EVENT PROCEEDINGS

A Gender-Intentional Framework for Evaluating AI-based Solutions in India's Development Sector

gLocal Evaluation Week 2026 | Theme: Evaluation, Evidence, and Trust in the Age of AI

Organised by the Gender and Digital hub (GxD hub), LEAD at Krea University–IFMR

4 June 2026 | 5:30–7:00 PM IST | Zoom Webinar

1. Event Details

Title	A Gender-intentional Framework for Evaluating AI-based solutions in India's Development Sector
Date & Time	4 June 2026 | 5:30–7:00 PM IST
Format	Panel + Practitioner Dialogue | 90 minutes | Online (Zoom Webinar)
Organiser	Gender and Digital hub (GxD hub), LEAD at Krea University–IFMR
Event Theme (gLocal 2026)	Evaluation, Evidence, and Trust in the Age of AI
Classification	Ethics, Standards & Human Judgement; Practical Applications of AI
Target Audience	Researchers, M&E practitioners, policymakers working in AI, gender, and development
Total Participants	*35*
Event Page	*https://glocalevalweek.org/glocal-events*

2. Background & Rationale

While this year's gLocal theme focuses on evaluation, evidence, and trust in the age of AI, this session offers a complementary lens: the evaluation of AI itself, from an equity and gender perspective. Evaluations of AI-powered solutions are largely trapped in a narrow technical frame - measuring accuracy and engagement metrics while missing gendered risks and blind spots in design, deployment, sustained adoption, and developmental impact of AI solutions. When evaluations fall short, flawed tools get scaled. This session introduced a gender-intentional conceptual evaluation framework: grounded in theory and practical to apply. This was followed by a panel discussion bringing together researchers and practitioners evaluating AI tools in the development sector, which are at different stages of

deployment and testing. Through a discussion grounded in actual experiences, the panel surfaced what current approaches are missing, what is at stake, and what addressing it would require.

This session recording will be relevant for anyone commissioning, conducting, or using AI evaluations and provide gender-intentional guidance for framing evaluative questions and a shared vocabulary for advocating for gender-intentional evaluation standards within their organisations.

3. Session Summary

Segment	Speaker(s)/Moderator	Description
Opening & Introduction **5:30 PM – 5:36 PM**	Mahima Taneja – Associate Director – Research & MLE, GxD hub, LEAD at Krea University (Moderator)	Introduced the session's framing: while gLocal 2026 asks how AI can improve evaluations, this session asks how AI solutions themselves should be evaluated, from an equity and gender perspective. Introduced the four panelists and set out the stakes: when evaluations of AI tools fall short, flawed tools get scaled.
Framework Presentation (Part 1): What is at Risk **5:36 PM – 5:50 PM**	Mahima Taneja – GxD hub, LEAD at Krea University	Presented four categories of gendered risk in AI: bias in training data and design; gendered harms (deepfakes, welfare exclusion, biased credit); epistemic harm from AI's optimisation for smoothness at the cost of friction where marginalised voices live; and deepening of the gender digital divide through assumptions of individual, uninterrupted device access.
Framework Presentation (Part 2): Gender-Intentional Evaluation Across Four Stages **5:50 PM – 6:07 PM**	Deepti – Research Associate, GxD hub, LEAD at Krea University	Presented the gender-intentional evaluation framework across four stages: Model (framing trap: accuracy assumed neutral; Dalit, Adivasi, OBC women under-represented); Product (formalism trap: solo literate user assumed; shared devices and constrained digital access invisible to engagement metrics); User (ripple

		effect trap: epistemic displacement when AI substitutes embodied knowledge); Impact (solutionism trap) : evaluation commissioned post-deployment; efficiency gains not equivalent to livelihood improvements). Argued for concurrent interdisciplinary evaluation from the model stage onwards.
Panellist Q&A: Round 1 **6:07 PM – 6:34 PM**	Urvashi Wattal, Nilakshi Biswas, Kalika Bali, Aarushi Gupta, Mahima Taneja (Moderator) Khushi Baby | NuSocia | Microsoft Research India | Digital Futures Lab | GxD hub	Each panelist reflected on their experience evaluating AI tools in practice: the difference between model validation and field-level user assessment (Wattal); lessons from evaluating nine AI organisations across sectors under the Pragati programme (Biswas); benchmark limitations and LLMs-as-judges in multilingual contexts (Bali); gender bias in Indian-language LLMs and the difficulty of standardised rubrics (Gupta).
Panellist Q&A: Round 2 & Moderator Follow-ups **6:34 PM – 6:50 PM**	Urvashi Wattal, Nilakshi Biswas, Kalika Bali, Aarushi Gupta, Mahima Taneja (Moderator) Khushi Baby | NuSocia | Microsoft Research India | Digital Futures Lab | GxD hub	Follow-up questions on adaptive evaluation timelines for iterating AI tools and measuring intermediary versus beneficiary outcomes; challenges of creating socioculturally grounded benchmarks; benchmark contamination and LLM-judge limitations; funding asymmetry between AI deployment and evaluation.
Audience Q&A & Closing **6:50 PM – 7:00 PM**	Urvashi Wattal, Nilakshi Biswas, Kalika Bali, Aarushi Gupta, Mahima Taneja (Moderator) Khushi Baby | NuSocia | Microsoft Research India | Digital Futures Lab | GxD hub	Audience questions on the theory of change for AI solutions; gender audits of mental health chatbots; governance of open-weight versus proprietary models. Closing summary and thanks.

4. Speakers and Contributors

Panelists

Urvashi Wattal
Country Lead – Evidence & Impact | Khushi Baby
Urvashi Wattal is a public health and digital development professional working at the intersection of AI, evidence generation, and health systems strengthening. At Khushi Baby, she leads implementation research, M&E, and the use of AI-enabled tools to improve last-mile healthcare delivery, with a focus on equitable and scalable approaches within public health systems in India.

Aarushi Gupta
Senior Research Manager | Digital Futures Lab
Aarushi Gupta leads research on gender bias in Indian-language LLMs across healthcare and agriculture. She has presented at ACM FAccT and advised AI developers in South Asia and Africa on responsible AI practices.

Nilakshi Biswas
Head, NuSocia Translational Research Centre (NTRC) | NuSocia
Nilakshi Biswas brings expertise in evidence synthesis, theory of change, and public health policy. She previously contributed to impact evaluation research at 3ie and holds an MPH from George Washington University.

Kalika Bali
Senior Principal Researcher | Microsoft Research India
Kalika Bali works on NLP, multilingual AI, and gender-intentional datasets for Indian languages. Named to TIME100 AI 2023, she advocates for culturally grounded, gender-aware AI development.

Organisers and Moderator

Mahima Taneja
Associate Director – Research & MLE, GxD hub | Moderator | LEAD at Krea University
Dr. Mahima Taneja leads research and evaluation for the Gender x Digital (GxD) hub at LEAD. She brings over 12 years of experience and has led several program evaluations and research studies in the domain of gender equality, women's livelihoods, digital inclusion, and urban governance. Prior to joining LEAD, she worked with the Gates Foundation and Sambodhi on multiple MLE

projects. Mahima is an active member of the Evaluation Community of India and EvalYouth Asia and has been engaging with critical evaluation methods and feminist evaluation approaches. She holds an M.A., M.Phil., and Ph.D. in Political Science from Jawaharlal Nehru University, New Delhi.

Deepti
Research Associate, GxD hub | Co-presenter | LEAD at Krea University
Deepti is an early-career researcher at the Gender x Digital (GxD hub), LEAD at Krea University, where her work focuses on gender, technology, and digital labor in India. She has previously worked on labour market research using large-scale administrative datasets and computing technology exposure and at the University of Michigan on computational research examining health misinformation in Indian digital media, resulting in a co-authored paper and presented at ICWSM '26. Her research interests span feminist policy analysis, AI governance, development economics, and computational social science. She holds an M.Sc. in Computational Social Science from IIT Jodhpur and a B.A. (Hons.) in Economics from Ambedkar University Delhi.

Vaidehi Sahasrabhojanee
Policy Associate, GxD hub | Co-presenter | LEAD at Krea University
Vaidehi Sahasrabhojanee is an early-career researcher at the Gender x Digital (GxD hub), LEAD at Krea University. As part of her role, she conducts policy research and qualitative analysis, translating research and evidence into actionable insights, with expertise in literature reviews. evaluations, and qualitative methods. Before joining LEAD, Vaidehi gained experience in government relations, research, and evaluation at the Quality Council of India. She has a Master's degree in Public Policy from Christ University, Bangalore, and a Bachelor's degree in Sociology from the Manipal Centre for Humanities, Manipal.

5. Panel Discussion

The session was structured around questions posed by the moderator to individual panelists, with follow-up rounds and audience Q&A. What follows is a summary of the key points made by each speaker in response to each question.

Question 1

Khushi Baby has been evaluating AI tools at two very different stages of maturity: the Asha Saheli chatbot, which is in the field testing phase, and the smartphone anaemia diagnostics tool, which is still at a model validation stage. Walk us through what evaluation has looked like at each stage, what the difference is between expert technical evaluation and community-level user assessment, and what a feminist evaluation lens has meant in practice.

Speaker: Urvashi Wattal (Country Lead – Evidence & Impact, Khushi Baby)

- ➔ Khushi Baby has spent a decade designing technology with ASHA workers and frontline health workers, with the aim of improving health system decision-making through data. The organisation's foundational platform, CHIP (Community Health Integrated Platform), was built through sustained field engagement with ASHAs, mapping the hierarchy from ASHAs to ANMs to CHOs.
- ➔ The anaemia diagnostics tool is at a nascent model validation stage. Using conjunctiva images taken by frontline workers, the ML algorithm predicts moderate or severe anaemia. Current evaluation questions focus on accuracy, sensitivity, specificity, robustness, and whether training data is representative enough for geographic expansion. This is the first AI-based anaemia diagnostics tool developed in the Global South context, and evaluation is expert-led with clinicians and technical partners.
- ➔ The Asha Saheli chatbot is an LLM integrated into WhatsApp, voice-enabled, and available in local vernacular languages. It was developed in collaboration with Microsoft Research, trained on government health protocols, and designed to answer ASHA workers' real-time queries during home visits. It was tested first with clinicians for response accuracy, then piloted with 20 ASHAs, and has since scaled to 11,000 ASHAs in Rajasthan and Maharashtra. Current evaluation asks whether it improves ASHA knowledge, confidence, and quality of service delivery.
- ➔ The feminist lens matters at both stages. For the anaemia tool: tribal populations have higher prevalence of sickle cell anaemia, making demographic representativeness of training data a critical equity question. For Asha Saheli: older ASHAs less comfortable with digital tools risk being left behind, affecting the populations they serve. The tool's design on WhatsApp with voice enablement was a deliberate response to gender-related digital divide constraints.
- ➔ A key insight from field implementation: digital literacy cannot be reduced to device familiarity. ASHA workers struggled to frame prompts effectively. Once guided on how to structure queries, chatbot response quality improved and adoption increased. This dimension of literacy is invisible to standard evaluation frameworks.
- ➔ The prior question raised during GxD hub's presentation, 'is there actually a problem that AI should solve?', resonated strongly and was identified as the first evaluative act before any solution-level assessment begins.

Question 2

You have been the M&E partner for the Pragati programme, evaluating AI solutions across health, education, justice, and agriculture. How has evaluating AI-powered solutions differed from conventional programme evaluation? What framework did you apply, and where did it fall short?

Speaker: Nilakshi Biswas (Head, NuSocia Translational Research Centre (NTRC))

- ➔ NuSocia serves as an M&E partner for Pragati, led by the Nudge Foundation and supported by Meta, which brought together nine organisations spanning edtech, health tech, legal tech, agritech, and livelihoods. Conventional M&E frameworks did not apply; a bespoke AI evaluation architecture was developed.
- ➔ The Pragati framework comprised four components: organisational evaluation (AI adoption and readiness); model evaluation (operational efficiency, social impact metrics, scalability); system and contextual evaluation (data privacy policies, infrastructure constraints); and user centricity (stakeholder satisfaction and feedback, which received relatively less weight).
- ➔ A central finding: the object and objective of AI solutions keep changing as the model learns, the project pivots, use cases expand, and the user base shifts. This made applying fixed baseline-midline-endline metrics deeply problematic. For some organisations, metrics were applied too early; for others, too late.
- ➔ Quarterly reporting cycles were mismatched to developmental pace, particularly in health tech where compliance requirements constrain iteration speed. One-size evaluation architectures did not fit nine organisations at very different maturity stages.
- ➔ A critical distinction emerged between AI-native organisations (built entirely around their AI solution, moving fast) and AI-layered organisations (embedding AI onto existing community-grounded interventions, often outsourcing technology development, which breaks the connection between user knowledge and model design).
- ➔ A second distinction: organisational AI readiness (capability to understand and scale the solution) versus solution AI validity (model capability itself). These two are frequently conflated, distorting evaluation questions and metrics.
- ➔ A third: the user is not always the same as the last mile beneficiary. In many deployments, the intervention passes through an intermediary – an ASHA worker, a health worker, a teacher – before reaching the person intended to benefit. We need to ask evaluation questions at the right level, for the right stakeholder group
- ➔ Acknowledged failures: hard-to-reach beneficiaries; coming in after deployment decisions were already made; being unable to track quality of scale as tools expanded (reach was measured, model quality degradation was not).
- ➔ The question 'Do you need AI, or do you need automation?' i.e., a relevance question which is present in frameworks like OECD-DAC is routinely skipped because organisations had already been funded to build AI solutions.

Question 3

From your work on NLP, multilingual AI, and gender audits: what does model-stage evaluation look like in practice? How do you create meaningful benchmarks that capture representational and cultural context, what are the limits of that approach, and what is the role of different stakeholders at this stage?

Speaker: Kalika Bali *(Senior Principal Researcher, Microsoft Research India)*

- ➔ Most model evaluations are conducted on standard benchmarks assembled by technologists, reflecting a narrow scope of what can be discovered. Many benchmarks are translated from English rather than created in the local languages - a practice that fundamentally misrepresents sociocultural knowledge. Translation-based benchmarks are only adequate for factual, context-free queries; they collapse entirely for culturally embedded, value-laden, or socially complex questions.
- ➔ The shift in her team's work has been toward contextually situated benchmarks: developed with domain practitioners, application builders, and end users who specify what the model should actually be tested for. Simplistic benchmarks still have value as early-stage milestones during model development, but they represent a very limited slice of what matters at deployment.
- ➔ Benchmark contamination is a structural problem: once a benchmark is public, models train on it and it becomes invalid as an evaluation instrument. This requires either continuous benchmark renewal or custodianship by a credible authority keeping benchmarks private.
- ➔ The LLMs-as-judges approach using language models to evaluate other language models because human evaluators are difficult to recruit is, at best, a stand-in. Emerging work (eg under Samiksha project) shows that LLM judges agree with humans on factual, clear-cut queries but diverge significantly when questions become socioculturally complex, multidimensional, or contested. Human judges themselves disagree in such cases, which is a signal of genuine complexity rather than a reason to collapse to a single metric.
- ➔ Gender bias in AI is not a binary or a fixed property. What constitutes bias varies by social context, economic structure, power dynamics, and geography. A bias framing appropriate to a tech professional in Bangalore is categorically different from one appropriate to a woman in a village outside Bangalore. Collapsing this into standardised occupational proxies (doctor = male, nurse = female) misses the structural, embodied, and contextual nature of patriarchal norms in India.
- ➔ Languages like Hindi require gendered grammatical assignment; a model trained on historically male-associated data will consistently output male gender for 'doctor'. Mandating a female output does not achieve equity; arbitrary frequency-based

solutions create incoherence. Understanding grammatical gender and its interaction with social power is a prerequisite for meaningful evaluation.

- ➔ Research on the Sanmati project engaging women in mobile-based data tasks showed that for men, participation decisions turned on pay. For women, the decision involved family permission, social sanction, and time availability. This complexity is entirely absent from benchmark-based model evaluation, yet it is central to understanding AI's impact on women's agency. Quantifying that agency requires going into the field, not running metrics.
- ➔ Populations are not homogeneous. 'Population scale' in practice often means 'one thing fits all'. Evaluation for women in Chhattisgarh must be substantively different from evaluation for women in Himachal Pradesh. The resources and effort required for contextually differentiated evaluation are rarely funded.

Question 4

As an AI safety researcher, your work looks directly at gender bias inside LLM models. What does gender bias actually look like in a model, how do you measure it, and why does this make building a standardised evaluation rubric so difficult?

Speaker: Aarushi Gupta *(Senior Research Manager, Digital Futures Lab)*

- ➔ Starting in 2023, the team examined gender bias in Indian-language LLMs. Existing flagship literature focused primarily on occupational term associations with male/female pronouns, the doctor-nurse framing, which captures a narrow and relatively superficial form of bias. The more structurally embedded forms of patriarchal norms in India are not captured by these prompts.
- ➔ Baseline model-level evaluation remains necessary: it surfaces what the base model has learnt in terms of associations, stereotypes, and propensity to generate denigrated portrayals of women. But the end user does not encounter the base model; they encounter a finished product that will generate context-specific harms differing substantially from what base-model benchmarks reveal.
- ➔ Standardisation is genuinely difficult. Some forms of gender bias command near-universal agreement (wage gaps). Others' norms around clothing, food, and mobility vary significantly across geographies, class positions, and religious communities. A rubric encoding the values of liberal, city-based women fails to reflect the lived realities of women in diverse settings. Whose values a rubric encodes is a political question, not a technical one.
- ➔ A more productive approach may be to align on shared evaluation values rather than standardised indicators. Values such as recognition, agency, actionability, and contextual relevance can be agreed upon at a general level and then localised into context-specific indicators by practitioners in each sector and deployment

environment. In healthcare, agency looks different from what it looks like in agriculture, welfare administration, or EdTech.

- ➔ Subtle forms of gender bias that recent research is beginning to examine: differential tonality when the model responds to female versus male versus non-binary users; differential assumptions about decision-making capacity (does the model assume a female farmer has land entitlements? Does it assume a woman can independently make reproductive health decisions? These are not captured by standard benchmarks.
- ➔ Tacit knowledge held by grassroots and civil society organisations – the kind of contextually embedded understanding of Does realities that formal datasets omit – is precisely what models need to learn from and what evaluation frameworks need to surface. Involving these organisations is not optional.
- ➔ Benchmarks are also gameable. Once public, models can be trained on them, which is why models now produce gender-corrected occupational outputs on benchmarked prompts without underlying structural change. Evaluation must shift toward approaches not reliant on static benchmark prompts.
- ➔ Open-weight models shift agency downstream: local organisations and civil society can fine-tune on their own sociocultural data and define their own safety parameters, democratising what was previously inherited risk. The countervailing risk is the absence of reputational or financial accountability mechanisms for downstream actors and the structural lag of governance relative to model proliferation.

Q&A: Round 2

To Urvashi and Nilakshi: Both of you are looking at AI solutions at or approaching deployment stage. AI tools are constantly evolving and learning. How do we design evaluation for continuously evolving tools? When do evaluators come in? And how do we measure community-level outcomes when an ASHA worker, banking sakhi, or agri-extension worker is the intermediary holding the tool?

Speaker: Urvashi Wattal *(Khushi Baby)*

- ➔ External evaluators should come in as early as possible, even before formal evaluation begins, because the initial stage is where critical questions about whether the solution is appropriate and whether objectives are correctly framed are most tractable. An evaluator's presence at ideation can prevent the solutionism trap from closing.
- ➔ External evaluators are most powerful at the last-mile impact stage: answering what lasting change the tool leaves on health systems, whether it genuinely solves the problem at scale, and what outcomes governments and funders need to see. These

questions, covering both intermediary users and final beneficiaries, are where conventional evaluation expertise is most directly applicable.

- ➔ Traditional 2–3-year evaluation timelines do not fit tools that are rapidly evolving. What is needed are early signals of adaptive evaluation that tracks whether the solution is on the right track as it iterates. This resembles A/B testing but with social impact as the primary outcome, not product usage metrics.

Speaker: Nilakshi Biswas *(NuSocia)*

- ➔ Setting the success matrix at the beginning before deployment decisions harden is the single most important evaluative act. In social development contexts, success is not always efficiency or speed. For a solution giving internet access to physically disabled children, the only metric that matters is whether access is achieved at all, regardless of how long it takes. Default efficiency frameworks miss this entirely.
- ➔ The evaluator's role is to prompt the question: what is the objective for the user, and what is the objective for the beneficiary? Tool developers systematically underask this question and overindex on product functionality.

To Kalika and Aarushi: Is it challenging to create benchmarks attuned to local context, dialects, gender contexts, and cultural contexts? What have those challenges been, and what is the workaround?

Speaker: Kalika Bali *(Microsoft Research India)*

- ➔ It is extremely challenging. The people who have the knowledge to specify what should be evaluated – those with deep cultural and sociocultural understanding – are generally not people who understand evaluation methodology or model behaviour. Annotators available for dataset work often lack the cultural context to annotate accurately. The gap between model specialists, cultural knowledge holders, and data workers is a structural obstacle that requires intensive facilitated interaction to bridge, and this is rarely resourced.

Speaker: Aarushi Gupta *(Digital Futures Lab)*

- ➔ Beyond the challenge of creation, benchmarks have inherent methodological limits: they are static test prompts and cannot capture sociocultural nuance or harms that manifest over repeated interactions rather than at a single point. For many Indian languages, there is insufficient data to construct a meaningful benchmark at all. The field needs to shift toward evaluation paradigms not anchored in benchmark performance.
- ➔ Benchmark contamination: once public, benchmarks are trained on. Models now produce superficially gender-corrected outputs on known benchmark prompts

without structural change in underlying representations. Credible benchmarking requires either continuous renewal or private custodianship by an authoritative body.

Question for all panelists:

In this enthusiasm for AI, is there appetite and funding for the depth of evaluation and research that is actually needed? If not, what can evaluators do or push for?

Speaker: Urvashi Wattal *(Khushi Baby)*

- ➔ Among practitioners, there is genuine appetite. Organisations like Khushi Baby want to test, not assume. But funding research on safety, impact, and user feedback is a different matter from funding solution deployment. Healthy scepticism about AI solutions needs dedicated resourcing.

Speaker: Nilakshi Biswas *(NuSocia)*

- ➔ Funders of AI solutions in social development tend to apply commercial ROI logic: what is the return? Social programme funders ask the why, the ground-level detail. The rush to fund AI in development, largely driven by large tech, has imported commercial evaluation logics that are poorly suited to social impact questions. What is being asked for in reporting shapes what gets evaluated.

Speaker: Kalika Bali *(Microsoft Research India)*

- ➔ The funding asymmetry is extreme. The ratio of funding available to build and deploy AI solutions versus funding available for their evaluation is not comparable. This structural imbalance is a constraint the evaluation community needs to name, advocate against, and seek to redress.

Audience Q&A

Audience Question 1

Can you share a theory of change for the Asha Saheli chatbot in terms of outcomes? Has having a theory of change been helpful, or have you had to pivot it as an AI solution evolves?

Speaker: Urvashi Wattal *(Khushi Baby)*

- ➔ Theories of change are valuable for any programme because they articulate objectives and surface testable assumptions. For Asha Saheli, a key assumption was that ASHA workers would trust the chatbot's responses consistently. In practice,

during onboarding, many ASHAs assumed a Khushi Baby staff member was responding, not an AI system. This created functional trust but raised questions about informed understanding of the technology.

- ➔ AI solutions require unusually careful attention to the assumptions embedded in the theory of change, particularly around how a new technology is understood and appropriated by users whose prior experience of the world may be very different from that of designers.

Audience Question 2

What does the training data of mental health chatbots not have, and has any gender audit been done of them?

Speaker: Kalika Bali *(Microsoft Research India)*

- ➔ Little systematic evaluation or gender audit of mental health chatbots exists. This is a genuinely concerning space, particularly at the intersection of AI and marginalised communities where emotional and psychological stakes are high.

Speaker: Aarushi Gupta *(Digital Futures Lab)*

- ➔ The question cannot be reduced to training data adequacy. The prior question is whether a conversational agent is the appropriate modality for mental health support. Models can be contextually fluent; having scraped social media and public forums, they can recognise distress, but the output, however well-crafted, cannot substitute for substantive mental health care infrastructure. In India, where mental health services are acutely under-resourced and young people are struggling, the solutionism trap is especially dangerous in this domain.

Audience Question 3

Models, once released, cannot be course-corrected the way proprietary models can. How should they be approached from a governance and evaluation perspective?

Speaker: Aarushi Gupta *(Digital Futures Lab)*

- ➔ Open-weight models afford local organisations and civil society the ability to fine-tune their own sociocultural data and define their own safety parameters. This pushes agency downstream, toward actors who have previously only inherited the assumptions and risks embedded in foundation models. From a contextual relevance and feminist evaluation standpoint, this democratisation has significant potential.

Speaker: Kalika Bali *(Microsoft Research India)*

- ➔ The countervailing risk is the absence of accountability. Proprietary developers have reputational and commercial incentives to address harms. Open-weight downstream actors do not. Regulation cannot keep pace with proliferation; governance frameworks lag structurally behind capability. The trade-off between democratisation and accountability is real and has no easy resolution.

6. Conclusions

- ➔ Gender-intentional evaluation of AI must be embedded across all four stages – model, product, user, impact – from the point of model development, by interdisciplinary teams that include feminist evaluators. And these four stages will often have to move concurrently as the model learns and pivots after deployment.
- ➔ No single evaluation methodology gives the full picture. Composite, concurrent evaluation strategies are required.
- ➔ The question 'is AI the right solution for this problem?' must precede deployment decisions. Evaluators should be engaged at the ideation stage before objectives and architectures harden.
- ➔ Solution maturity matters. Evaluation questions, methods, and indicators differ substantially depending on whether a tool is at the model validation, product piloting, or scale-up stage.
- ➔ Organisational AI readiness and model validity are conceptually distinct but routinely conflated, with distorting effects on what gets measured.
- ➔ Benchmarks are necessary but insufficient. The sector needs a shift toward shared evaluation values that can be localised into context-specific indicators.
- ➔ The funding asymmetry between AI deployment and AI evaluation is structural. The evaluation community needs to name and advocate against it.
- ➔ Evaluators must move from external judges to embedded partners, co-defining success criteria with practitioners and communities from the outset.

7. About the Organiser

The Gender and Digital hub (GxD hub) is housed at LEAD, the action-orientated research centre at Krea University–IFMR. Strategically situated within a dynamic academic ecosystem, the GxD hub draws on cross-disciplinary expertise to co-create gender-intentional digital solutions and policies, bridging research, practice, and evaluation across India's development sector.