.

Table 1 situates the *nexbax* dimensions within the current measurement landscape. It is indicative rather than exhaustive: mature benchmarks can inform some dimensions, while others require proxy-based evidence and situated human interpretation. As Table 1 illustrates, the measurement landscape is uneven. Some dimensions, such as efficiency, robustness, and multilingual performance, benefit from established benchmarks, while others—particularly adaptability, usability, and empowerment—remain under-specified in existing evaluation frameworks. This unevenness is not incidental but reflects the inherently context-dependent nature of utility in next-billion settings, motivating *nexbax*’s hybrid measurement approach.

### 3.4 Measurement Philosophy

While *nexbax* emphasizes measurable system properties, not all dimensions of utility can be fully captured through automated benchmarks or standardized metrics. In next-billion contexts, value is often emergent, context-dependent, and mediated by human judgment—particularly in environments characterized by infrastructural variability, informal workflows, and diverse user capabilities.

*Nexbax* therefore adopts a hybrid measurement approach, combining:

- Objective signals, such as benchmark performance and observable system properties, and
- Situated evaluations, including interpretation by practitioners, public-sector implementers, civil society organizations, community intermediaries, and affected users

grounded in real deployment contexts.

This reflects a fundamental asymmetry: while model capability can be evaluated in controlled environments, utility under constraint must be interpreted in context. Fully automated evaluation risks privileging what is easily measurable—such as latency, accuracy, or cost—over what is consequential but context-dependent, including ease of integration into informal workflows, cultural appropriateness, or accessibility for developers in low-resource environments. This issue is especially acute for cultural appropriateness. Recent work on cultural benchmarks argues that evaluation cannot be reduced to static question sets or language coverage alone: the questions, tasks, and interpretation criteria themselves must be developed with attention to the communities and contexts being represented (Chiu et al. 2024). *Nexbax* therefore treats cultural fit as a participatory measurement problem, not merely a localization feature.

*Nexbax* measures not only what AI systems are, but how they are experienced in use. We do not treat subjectivity as noise to be eliminated, but as signal that reflects variation in real-world applicability across settings. In practice, *nexbax* complements local validation processes such as stakeholder consultation, domain-specific testing, institutional risk assessment, and deployment-specific evaluation rubrics. The index therefore functions as a comparative and diagnostic framework that is intended to be paired with context-sensitive interpretations of utility, risk, and institutional fit.

### 3.5 Rubrics and Initial Scoring Exercise

The measurement framework above defines what *nexbax* seeks to capture in principle. To support comparative and formative assessment across diverse AI systems, we define broad rubrics that translate each dimension into a three-level scoring system (weak, moderate, strong). These rubrics simplify a multidimensional measurement space into interpretable evaluation criteria that can be applied consistently across systems and deployment settings. Rather than exhaustively capturing every aspect of a dimension, the rubrics provide a shared evaluative baseline that can be extended or adapted for particular organizational and domain needs.

Because usefulness depends on who is adopting or adapting a technology, the first version focuses on **developer-facing system properties**: how developer choices and capabilities shape the preconditions for AI utility and inclusion. This is an intentionally limited stakeholder layer, and future iterations should develop community- and institution-facing rubrics. We summarize the rubric in Table 2. Given the hybrid measurement philosophy, the initial use of *nexbax* relies on expert-driven formative assessment rather than fully automated scoring. Practitioners interpret rubric criteria in relation to their own deployment contexts. This first exercise assesses the clarity and perceived relevance of the index for technical and product stakeholders; it does not by itself validate usefulness for end users or communities.

## 4 Formative Study

For the formative exercise, we applied the rubrics in Table 2 to three common AI stack configurations:

open-weight models, closed models, and model-plus-orchestration/application layer systems. We selected these categories because they give practitioners a recognizable basis for discussion across architecture and governance choices. They are not intended to define the full scope of *nexbox*; the same diagnostic framework can be adapted to other AI technologies, sectors, and deployment arrangements as the evidence base matures.

#### 4.1 Methods

We conducted a formative expert evaluation of the first version of the index through one-hour, semi-structured user interviews with eleven subject matter experts working on AI technology catered to next billion markets, including founders, developers, and product leaders. The goal of this evaluation was to assess the clarity, actionability, and perceived accuracy of the dimensions for technical and product stakeholders, and to iterate on them based on expert feedback. Participants tested the index by rating three illustrative technology configurations (open weight models, closed weight models, and open weight model + orchestration layer) across the ten dimensions using the rubrics in Table 2. They shared feedback on their experience using the dimensions and rubrics, including where they saw value, points of confusion, and suggestions for improvement. While this sample is appropriate for probing developer-facing system properties, we note that it does not represent the full set of stakeholders needed to assess usefulness or cultural appropriateness in next-billion contexts. We summarize participants’ job roles, use cases, and geographies in Table 3.

Table 3: Participants of the formative expert evaluation.

#	Job Role	Use Case	Geography
1	Founder	Visual analytics	India
2	Founder	Content authentication	India
3	Engineering lead	Content authentication	India
4	Head of AI	Education	India
5	Operations & analytics	Education	India
6	Product manager	Customer service	United States/India
7	Founder & CTO	Visual analytics	India
8	Developer	Information technology	Kenya
9	Engineering lead	Information technology	Kenya
10	Data scientist	Information technology	Kenya
11	CTO	AgTech	Ghana

#### 4.2 Results

The outcomes of this evaluation were (1) formative support for the index’s perceived value and ease of use among technical stakeholders, and (2) refinements to the dimensions and rubrics. Overall, the experts we interviewed felt that the index was useful for reasoning about AI utility and adoption in next-billion settings from a developer and product perspective. As one participant explained, “*all the tools that we use to evaluate our customers beyond localization, you hit all of them, like ease of use, reliability, the cost, the set up time — all those things are hit. Those are the same types of things that we use to pick our vendors.*” In particular, our findings indicate that cost effectiveness, usability & automation, and trustworthiness & ethics were the top three considerations for this set of experts. When asked to identify

the most important decision-making considerations in their work, ten participants chose cost effectiveness, five chose usability & automation, and four chose trustworthiness & ethics. It is worth noting that these dimensions span all three themes, providing early signal that each theme captures distinct and meaningful aspects of real-world decision making. At the same time, specific dimensions will require different prioritization across markets and domains, and emphasizing one may come with trade-offs for another.

Participants also saw tangible value for their own work. For example, one participant envisioned the ratings serving as an “*elimination mechanism*” to rule out technologies that don’t score well across dimensions; another speculated that “*this framework would help me choose the right models to use.*” Similarly, one product leader described how the index would simplify their decision-making process:

*“This kind of gives a summary or a digest...If you don’t have something like this that...has already done the research for you, right? And it’s presenting to you various options...you would have to sit down and do all of these things by yourself.”*

A developer similarly spoke to the value of seeing expert evaluations based on the index:

*“Within the community, I think there’s no consensus on what works for us. So we don’t have our models, so that’s for sure, but then in the plethora of models that are accessible to us, which ones actually work for us? ...[The index] would be for us to, as a community, figure out what’s best for us beyond just building policy documents, which is what countries have been doing. I think I would be very keen to see what even more experienced people, who’ve done a lot of work evaluating this, what they think.”*

At the same time, a few participants felt that the index was limited in actionability and shared suggestions for improvement. Some felt that the dimensions were too high level for their specific use cases, and others felt that a numerical rating would be insufficient to capture nuances within a dimension. For example, within cost effectiveness, one participant differentiated between upfront cost (e.g., setting up a local service) vs. long-term costs (e.g., price hikes over time). Another participant gave similar examples: “*education, I mean, is it materials for me as a dev [or] for me as...a non-tech user? ...Say, like, domain...when you talk about resource efficiency, there could be some groupings.*” To address these concerns, we recommend including explanations with ratings, as well as indicating the domain in which evaluations were conducted to support consumers in identifying evaluations relevant to them. As one participant suggested,

*“It will be very useful for sure, but I would find myself trying to go beyond this and looking for some more information. Let’s say, resource efficiency...it’s rated as a two or three but why? And for my case, in that context...what does it mean at three?”*

From a usability standpoint, the experts generally found the dimensions and rubrics easy to apply, although some felt

that it would have been difficult without the guidance and clarifications they received from the researchers during the study session. In response to participants' confusions, we made several clarifications that we anticipate will facilitate independent evaluation: adding a definition for each dimension, providing examples for a wider variety of use cases (e.g., vision and video analytics), and clarifying unfamiliar or confusing terms within dimension names and rubrics. For example, after five participants confused "robustness" with output accuracy, we de-emphasized it from the dimension name, renaming it from "Robustness (Resilience and Reliability)" to "Resilience, Reliability, and Robustness."

Finally, since our goal was to evaluate the index itself rather than collect accurate ratings for the three illustrative configurations, we omit the ratings collected across participants. A future iteration of this work will include aggregated, system-specific ratings from a larger and more diverse stakeholder sample, including public-sector implementers, civil society organizations, local domain experts, and affected users.

## 5 Limitations and Future Work

*Nexbax* should not be read as a universal measure of usefulness for the global majority. The present framework prioritizes system-level and developer-facing signals, such as affordability, deployment requirements, documentation, interoperability, and openness. These signals are important because they shape what kinds of AI systems can plausibly be adapted and sustained under constraints, but cannot determine whether a system is locally useful, culturally appropriate, or socially beneficial. Accordingly, the measurement signals in Table 1 should be read as indicative rather than definitive.

The formative evaluation is similarly limited. Participants brought relevant technical and product expertise, but they were primarily founders, developers, and product leaders. They therefore represent only one fraction of the stakeholder landscape for next-billion AI deployment. Future evaluation must include government officials, public-sector implementers, civil society organizations, community-based groups, domain workers, and affected users who can speak to local practices, institutional realities, and cultural meanings that technical experts may not observe.

Cross-context comparison also creates an unavoidable risk of flattening difference across countries, regions, languages, classes, and communities. *Nexbax* should therefore be used with explanatory evidence and deployment context, not as a standalone scalar ranking. Future work will refine the mapping between dimensions, indicators, and scoring evidence through participatory, locale-specific evaluation protocols in which local stakeholders help define the questions, evidence standards, and interpretation criteria used to assess cultural appropriateness and practical usefulness.

More fundamentally, we do not claim to provide a complete or universal rubric for assessing usefulness across the extremely diverse populations, cultures, and institutional realities that make up the global majority. Rather, we see *nexbax* as the beginning of a broader, grassroots effort to give greater voice and evaluative agency to stakeholders who

are often underrepresented in dominant AI conversations. We therefore encourage researchers, practitioners, governments, and community organizations to adapt, extend, and reinterpret the framework for their own domains, deployment settings, and local priorities. The long-term value of *nexbax* will depend not on enforcing a single canonical definition of useful AI, but on enabling locally grounded and participatory approaches to defining usefulness itself.

## 6 Conclusion

This paper began from a simple motivation: the systems most celebrated by current AI benchmarks are not necessarily the systems most likely to be useful for the next billion users. In next-billion contexts, adoption is shaped by affordability, unreliable infrastructure, linguistic diversity, informal and small-enterprise workflows, public-sector delivery channels, and uneven institutional capacity. Under these conditions, progress cannot be measured only through peak model capability. It must also be measured through the practical conditions that allow AI to be deployed, adapted, trusted, and sustained.

*Nexbax* responds to this measurement gap by offering a first-stage diagnostic for artificial *useful* intelligence. It brings together 10 dimensions under three themes: Effective Efficiency, Operational Practicality, and Societal Integrity. In doing so, the index translates broad goals of inclusive AI into concrete system-level properties, including cost-effectiveness, resource efficiency, adaptability, interoperability, robustness, usability and automation, education and empowerment, trustworthiness and ethics, multicultural alignment, and openness and collaboration. The accompanying rubrics make these properties discussable and comparable across models, orchestration frameworks, and application-layer systems, while preserving the central claim that next-billion usefulness depends on more than benchmark accuracy or model scale.

Our formative expert evaluation provides early support for this framing. Across eleven semi-structured interviews with founders, developers, product leaders, and technical practitioners working in next-billion AI markets, participants generally found the dimensions and rubrics relevant and useful for reasoning about technology choice, adoption trade-offs, and system fit. Their feedback also clarified where the framework can become more actionable: ratings should be accompanied by explanations, evaluations should specify deployment domain and context, and future versions should support more granular evidence within each dimension.

In that sense, the value of *nexbax* is not only that it produces an index, but that it shifts the object of evaluation itself: from asking which AI systems are most capable in idealized settings to asking which systems create the conditions for useful, inclusive, and sustainable adoption where the next billion users live and work.

<i>Nexbax</i> dimension	Measurement direction	Existing benchmarks / signals	Measurement gap
<b>Cost-effectiveness and Price-performance</b>	Utility per dollar; cost per successful task; total cost of ownership; availability of low-cost or community-access modes	Pricing primitives/disclosures; subscription tiers; efficiency-aware evaluation suites such as HELM (Liang et al. 2023); deployment cost estimates	No widely adopted benchmark links task utility, local affordability, and long-term cost of ownership
<b>Resource efficiency</b>	Utility per watt; tokens or outputs per joule; latency; memory footprint; bandwidth requirements; offline or edge viability	MLPerf (Reddi et al. 2020)Inference; AI Energy Score (Lucioni et al. 2024); Intelligence per watt (Saad-Falcon et al. 2025); hardware and deployment footprint disclosures	Energy and compute benchmarks rarely capture low-connectivity, mobile-first, or intermittent-infrastructure deployment conditions
<b>Adaptability and Customizability</b>	Fine-tuning support; adapter or LoRA support; prompt and configuration flexibility; domain adaptation; multimodal extensibility	Evidence from model documentation, fine-tuning APIs, adapter ecosystems, and deployment templates, tools/agent-stack	No standardized benchmark captures how easily systems can be adapted to local domains, workflows, or constraints
<b>Interoperability</b>	Standards support; API and SDK availability; connectors; data portability; ease of integration with existing tools and workflows	OpenAPI-compatible APIs; SDKs; platform integrations (LangChain/LlamaIndex/Ollama); vector database connectors; webhook and export support	Existing signals are ecosystem-specific and do not provide a unified interoperability score
<b>Resilience, Reliability, and Robustness</b>	Performance under noisy inputs, poor connectivity, outages, out-of-distribution prompts, dialectal variation, and adversarial behavior	HELM robustness evaluations; DecodingTrust (Wang et al. 2023); stress tests for safety, bias, privacy, and adversarial behavior; red-teaming; bug bounty programs; API reliability/graceful degradation	Lab robustness tests only partially reflect hostile or resource-constrained deployment environments
<b>Usability and Automation</b>	Setup time; number of steps to first working deployment; low-code/no-code support; agentic workflow reliability; documentation clarity	Developer benchmarks such as SWE-Bench (Jimenez et al. 2024) and BigCodeBench (Zhuo et al. 2025); onboarding flows; deployment tutorials; human task-completion studies	Usability is highly context-dependent and often requires practitioner judgment rather than automated scoring alone
<b>Education and Empowerment</b>	Quality of documentation; tutorials; local-language help; training materials; community support; examples for low-resource settings	Documentation audits; community forum activity; localized tutorials; public training programs; NGO or public-sector enablement materials	Few mature benchmarks assess whether users or developers are empowered to understand, adapt, and maintain systems
<b>Trustworthiness and Ethics</b>	Safety behavior; bias evaluation; privacy protections; transparency artifacts; user controls; incident reporting channels	MLCommons AILuminate (Ghosh et al. 2025); DecodingTrust (Wang et al. 2023); model cards; system cards; safety and bias evaluation reports	General safety benchmarks may miss context-specific harms, local norms, and institutional trust requirements
<b>Multiculturalism, Inclusivity, and Pluralism</b>	Language and dialect coverage; low-resource language performance; cultural appropriateness; low-literacy interaction; locale-sensitive UX	FLORES-200 (Goyal et al. 2022); MTEB/MMTEB (Muennighoff et al. 2023; Enevoldsen et al. 2025); multilingual and low-resource language benchmarks; localization documentation	Language benchmarks do not fully capture cultural fit, pluralism, dialectal nuance, or accessibility for low-literacy users
<b>Openness and Collaboration</b>	Open weights, code, and data; licensing; reproducibility; public roadmap; community governance; contribution pathways	Open-weight releases; open-source repositories; license terms; issue trackers; model cards; public roadmaps	Openness is multidimensional and cannot be inferred from weight availability alone; no feedback loops (like issues/PRs) exist for model repositories

Table 1: Indicative measurement signals for *nexbax* dimensions. Existing benchmarks provide useful starting points for some dimensions, but many aspects of next-billion utility require proxy-based evidence and situated human evaluation.

Table 2: Rubrics for the Next Billion AI Index

<b>Dimension</b>	<b>1 – weak</b>	<b>2 – okay</b>	<b>3 – strong</b>
<b>Cost-Effectiveness and Price-Performance</b>	high TCO (usage cost + infra) for target users; no low-cost mode	Some free/low-cost tier or decent on-device/edge efficiency	clear low-cost path (free/community tier, metered API at local-friendly prices, runs on low-end devices/edge)
<b>Resource Efficiency</b>	heavy compute/infra, high latency, bandwidth hungry	moderate footprint (e.g. via quantization/pruning); decent latency	runs on low-spec edge/mobile; offline mode; optimized IO/latency
<b>Adaptability and Customizability</b>	little/no finetune or UI/config; one modality	configurable; accepts adapters; multi-modal options	modular and extensible: finetuning, plug-in tools, multi-modal, rapid pivots
<b>Interoperability</b>	closed garden; limited standards/APIs	some SDKs; limited adapters	open standards plus rich connectors (tools, packages, webhooks); easy data in/out
<b>Resilience, Reliability, and Robustness</b>	fails on flaky networks/dirty inputs; crashes; no retries	survives short outages; tolerates moderate input noise; onus on infra provider	designed for hostile environments: long offline queues/sync, power-safe modes, noise/dust tolerant pipelines, graceful degradation
<b>Usability and Automation</b>	expert-only; steep setup; code-only	guided wizards/templates; some autoconfig; reasonable documentation	no/low-code flows; opinionated templates; agentic automation; strong onboarding
<b>Education and Empowerment</b>	sparse documentation; no local-language help	documentation and videos; community forum; some localization	tutorials, local-language/ contextual help, community programs and NGO partners
<b>Trustworthiness and Ethics</b>	opaque systems; unknown data; weak safety; no bias evaluation	some documentation; basic filters; incident channel	model/system cards; bias and safety evaluations across locales; user controls
<b>Multiculturalism, Inclusivity, and Pluralism</b>	English-only; Western defaults	major languages; partial localization; limited review of cultural assumptions	broad language and dialect support; documented cultural assumptions; local stakeholder review; low-literacy UX and adaptation pathways
<b>Openness and Collaboration</b>	no weights or code; no roadmap or community	some APIs or open SDKs; limited collaboration	open weights or strong open standards; active community and public roadmap

## 7 Positionality Statement

The authors of this paper bring interdisciplinary and cross-regional perspectives shaped by work in AI systems, HCI, trustworthy AI, product development, and AI governance across Africa, Asia, Europe, North America, and South America. Several authors have worked closely with communities and organizations operating in low-resource and next-billion-user contexts, including settings characterized by infrastructural constraints, linguistic diversity, and unequal access to digital technologies. These experiences informed the paper's focus on affordability, deployability, inclusivity, and contextual utility as central dimensions of AI adoption.

At the same time, the authors recognize that no single research team can fully represent the diversity of experiences and priorities across the global majority. The proposed framework is therefore intended not as a universal definition of useful AI, but as a starting point for participatory, context-sensitive evaluation and dialogue with local stakeholders and affected communities.

## 8 Acknowledgments

We thank Varun Aggarwal, Anthony Annunziata, Daksh Chawla, Nahuel Defosse, George K. Githae, Sudeep Gowrishankar, Raghav Gupta, Jaikrishnan Hari, Ronak Khandewal, Jordan McAfoose, Rohan Sahu, Siddhant Sachdeva, Het Shah, and Amith Singhee for their assistance with the survey process, feedback, and other contributions during the development of this work. Subhabrata Majumdar's research is supported by the Indian Institute of Management Bangalore Young Faculty Research Grant.

## References

2025. Localizing AI in the global south. *Nature Machine Intelligence*, 7: 675. Editorial.
- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Adams, R.; Adeleke, F.; Florido, A.; de Magalhães Santos, L. G.; Grossman, N.; Junck, L.; and Stone, K. 2024. Global Index on Responsible AI 2024 (1st Edition). Technical report, South Africa: Global Center on AI Governance.
- Adler, B.; Agarwal, N.; Aithal, A.; Anh, D. H.; Bhat-tacharya, P.; Brundyn, A.; Casper, J.; Catanzaro, B.; Clay, S.; Cohen, J.; et al. 2024. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*.
- African Union. 2024. Continental Artificial Intelligence Strategy. <https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy>.
- Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic.
- Anthropic. 2026. Anthropic Economic Index. <https://www.anthropic.com/research/economic-index-primitives>.
- Bailey, G.; Kalarikalayil Raju, D.; Pearson, J.; Robinson, S.; and Jones, M. 2026. Street Scenes: Public Appliances for GenAI Video in Informal Settlements. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–22.
- Banerjee, A. V.; and Duflo, E. 2009. The experimental approach to development economics. *Annu. Rev. Econ.*, 1(1): 151–178.
- Banerjee, A. V.; and Duflo, E. 2026. *Poor economics: Rethinking poverty & the ways to end it*. Penguin Random House India Private Limited.
- Belcak, P.; Heinrich, G.; Diao, S.; Fu, Y.; Dong, X.; Muralidharan, S.; Lin, Y. C.; and Molchanov, P. 2025. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*.
- Borning, A.; and Muller, M. 2012. Next Steps for Value Sensitive Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- Casper, S.; O'Brien, K.; Longpre, S.; Seger, E.; Klyman, K.; Bommasani, R.; Nrusimha, A.; Shumailov, I.; Mindermann, S.; Basart, S.; et al. 2025. Open technical problems in open-weight AI model risk management. *Transactions on Machine Learning Research*.
- Chandiramani, A.; Blakeman, A.; Olaoye, A.; Gupta, A.; Somasamudramath, A.; Khattar, A.; Adesoba, A.; Renduchintala, A.; Asif, A.; Agrawal, A.; et al. 2026. Nemotron 3 Super: Open, Efficient Mixture-of-Experts Hybrid Mamba-Transformer Model for Agentic Reasoning. *arXiv preprint arXiv:2604.12374*.
- Chiu, Y. Y.; Jiang, L.; Lin, B. Y.; Park, C. Y.; Li, S. S.; Ravi, S.; Bhatia, M.; Antoniak, M.; Tsvetkov, Y.; Shwartz, V.; et al. 2024. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring (the Lack of) Cultural Knowledge of LLMs.
- Clancy, R. F.; Zhu, Q.; and Majumdar, S. 2025. Exploring AI ethics in global contexts: a culturally responsive, psychologically realist approach. *AI and Ethics*, 5(6): 6329–6338.
- Eiras, F.; Petrov, A.; Vidgen, B.; Schroeder, C.; Pizzati, F.; Elkins, K.; Mukhopadhyay, S.; Bibi, A.; Purewal, A.; Botos, C.; et al. 2024. Risks and opportunities of open-source generative ai. *arXiv preprint arXiv:2405.08597*.
- Enevoldsen, K.; Chung, I.; Kerboua, I.; Kardos, M.; Mathur, A.; Stap, D.; Gala, J.; Siblinski, W.; Krzemiński, D.; Winata, G. I.; et al. 2025. Mmtex: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.
- FMCIDE. 2025. National Artificial Intelligence Strategy (2025–2029). <https://fmcide.gov.ng/initiative/nais/>.
- Ghosh, S.; Frase, H.; Williams, A.; Luger, S.; Röttger, P.; Barez, F.; McGregor, S.; Fricklas, K.; Kumar, M.; Bollacker, K.; et al. 2025. Ailuminat: Introducing v1. 0 of the ai risk and reliability benchmark from mlcommons. *arXiv preprint arXiv:2503.05731*.
- Gibbons, S. 2018. Empathy Mapping: The First Step in Design Thinking. Nielsen Norman Group.
- Government of India. 2026a. AI Impact Summit 2026. <https://impact.indiaai.gov.in/>.
- Government of India. 2026b. IndiaAI Mission. <https://indiaai.gov.in/>.

- Government of Kenya. 2025. Kenya Artificial Intelligence Strategy 2025–2030.
- Goyal, N.; Gao, C.; Chaudhary, V.; Chen, P.-J.; Wenzek, G.; Ju, D.; Krishnan, S.; Ranzato, M.; Guzmán, F.; and Fan, A. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10: 522–538.
- ISO and IEC. 2023. ISO/IEC 42001:2023 Information Technology – Artificial Intelligence – Management System. <https://www.iso.org/standard/42001>.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Lavallée, B.; Douillard, X.; El-Sayed, W.; Flagel, A.; Galtier, D.; Gombard, J.; et al. 2024. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.
- Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, 54107–54157.
- Karnani, A. 2007. The mirage of marketing to the bottom of the pyramid: How the private sector can help alleviate poverty. *California management review*, 49(4): 90–111.
- Kim, R. M.; Kuehnert, B.; Lazar, S.; Singh, R.; and Heidari, H. 2025. The AI Power Disparity Index: Toward a Compound Measure of AI Actors’ Power to Shape the AI Ecosystem. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 1453–1464.
- Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344.
- KPMG. 2025. Trust, Attitudes and Use of Artificial Intelligence. <https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html>.
- Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C. A.; Manning, C. D.; Re, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R. A.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*.
- Luccioni, S.; Gamazaychikov, B.; Strubell, E.; Hooker, S.; Jernite, Y.; Mitchell, M.; and Wu, C.-J. 2024. AI Energy Score. <https://huggingface.co/AIEnergyScore>.
- Melville, P.; Chenthamarakshan, V.; Lawrence, R. D.; Powell, J.; Mugisha, M.; Sapra, S.; Anandan, R.; and Assefa, S. 2013. Amplifying the voice of youth in africa via text analytics. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1204–1212.
- METR. 2025. Resources for Measuring Autonomous AI Capabilities. <https://metr.org/measuring-autonomous-ai-capabilities/>.
- METR. 2026. Task-Completion Time Horizons of Frontier AI Models. <https://metr.org/time-horizons/>.
- Microsoft. 2025. Microsoft New Future of Work 2025. <https://www.microsoft.com/en-us/research/project/the-new-future-of-work/>.
- Microsoft Research Africa. 2025. Atlas: A Playbook for Cross Cultural AI. <https://github.com/microsoft/Atlas>. A practical guide conducting cross-cultural AI Research for Development and Deployment.
- Microsoft Research Africa, N. 2026. PazaBench: A Benchmark for Automatic Speech Recognition on Low Resource Languages. <https://www.microsoft.com/en-us/research/project/project-gecko/>. Alpha version. Part of Project Gecko - Equitable Generative AI for the Global Majority.
- Miranda, L. J. V.; Hu, S.; Reichart, R.; and Korhonen, A. 2026. Multilinguality at the Edge: Developing Language Models for the Global South. *arXiv preprint arXiv:2604.21637*.
- MIT Media Lab. 2026. ImpactBench. <https://impactbench.media.mit.edu/>.
- Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, 33(4): 659–684.
- Muchai, M.; Chege, K.; Mumbero, N.; and Nyairo, S. 2026. Paza: Introducing automatic speech recognition benchmarks and models for low-resource languages. <https://www.microsoft.com/en-us/research/blog/paza-introducing-automatic-speech-recognition-benchmarks-and-models-for-low-resource-languages/>.
- Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037.
- NITI Aayog. 2025. AI for Viksit Bharat: The Opportunity of Accelerated Economic Growth. <https://niti.gov.in/whats-new/ai-viksiti-bharat-opportunity-accelerated-economic-growth>.
- Odame-Darkwah, K.; and Tsado, A. 2025. ISF Voices 2025: Africa’s Playbook. Substack article.
- OECD. 2026. The OECD.AI Index. <https://doi.org/10.1787/32c01014-en>.

- Oluwatuyi, R.; Pillay, V.; Mazwi, J.; Castro, A.; Farao, J.; Srivatsa, S.; Bagalkot, N.; and Densmore, M. 2026. Collectively Reimagining Artificial Intelligence With Marginalized Communities. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–21.
- Oxford Insights. 2025. Government AI Readiness Index. <https://oxfordinsights.com/ai-readiness/>.
- Prabhakaran, V.; Mitchell, M.; Gebru, T.; and Gabriel, I. 2022. A Human Rights-Based Approach to Responsible AI. *arXiv preprint arXiv:2210.02667*.
- Prahalad, C. K. 2006. *The Fortune at the Bottom of the Pyramid: Eradicating Poverty Through Profits*. Upper Saddle River, NJ: Prentice Hall.
- Prahalad, C. K. 2012. Bottom of the Pyramid as a Source of Breakthrough Innovations. *Journal of Product Innovation Management*.
- Reddi, V. J.; Cheng, C.; Kanter, D.; Mattson, P.; Schmuelling, G.; Wu, C.-J.; Anderson, B.; Breughe, M.; Charlebois, M.; Chou, W.; et al. 2020. MLPerf inference benchmark. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 446–459. IEEE.
- Rismani, S.; Shelby, R.; Davis, L.; Rostamzadeh, N.; and Moon, A. 2025. Measuring What Matters: Connecting AI Ethics Evaluations to System Attributes, Hazards, and Harms. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2199–2213.
- Roberts, H. 2024. Digital Sovereignty and Artificial Intelligence: A Normative Approach. *Ethics and Information Technology*, 26(70).
- Saad-Falcon, J.; Narayan, A.; Akengin, H. O.; Griffin, J.; Shandilya, H.; Lafuente, A. G.; Goel, M.; Joseph, R.; Natarajan, S.; Guha, E. K.; et al. 2025. Intelligence per watt: Measuring intelligence efficiency of local ai. *arXiv preprint arXiv:2511.07885*.
- Sambasivan, N.; Arnesen, E.; Hutchinson, B.; Doshi, T.; and Prabhakaran, V. 2021. Re-imagining Algorithmic Fairness in India and Beyond. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Seger, E.; Dreksler, N.; Moulange, R.; Dardaman, E.; Schuett, J.; Wei, K.; Winter, C.; Arnold, M.; hÉigeartaigh, S. Ó.; Korinek, A.; et al. 2023. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.
- Shoham, Y.; Perrault, R.; Brynjolfsson, E.; and Clark, J. 2017. Artificial Intelligence Index Report. Technical report, Stanford University, Human-Centered Artificial Intelligence.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- UNDP. 2025. Human Development Report 2025: A Matter of Choice: People and Possibilities in the Age of AI. Technical report, UNDP, New York.
- UNESCO. 2023. Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO.
- Varshney, K. R. 2023. Foundation Model Platforms and Bottom-of-the-Pyramid Innovation. In *ICLR Workshop on Practical Machine Learning for Developing Countries*.
- Varuvel Dennison, D.; Jain, M.; Ganu, T.; and Vashistha, A. 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*, 1–22.
- Vibhasha Team. 2025. Vibhasha: The Multilingual Playbook for Large Language Models. <https://github.com/microsoft/vibhasha-playbook>. A practical guide for developing multilingual and culturally-aware LLM systems.
- Wan, A.; Kyleman, K.; Kapoor, S.; Maslej, N.; Longpre, S.; Xiong, B.; Liang, P.; and Bommasani, R. 2023. Foundation Model Transparency Index.
- Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in {GPT} Models. *Neural Information Processing Systems Datasets; Benchmarks Track*.
- Zhuo, T. Y.; Vu, M. C.; Chim, J.; Hu, H.; Yu, W.; Widyasari, R.; Yusuf, I. N. B.; Zhan, H.; He, J.; Paul, I.; et al. 2025. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations*, volume 2025, 66602–66656.