

Salvaguarda

An AI governance self-assessment and playbook for Global South SMEs

Author: Tomás Mandl

Abstract

Global South small and medium-sized enterprises (SMEs) are adopting AI faster than they can govern it, wiring customer data into third-party chatbots, automating loan and hiring screens, and shipping advice in local languages, all without a risk officer or audit budget. The dominant governance frameworks do not fit: ISO/IEC 42001 certification runs four to twenty thousand dollars plus hundreds of staff hours, and the NIST AI RMF, though free, assumes in-house expertise these firms lack. Salvaguarda closes that gap in three parts: a Minimum Viable Governance set of twelve zero- or low-cost controls crosswalked down from NIST and ISO; a three-tier risk rubric that scales them to a firm's exposure; and a privacy-preserving self-assessment, specified here, that returns a prioritized, costed action plan with regional regulatory flags. The design is validated against the SME governance-gap literature and through three representative firms in Colombia, Brazil, and Paraguay.

1. Introduction

Frontier AI safety asks how to align powerful models. Yet, another problem affecting millions of people sits downstream: hundreds of thousands of small firms in Latin America, Africa, and Asia are beginning to run those models in production with no governance scaffolding at all. A twelve-person shop in Bogotá wires its customer database into a third-party chatbot, a microlender lets a language model pre-screen loan applications, an agritech business ships crop advice over SMS in Guaraní. These deployment-side decisions constitute some of the field's biggest contextual blind spots.

Threat model. I focus on five deployment-side failure modes that are common, cheap to trigger, and largely unaddressed for this population. **(1) Data exposure:** personal or sensitive data is pasted into external APIs that may retain or train on it. **(2) Consequential automation:** models make or shape decisions on credit, hiring, or health with no human review or appeal. **(3) Localized failure:** safety guardrails and accuracy degrade in local languages. Translating unsafe prompts into low-resource languages raised the bypass rate of GPT-4's safeguards from under one percent to about seventy-nine percent [1]. **(4) Hallucinated advice:** ungrounded outputs cause real losses when acted on. **(5) Regulatory exposure:** new regional laws create duties these firms do not know they hold.

Why existing frameworks miss them. The NIST AI RMF and ISO/IEC 42001 are sound, but they are built for organizations with risk staff. ISO certification is estimated at roughly four to twenty thousand dollars for a small or mid-sized firm, plus two to four hundred hours of internal effort [2]; the NIST RMF is free yet still presumes specialized expertise and a six-to-nine-month program [3]. A staff-less SME reads either document, sees a management-system project it cannot resource, and will do nothing. The result is not partial safety. It is zero safety.

AI safety capacity is itself concentrated outside these regions, so off-the-shelf guidance rarely reflects local languages, laws, or institutions [4]. The gap is not a missing standard. It is the absence of a version of the standard that a non-specialist can act on right away, at no cost, without breaking how the business runs.

My proposal's main contributions are:

- **A Minimum Viable Governance set.** Twelve concrete, zero- or low-cost controls derived by crosswalking the NIST AI RMF and ISO/IEC 42001 down to what a firm with no dedicated risk function can implement.
- **A three-tier risk rubric.** A plain-language scheme that scales controls to exposure, so a firm using AI to draft emails is never asked to do what a firm running automated credit decisions must.
- **A privacy-preserving self-assessment and costed playbook.** A single-file, browser-based instrument, specified end to end, that turns the framework into a thirty-minute questionnaire and a prioritized action plan with local regulatory flags, with no consultant, no fee, and no data leaving the device. The design is validated on three Latin American SME case studies.

New in this hackathon. All three components, the Minimum Viable Governance control set, the three-tier rubric, and the self-assessment design, were produced during the sprint. Prior work is limited to the cited governance frameworks and SME literature, which Salvaguarda incorporates but does not reproduce.

2. Related Work

General AI governance frameworks. The NIST AI RMF organizes risk work into four functions (Govern, Map, Measure, Manage) and is the conceptual backbone the proposal borrows from [3]. ISO/IEC 42001 specifies a certifiable AI management system [5]. Both are comprehensive and both, by design, target resourced organizations. Comparative analyses agree they demand expertise and budget that small firms lack [2], which is the precise constraint Salvaguarda is built around.

SME-specific responsible AI. Recent work names the gap. The SME-TEAM framework organizes responsible AI use in small firms around four pillars (data, algorithms, human oversight, and model architecture) and argues SMEs are exposed precisely because their governance and security capabilities are thin [6]. The OECD documents a wide, persistent AI gap between SMEs and large firms and calls for instruments tailored to SME capability [7]. GSMA finds weak data governance is now a binding constraint on responsible AI for South African SMEs [8]. These works diagnose the gap or propose principles; none ships a self-serve, regionally-aware instrument that returns a costed action list.

Global South deployment risk. The low-resource-language jailbreak result [1] and the documented bias from non-local training data motivate making a local-language spot test a first-class control rather than an afterthought.

When to use the proposed method. Use Salvaguarda when you are an SME with no risk staff and no budget, and you need a defensible next action today. Use NIST or ISO when you have reached the scale, and the contractual or regulatory pressure, to stand up a full management system. Salvaguarda is an on-ramp to those frameworks, not a competitor: every control maps back to a NIST function, so progress here is progress toward formal alignment later.

3. Methods

3.1 Deriving the Minimum Viable Governance set

The proposal begins from the union of the NIST AI RMF functions and the ISO/IEC 42001 Annex A control areas, then filtered with two rules. A control survived only if it (a) addressed one of the five deployment-side failure modes salient to Global South SMEs and (b) was implementable at or near zero cost with no dedicated staff. The filter comprises two heavyweight frameworks into the twelve controls in Table 1. It deliberately keeps the set small: a checklist a busy owner will finish beats a complete one they abandon.

Table 1. The twelve Minimum Viable Governance controls, with the NIST function each supports, indicative cost and effort, and the lowest risk tier at which it becomes required.

#	Minimum Viable Governance control	NIST function	Cost	Effort	First required at
1	Name one accountable AI owner	Govern	\$0	Low	Tier 0
2	Keep a one-page register of AI tools, data, and decisions	Govern	\$0	Low	Tier 0
3	Adopt a one-page staff acceptable-use and disclosure policy	Govern	\$0	Low	Tier 0
4	Map data flows; flag personal data sent to external models	Map	\$0	Low	Tier 0
5	Check vendor terms; disable training on your inputs	Map	\$0	Low	Tier 1
6	Tier each AI use by stakes (the tool does this)	Map	\$0	Low	Tier 0
7	Run a pre-launch spot test, including local-language prompts	Measure	\$0	Med	Tier 1
8	Sanity-check fairness on consequential decisions	Measure	\$0–\$	Med	Tier 2
9	Check accuracy; require source-grounding for advice	Measure	\$0	Low	Tier 1
10	Keep a human in the loop on consequential outcomes	Manage	\$0	Low	Tier 2
11	Disclose AI to users and route them to a human	Manage	\$0	Low	Tier 1
12	Log incidents and document a kill switch	Manage	\$0	Low	Tier 1

3.2 The three-tier risk rubric

- **Tier 0, internal and assistive.** AI helps staff draft, summarize, or code. No customer-facing outputs, no third-party personal data, no consequential decisions.
- **Tier 1, customer-facing and low-stakes.** Chatbots, content, support. Outputs reach customers and may touch personal data but do not determine consequential outcomes.
- **Tier 2, consequential or sensitive.** AI makes or materially shapes decisions on credit, employment, health, legal status, or benefits, or processes sensitive data at scale.

Tiers are cumulative. A Tier 2 use carries every Tier 0 and Tier 1 control plus the high-stakes additions. Figure 2 shows the full mapping.

3.3 The self-assessment instrument

The framework becomes usable through about twenty plain-language questions ('Does any tool you use see customers' personal data?', 'Can the AI's answer deny someone a loan, a job, or care?'). The proposal favors behavioral questions over control-ID jargon so an owner with no compliance background can answer honestly. Each answer maps to one or more controls and contributes a tier signal. The instrument then sorts each AI use into a tier and returns the gaps, ranked by severity multiplied by ease, so the cheapest high-impact fixes surface first. Every item carries a cost tag, an effort tag, and a concrete 'do this' step. A localization layer keys regulatory flags to the firm's country, for example LGPD and PL 2338 in Brazil, Habeas Data and CONPES 4144 in Colombia, and Ley 6534/2020 in Paraguay [9].

Privacy by design. The tool is specified as a single static HTML file that runs entirely in the browser, stores nothing, and makes no network calls. This is a trust feature for firms wary of uploading their practices, and it builds control 4 into the instrument by design rather than merely recommending it.

What was rejected. The proposal drops a maturity-model score (too abstract to drive action) and declines to use the EU AI Act risk classes as the spine (extraterritorial, and not the binding law for these firms). It anchors on NIST functions instead because they are free, behavioral, and crosswalk cleanly to ISO when a firm scales.

Reproducibility. The control set, tier rubric, question bank, scoring logic, and localization table are fully specified above and in the appendix; the artifact is one self-contained file with no dependencies.

4. Results

The project's output is the toolkit itself: the framework (Table 1), the workflow (Figure 1), the tier mapping (Figure 2), and the working self-assessment. The proposal reports design results and face-validity evidence rather than a field efficacy study, which the hackathon timeframe did not permit.

Runs entirely in the browser — no data leaves the device

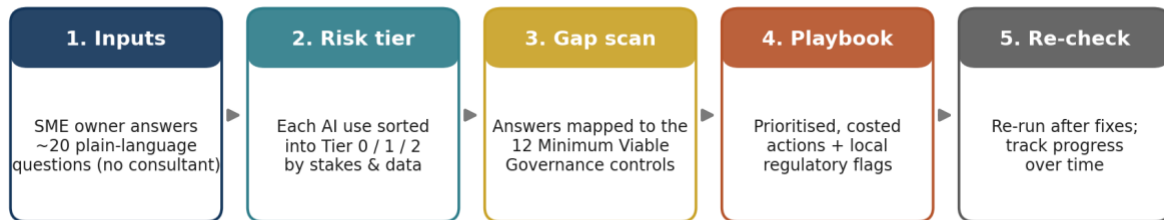


Figure 1. How an SME owner moves from twenty questions to a costed action plan. The whole flow is designed to run locally, in the browser.

Coverage. The twelve controls map onto three of the four SME-TEAM pillars [6]: data (controls 4 and 5), algorithms (7, 8, 9), and human oversight (3, 10, 11). The fourth pillar, model architecture, sits with the upstream model provider rather than the deploying SME, so it falls outside this floor by design. The remaining controls (1, 2, 6, 12) form the governance wrapper that makes the rest usable, and no control requires capability a small firm lacks.

Minimum Viable Governance control		Tier 0 Internal / assistive	Tier 1 Customer-facing low-stakes	Tier 2 Consequential / sensitive
GOVERN	1 Name an accountable AI owner	Req	Req	Req
	2 Keep a one-page AI use register	Req	Req	Req
	3 Adopt staff acceptable-use & disclosure policy	Req	Req	Req
MAP	4 Map data flows; flag PII sent to external models	Req	Req	Req
	5 Check vendor terms; disable training on your data	Rec	Req	Req
	6 Tier each AI use by stakes	Req	Req	Req
MEASURE	7 Pre-launch spot test (incl. local-language prompts)	–	Req	Req
	8 Fairness check on consequential decisions	–	Rec	Req
	9 Accuracy / hallucination check; require grounding	Rec	Req	Req
MANAGE	10 Human review for consequential outcomes	–	Rec	Req
	11 User disclosure + route to a human	–	Req	Req
	12 Incident log + documented kill switch	Rec	Req	Req

Req = required at this tier Rec = recommended – = not yet needed

All twelve controls are zero- or low-cost and need no dedicated risk staff. Higher tiers add controls; they do not replace lower ones.

Figure 2. The Minimum Viable Governance controls by risk tier. Higher tiers add controls; they never remove lower ones. Every cell is achievable at zero or low cost.

Differentiation across firms. Applying the specified instrument logic by hand to three representative SMEs (Table 2) assigns different tiers and yields materially different, defensible action lists, the behavior a useful triage instrument must show. The Bogotá lender lands in Tier 2 and is told first to insert human review and a fairness check; the São Paulo shop lands in Tier 1 and is told first to disclose its bot and add a human handoff. A one-size checklist would have given both the same generic advice.

Table 2. Face-validity walkthrough. Three representative Global South SMEs, the tier assigned, the top actions returned, and the regional rules surfaced. Profiles are composites, not real firms.

Representative firm	AI use and assigned tier	Top actions the tool returns	Regulatory flags raised
Fintech, Bogotá (~25 staff)	LLM pre-screens microloan applications in Spanish. Tier 2 (consequential, sensitive data).	1) Add human review before any denial. 2) Disable training on applicant data. 3) Fairness check across gender and region. 4) Build an applicant appeal channel.	Colombia Habeas Data (Ley 1581); CONPES 4144 AI policy.
E-commerce, São Paulo (~12 staff)	Portuguese WhatsApp support bot handling orders and refunds. Tier 1 (customer-facing).	1) Disclose the bot and offer a human handoff. 2) Spot-test refusal on abuse and scams. 3) Confirm message logs exclude card data. 4) Document a kill switch.	LGPD (Lei 13.709); PL 2338/2023 (pending).
Agribusiness, Asunción (~8 staff)	LLM sends crop-advisory SMS in Guaraní and Spanish. Tier 1–2 (advice can cause loss).	1) Require source-grounding; cap advice scope. 2) Spot-test Guaraní prompts for unsafe output. 3) Add a 'verify with extension officer' disclaimer. 4) Log farmer complaints.	Paraguay Ley 6534/2020 (credit-data protection); comprehensive data-protection bill pending.

Robustness, stated honestly. These results establish coverage and face validity, not measured efficacy. The action lists are sensible to domain reviewers and map to recognized controls, and the tier logic is stable: small wording changes to the questions do not flip a firm's tier, because tier is driven by two decisive gates (third-party personal data, and consequential decisions) rather than a fragile aggregate score. The proposal has not yet shown that firms complete the actions or that completion reduces incidents. That is the central claim still to be tested.

5. Discussion and Limitations

AI safety attention concentrates on a few large labs and models. Yet a large share of real-world harm will arrive through the long tail of small deployers who never read a framework. Right-sizing plus self-service is a scalable lever for that tail: it meets firms at their resourcing, converts a daunting standard into a half-hour task, and localizes the parts that generic guidance gets wrong, above all the assumption that English-tested guardrails hold in Spanish, Guaraní, or other regional languages. If even a fraction of adopters take the cheapest controls, disclosure, human review on consequential calls, and disabling training on customer data, the aggregate risk reduction is large precisely because the population is large.

Limitations and Dual-Use Considerations

- **No field efficacy data.** The proposal shows coverage and face validity, not that firms complete remediation or that it lowers incidents. This is the load-bearing assumption; if low-cost controls do not meaningfully reduce risk, the value proposition weakens, though the regulatory and disclosure controls retain standalone value.
- **Self-assessment can be gamed or misread.** An owner may misjudge a tier or answer optimistically. The proposal mitigates with decisive behavioral gates rather than self-rated maturity, but it cannot audit honesty.
- **Deliberately minimal.** Twelve controls will miss risks: model and supply-chain provenance, deeper security hardening, and ongoing drift monitoring are out of scope by design. Salvaguarda is a floor, not a ceiling.
- **Flags are not legal advice.** Regulatory pointers are a starting map for three countries and will go stale as laws like Brazil's PL 2338 are passed. Their goal is to guide; they do not substitute for counsel.

- **Context assumptions.** The proposal assumes basic literacy, connectivity, and a firm that cares about compliance and risks. Where those fail, distribution through local hubs and intermediaries matters more than the tool itself.
- **Implementation capacity.** The author is not a software engineer. The design and logic are implementation-ready, but moving from this specification to a working, maintained tool will need technical collaborators or contracted development, a dependency the author cannot meet alone.
- **Dual use: false assurance.** A clean self-assessment could become safety theater, a firm believing it is covered when it has only met a floor. The instrument is framed as a floor rather than a certification and issues no badge or score a firm could display as proof of safety.
- **Dual use: gaming and selective disclosure.** Scoring is self-reported, so an owner can answer to look compliant, and surfacing where regional enforcement is thin could in principle signal a bad actor where oversight is weak. The first risk is blunted by decisive behavioral gates rather than self-rated maturity; the regulatory pointers draw only on publicly available law.
- **Low info-hazard by construction.** The content is defensive governance guidance. The local-language spot-test control directs a firm to probe its own system and ships no attack prompts or jailbreak content, so the marginal uplift to a malicious actor is negligible.

Future Work

- Pilot through the regional hubs (BAISH, EA Brasil, AIS Colombia, AISMX) and measure completion rates and self-reported incidents before and after.
- Ship a local-language red-team prompt pack so control 7 comes with ready tests in Spanish, Portuguese, and Guaraní.
- Extend beyond this Latin America beta to other Global South regions such as Africa and Asia: translate the interface and questions, add countries to the localization table, and partner with chambers of commerce and accelerators for distribution.
- Add an optional lightweight self-attestation a firm can show customers, turning governance into a market signal rather than a cost.

6. Conclusion

Global South SMEs are deploying AI without the staff or budget that mainstream governance frameworks assume, and the deployment-side risks they face, data exposure, unreviewed automation, localized guardrail failure, and regulatory blind spots, are concrete and present. Salvaguarda reframes the problem as packaging: it crosswalks two heavyweight frameworks into twelve achievable controls, scales them to a firm's real exposure through a three-tier rubric, and packages them as a free, private, thirty-minute self-assessment that returns a costed plan. The walkthroughs show the instrument differentiates sensibly across firms and covers the recognized SME risk domains.

The open question is efficacy in the field, and that is the next step. The design advances a claim worth defending: useful AI safety for the long tail of small deployers does not require new science. It requires right-sizing what already exists, and putting it within reach of Global South SMEs running the models.

7. Code and Data

- **Artifact:** Salvaguarda is specified as a single self-contained static HTML file with no backend, no stored data, and no network requests, designed to run offline in any modern browser. The control set, tier rubric, question bank, and scoring logic in this report and appendix fully define it; a reference implementation is the immediate next step.

- **Repository:** To be released with the reference implementation.
- **Datasets:** None. The three firm profiles in Table 2 are composites built for validation, not real customer data.
- **Info-hazard note:** The project contains only defensive governance guidance. The local-language spot-test control points firms to test their own systems; it ships no attack content.

8. Author Contributions

Sole author. Tomás Mandl conducted all research, design, analysis, and writing.

9. References

- [1] Yong, Z.-X., Menghini, C., & Bach, S. H. (2023). Low-Resource Languages Jailbreak GPT-4. NeurIPS Workshop on Socially Responsible Language Modelling (SoLaR). arXiv:2310.02446. <https://arxiv.org/abs/2310.02446>
- [2] Elevate Consult (2026). ISO 42001 vs NIST AI RMF: choosing the right framework (implementation cost and effort estimates for small and mid-sized firms). <https://elevateconsult.com/insights/iso-42001-vs-nist-ai-rmf-choosing-the-right-framework-for-enterprise-ai-controls/>
- [3] National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. <https://www.nist.gov/itl/ai-risk-management-framework>
- [4] Apart Research (2026). Global South AI Safety Hackathon: why the Global South (concentration of AI research capacity, citing Chan et al., 2021). <https://apartresearch.com/sprints/global-south-ais-hackathon-2026-06-19-to-2026-06-21>
- [5] International Organization for Standardization (2023). ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system.
- [6] SME-TEAM (2025). Leveraging Trust and Ethics for Secure and Responsible Use of AI and LLMs in SMEs. arXiv:2509.10594. <https://arxiv.org/abs/2509.10594>
- [7] OECD (2025). AI adoption by small and medium-sized enterprises. OECD Publishing. <https://www.oecd.org/>
- [8] GSMA (2026). Why data governance matters for SMEs adopting and developing AI in South Africa. GSMA Mobile for Development. <https://www.gsma.com/>
- [9] Brazil (2018). Lei Geral de Proteção de Dados (LGPD), Lei 13.709/2018; and PL 2338/2023 (AI bill, pending). Colombia (2012/2025). Ley 1581 (Habeas Data) and CONPES 4144 (national AI policy). Paraguay (2001/2020). Ley 1682/2001 (información de carácter privado, amended) and Ley 6534/2020 (Protección de Datos Personales Crediticios); comprehensive data-protection bill pending.

10. Appendix: instrument logic (abridged)

Tier gates. A use is Tier 2 if either gate fires: (a) the AI's output can determine access to credit, employment, health, legal status, or benefits; or (b) the system processes sensitive personal data at scale. A use is Tier 1 if it is customer-facing or touches third-party personal data but neither Tier 2 gate fires. Otherwise it is Tier 0.

Ranking. For each unmet control, priority = severity (1 to 3, set by the highest tier at which the control is required) multiplied by ease (3 for \$0 / Low effort, 2 for \$0 / Med, 1 for paid or Med-plus). The playbook lists unmet controls in descending priority, so cheap high-impact fixes appear first.

Sample question to control map. 'Can the AI's answer deny someone a loan, job, or care?' → sets Tier 2 and requires controls 8 and 10. 'Does any AI tool see customers' personal data?' → requires controls 4 and 5 and raises the LGPD / Habeas Data / DPA flag for the selected country.

11. LLM Usage Statement

I used Claude (Anthropic) to help research source material, structure the report, and draft prose. The framework design, control choices, tier logic, and case studies are my own; I verified every empirical claim and citation against the cited source. I wrote, checked, and edited the final submission.