Beyond the cloud: Unlocking edge AI for LMICs – quantitative study

Term of Reference

1. Introduction

The GSMA is a global organisation unifying the mobile ecosystem to discover, develop and deliver innovation foundational to positive business environments and societal change. Our vision is to unlock the full power of connectivity so that people, industry, and society thrive. Find more at gsma.com.

The GSMA Mobile for Development (M4D) foundation operates at the intersection of the mobile ecosystem and the development sector. Our aim is to stimulate digital innovation and deliver both sustainable business and large-scale socio-economic impact. Our research and insights platform, in-market expertise and community of partners push forward digital innovations and implementations that empower underserved populations. Find out more at gsma.com/solutionsand-impact/connectivity-for-good/mobile-for-development/.

The Central Insights Unit (CIU) sits at the core of GSMA Mobile for Development (M4D) and produces in-depth research on the role and impact of mobile and digital technologies in advancing sustainable and inclusive development. The CIU engages with public and private sector practitioners to generate unique insights and analysis on emerging innovations in technology for development. Through our insights, we support international donors to build expertise and capacity as they seek to implement digitisation initiatives in low- and middle-income countries through partnerships within the digital ecosystem

2. Background, scope and objectives

Background

The global AI ecosystem today is largely built on centralised, cloud-based infrastructure. Most LMICs face structural barriers to using that infrastructure at scale including costs, latency, foreign exchange access, data sovereignty and unstable power. Edge AI – running inference on or near where the data is generated – has become technically more viable due to rapidly improving smartphone chips, on-device microcontrollers, and model-efficiency techniques. In Africa in particular, mobile is the dominant digital channel, and smartphones projections are projected to keep rising, widening the base for on-device capability.

Despite the growing emergence of edge AI, there is still limited evidence of the potential and viability of edge AI in resource-constrained environments, where economic, technical and market conditions differ significantly from high-income contexts. Existing studies, including recent GSMA research, have examined the demands of model training, dominant compute pathways, and large-scale cloud infrastructure, but have not assessed the market readiness or operational feasibility of edge AI in these settings. At the same time, an increasing number of innovators are developing edge AI solutions to enable scalable AI applications. However, significant gaps remain in understanding the technology needs, economic feasibility, operational models, priority use cases, and policy frameworks required to scale edge AI sustainably in LMICs.

Scope

The GSMA is undertaking a mixed-methods study to address evidence gaps on the feasibility, economics, and enabling conditions for edge AI in LMICs. The research aims to quantify market readiness and unit economics, document real-world deployments, and generate actionable guidance for industry, investors, and policymakers. Findings will feed into a public GSMA report.

This ToR covers the quantitative workstream. The GSMA is seeking a supplier to generate robust, policy and investment-grade estimates of the current and near-term (3–5 year) availability, affordability, and capability of edge-AI devices and enabling infrastructure in LMICs. The work should quantify the economics of edge vs. cloud (e.g., TCO/unit economics), project market readiness under plausible scenarios, and identify readiness gaps by market segment. The quantitative analysis will complement the qualitative workstream by grounding case studies and interview insights in numerical estimates.

For the purposes of this study, edge AI refers to the deployment of AI models on or near the device where data is generated. This includes smartphones (across capability tiers), embedded and IoT devices (e.g. sensors, single-board computers, microcontrollers), as well as optimisation techniques (e.g. TinyML, quantisation, model compression) that enable AI inference on constrained devices. The study will consider this full spectrum but place particular emphasis on smartphones and mobile devices, given their central role in LMIC digital ecosystems.

The primary audience for this project includes industry, i.e. startups, MNOs, device manufacturers, infrastructure providers, and sources of capital, i.e. VCs, DFIs, corporate investors. The secondary audience includes governments and policy makers, donor organisations and academia and training providers.

Objectives

The quantitative stream of the study aims to:

- Quantify the economic case for edge AI, including cost savings, reduced cloud reliance, productivity gains
- Project market readiness over the next 3-5 years: device penetration, affordability and pricing, processing capability by device class and tier (e.g. smartphone tiers, IoT, SBCs, MCUs where feasible)
- Assess the long-run feasibility of edge inference as higher-demand models (e.g. reasoning) spread – estimating compute/latency/power envelopes and the role of optimisation (compression, quantisation, TinyML)
- Map enabling conditions quantitatively: mobile broadband coverage/latency, power reliability, and availability of edge facilities (e.g. micro-data centres, cell-site compute)
- Translate findings into practical recommendations for industry and policy.

Research questions

It will seek to answer the following questions:

Economic feasibility	- What are the per-device and per-deployment costs and returns of edge AI solutions across different sectors, and how do these economics vary by device type?

	- What is the overall economic case for edge AI in LMICs, including cost savings, reduced cloud dependency, productivity gains, and potential for new market creation?
Technical feasibility	- How feasible is edge AI for higher-demand models (e.g., reasoning models), and what optimisation techniques (quantisation, compression, TinyML) are most promising? - How do different categories of AI models (e.g., classification, generative, reasoning) vary in their suitability for on-device inference, and what trade-offs exist in latency, memory, and compute? - How might advances in hardware, optimisation (e.g., TinyML, model compression, quantisation), and distributed approaches (e.g., federated learning) improve the feasibility of edge AI in LMICs?
Market readiness	- What types of edge AI-capable devices and supporting software ecosystems (e.g., frameworks, tools, optimisation libraries) are currently accessible in LMICs, and how is this landscape likely to evolve over the next 3–5 years?
Impact and lessons	- What evidence exists of measurable social or economic impact from early edge AI deployments, and what lessons can be drawn for scaling?

Geography

This research project will focus on primarily on Africa, due to the increasing need for alternative compute approaches in the region and comparatively lower connectivity environments, but also consider other regions (e.g. South/east Asia, Latin America). The quantitative component will provide a regional outlook with – subject to data availability – country level breakdowns for selected priority markets (e.g. the top 5 projected smartphone markets in Sub-Saharan Africa). All results should, where feasible, be disaggregated by gender to reflect device affordability/ownership gaps.

3. Anticipated approach and delivery timeline

We are particularly interested in modelling approaches that could deliver outputs such as:

1. Cost comparison (edge vs. cloud): A simple model that shows under what conditions running AI locally (on-device/edge) becomes cheaper or more reliable than cloud, factoring in device cost, data use, power, and cloud API fees.
2. Adoption scenarios: Projections of how many users could realistically access edge AI over the next 3–5 years, segmented by device class (basic/advanced smartphones, IoT), affordability, and gender.
3. Inclusion impact estimates: Quantitative illustrations of how cost reductions or lower latency could expand adoption among specific groups (e.g. women, rural users, low-income households).

These are indicative examples. We welcome suppliers' own suggestions on the most appropriate methods and tools to meet the study's objectives.

The table below provides a suggested high-level approach and delivery timeline.

Phase	Activities and deliverables	Timeline (indicative)
Phase 1: Inception	**Key activities** - Kick off call to discuss methodology, scope, timelines and methodological approach - Rapid review of literature on edge AI feasibility/economics in LMICs - Data landscape mapping - Agree country sample, device taxonomy/tiers, anchor use cases for economics (edge vs cloud)	Mid-October
	Deliverables - Inception deck summarising state of knowledge, gaps, and modelling implications. - Data inventory and access plan (sources, coverage, quality notes).	
Phase 2: Modelling and scenarios	**Key activities** - Design the modelling method, propose country sample and device categories - Build scenarios for penetration, device capability and prices, applying gender lens where possible - Simple edge vs cloud TCO comparison for 2–3 anchor use cases, with light sensitivity checks (data, power, FX). - Enablers snapshot by market: coverage/latency, power reliability, local edge facilities.	November-December
	Deliverables - Model and data pack - Interim report/slide deck with preliminary findings, scenarios, and sensitivities	
Phase 3: Synthesis and final report	**Key activities** - Finalise numbers, QA and documentation; incorporate GSMA feedback. - Produce figures/tables for integration in GSMA's public report. - Translate findings into practical recommendations for industry and policy.	January
	Deliverables - Final report (slide deck) with findings, implications and recommendations; detailed methodology note - Handover pack with clean datasets, model files, data dictionary	

4. Supplier requirements

Key requirements

The GSMA is searching for a partner to deliver analysis responding to the outlined objectives. Ideally, they will have:

Essential:

- Proven experience in quantitative modelling for emerging markets (market sizing, scenarios, TCO/unit economics, sensitivity analysis).
- Strong familiarity with the AI/edge ecosystem, devices, and infrastructure.
- Understanding of LMIC contexts (affordability, FX/tariffs, coverage/latency, power), ideally with Africa experience.
- Capability in device and network analytics (smartphone tiers, chipsets/NPU/GPU, RAM/flash, network KPIs).

Desirable:

- Experience working with MNOs, device makers, or big tech.
- Track record with donors/DFIs and policy audiences; ability to translate technical findings into investment/policy implications.
- Experience embedding gender-aware affordability and inclusion analysis.

GSMA requires the appointed supplier to be fully transparent about subcontractors they intend to use and GSMA has the power to veto selection. The supplier is expected to comply with GDPR in data collection and processing. The supplier will need to have and adhere to research approvals, as required. They are also expected to establish and review assessment risks, challenges and limitations and recommend how these will be managed. This should include:

- Methodology limitations
- Insufficient capacity/availability/interest of chosen scope
- Reputational risk for the GSMA (in the event of damaging findings)

Proposals should include a technical and financial proposal:

Technical proposal

1. A short (1 page) statement of suitability, highlighting recent relevant experience.
2. A short (2-4 page) discussion of the proposed approach including: the analytical frameworks to be used, identified data sources, and initial proposals on case studies.
3. Any proposed changes to the ToR.
4. Details of relevant firm project experience.
5. Gantt chart outlining major project stages and timelines
6. CVs, and location of team members.

Financial proposal

1. Level of effort (person-day) by activity.
2. Fee rates (per day in GBP).
3. Total project cost (GBP), without VAT[1].
4. The Respondent's Total Price is inclusive of all costs, insurances, fees, costs, expenses, liabilities, obligations, risks, and all financial requirements for the performance of Services and provision of Deliverables.

5. Any charge not stated in this Proposal, which extends above to the Total Price, is not permitted.

Due to GSMA compliance requirements, exact project budgets cannot be provided at this stage. You are, however, able to provide a few implementation/budget options that can help us assess value for money and we can align our project scope to the relevant budget after a consultant has been selected.

Proposal assessment and selection process

The proposal will be scored on the following set of criteria:

Criteria	**Importance**	**Weighting**
Cost	Proposal's value for money	20%
Quality	Quality of the research approach outlined in the proposal, including degree to which it addresses the outlined research questions and proposal elements	35%
Bidder's capacity to manage the project on time and on budget	Selection of experienced high-quality research partner(s) and ability to manage the project on time and on budget	30%
Relevant experience	Bidder's experience in successfully conducting similar projects	15%

- Proposals are to be submitted no later than **17:30 BST, Wednesday 24th September 2025** for this work to Daisy Macaskie (dmacaskie@gsma.com) and Eugénie Humeau (ehumeau@gsma.com).
- Clarification questions can be sent to Daisy Macaskie (dmacaskie@gsma.com) and Eugénie Humeau (ehumeau@gsma.com).
- Shortlisted consultants may be contacted for an interview **w/c 29th September**.