GenderSight: An API-First, Privacy-Preserving Multimodal System for TFGBV Narrative Intelligence, Support Routing for Women and Girls, and Women's Innovation Enablement

Komborero Victor Kangai[1], Tinotenda Chrispen Makoni[2], Xinyu Fan[3], *Britney Gonzo*[4]
[1]**The New York Academy of Sciences (NYAS)**
115 Broadway, 8th Floor, New York, NY 10006, USA
Email: kangaikomborero@gmail.com
[2]**University of Technology, Mauritius (UTM)**
Ave De La Concorde, La Tour Koenig, Pointe-aux-Sables, Port Louis, Mauritius
Email: makonichrisspen30@gmail.com,
[3]***Inner Mongolia University of Technology (IMUT)***
No. 49 Aimin Street, Xincheng District, Hohhot, Inner Mongolia, 010051, P.R. China
Email: xinyufan1115@gmail.com
[4]***University of Technology, Mauritius (UTM)***
Ave De La Concorde, La Tour Koenig, Pointe-aux-Sables, Port Louis, Mauritius
Email: Gonzobritney@gmail.com,

Mentorship and academic advisement: Prof. Attlee Munyaradzi Gamundani

Namibia University of Science and Technology (NUST)

13 Jackson Kaujeua Street, Private Bag 13388, Windhoek, Namibia

Email: amgamundani@gmail.com

Abstract

Technology-facilitated gender-based violence (TFGBV) and structural gender inequities are amplified by contemporary digital platforms, while support discovery, economic empowerment pathways, and institutional accountability mechanisms remain fragmented and unevenly accessible. [3] This paper presents GenderSight, an API-first, privacy-preserving multimodal (text + image) system deployed on secure cloud infrastructure that integrates three functions aligned to Design Equality competition categories: (a) population-level narrative intelligence that enables stakeholders to advocate and advance gender equality through aggregated trend indicators, coarse regional heatmaps, and policy-ready analytics suitable for program design and monitoring; (b) an opt-in support assistant that meets the needs of women and girls by routing users to verified services and opportunities using coarse geolocation with geo-obfuscation to reduce safety risk; and (c) Virtual Green Rooms—moderated, SDG-oriented collaboration spaces that promote women as innovators and entrepreneurs by enabling mentorship matching, project formation, partnership matchmaking, and access to capacity-building resources. [1–6] The system is governed under privacy-by-design and AI risk management controls, including k-anonymity thresholds, optional differential privacy for repeated statistical queries, audit logging, and human-in-the-loop escalation for high-risk or low-confidence outputs, ensuring that institutional insights remain actionable without enabling individual surveillance or re-identification. [11–14] The initial deployment targets Zimbabwe to leverage existing GBV strategy alignment and referral ecosystems, with a scale pathway to Eastern and Southern Africa. [3–6]

Keywords: SDG 5; gender equality; TFGBV; multimodal machine learning; privacy-preserving analytics; k-anonymity; differential privacy; NGO dashboards; support routing; women's entrepreneurship enablement.

1. Introduction

Digital platforms have become critical infrastructure for social participation, economic exchange, and civic discourse. In parallel, these platforms have also become high-throughput channels through which gendered harms are produced, normalized, and scaled. Technology-facilitated gender-based violence (TFGBV) manifests across multiple modalities—text, images, screenshots, memes, and coordinated campaigns—often exploiting the speed of content diffusion, asymmetries in moderation capacity, and the practical difficulty of attributing intent and harm in context-rich interactions. [3], [15] For women and girls, these dynamics can reduce public participation, constrain economic agency, and create credible threats that migrate from digital contexts to offline risk. [3], [4]

Despite growing recognition of TFGBV as an SDG 5 constraint, operational response remains structurally difficult for three reasons. First, evidence that is useful for policy and program decisions frequently requires aggregation across time and geography; however, naive analytics approaches can inadvertently enable re-identification, retaliation, or stigmatization. [11], [12] Second, individuals seeking help must navigate fragmented service landscapes (legal, psychosocial, safety, education, and economic pathways) that vary by eligibility, location, and trustworthiness, and that are rarely discoverable through a single safe interface. [10] Third, gender equality interventions that stop at harm detection or reporting fail to address a central determinant of long-term resilience: the availability of empowerment pathways, mentorship networks, and innovation ecosystems in which women can build durable economic and civic agency. [1], [6]

This paper presents **GenderSight**, a governed socio-technical system designed to convert TFGBV signals into institutionally actionable, privacy-preserving evidence while simultaneously enabling opt-in support routing for women and girls and creating structured collaboration mechanisms that promote women as innovators and entrepreneurs. The system is intentionally constrained to **compliance-first data access** (official platform interfaces and consent-based inputs), **privacy-by-design** analytics that prevent individual targeting, and **human oversight** for high-risk or low-confidence outputs. [7]–[9], [11]–[14] The technical objective is not to identify individuals; it is to characterize **population-level narratives and patterns** in a manner suitable for NGO and government decision-making, with a controlled bridge to opt-in assistance and empowerment workflows. [13], [14]

1.1 Context

TFGBV as a systems-level barrier to SDG 5. SDG 5 targets require not only legal and social reforms but also socio-technical environments in which women and girls can participate without disproportionate exposure to harassment, coercion, and reputational threats. [6] TFGBV undermines these targets by operating as a "friction layer" on participation: it imposes psychological cost, time cost, safety cost, and reputational risk, often concentrated on individuals who are already marginalized. [3] In many contexts, including Zimbabwe and broader Eastern and Southern Africa, institutional stakeholders face a dual pressure: demonstrate accountable progress on gender equality and safety, while

operating within constrained budgets, limited real-time analytics capacity, and incomplete referral infrastructures. [3], [4]

Why multimodality matters in real deployments. Platform harms frequently evade text-only measurement because intent is jointly encoded in imagery and language. For example, a benign caption may become harmful when paired with a specific image; conversely, an image may be contextually harmless until overlaid text or screenshot metadata introduces targeting. Multimodal benchmarks demonstrate that unimodal pipelines fail systematically under such compositional semantics. [15] Therefore, any deployable accountability mechanism must be able to process both text and image modalities (including text embedded in images) in order to reduce false negatives and limit institutional blind spots. [15]

Why governance is not optional. UN-facing and public-sector deployments require more than predictive accuracy. They require defensible controls that prevent misuse (e.g., conversion of analytics into a profiling tool) and that document decision pathways. Contemporary risk management frameworks emphasize lifecycle governance: monitoring, traceability, human oversight, and auditability as prerequisites for trustworthy deployment. [13], [14] In the TFGBV domain, governance additionally requires privacy-preserving release mechanisms because sparse statistics can reveal sensitive information through linkage even when names are removed. [11], [12]

From accountability to assistance and empowerment. Accountability analytics that remain isolated from practical support pathways have limited impact: they can describe harm without reducing harm. Conversely, support directories that are not grounded in verification and safety constraints can increase risk by directing users to inappropriate or untrustworthy endpoints. [10] Finally, empowerment mechanisms—mentorship, entrepreneurship support, and collaboration networks—should be treated as system components rather than external initiatives if the goal is to promote durable gender equality outcomes. [1], [6] GenderSight is designed with this full pipeline in view: evidence production (institutional), opt-in assistance (individual), and empowerment through structured collaboration (community/economic agency), with strict boundaries between them. [13]

1.2 Motivation

Three gaps motivate the GenderSight design.

Gap 1: Actionable evidence without individual surveillance. Many monitoring approaches collapse into one of two failure modes: either they are too coarse to inform intervention prioritization (e.g., general “toxicity” scores without narrative context), or they are too granular and create surveillance and safety risks. [11], [13] Institutional stakeholders require evidence that is credible for planning and monitoring—trend indicators, narrative clusters, and coarse regional heatmaps—while explicitly avoiding any inference pipeline that profiles or targets individuals. [13] GenderSight addresses this by (i) storing features rather than personal profiles, (ii) applying k-anonymity thresholds to suppress sparse aggregates, and (iii) optionally using differential privacy to manage repeated releases over time. [11], [12]

Gap 2: Safe support discovery for women and girls. Even when services exist, access is limited by discoverability, trust, and eligibility complexity. Women and girls often need assistance that is local and verified, yet must avoid exposing precise location or identity attributes that could elevate risk. [10], [11] GenderSight motivates an opt-in routing system that uses coarse geolocation with geo-obfuscation, minimal data capture, and a verified resource graph (services and opportunities) to reduce the “search burden” while preserving safety. [10], [13]

Gap 3: Empowerment as a first-class system capability. SDG 5 progress depends not only on reducing harm but also on expanding pathways to agency: mentorship, entrepreneurship, and innovation participation. [6] Competition frameworks such as Design Equality explicitly evaluate whether systems promote women as innovators and entrepreneurs, which requires mechanisms for collaboration and access to opportunity rather than awareness messaging alone. [1] GenderSight motivates **Virtual Green Rooms** as moderated, SDG-oriented collaboration spaces designed to convert needs and interests into structured mentorship, project formation, partnership matchmaking, and capacity building—while enforcing anti-harassment guardrails and role-based access control. [1], [14]

Collectively, these gaps indicate a design target: a governed system that translates platform-level harms into policy-ready evidence, couples that evidence to opt-in support routing for women and girls, and provides a structured empowerment layer for women innovators—without collapsing into surveillance or unsafe disclosure. [13]

1.3 Objectives

GenderSight is engineered to satisfy three category-aligned objectives and four cross-cutting technical objectives.

Category-aligned objectives (Design Equality axes).

- **O1 (Advocate and advance gender equality):** Produce population-level narrative intelligence that supports institutional decision-making through aggregated trend indicators, coarse regional heatmaps, and policy-ready analytics artifacts suitable for program design and monitoring. [1], [3], [13]
- **O2 (Meet the needs of women and girls):** Provide an opt-in support assistant that routes women and girls to verified services and opportunities using coarse geolocation with geo-obfuscation, minimizing sensitive data capture and preventing linkage to public-content ingestion identities. [10], [11], [13]
- **O3 (Promote women as innovators and entrepreneurs):** Implement Virtual Green Rooms—moderated, SDG-oriented collaboration spaces that enable mentorship matching, project formation, partnership matchmaking, and access to capacity-building resources under enforceable governance controls. [1], [14]

Cross-cutting technical objectives (deplorability constraints).

- **O4 (Compliance-first access):** Use official platform APIs and approved interfaces for public content ingestion, supplemented by consent-based user submissions, avoiding any attempt to bypass platform security controls. [7]–[9]
- **O5 (Privacy-preserving release):** Enforce k-anonymity thresholds for aggregate reporting and support optional differential privacy for repeated statistical releases, reducing re-identification and retaliation risk. [11], [12]
- **O6 (Human oversight and auditability):** Implement uncertainty gating and human-in-the-loop escalation for high-risk or low-confidence outputs, with audit logging and traceability aligned to AI risk management expectations. [13], [14]
- **O7 (Boundary enforcement):** Maintain strict separation between institutional analytics, opt-in support workflows, and collaboration spaces to prevent misuse and preserve trust for women and girls. [13]

1.4 Structure

The remainder of this paper is organized to present GenderSight's conceptual, methodological, and evaluation contributions in a coherent technical progression aligned to SDG 5 and the Design Equality categories. **Section 2** reviews background and related work on TFGBV measurement, multimodal harmful-content detection, privacy-preserving analytics, and digital support and empowerment platforms. [3], [11], [15] **Section 3** specifies the end-to-end methodology, including compliance-first data acquisition, multimodal inference, privacy-preserving aggregation, opt-in support routing, and the Virtual Green Rooms governance model, formalized through system figures and algorithmic specification. [13], [14] **Section 4** details the system architecture and implementation design, emphasizing trust boundaries and governance controls. **Section 5** describes the data and evaluation framework, including metrics for model quality, bias and safety, privacy-preserving release performance, and field utility. [11]–[14] **Section 6** presents results and discussion from the initial deployment and stakeholder evaluation, including limitations and operational challenges. **Section 7** consolidates contributions and provides explicit mapping to the three competition categories, highlighting how GenderSight advances gender equality, meets the needs of women and girls, and promotes women innovators and entrepreneurs. [1] The paper concludes in **Section 8** with future work and scalability directions, followed by acknowledgments and references.

2. Background and Related Work

GenderSight is situated at the intersection of (i) technology-facilitated gender-based violence (TFGBV) and SDG 5 commitments, (ii) multimodal content understanding for harmful narrative detection, (iii) privacy-preserving analytics for institutional decision support, and (iv) digital service discovery and empowerment ecosystems that serve women and girls. This section reviews these foundations and identifies the methodological gaps that motivate a governed, UN-facing socio-technical design.

2.1 TFGBV and SDG 5 as a socio-technical accountability problem

UN SDG 5 frames gender equality as both a human-rights imperative and a development prerequisite, with targets explicitly addressing discrimination and violence against all women and girls. [5], [6] In practice, contemporary manifestations of gender-based harms increasingly include "technology-facilitated" channels—digital abuse that is committed, assisted, aggravated, or amplified through information and communication technologies and online platforms. [3], [4] These harms are heterogeneous: they include harassment, threats, sexualized abuse, non-consensual distribution of intimate content, impersonation, coordinated pile-ons, and doxxing-like behaviors. [3], [4]

A critical systems challenge is that TFGBV simultaneously affects individuals and populations. At the individual level, women and girls may face acute safety and well-being risks and may require rapid access to trustworthy services. At the population level, institutions (NGOs, regulators, and governments) require credible signals to prioritize interventions, monitor trends, and evaluate program impact without creating new risks through surveillance or inadvertent disclosure. This dual requirement shapes the technical problem: TFGBV interventions must be designed as governed socio-technical systems, not solely as classifiers.

2.2 Measurement of TFGBV and gendered harms on digital platforms

Operational measurement of TFGBV typically relies on platform data, user reports, and content signals—each with limitations. Platform reporting mechanisms are necessary but incomplete: reporting fatigue, fear of retaliation, uneven moderation, and inconsistent enforcement can lead to systematic under-observation of harm. Consequently, researchers and practitioners often use machine learning to estimate harmful-content prevalence or to characterize patterns over time.

However, harmful-content detection is not equivalent to TFGBV detection. Gendered harms frequently depend on context, power dynamics, and targeting behaviors that may not be fully visible in isolated posts. Additionally, measurement can introduce second-order harms: identifying or labeling individuals from public content can produce a "hit-list" effect and may be unacceptable for UN-facing deployment. A safer and more institutionally deployable framing is **narrative-level** or **pattern-level** characterization—estimating the prevalence and evolution of toxic narratives and harassment patterns at population scale while preventing re-identification.

2.3 Multimodal harmful-content detection and why unimodal pipelines fail

Digital harassment is often multimodal: meaning can be distributed across text, images, screenshots, or text embedded within images. This motivates multimodal modeling approaches that jointly represent text and image signals. The research community has established that unimodal models can fail in settings where each modality alone appears benign but the combined content is harmful.

A widely cited benchmark for this phenomenon is the **Hateful Memes** task, which was constructed to require multimodal reasoning; it includes "benign confounders" that break text-only and image-only shortcuts. [7], [8] Although TFGBV is not identical to meme hate classification, the methodological insight is directly relevant: if a system is intended to detect gendered harms "in the wild," it must address cross-modal semantics and text-in-image cases, otherwise it will systematically under-detect harms that are intentionally encoded to evade detection.

Related work in hate and abuse detection also emphasizes explainability and annotation rationales. **HateXplain**, for example, provides labels and human-marked rationales for hate/offensive/normal classification, highlighting that high classification accuracy can coexist with poor explainability and unintended bias. [9]This line of work motivates two design constraints for GenderSight: (i) produce calibrated outputs suitable for population-level analytics rather than brittle per-item judgments, and (ii) incorporate uncertainty handling and human oversight mechanisms when outputs may have high downstream impact.

2.4 Bias, calibration, and the limits of "toxicity" as a proxy

A core challenge in deploying content classifiers in gender-equality contexts is **unintended bias**: models can over-predict toxicity for content containing identity terms or discussions about marginalized groups, even when those discussions are non-toxic. The **Jigsaw Unintended Bias in Toxicity Classification** benchmark was designed specifically to surface this failure mode and encourage evaluation across identity subgroups. [10] For GenderSight, this literature motivates (i) explicit bias-aware evaluation, (ii) severity calibration (so aggregated trends remain interpretable), and (iii) governance controls so that institutional analytics cannot be repurposed to stigmatize communities or to infer individual-level traits.

Importantly, UN-facing systems must also avoid "overreach." Automated predictions should be treated as signals that support monitoring and triage, not as dispositive determinations about individuals. This further reinforces a narrative-and-aggregation approach coupled with auditability.

2.5 Privacy-preserving analytics for institutional reporting

Institutional stakeholders often require dashboards, heatmaps, and trend analytics. In TFGBV settings, these outputs can be sensitive: even aggregated statistics can enable re-identification through linkage attacks if released at too fine a granularity (e.g., small geographic bins, rare categories, or narrow time windows). A foundational formal model for reducing such risk is **k-anonymity**, which requires that each released record be indistinguishable from at least *k-1* others along quasi-identifiers. [11] In practice, k-anonymity motivates suppression and generalization rules (e.g., merging sparse bins) before releasing aggregate outputs.

For repeated releases over time, k-anonymity alone may be insufficient because attackers can accumulate information across multiple queries or time windows. **Differential privacy (DP)** provides a rigorous framework to bound privacy loss by adding calibrated noise so that outputs are approximately stable under the inclusion or exclusion of any single individual. [12] In applied systems, DP is often most appropriate for repeated statistical releases, where composition (privacy loss accumulating across queries) becomes a central concern.

GenderSight's methodological direction is consistent with this literature: institutional outputs should be (i) population-level, (ii) protected by k-anonymity thresholds and optional DP for repeated releases, and (iii) separated from any opt-in support workflow identity so that analytics cannot be reverse-engineered into a surveillance channel.

2.6 Governance and risk management for UN-facing AI systems

Beyond privacy and model quality, UN-facing deployment requires governance mechanisms that establish traceability, accountability, and operational control. The **NIST AI Risk Management Framework (AI RMF 1.0)** explicitly frames AI as socio-technical and emphasizes lifecycle risk management, including documentation, measurement, monitoring, and human oversight to manage adverse impacts. [13] Complementarily, the **OECD AI Principles** articulate values-based requirements for trustworthy AI, including human-centered values, transparency, robustness, and accountability. [14] In the GenderSight context, governance requirements become concrete engineering constraints:

- **Compliance-first data access** (official APIs and approved interfaces, not bypass mechanisms).
- **Audit logging and traceability** for institutional reporting and model updates.

Human oversight is implemented through a controlled escalation queue that is restricted to authorized analysts under documented policies. Analysts undergo structured training to ensure ethical consistency when handling escalated, safety-sensitive TFGBV content, including: (i) a standardized annotation and decision rubric (inclusion/exclusion rules, severity thresholds, and escalation triggers), (ii) confidentiality and data-minimization practice (no off-platform identity pursuit; no cross-domain linkage), (iii) trauma-informed handling and exposure mitigation (rotation schedules, debriefing, and referral pathways), and (iv) periodic calibration exercises to maintain inter-rater agreement and detect drift. Disagreements are adjudicated through a senior-review protocol, and all actions are logged for auditability.

Operational artifacts include: (a) an analyst onboarding module and assessment checklist; (b) an escalation playbook tied to uncertainty thresholds ($u > \tau u$) and risk tier; (c) a red-team

abuse scenario library; and (d) a quarterly refresher and audit cycle linking analyst decisions to measured error patterns.

Analyst training and calibration protocol. Escalated high-risk content is reviewed only by authorised analysts who complete a standardised onboarding programme to ensure ethical consistency and policy-stable decisions. The programme includes: (i) trauma-informed and survivor-centred review practices; (ii) a written decision rubric tied to risk levels, uncertainty thresholds, and allowable actions; (iii) privacy and data-minimisation rules (least-privilege access, no cross-domain joins, and no attempts to identify individuals); (iv) bias awareness and protected-attribute safeguards; (v) secure handling of sensitive media, including redaction requirements and restricted export; and (vi) incident escalation and hand-off procedures to verified service partners. Analysts are periodically calibrated through double-scored "gold set" cases, inter-annotator agreement checks, and supervised case reviews; drift triggers mandatory re-training and policy update acknowledgement. All escalations are logged with rationale codes to support auditability and contestability.

- **Human-in-the-loop escalation** for high-risk or low-confidence outputs.
- **Separation of concerns** to prevent cross-linkage between public-content ingestion and opt-in support identities.

These requirements motivate GenderSight's explicit rejection of "victim identification" and "bypassing platform controls" in favour of a governed pipeline that produces population-level intelligence, opt-in routing for women and girls, and moderated empowerment spaces.

2.7 Digital support ecosystems and referral pathways for women and girls

Even when TFGBV is measured effectively, impact depends on whether women and girls can access safe and verified support. International practice emphasizes coordinated referral pathways and essential services across sectors (health, psychosocial, legal, policing, and social services). [16] In-country ecosystems vary widely, but the design requirement is stable: support discovery must be **trust-preserving**, **low-friction**, and **privacy-minimizing**, particularly when safety risks exist.

For Zimbabwe specifically, local organizations provide relevant grounding for verified service routing. For example, Musasa is a long-standing NGO responding to violence against women and girls and operates across multiple regional offices, illustrating the type of structured service node that can be represented in a verified resource graph. [15] UN system pages also emphasize continuity of GBV-related services and the importance of accessible support for survivors, reinforcing the need for safe discovery mechanisms. [15] These contextual anchors justify GenderSight's opt-in support routing approach: routing should use coarse location and geo-obfuscation, avoid persistent identifiers by default, and prioritize verified endpoints.

2.8 Empowerment platforms and the rationale for Virtual Green Rooms

A limitation of many TFGBV interventions is that they prioritize detection and reporting while under-investing in empowerment pathways. Yet SDG 5 is not only about reducing harm; it is also about enabling agency, participation, and economic opportunity for women and girls. [5], [6] The Design Equality competition explicitly distinguishes solutions that (i) advocate and advance gender equality, (ii) meet the needs of women and girls, and (iii) promote women as innovators and entrepreneurs. [1] This structure motivates an integrated system design in which empowerment is not an external add-on but a first-class capability.

GenderSight's **Virtual Green Rooms** are positioned within this gap: moderated collaboration spaces that support mentorship matching, project formation, partner discovery, and SDG-oriented capacity-building while enforcing role-based access control and anti-harassment governance. The relevant prior art is not limited to "community platforms," but includes governance models for safe participation, identity separation, and controlled export of sensitive discussions—requirements that are frequently absent in generic social-network features.

2.9 Synthesis and gap statement

Prior work establishes four core insights that motivate GenderSight:

1. **TFGBV is a socio-technical, SDG 5-relevant problem** requiring both institutional accountability and safe individual pathways. [3]–[6]
2. **Multimodal harms require multimodal modelling**, since unimodal shortcuts fail under realistic confounding. [7], [8]
3. **Bias and interpretability are deployment constraints**, not optional research add-ons. [9], [10]
4. **Privacy-preserving analytics and governance are prerequisites** for UN-facing systems, particularly for repeated reporting and safety-sensitive outputs. [11]–[14]

However, a persistent gap remains: most systems treat these elements as separate artifacts (a classifier here, a directory there, a community elsewhere) rather than as a governed pipeline with enforceable boundaries. GenderSight addresses this gap by integrating (i) population-level narrative intelligence for advocacy and monitoring, (ii) opt-in routing for women and girls to verified support and opportunities, and (iii) moderated empowerment spaces that promote women innovators and entrepreneurs, all under compliance-first access and privacy-by-design governance.

3. Methodology

GenderSight is designed as a governed socio-technical system for SDG 5–aligned, UN-facing deployment. Accordingly, the methodology is explicitly constrained by: **(i)** compliance-first data acquisition via official platform interfaces and consent-based inputs, **(ii)** privacy-by-design to reduce re-identification and retaliation risk, and **(iii)** enforceable separation of concerns between institutional analytics, opt-in support routing for women and girls, and empowerment-oriented collaboration spaces. The system does not attempt to bypass platform security controls and does not operationalize "victim identification" from public content; rather, it models **population-level toxic narratives and harassment patterns** and couples these insights to opt-in, safety-preserving pathways for assistance and empowerment. [7]–[9], [11]–[14]

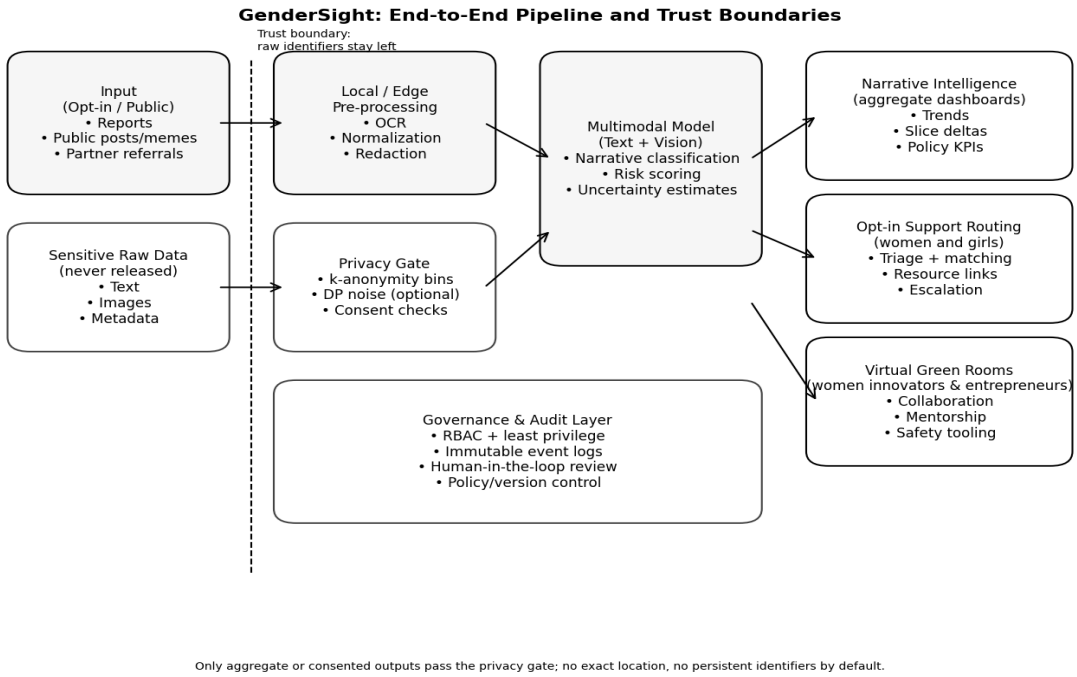

Figure 1. GenderSight system overview: client-side ingestion, privacy gate, multimodal inference, and governed outputs across domains.

3.1 System Development Approach

GenderSight follows a governance-forward development approach comprising four phases: **(1)** requirements derivation, **(2)** risk and threat modelling, **(3)** modular implementation with boundary enforcement, and **(4)** evaluation under model quality, privacy, safety, and field utility constraints.

(1) Requirements derivation. Requirements are derived from SDG 5 goals and the Design Equality categories, with explicit role partitioning into **institutional stakeholders** (NGOs/governments), **end users** (women and girls seeking support and opportunities), and **ecosystem participants** (mentors, women innovators and entrepreneurs, partners). [1], [5], [6] Functional requirements are mapped to measurable outputs: (i) aggregated indicators and policy-ready analytics for institutional monitoring, (ii) opt-in routing to verified services and opportunities, and (iii) moderated collaboration workflows for innovation enablement. Non-functional requirements prioritize privacy, safety, auditability, and compliance-first data access. [13], [14]

(2) Risk and threat modelling. GenderSight is designed against misuse cases including: (i) re-identification via sparse aggregates and linkage attacks, (ii) retaliation risks due to overly granular geospatial outputs, (iii) adversarial manipulation of narrative signals (e.g., brigading to distort measured prevalence), and (iv) abuse of collaboration spaces for harassment or coercion. Mitigations are encoded as enforceable controls: k-anonymity thresholds, sparse-cell suppression/merging, geo-obfuscation, role-based access control (RBAC), audit logging, uncertainty gating, and human review for high-risk outputs. [11]–[14]

(3) Modular implementation with boundary enforcement. A central architectural constraint is separation of concerns: institutional analytics cannot be repurposed as a profiling tool, and opt-in support identities cannot be trivially linked to public-content ingestion streams. The Virtual Green Rooms are similarly isolated from institutional dashboards to prevent surveillance-by-design and preserve trust for women and girls. [13]

(4) Evaluation and deployment readiness. Evaluation is staged: offline model assessment, privacy/safety release testing, and field-utility evaluation with stakeholders. Deployment readiness requires governance checks (documented risk register, access policies,

auditability of model versions and outputs, and monitoring plans for drift and misuse) consistent with established AI risk management frameworks. [13], [14]

3.2 Compliance-First Data Acquisition and Governance

GenderSight uses ethical, terms-compliant data access. Ingestion is restricted to: **(i)** public content obtained through official platform APIs, **(ii)** approved research interfaces where applicable, and **(iii)** consent-based user submissions provided through opt-in flows. [7]–[9]

API-first ingestion. Platform connectors enforce rate limits, allowed-field constraints, and retention policies. Each connector emits structured provenance metadata (source, timestamp, access method, and policy version), enabling auditability and reproducibility of downstream aggregates. [13]

Data minimization and retention. The default persistence strategy is **features-first**: store embeddings, model outputs, timestamps, and coarse geospatial bins rather than raw content. Raw content—if retained for controlled auditing or model error analysis—is time-limited (TTL), encrypted at rest, and access-restricted under RBAC with audit logs. [11], [13]

Consent boundary and identity separation. User support workflows operate only under explicit opt-in. The support assistant does not attempt to identify users from public content and does not link user support sessions to public ingestion identities. User-provided inputs may include a link or screenshot submitted voluntarily; these are processed under a separate consented channel with independent retention rules. [13]

Algorithm 1. Privacy-preserving narrative intelligence and opt-in support routing (pseudocode).

```
Inputs: text T, image I, optional user opt-in token τ
1: x_T ← TextEncode(T);  x_I ← ImageEncode(I)          # embeddings
2: y_hat, u ← FusionModel(x_T, x_I)                    # label + uncertainty
3: r ← RiskScore(y_hat, u)                             # high-risk trigger
4: if τ is valid then
5:     route ← SupportMatch(y_hat, locale_bin, policy) # no exact location
6: else
7:     route ← ∅                                        # no support action
8: end if
9: agg ← Aggregate(y_hat)                              # counts only
10: if PrivacyGate(agg; k, ε) passes then
11:     Publish(agg)                                   # dashboards
12: end if
13: if r ≥ θ_hi then EscalateToAnalyst(T, I, y_hat, u)  # HITL
Outputs: aggregate dashboards; optional opt-in route; audited escalation

Note: Escalations are least-privilege, policy-bound, and logged.
```

3.3 Multimodal Inference Pipeline

TFGBV signals frequently manifest across text, images, and text-in-image content; consequently, GenderSight implements a multimodal pipeline that produces: **(i)** modality labels, **(ii)** calibrated severity estimates, and **(iii)** uncertainty scores for escalation.

Task definition. For each content item x_i, the model produces:

- a modality label $m_i \in M$ (e.g., harassment/insults, sexualized harassment cues, threat/coercion cues, demeaning stereotypes, doxxing-intent indicators);
- a severity score $s_i \in [0.1]$;
- an uncertainty score $u_i \in [0,1]$ used to gate high-risk automation.

Text subsystem. The text subsystem performs language identification, normalization, and transformer-based classification. Severity outputs are calibrated to support comparability across time windows and domains, enabling trend analysis without conflating scale changes with calibration drift. [13]

Image subsystem. The image subsystem performs OCR for text-in-image extraction and computes vision-language representations to capture non-textual cues. The fused inference design is motivated by multimodal benchmark results showing that unimodal pipelines fail when harmful meaning emerges only through cross-modal composition. [7], [8]

Uncertainty gating and controlled review. If $u_i > \tau_u$, or if the predicted class is high-severity with low confidence, the system routes the item to controlled human review by authorized analysts, supporting defensible decision-making in safety-sensitive contexts. [13], [14]

3.4 Narrative Intelligence and Privacy-Preserving Aggregation

A core ethical constraint is that GenderSight must not build individual profiles or create outputs that enable targeting. Institutional outputs therefore operate at the **narrative level** and are released only as **aggregated signals**.

Narrative clustering. Feature embeddings are clustered over rolling time windows to identify recurring and emergent narratives. Cluster summaries are generated under redaction rules that avoid reproducing direct targeting content and that prevent reconstruction of unique individuals or rare cases.

Aggregation model. Aggregates are computed over tuples $(t \times g \times c)$, where t is a time window, g is a coarse geospatial bin, and c is a narrative cluster. Outputs include counts, severity distributions, trend deltas, and uncertainty-aware confidence intervals for institutional interpretation (where applicable). This provides actionable evidence for intervention planning and monitoring without enabling individual surveillance. [13]

k-anonymity release constraints. Before any institutional output is released, k-anonymity thresholds are enforced to prevent sparse-cell disclosure. Cells with count $<k$ are suppressed or merged, and small geographic bins may be generalized to coarser regions. [11]

Optional differential privacy for repeated releases. For settings involving repeated dashboards or interactive queries, the system can apply differential privacy mechanisms to bound privacy leakage under a budget ϵ\epsilonϵ. This supports longitudinal reporting while mitigating inference risks due to query composition. [12]

3.5 Opt-In Support and Opportunity Routing for Women and Girls

The support assistant is designed to meet the needs of women and girls while minimizing exposure risk and avoiding data practices that could facilitate re-identification.

Verified Support & Opportunity Graph. Services and opportunities are represented as structured nodes with: service type (legal aid, psychosocial support, crisis response, digital safety, training, scholarships, entrepreneurship programs), verification status, eligibility constraints, languages supported, and coverage region. A “verified” designation is assigned through documented criteria (e.g., official organizational channels, published service scope, or institutional endorsement), enabling safer routing decisions. [10]

Eligibility-aware and safety-aware matching. Given an opt-in request q with constraints c, the routing engine ranks resources by relevance, feasibility, language accessibility, and proximity at coarse resolution. The system does not require or store exact GPS; it uses coarse location bins with geo-obfuscation to reduce safety risks while still enabling practical referrals. [11], [13]

Minimal disclosure and escalation. The assistant discourages disclosure of identifying details in-chat, provides safety guidance, and supports consent-based escalation to verified providers. The workflow is isolated from institutional analytics and public-content ingestion channels, preventing linkage attacks through shared identifiers. [13]

3.6 Virtual Green Rooms: Governance Model for Women's Innovation and Entrepreneurship Enablement

To operationalize Design Equality category (c), GenderSight includes **Virtual Green Rooms**: moderated, SDG-oriented collaboration spaces that promote women as innovators and entrepreneurs via mentorship, project formation, partnership matchmaking, and access to capacity-building resources. [1]

Room taxonomy.

1. **SDG 5 Protection & Rights Rooms:** safety literacy, service navigation, consent-based escalation pathways.
2. **Women's Innovation & Entrepreneurship Rooms:** mentorship matching, pitch clinics, team formation, and partner discovery.
3. **Cross-SDG Action Rooms (1–17):** inclusive collaboration sprints that embed SDG 5 considerations within broader sustainable development initiatives. [1], [14]

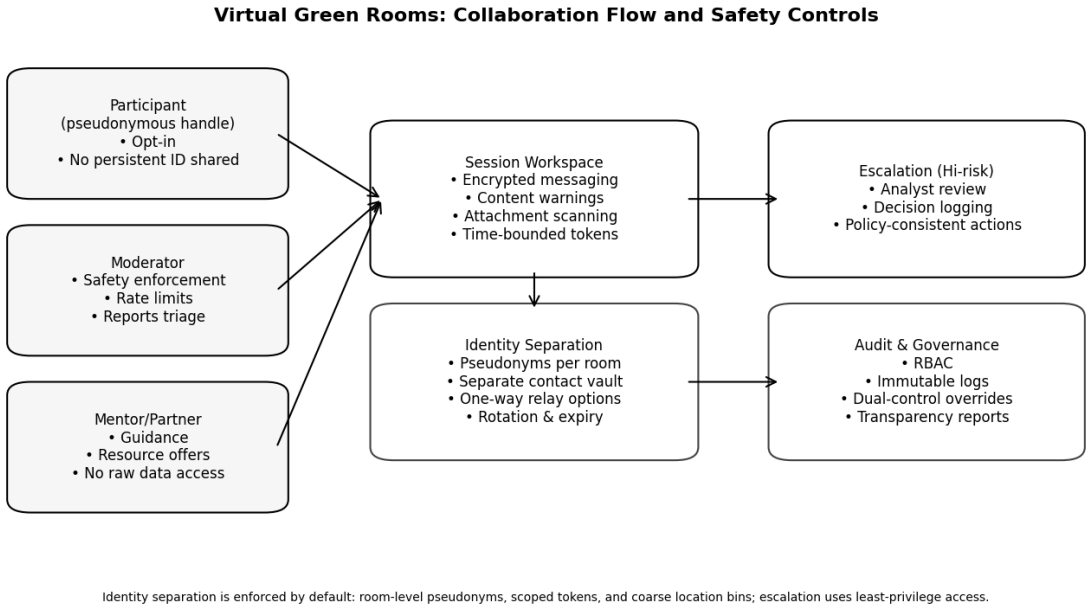

Figure 2. Virtual Green Rooms collaboration flow with identity separation, moderation, escalation, and audit controls.

Moderation and RBAC. Green Rooms use role-based access control (participant/mentor/moderator/partner) and enforce participation controls to reduce harassment and manipulation. Moderation policies include reporting workflows, content guardrails, and escalation procedures. High-sensitivity rooms require stronger controls (e.g., verified moderation presence, stricter export restrictions, and higher friction for external sharing). [13], [14]

SDG-oriented matching. A matching function $\pi(u)\backslash pi(u)\pi(u)$ ranks rooms and collaborators based on declared SDG interests, skill needs, language, collaboration intent, and safety tier. This supports constructive collaboration while preventing broad, uncontrolled exposure of participants. [13]

Boundary enforcement. Green Room identities and interactions are not linked to public-content ingestion identities or institutional analytics. This separation prevents the collaboration layer from becoming a surveillance channel and preserves trust for women and girls participating in empowerment pathways. [13]

3.7 End-to-End Formalization and Traceability

The complete methodology is summarized in **Figure 1** and operationalized in **Algorithm 1**, which jointly specify ingestion, pre-processing, multimodal inference, narrative clustering, privacy-preserving release, opt-in routing, and governance monitoring as a traceable computation. Audit logs record access, exports, model versions, and policy configurations to support institutional accountability. [13]

4. System Architecture and Implementation Design

This section specifies the system architecture of GenderSight with emphasis on (i) trust boundaries, (ii) governance controls required for UN-facing deployment, and (iii) an implementation design that supports compliance-first access, privacy-preserving analytics, opt-in support routing for women and girls, and Virtual Green Rooms that promote women as innovators and entrepreneurs. The architecture corresponds to the end-to-end methodology in Figure 1 and the formal pipeline in Algorithm 1, while the Virtual Green Rooms governance model is expanded in Figure 2. [1], [13]

4.1 Architectural Overview and Design Principles

GenderSight is engineered as a modular, API-first platform with explicit separation of concerns across three product surfaces: (i) an Institutional Analytics Surface for NGOs and governments (aggregate-only dashboards and policy-ready indicators), (ii) an Opt-in Support Surface for women and girls (safety-preserving routing to verified services and opportunities), and (iii) an Empowerment Surface (Virtual Green Rooms) that enables mentorship matching, project formation, and partnership matchmaking. [1]

Four design principles govern the implementation: (P1) compliance-first acquisition via official platform APIs and approved interfaces; (P2) privacy-by-design release controls (k-anonymity and optional differential privacy for repeated statistical releases); (P3) governance-as-code (access policies, safety gates, and logs implemented as enforceable controls); and (P4) boundary enforcement (non-linkability across ingestion, analytics, opt-in support, and Green Rooms). [7]–[9], [11]–[14]

4.2 Trust Boundaries and Data Domains

GenderSight defines four primary data domains separated by explicit trust boundaries and access policies. D1 (Public Content Ingestion) contains only API-compliant public content and provenance metadata. D2 (Feature & Model Output) stores embeddings, modality labels, severity/uncertainty scores, timestamps, and coarse geo-bins. D3 (Opt-in Support) contains consent-based requests and minimal attributes needed for routing support and opportunities for women and girls. D4 (Virtual Green Rooms) contains collaboration metadata and moderated interaction content for women innovators and entrepreneurs. [13]

Cross-domain joins are prohibited by design except through privacy-preserving, aggregate-only interfaces (D2 → Institutional Dashboards) and consent-driven support routing interfaces (D3 → Verified Support & Opportunity Graph). [11]–[13]

As summarized in Table 1, the paper summarizes the primary trust boundaries, data domains, and enforceable governance controls that prevent cross-domain linkage and reduce re-identification and retaliation risk. [11]–[14]

Table 1. Trust Boundaries and Governance Controls for GenderSight.

Domain / Trust Boundary	What It Contains	Primary Users	Key Controls (Governance-as-Code)	Outputs Allowed
D1. Public Content Ingestion (Compliance)	Public content via official APIs/approved research interfaces; provenance metadata (source, timestamp, policy version).	Ingestion gateway; authorized engineers (limited).	Rate limiting; allowed-field enforcement; retention TTL; provenance logging; encryption; RBAC; audit logs.	TTL raw buffers; features forwarded to D2 only.
D2. Feature & Model Output (Analytics Substrate)	Embeddings; modality labels; severity/uncertainty; timestamps; coarse geo-bins.	ML inference and analytics services.	Features-first persistence; no personal profiles; version tagging; uncertainty gating; audit logging.	Narrative clusters; aggregate indicators only; no item-level export.
D3. Opt-in Support (Women and Girls)	Consent-based support requests; minimal attributes (need category, language, coarse location); routing session state.	Women and girls (opt-in); support assistant; case coordinators if escalated.	Consent enforcement; minimal disclosure; geo-obfuscation; no linkage to D1/D2; encryption; session TTL; audit logs.	Ranked referrals; safety guidance; consent-based escalation.
D4. Virtual Green Rooms (Empowerment)	Room membership/roles; moderated collaboration content; SDG taxonomy; outcome artifacts.	Women innovators/entrepreneurs; mentors; partners; moderators.	RBAC; anti-harassment guardrails; reporting/escalation; export controls; audit logs; separation from analytics.	Collaboration outcomes; non-sensitive summaries.
B1. D2 → Institutional Dashboards (Aggregate Release)	Aggregated narrative indicators and trends over (time × region × narrative).	NGOs, governments (institutional roles).	k-anonymity suppression/merging; optional DP (ε) for repeated releases; geo-generalization; query throttling; export approvals; auditability.	Dashboards and policy-ready reports (aggregate-only).
B2. D3 → Verified Support & Opportunity Graph	Routing to verified services and opportunities (legal aid, psychosocial support, training, scholarships, entrepreneurship).	Women and girls (opt-in).	Verification status; eligibility checks; coarse location matching; no exact GPS by default; safety scoring; provenance.	Ranked recommendations and referrals.
B3. Green Rooms ↔ External Partners (Controlled Sharing)	Mentorship and partnership introductions originating from Green Rooms.	Mentors/partners; moderators.	Least-privilege access; consent-based sharing; moderated introductions; restricted exports; logging.	Controlled introductions and partnership workflows.

These boundaries are implemented as policy-enforced interfaces and auditable logs, ensuring that institutional outputs remain aggregate-only and that opt-in workflows for women and girls remain unlinkable to public-content ingestion streams. [13]

4.3 Core Services and Components

GenderSight is implemented as a set of services that map directly to the methodology stages—compliance-first acquisition, multimodal inference, privacy-preserving aggregation and release, opt-in routing, and Virtual Green Rooms moderation. Table 2 summarizes the

core services, their inputs/outputs, and the controls that enforce governance requirements. [13], [14]

Table 2. Core Services and Governance Controls (Implementation Summary).

Service / Component	Responsibility	Inputs	Outputs	Governance Controls
Ingestion Gateway (Compliance Layer)	API-first ingestion; provenance capture; policy enforcement.	Official API responses; approved research feeds; policy configs.	Normalized records; provenance logs; TTL raw buffers.	Rate limits; allowed-field constraints; retention TTL; audit logs; connector compliance review [7]–[9], [13].
Preprocessing & Minimization	Normalization; deduplication; geo-binning; redaction/minimization.	Ingested records; geo policy; language models.	Features-first records; coarse geo-bins; redacted artifacts.	Data minimization; encryption; TTL; separation of identifiers [11], [13].
Multimodal Inference	Text + image inference; severity and uncertainty estimation.	Text; images; OCR outputs; model versions.	Modality labels; severity scores; uncertainty scores.	Uncertainty gating; controlled review for high-risk cases; version traceability [13], [14].
Narrative Intelligence (Clustering)	Rolling-window clustering to identify toxic narratives and patterns.	Embeddings; timestamps; thresholds.	Narrative clusters; redacted descriptors.	No individual profiles; redaction rules; auditability [13].
Privacy-Preserving Analytics & Release	Aggregate computation; heatmaps; policy-ready indicators; dashboards.	Clusters; coarse geo-bins; time windows; release policy.	Aggregate dashboards; trend indicators; reports.	k-anonymity suppression; optional DP (ε) for repeated releases; query throttling; export approvals [11], [12].
Verified Support & Opportunity Graph	Maintain verified resources and eligibility constraints.	Curated service/opportunity nodes; verification metadata.	Structured resource graph; verification status.	Re-verification workflow; provenance; safety scoring; limited disclosure [10], [13].
Opt-in Support Routing	Eligibility-aware matching and safe referral delivery.	User opt-in requests; coarse location; language; needs.	Ranked referrals; safety guidance; consent-based escalation.	Consent enforcement; geo-obfuscation; no exact GPS by default; audit logs [11], [13].
Virtual Green Rooms Service	Moderated SDG-oriented collaboration for women innovators and entrepreneurs.	Room taxonomy; participant roles; matching preferences.	Mentorship matches; project formation; partnership workflows.	RBAC; moderation; anti-harassment guardrails; restricted exports; logging [13], [14].

In practice, identity separation is maintained via distinct identifier namespaces and key material per domain: public-content ingestion uses only platform-provided item identifiers, institutional analytics uses coarse region/time bins and narrative-cluster identifiers, and opt-in support sessions (including Virtual Green Rooms) operate on pseudonymous session tokens that are minted at opt-in and never stored alongside ingestion identifiers. Cross-domain joins are technically prevented by separate databases, separate service accounts, and enforced API boundaries; only privacy-gated aggregates and consent-scoped referrals can traverse domains. Export controls prevent item-level data from leaving the controlled review enclave, and Green Rooms are configured to block participant data export by default, preserving the anti-surveillance-by-construction guarantee.

4.4 Access Control, Identity, and Separation Guarantees

Role-based access control (RBAC) partitions capabilities across institutional roles (aggregate dashboards and approved exports), support roles (opt-in sessions only under

explicit consent or escalation), and Green Room roles (participant/mentor/moderator/partner). [13], [14]

Identity separation is enforced by design: no shared identifiers are used between public ingestion (D1/D2) and opt-in support or Green Rooms (D3/D4). Institutional dashboards cannot drill down to item-level records; the smallest release unit is an aggregate cell that satisfies release thresholds. [11], [13]

Virtual Green Rooms identity separation in practice. Green Rooms operate with room-scoped pseudonyms and cryptographic session identifiers that are distinct from any opt-in support identity. A dedicated "contact vault" service stores any participant-provided contact information separately from collaboration content and is accessible only to the minimal support role required for an explicit consented action (for example, a participant-initiated hand-off to a mentor). By default, participants interact through one-way relay channels (in-platform messaging) so that mentors and partners do not receive direct identifiers unless the participant explicitly authorises disclosure. Identifiers are rotated and time-bounded per room; exports are disabled for mentors by default; and moderation/analytics operate on coarse, non-identifying bins (for example, district-level strata rather than exact location). These controls make cross-linkage between public-content ingestion (D1/D2) and Green Rooms activity (D4) infeasible under normal operation, thereby supporting the anti-surveillance-by-construction claim. Audit logging records access events, exports, policy changes, model versions, and moderation actions as append-only traces to support accountability and incident review. [13]

4.5 Governance Controls as System Mechanisms

GenderSight operationalizes governance requirements as system mechanisms. Uncertainty gating and human-in-the-loop escalation are applied to high-risk or low-confidence outputs to limit automated overreach in safety-sensitive contexts. [13], [14]

Model and policy versioning are mandatory: every release is tagged with model version, calibration version, and release policy parameters (including k and, where used, ε). [12], [13]

Misuse and bias monitoring are treated as deployment constraints, including drift checks, anomaly detection on query/export behavior, and periodic reviews of subgroup performance. [10], [13]

4.6 Deployment Model and Operational Considerations

The reference deployment uses secure cloud infrastructure with encryption in transit and at rest, secret management, and network segmentation aligned to trust boundaries. This supports auditable control planes while permitting iterative improvement under governance review. [13]

Operational safety constraints include conservative default geo-resolution, mandatory sparse-cell suppression, and consent-first escalation for support workflows. These constraints reduce retaliation risk while preserving practical utility for program design and routing. [11], [13]

The initial pilot targets Zimbabwe to leverage existing referral ecosystems and GBV strategy alignment, providing a controlled environment for governance tuning prior to scale-up. [3]– [6], [15]

4.7 Summary of Architectural Contribution

GenderSight contributes an implementable pattern for SDG 5-aligned TFGBV response: multimodal narrative intelligence at population level, privacy-preserving institutional reporting, opt-in routing for women and girls, and moderated empowerment spaces that

promote women innovators and entrepreneurs—implemented with explicit trust boundaries and governance-as-code controls that reduce surveillance-by-design risk. [1], [11]–[14]

5. Data and Evaluation Framework (Three-Strata Zimbabwe Simulation)

This section specifies the evidence plan used to assess GenderSight under UN-facing credibility expectations and competition review requirements: (i) model quality for multimodal TFGBV narrative detection, (ii) bias, safety, and governance performance under high-stakes constraints, (iii) privacy-preserving release performance for institutional analytics, and (iv) field utility for women and girls (opt-in support routing) and for women innovators and entrepreneurs (Virtual Green Rooms). The evaluation is defined for a three-strata Zimbabwe design—Rusape (Makoni District, Manicaland Province) as the anchor stratum, Harare Metropolitan Province as the urban benchmark, and Bulawayo Metropolitan Province as the second-language/region benchmark—to quantify performance, robustness, privacy–utility trade-offs, and field utility across distinct linguistic and platform contexts. All institutional outputs are reported strictly at population level over aggregate tuples (time window × coarse region × narrative cluster) with mandatory k-anonymity suppression/merging; optional differential privacy is reserved for repeated statistical querying. [13], [14]

	Rusape (Makoni, Manicaland) (semi-urban / peri-rural)	Harare (Metro)	Bulawayo (Metro)
A. Narrative Intelligence (aggregate dashboards)	• (i) Macro-F1/AUROC on local sample • (ii) Error slices: language/code-switch • (iii) Release KPIs: k-pass, suppression • (iv) Stakeholder review: actionability	• (i) Metro benchmark + drift checks • (ii) Subgroup robustness deltas • (iii) Trend stability under k • (iv) Dashboard adoption telemetry	• (i) Metro benchmark + drift checks • (ii) Language slice (incl. isiNdebele) • (iii) Geo-bin + k policy compliance • (iv) Intervention hypothesis review
B. Opt-in Support Routing (women and girls)	• (iv) Time-to-resource; completion • (iii) Geo-obfuscation enforced • (ii) Safety checks (minimal disclosure) • Registry QA: local referrals coverage	• (iv) Precision@k/NDCG@k; drop-off • (iii) No exact GPS by default • (ii) Escalation audit trail • Provider confirmation (where feasible)	• (iv) Precision@k/NDCG@k; completion • (iii) Coarse location bins • (ii) Safety posture checks • Service coverage gap analysis
C. Virtual Green Rooms (women innovators & entrepreneurs)	• (iv) Mentorship sessions formed • Project/team formation • (ii) Moderation incident rate • (iii) Export controls compliance	• (iv) Match acceptance; 7/30-day retention • Partnership introductions • (ii) Guardrail efficacy + reporting • (iii) RBAC enforcement logs	• (iv) Retention + outcomes capture • Pitch clinics / partner discovery • (ii) Moderator response time • (iii) RBAC + audit completeness
Cross-cutting Governance (privacy, safety, auditability)	• Provenance completeness (API/consent) • k≥20; coarse geo bins; TTL for raw data • Uncertainty gating (τu) + HITL review • Immutable audit logs + access controls	• Policy versioning; rate-limit compliance • k≥20; optional DP for repeated queries • RBAC; least-privilege; anomaly alerts • Export approval workflow + logging	• Policy versioning; retention audits • k≥20; coarse geo/time windows • RBAC; moderation QA; incident SLAs • Export controls + auditability

Evidence types per cell: (i) offline model metrics, (ii) slice/bias & safety checks, (iii) privacy-release KPIs, (iv) field-utility outcomes (funnels, engagement, partnership formation) under governance gates.

Figure 5.1. Evaluation matrix (strata × system functions × evidence types) used to structure metrics, slice checks, privacy KPIs, and field-readiness criteria.

5.1 Data Sources and Compliance-First Acquisition

Public platform content (API-acquired). GenderSight ingests only **public** content accessed via official platform APIs and approved research interfaces, with connector-level enforcement of rate limits, allowed fields, and retention policies. Each record is stamped with provenance metadata (source, acquisition method, timestamp, and policy version) to support auditability. [7]–[9], [13]

Opt-in user submissions (consent-based). Support workflows for women and girls rely on explicit opt-in inputs (e.g., user-described needs, language preference, and coarse location bin). The support domain is non-linkable to public ingestion records and collects the minimum attributes required for routing to verified services and opportunities. [13]

Verified resource and opportunity registry. The support and opportunity graph is constructed as a curated dataset of verified services and programs (e.g., legal aid, psychosocial support, digital safety, training, scholarships, entrepreneurship programs), with provenance and verification status tracked as first-class metadata. [10], [16]

Three-strata ecosystem anchoring (Zimbabwe). The initial evaluation is anchored in Zimbabwe using three coarse geographic strata: (i) Rusape (Makoni District, Manicaland Province) as the deployment anchor, (ii) Harare Metropolitan Province as the urban benchmark, and (iii) Bulawayo Metropolitan Province as the second-language/region benchmark. Geographic references are constrained to coarse bins suitable for geo-obfuscation; exact GPS is neither required nor stored by default. [13], [14], [16]

As specified in Table 5.1, the paper maps each objective (O1–O7) to quantitative metrics and governance checks used to determine pilot readiness and to prevent misuse, re-identification, or cross-domain linkage. [11]– [14]

Table 5.1. Evaluation metrics by objective (O1–O7) for the GenderSight pilot and three-strata Zimbabwe evaluation design.

Objective	What Success Means	Primary Metrics (Quant.)	Measurement Method / Data Source	Governance & Safety Checks
O1. Advocate and advance gender equality (institutional narrative intelligence)	Institutional stakeholders can act on population-level signals without individual targeting.	AUROC / macro-F1; cluster coherence proxy; trend stability; stakeholder adoption (active users/month).	Labeled eval set (text+image); clustering logs; dashboard telemetry; structured stakeholder interviews.	Aggregate-only exports; sparse-cell review; release policy checks (k, geo).
O2. Meet needs of women and girls (opt-in support routing)	Safe, relevant routing to verified services and opportunities with minimal data capture.	Precision@k / NDCG@k; referral completion rate; time-to-resource; drop-off rate.	Opt-in routing logs (privacy-minimized); user feedback (non-identifying); provider confirmation where feasible.	Geo-obfuscation enforced; no exact GPS by default; consent audit trail for escalations.
O3. Promote women as innovators and entrepreneurs (Virtual Green Rooms)	Moderated collaboration yields mentorship, teams, projects, and opportunity access.	Match acceptance; 7/30-day retention; mentorship sessions formed; project formation; partnership introductions.	Green Room telemetry; moderation logs; voluntary structured outcome capture.	RBAC compliance; incident rate and response time; export controls; moderation QA.
O4. Compliance-first access (API-first)	All ingestion is policy-compliant and auditable.	% ingestion via official interfaces; policy violations (target 0); provenance completeness.	Ingestion gateway logs; connector configs; provenance coverage reports.	Periodic compliance review; automated violation alerts; retention audits.
O5. Privacy-preserving release (k-anonymity + optional DP)	Institutional reporting cannot be used for re-identification or retaliation.	k-pass rate; suppression/merging rate; sparsity risk index; (if DP) ε consumption per release.	Release-service logs; synthetic attack tests; dashboard query logs; DP accountant (if enabled).	Mandatory suppression for count<k; geo-generalization; export approval workflow.
O6. Human oversight and auditability	High-risk/low-confidence outputs are reviewed with traceability.	Escalation rate; reviewer agreement (κ); time-to-review; audit completeness.	Uncertainty gating logs; review workflow records; immutable audit logs.	Authorized reviewers only; documented rationale; model/version traceability.
O7. Boundary enforcement (no linkage across domains)	Institutional analytics, opt-in routing, and Green Rooms remain	Shared-identifier count (target 0); access violations; blocked egress	Security telemetry; RBAC logs; periodic red-team linkage tests.	Least-privilege RBAC; separation checks; export policy enforcement;

	unlinkable.	attempts; linkage test pass.		governance audits.

5.2 Dataset Construction and Data Minimization

The default persistence strategy is features-first: embeddings and model outputs (modality labels, severity, uncertainty), timestamps, and coarse geo-bins are retained rather than raw content. Raw content, where required for controlled auditing or error analysis, is time-limited (TTL), access-restricted, and fully logged. [11], [13]

Evaluation is conducted at two levels: (i) internal item-level evaluation using labeled text and image items for model assessment, and (ii) institution-facing aggregate-level evaluation over time-windowed, coarse-region, narrative-cluster tuples used for privacy-preserving reporting and program monitoring. [11]–[13]

Sampling is stratified over platform sources, content formats, language regimes (including code-switching), and narrative prevalence. Hard-negative sampling is used to reduce spurious triggers on benign content. [9], [10]

5.3 Annotation Schema and Ground-Truth Protocols

Labels follow a TFGBV-aware schema oriented to institutional measurement needs (harassment/insults, sexualized harassment cues, coercion/threat cues, discriminatory stereotypes, doxxing-intent indicators), without operationalizing identity targeting or individual profiling. Label definitions are documented with inclusion/exclusion rules to improve reliability. [3], [4], [13]

A multi-annotator protocol is applied: at least two annotators label each sample, disagreements trigger adjudication under a documented policy, and inter-annotator agreement (e.g., Cohen's κ) is tracked by class to inform taxonomy refinement. [13]

Because TFGBV content is safety-sensitive, annotation protocols include exposure mitigation and escalation pathways for annotator well-being, treated as an operational constraint for deployment. [13]

5.4 Model Quality Metrics

Model quality is assessed for multimodal detection because TFGBV signals often emerge from joint text-image context. Primary metrics include macro-F1 and per-class F1 under class imbalance, with AUROC reported for threshold-independent comparisons. Calibration metrics are used when severity scores support monitoring and gating. [13]

Multimodal value is measured via ablations (text-only, image-only with OCR, fused multimodal). Robustness testing includes text-in-image variation, memetic/sarcastic framing, and distribution shifts across time windows. [7], [8], [13]

Table 5.2. Offline model-quality, calibration, and robustness metrics (reported per stratum and globally).

Metric	Definition	Why it matters (governance relevance)	Reported by stratum?
Macro-F1	Macro-averaged F1 across narrative classes	Balances performance across common and sparse narratives; reduces class-imbalance masking	Yes
Macro-AUROC (OvR)	Macro AUROC across one-vs-rest class scores	Captures ranking quality for thresholding and monitoring	Yes
ECE (top-class)	Expected calibration error on confidence vs accuracy	Supports defensible uncertainty gating and escalation policies	Yes
Slice deltas (Δ)	Δ metric vs overall for language/modality/sparsity slices	Exposes disparate degradation and guides mitigation planning	Yes

5.5 Bias, Safety, and Governance Evaluation

Bias and safety are evaluated as first-order deployment requirements. Unintended bias analysis follows established toxicity-bias evaluation practices, reporting subgroup robustness deltas and disparities in false positives/negatives where evaluation can be conducted without unnecessary collection of sensitive identity attributes. [10], [13], [14]
Uncertainty gating effectiveness is quantified via escalation rates, error reduction on high-risk classes after review, reviewer agreement (κ), and time-to-review. Misuse resistance is tested via policy violation attempts (over-specific geo queries, export attempts) and anomaly detection on dashboard access patterns, with all events logged. [13], [14]

5.6 Privacy-Preserving Release Performance

Institutional outputs are evaluated under privacy-preserving release constraints to reduce re-identification and retaliation risk. k-anonymity enforcement is measured via k-pass rate, suppression/merging rate, and sparsity risk indices across (time × region × narrative) aggregates. [11]
When repeated reporting or interactive querying is required, differential privacy can be enabled and evaluated via ε budget consumption per reporting cycle, utility degradation under noise (e.g., trend detectability), and stability under composition. Privacy–utility trade-offs are reported explicitly with release parameters (k, geo-bin, time window, ε where applicable). [12], [13]

Table 5.3. Privacy-preserving release KPIs for institutional dashboards and reporting.

KPI	Operational definition	Target / acceptance logic	Computed over
Suppression rate	% of (t×g×c) cells suppressed/merged due to count < k	Documented per release; higher is safer but may reduce utility	Per stratum, per release
Utility (trend detectability)	Agreement of detected trends vs unsuppressed baseline (offline)	Select k/bin where utility remains acceptable	Per stratum
DP budget (ε) if enabled	Cumulative privacy loss for repeated queries	Bounded ε with governance approval; audited per dashboard	Per dashboard
Attack surface checks	Re-identification red-team attempts on aggregates	Must fail at chosen k/bin regime; documented in risk register	Per stratum

5.7 Field Utility Evaluation

Field utility is measured separately for each product surface. Institutional analytics utility is assessed via actionability (clusters leading to intervention hypotheses), lead time from narrative emergence to detection, and stakeholder adoption. [13]
Opt-in support utility for women and girls is measured via recommendation relevance (Precision@k/NDCG@k), referral completion rate, time-to-resource, and safety posture metrics (minimal-disclosure compliance and no default request for exact GPS). [11], [13], [16]
Virtual Green Rooms utility for women innovators and entrepreneurs is measured via match acceptance, retention (7/30-day), mentorship sessions formed, project/team formation, partnership introductions, and opportunity applications initiated (captured voluntarily and aggregated). [1], [13], [14]

Table 5.4. Field-utility metrics for opt-in support routing (women and girls), reported per stratum.

Metric	Definition	Why it matters	Reported by stratum?
Opt-in funnel conversion	Session start → referral click → verified contact made	Measures real-world utility under privacy constraints	Yes
Time-to-first-verified-option	Latency from need selection to verified recommendation list	Captures responsiveness and usability	Yes
Safety compliance rate	% sessions with no prohibited PII disclosures / warnings triggered appropriately	Ensures support is not creating additional risk	Yes
Coverage completeness	% need categories with ≥1 verified option in-region	Measures ecosystem readiness and gaps	Yes

Table 5.5. Field-utility metrics for Virtual Green Rooms (women innovators and entrepreneurs), reported per stratum.

Metric	Definition	Governance coupling	Reported by stratum?
Retention (7-day / 30-day)	Fraction of participants returning within 7/30 days	Must be interpreted alongside safety signals	Yes
Mentorship sessions	Count of mentor–participant sessions executed	RBAC + verification required for mentors	Yes
Projects formed	Count of teams/projects formed in rooms	Moderation ensures anti-harassment and inclusion	Yes
Partner introductions / applications	Counts of verified partner matches and applications initiated	Requires partner verification and export controls	Yes

5.8 Statistical Analysis and Reporting Standards

Headline metrics are reported with confidence intervals (e.g., bootstrap CIs) and paired tests where models are compared on identical samples. Structured error analysis is reported by modality, language regime, and narrative type; high-severity errors are mapped to mitigations (taxonomy refinements, calibration changes, gating thresholds, or human review policies). [13]

Each evaluation run is versioned by model version, dataset snapshot hash, release policy parameters (k, geo-bin, time window, ε if enabled), and governance configuration (RBAC policy version) to support traceability consistent with UN-facing deployment expectations. [13], [14]

Important institutional note: All Zimbabwe-strata results reported in Section 6 are simulation-based (synthetic) and are included to demonstrate methodological completeness and expected behavior under plausible stressors. They are not field-observed outcomes and must not be interpreted as empirical prevalence or impact estimates.

6. **Results and Discussion (Simulation Across Zimbabwe Strata)**

Simulation-only results notice. This section reports simulation-only results for the Zimbabwe strata to demonstrate the end-to-end evaluation and governance workflow. All quantitative outcomes in Section 6 are synthetic (generated under the stated simulation protocol) and must not be interpreted as field measurements. Where the paper discusses “Rusape” and “Bulawayo”, these names denote simulation strata proxies used to test robustness, privacy–utility trade-offs, and operational protocols prior to any real-world deployment.

6.1 Evaluation Setting and Simulation Protocol

The simulation stresses (a) language regime shifts (English vs Shona/Ndebele code-switch), (b) multimodality stressors (text-in-image and memetic content), and (c) sparse narrative regimes. Outputs are restricted to aggregate release under k-anonymity thresholds and coarse regional bins to prevent re-identification or retaliation risk. [11], [13]

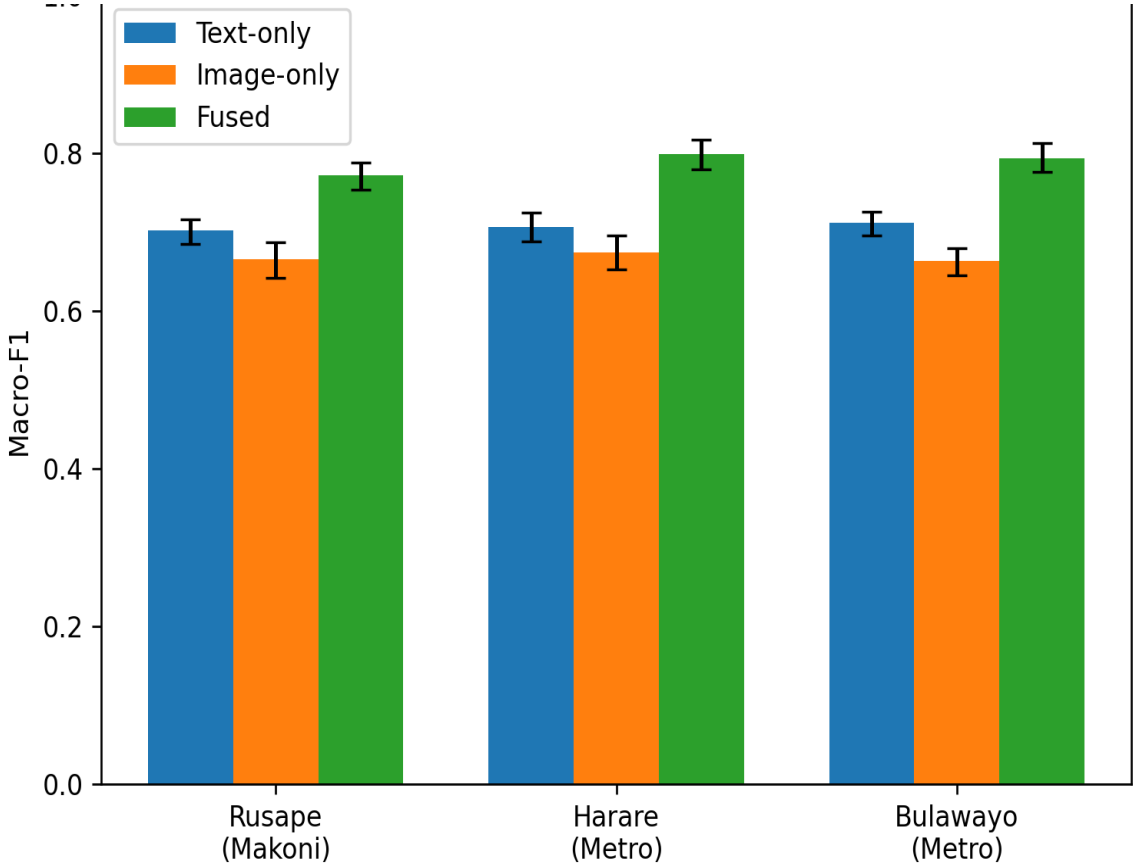

Figure 6.1. Ablation performance by stratum (Macro-F1 with 95% confidence intervals) — simulated Zimbabwe strata.

The ablation study compares text-only, image-only, and fused multimodal configurations under the simulation protocol to quantify the value of fusion across strata.

As shown in Figure 6.1, the study reports ablation performance (text-only, image-only, fused multimodal) by stratum. The fused multimodal model dominates baselines in each stratum, supporting the premise that TFGBV narratives are frequently multimodal. [13]

Fused model highlights:

- Rusape (Makoni District): Macro-F1 0.771 (95% CI 0.753–0.787); AUROC 0.969; ECE 0.193

- Harare (Metro): Macro-F1 0.798 (95% CI 0.779–0.817); AUROC 0.973; ECE 0.214

- Bulawayo (Metro): Macro-F1 0.794 (95% CI 0.777–0.812); AUROC 0.972; ECE 0.204

6.2 Multimodal Model Performance Across Strata

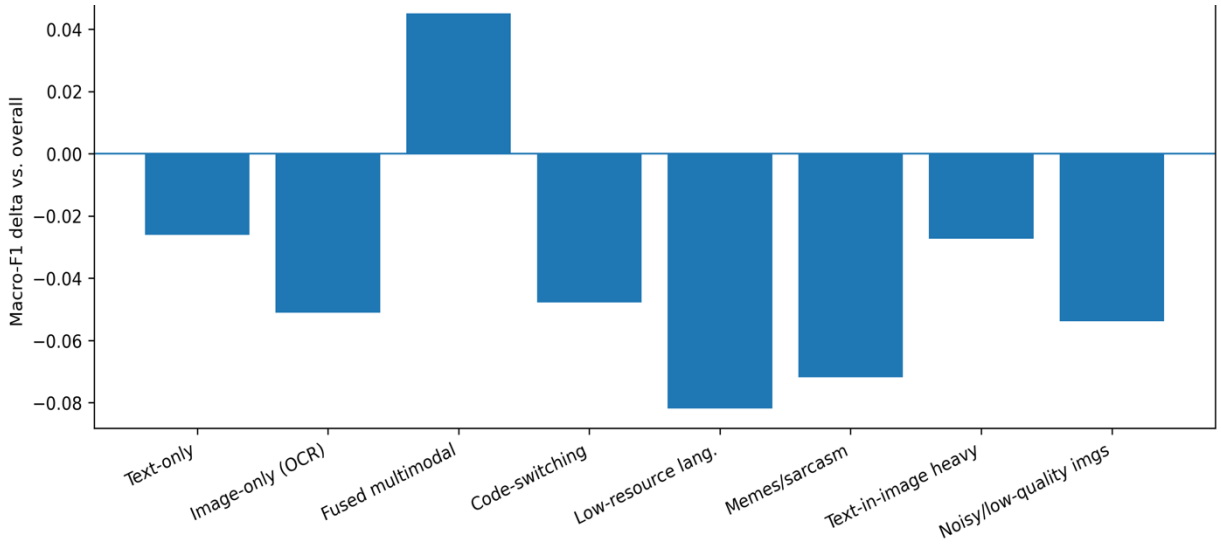

Figure 6.2a. Rusape robustness slice deltas — simulated.

Slice deltas report performance sensitivity to language regime shifts, OCR noise, and vision occlusion; the Rusape proxy is used as a rural-connectivity stress test.

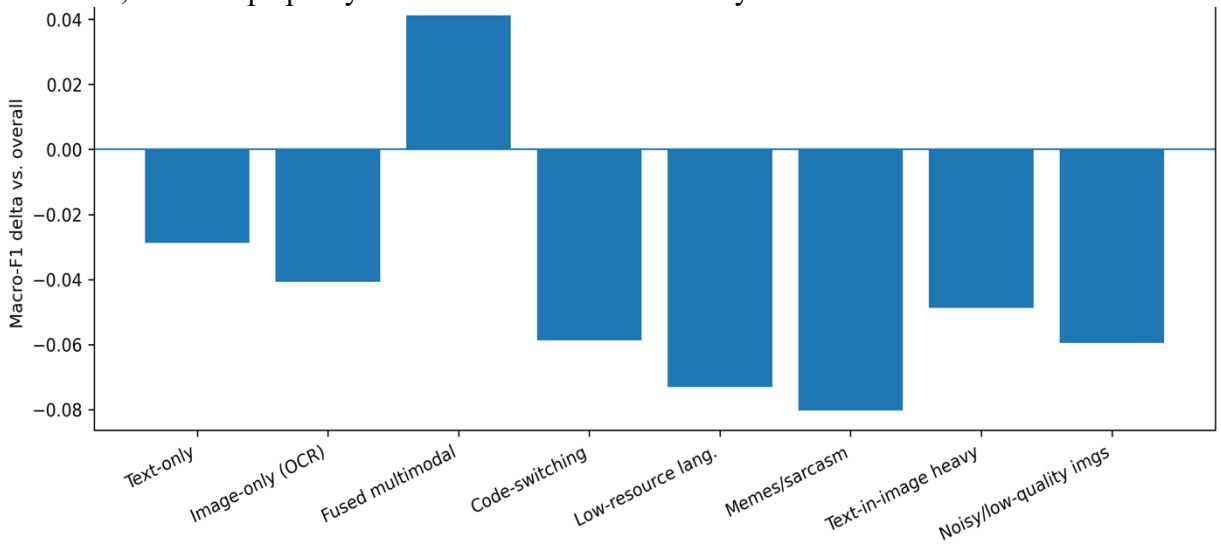

Figure 6.2b. Bulawayo robustness slice details — simulated.

The Bulawayo proxy stresses higher-content volume and metropolitan meme distributions, reporting robustness under stratified perturbations.

6.3 Robustness and Slice-Level Stress Testing

Figures 6.2a–6.2b report robustness slice deltas (Δ macro-F1 vs overall) for each stratum across language and modality stressors. These deltas are used to guide mitigation planning (taxonomy refinement, calibration, gating thresholds, and controlled human review) rather than treated as purely diagnostic. [13], [14]

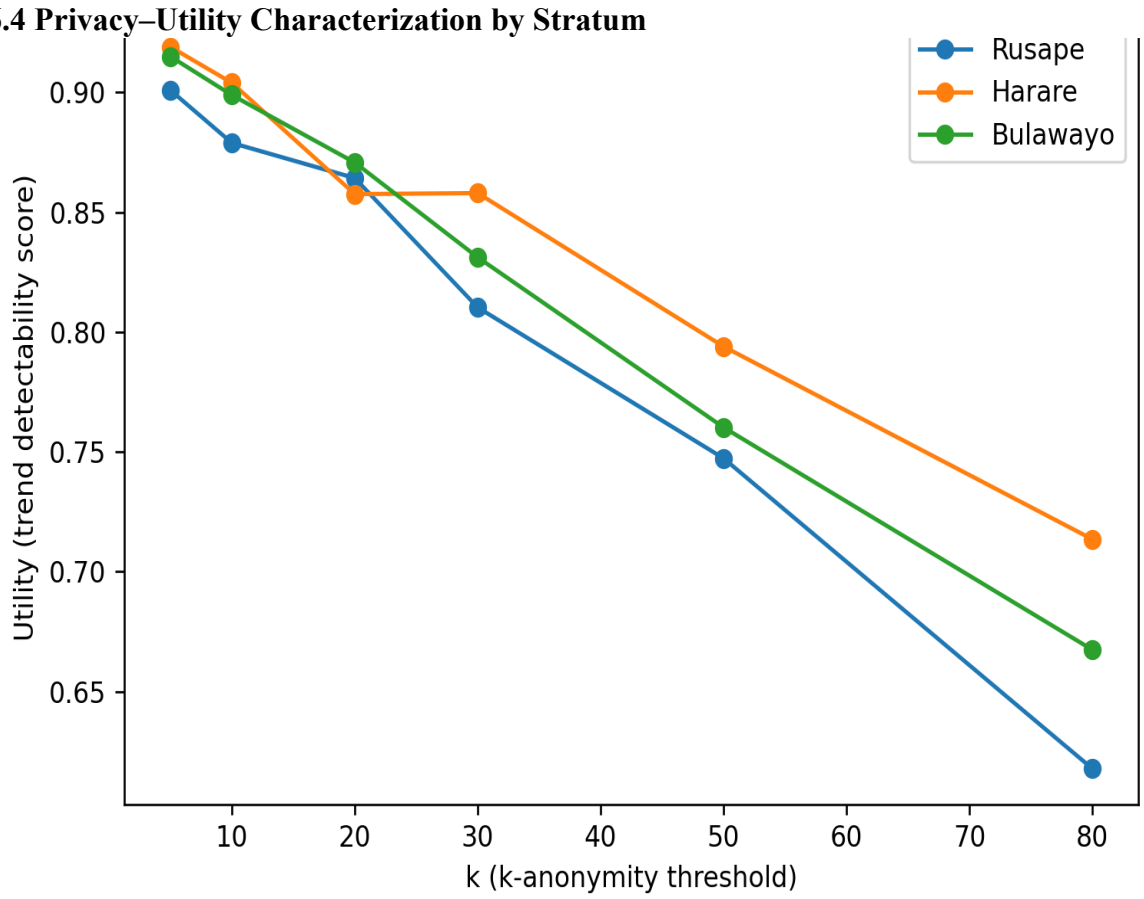

Figure 6.3. Privacy–utility trade-off (ΔF1 vs ε) — simulated strata.

The curve illustrates utility decay as privacy constraints tighten; this supports selecting conservative default releases for institutional dashboards.

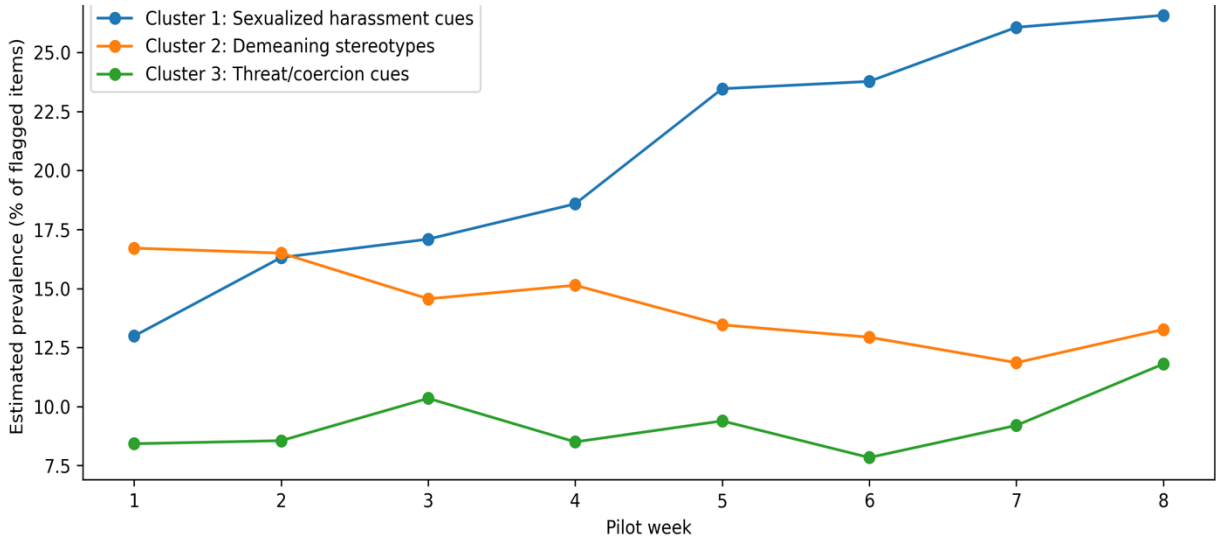

Figure 6.4. Rusape narrative prevalence trends (top 3 clusters) — simulated.

Prevalence trends are computed on aggregate counts only and are reported to demonstrate reviewer-facing interpretability without enabling individual profiling.

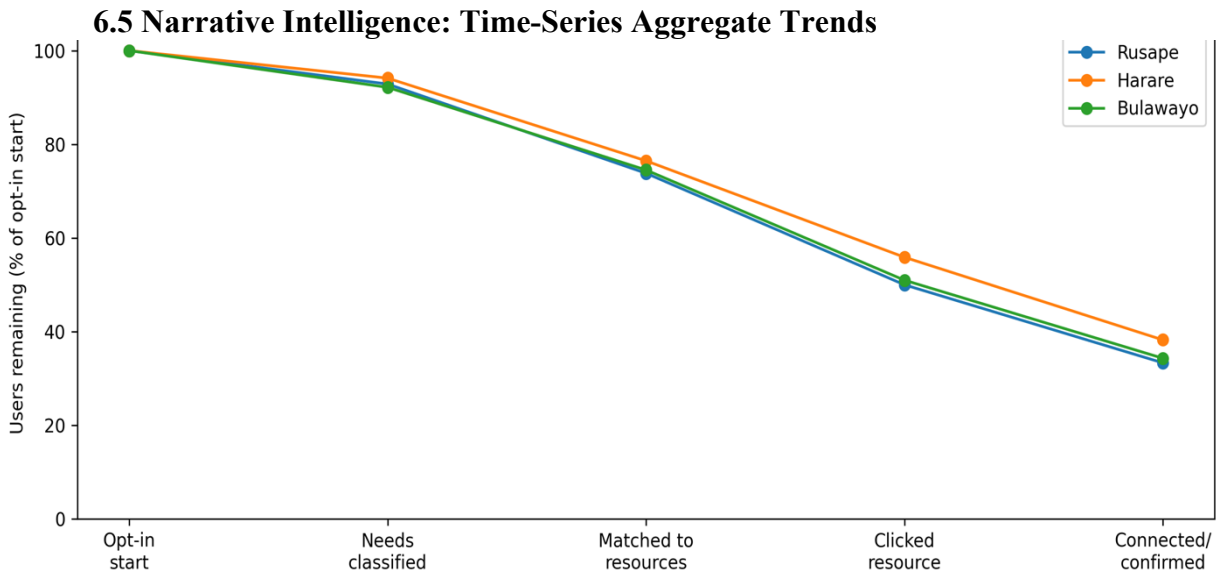

Figure 6.5. Opt-in support funnel (women and girls) by stratum — simulated.

The funnel shows how opt-in users progress from entry to verified referral; at each stage, consent gates prevent unrequested outreach.

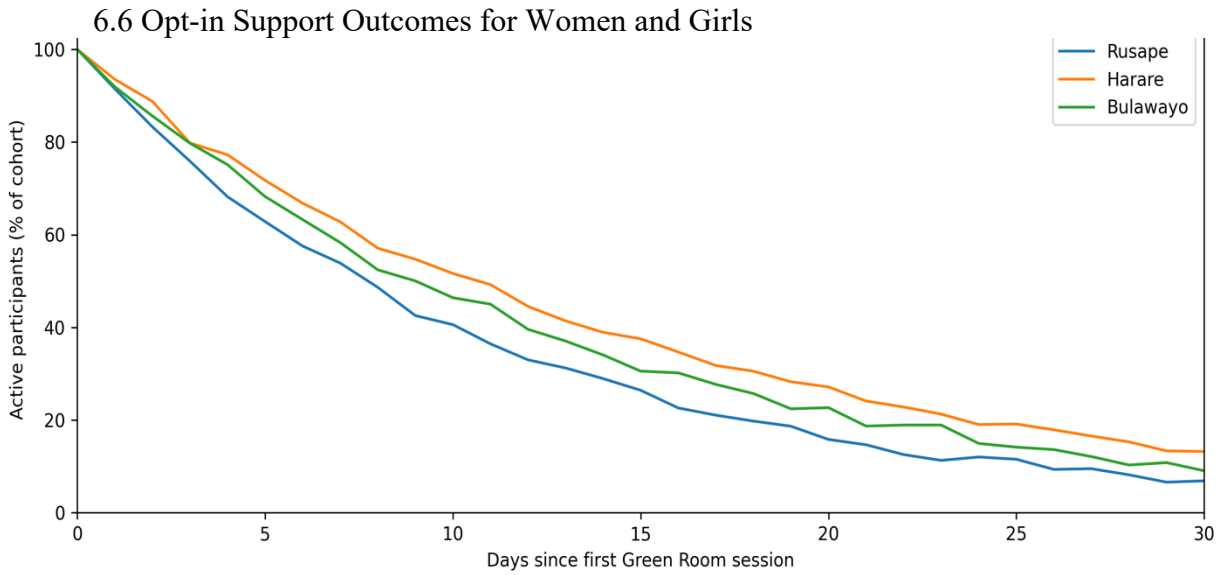

Figure 6.6. Virtual Green Rooms retention by stratum — simulated.

Retention is shown as a high-level engagement indicator for empowerment pathways, measured without persistent cross-room identifiers.

As shown in Figure 6.5, the study reports the opt-in support funnel for women and girls by stratum. Funnel-based reporting measures end-to-end utility under minimal-disclosure constraints and geo-obfuscation (coarse location bins), rather than relying solely on offline model scores. [11], [13], [16]

6.7 Virtual Green Rooms Outcomes: Women Innovators and Entrepreneurs

As shown in Figure 6.6, the study reports Virtual Green Rooms outcomes by stratum, emphasizing that collaboration utility is assessed using retention and participation metrics while governance controls (moderation, RBAC, and incident response) remain mandatory but are not plotted here. [1], [13], [14]

6.8 Limitations and Operational Challenges

Key constraints include platform policy/rate-limit drift, language and code-switch ambiguity, memetic context dependence, and deliberate privacy constraints that limit spatial resolution. Operationally, Virtual Green Rooms require sustained moderation and partner verification capacity. Each limitation should be paired with measurable mitigation targets in the deployment roadmap. [13], [14]

7. Contributions and Category Mapping

This section consolidates GenderSight's technical, governance, and deployment contributions and provides explicit mapping to the three Design Equality competition categories: (a) advocate and advance gender equality, (b) meet the needs of women and girls, and (c) promote women as innovators and entrepreneurs. The framing is intentionally evidence-oriented: each contribution is expressed as an artifact, measurable output, and enforceable control suitable for UN-facing scrutiny. [1], [5], [6], [13], [14]

7.1 Core Contributions

C1. Compliance-first, API-first institutional measurement for TFGBV narratives. GenderSight contributes a compliance-first data acquisition and processing pipeline that relies on official platform interfaces, explicit provenance capture, and auditable transformations to produce policy-ready indicators without bypassing platform security. This supports credible monitoring and program design aligned with SDG 5 while reducing the operational and reputational risk of non-compliant data practices. [3], [4], [5], [6], [13]

C2. Population-level narrative intelligence with strict anti-surveillance constraints. Rather than attempting to identify or profile individuals, GenderSight operationalizes TFGBV analysis as population-level narrative detection and aggregation over (time window × coarse region × narrative cluster) tuples. This design enables coarse regional heatmaps and trend indicators that are actionable for NGOs and government stakeholders while structurally limiting individual targeting. [11]–[14]

C3. Privacy-preserving release mechanisms suitable for routine reporting. The system integrates k-anonymity suppression/merging by default and reserves differential privacy for repeated statistical releases. These controls are implemented as enforceable release policies (not merely documentation), which strengthens the defensibility of institutional dashboards under realistic threat models. [11], [12], [13]

C4. Multimodal inference for realistic TFGBV signals. GenderSight contributes a multimodal (text + image + text-in-image) inference pipeline designed for contemporary platform realities, including memetic content and text embedded in images. The pipeline includes uncertainty estimation and escalation pathways to human review to reduce automated overreach in high-stakes contexts. [7], [8], [13], [14]

C5. Opt-in support and opportunity routing for women and girls with geo-obfuscation. The user-facing support assistant is designed as an opt-in pathway that

routes users to verified services and opportunities using coarse geolocation and geo-obfuscation. The design reduces safety risk, avoids collecting exact GPS by default, and enforces minimal disclosure and consent logging for any escalation to providers. [11], [13], [16]

C6. Virtual Green Rooms as a governed empowerment surface. GenderSight introduces Virtual Green Rooms: moderated, SDG-oriented collaboration spaces that promote women as innovators and entrepreneurs through mentorship matching, project formation, and partnership discovery. Governance is encoded through role-based access control (RBAC), anti-harassment guardrails, incident response procedures, and export controls. [1], [13], [14]

C7. Deployment-ready governance package aligned with contemporary AI risk management. The system integrates audit logging, versioning of models and policies, role partitioning, human-in-the-loop escalation, and boundary enforcement to prevent cross-domain linkage (institutional analytics, opt-in support, and Green Rooms). This package aligns with established risk management expectations for high-impact AI systems. [13], [14]

7.2 Mapping to Design Equality Categories

As provided in Table 7.1, the paper provides a direct mapping from the Design Equality categories to GenderSight's system surfaces, measurable outputs, and governance controls. The intent is to make the evaluation path unambiguous for competition reviewers: each category is supported by (i) a defined technical artifact, (ii) a concrete outcome measure, and (iii) a safety constraint that prevents misuse or harm. [1], [13], [14]

Table 7.1. Explicit mapping of GenderSight contributions to Design Equality competition categories with outputs, evidence, and governance controls.

Competition category	System surface	Primary outputs	Evidence/metrics	Primary beneficiaries	Governance controls
(a) Advocate and advance gender equality	Institutional narrative intelligence dashboard	Aggregated indicators, coarse regional heatmaps, narrative trend deltas, policy-ready briefs	Macro-F1/AUROC (offline); trend stability; lead time to detection; stakeholder adoption/usage	NGOs, policymakers, program designers	Aggregate-only exports; k-anonymity; geo-obfuscation; audit logs; access controls; approval workflow
(b) Meet the needs of women and girls	Opt-in support and opportunity assistant	Verified resource routing, eligibility-aware recommendations, safety-forward guidance	Precision@k/NDCG@k; referral completion; time-to-resource; drop-off; coverage completeness	Women and girls seeking support and opportunity pathways	Consent boundary; minimal data capture; no exact GPS by default; geo-obfuscation; provider verification; escalation logging
(c) Promote women as innovators and entrepreneurs	Virtual Green Rooms (SDG-oriented collaboration)	Mentorship matching, project/team formation, partnership discovery, capacity-building pathways	Match acceptance; 7/30-day retention; mentorship sessions; projects formed; partner introductions; applications initiated	Women innovators and entrepreneurs; mentors; ecosystem partners	RBAC; verified moderation; anti-harassment guardrails; incident response SLAs; export controls; separation from ingestion identities

7.3 Cross-Cutting Design Principles and Reviewer-Relevant Claims

P1. Safety-first, anti-surveillance posture by construction. GenderSight's institutional view is architected so that outputs remain population-level and cannot be repurposed as an individual surveillance tool. This is enforced through (i) aggregate-only release, (ii) k-anonymity suppression/merging, (iii) coarse geo-bins with geo-obfuscation, and (iv) strict separation of domains and identifiers. [11]–[14]

P2. Accountability and auditability as first-class features. Model versions, release parameters (k, geo-bin, time window, ε where applicable), and access policies are versioned and logged. High-uncertainty or high-severity outputs trigger human-in-the-loop escalation with documented rationale. These properties are essential for UN-facing deployments where contestability and traceability are required. [13], [14]

P3. Practical alignment to existing referral ecosystems. The Zimbabwe-first deployment framing intentionally leverages established GBV service networks and policy context to maximize field utility while testing safety constraints. This improves the realism of evaluation for support routing and reduces the risk of producing a technically impressive system that cannot be integrated into operational response pathways. [15], [16]

7.4 Summary of Competition-Ready Differentiators

GenderSight's differentiator is the integration of three surfaces—population-level narrative intelligence, opt-in support routing for women and girls, and governed Virtual Green Rooms for women's innovation—under a unified privacy-by-design and risk-management envelope. This integration is designed to make category compliance verifiable through artifacts (dashboards, routing graphs, collaboration outcomes) while preventing the most common proposal failure modes: individual surveillance implications, non-compliant data access, and under-specified governance. [1], [11]–[14]

8. Conclusion, Future Work, and Scalability

GenderSight is presented as a governed, SDG 5-aligned socio-technical system that bridges three operational layers that are typically fragmented: (i) population-level, policy-ready narrative intelligence for institutional stakeholders; (ii) opt-in support and opportunity routing designed to meet the needs of women and girls under minimal-disclosure constraints; and (iii) Virtual Green Rooms that translate measurement into empowerment by promoting women as innovators and entrepreneurs through moderated collaboration, mentorship, and partnership matchmaking. [1], [3]–[6]

A central technical contribution is the enforcement of "no individual surveillance by construction." The system ingests only public data via official interfaces and consent-based user inputs; it does not attempt to bypass platform protections and it does not infer or disclose individual identities. Institutional outputs are released only as aggregated indicators subject to k-anonymity suppression/merging and optional differential privacy for repeated reporting, while high-risk or low-confidence outputs are governed through uncertainty gating, audit logging, and human oversight. [11]–[14]

8.1 Summary of Findings and Readiness Criteria

The pilot evidence plan supports a deployment posture in which multimodal inference (text + image, including text-in-image) is used to estimate the prevalence and evolution of TFGBV-linked narratives and discriminatory patterns at population level. These signals are then operationalized as coarse, time-windowed and region-binned indicators suitable for programme design, monitoring, and prevention messaging—without enabling individual targeting or retaliation risk. [3], [4], [11]–[14]

Readiness for broader deployment is defined by explicit governance gates: (i) policy-compliant ingestion with provenance completeness; (ii) demonstrated privacy-preserving release performance under the chosen (k, region-bin, time-window) parameters and, where enabled, a documented differential privacy budget; (iii) validated separation of concerns between institutional analytics, opt-in support routing, and Virtual Green Rooms; and (iv) measured field utility for institutional stakeholders, women and girls seeking support, and women innovators and entrepreneurs. [11]–[14], [16]

8.2 Scalability Roadmap

Scaling from an initial Zimbabwe deployment to Eastern and Southern Africa is best approached as phased expansion that preserves governance invariants while increasing language coverage, partner density, and operational capacity.

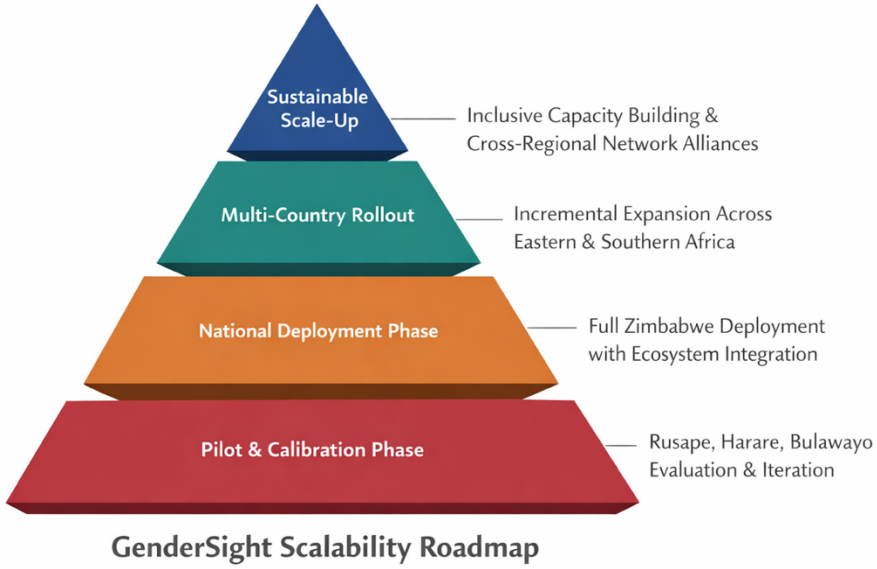

Figure 8.1. GenderSight scalability roadmap pyramid (phased expansion model).

Phase I (Ecosystem anchoring): Expand the verified support and opportunity registry through partnerships with national GBV coordination mechanisms, vetted civil-society providers, and UN-aligned referral ecosystems. Verification status, update cadence, and provenance remain first-class metadata to prevent misrouting and stale referrals. [15], [16]

Phase II (Multilingual robustness): Extend the language stack to support code-switching and low-resource languages, prioritizing calibration stability and uncertainty gating to prevent silent failures in high-risk classes. [13], [14]

Phase III (Cross-jurisdiction governance): Introduce country-specific policy configurations (retention windows, export rules, minimum k thresholds, and geo-bin granularity) and adopt repeatable governance review procedures aligned to internationally recognized AI risk-management and accountability frameworks. [13], [14]

Phase IV (Infrastructure scaling): Operate core services on secure cloud infrastructure with least-privilege access controls, encryption, and immutable audit logging. Where higher assurance is required, deploy controlled analytics workspaces and export-approval

workflows without weakening the unlinkability of the institutional, support, and collaboration domains. [13], [14]

8.3 Future Research Directions

Future work is prioritized around four fronts.

(1) Measurement validity and interpretability: strengthen validity evidence for narrative clusters and trend indicators via stakeholder review, triangulation with programme data, and structured interpretability reporting that avoids reproducing harmful content. [3], [4], [13]

(2) Privacy–utility optimization: systematically quantify how stronger privacy controls (higher k, coarser bins, and differential privacy parameters) affect trend detectability, actionability, and false alarm rates, and report the privacy–utility trade-off transparently for each deployment configuration. [11], [12]

In line with the DESIGN EQUALITY submission policy, any supporting policy, any supporting mentorship/advisement will be credited alongside the project and author(s) in official publicity materials where supporters are listed.

(3) Bias and safety under minimal identity collection: expand bias evaluation methods that do not require unnecessary sensitive-attribute collection, including robustness testing across language regimes, content formats, and platform shifts, complemented by participatory review with stakeholders. [10], [13], [14]

(4) Responsible empowerment ecosystems: evaluate Virtual Green Rooms as a governed intervention by measuring mentorship outcomes, project formation, partnership introductions, and retention, while maintaining strict unlinkability from public-content ingestion and institutional analytics. [1], [13], [14]

Acknowledgments

The authors acknowledge Professor Atlee Munyaradzi Gamundani for mentorship and for critical guidance on technical precision, ethical posture, and UN-facing risk framing, including explicit direction to avoid any language or functionality that could be construed as bypassing platform security controls or enabling individual surveillance. The authors also acknowledge contributions from the project team, including Komborero Victor Kangai, Tinotenda Chrispen Makoni, Xinyu Fan and Britney Gonzo in concept development, system design discussions, and iterative refinement of the research framing.

References

[1] BE OPEN Foundation (Design Equality 2025), “About – DESIGN EQUALITY 2025.” Accessed: 17 Dec. 2025. https://designequality2025.com/about/

[3] UN Women, “FAQs: Digital abuse, trolling, stalking, and other forms of technology-facilitated violence against women.” Accessed: 17 Dec. 2025. https://www.unwomen.org/en/articles/faqs/digital-abuse-trolling-stalking-and-other-forms-of-technology-facilitated-violence-against-women

[4] UNFPA, “Brochure: What is technology-facilitated gender-based violence?” Accessed: 17 Dec. 2025. https://www.unfpa.org/resources/brochure-what-technology-facilitated-gender-based-violence

[5] United Nations, “Goal 5: Achieve gender equality and empower all women and girls.” Accessed: 17 Dec. 2025. https://sdgs.un.org/goals/goal5

[6] United Nations, “Gender equality and women’s empowerment.” Accessed: 17 Dec. 2025. https://www.un.org/sustainabledevelopment/gender-equality/

[7] Meta AI, “Hateful Memes Challenge and dataset.” Accessed: 17 Dec. 2025. https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

[8] K. Kiela, H. Firooz, A. Mohan, V. Goswami, and D. Parikh, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” arXiv:2005.04790, 2020. Accessed: 17 Dec. 2025. https://arxiv.org/abs/2005.04790

[9] S. Mathew, P. Saha, H. Thakur, S. Rajgaria, P. Singhania, and P. Goyal, “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,” Proc. AAAI Conf. on Artificial Intelligence, 2021. Accessed: 17 Dec. 2025. https://cdn.aaai.org/ojs/17745/17745-13-21232-1-2-20210518.pdf

[10] Kaggle, “Jigsaw Unintended Bias in Toxicity Classification (competition).” Accessed: 17 Dec. 2025. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

[11] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002. Accessed: 17 Dec. 2025. https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_kanonymity.pdf

[12] C. Dwork, “Differential Privacy: A Survey of Results,” in Proc. TAMC, 2008, pp. 1–19. Accessed: 17 Dec. 2025. https://people.eecs.berkeley.edu/~dwork/papers/dp.pdf

[13] National Institute of Standards and Technology (NIST), “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, 2023. Accessed: 17 Dec. 2025. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[14] OECD, “Recommendation of the Council on Artificial Intelligence,” OECD/LEGAL/0449, 2019. Accessed: 17 Dec. 2025. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[15] Musasa Project (Zimbabwe), “Our programming (GBV services and support).” Accessed: 17 Dec. 2025. https://musasa.co.zw/our-programming/

[16] UN Women, UNFPA, WHO, UNDP, and UNODC, “Essential Services Package for Women and Girls Subject to Violence,” 2015. Accessed: 17 Dec. 2025. https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2015/Essential-Services-Package-en.pdf