Project Swayam: The Unbreakable Municipality

Building Nepal's First Fully Offline, Sovereign AI Governance System

"जब विपत्तिले तार चुँडाउछ, 'स्वयं' ले नगर जोड्छ।"

(When disaster snaps the wires, 'Swayam' reconnects the city.)

Sushant Poudel
Dept. of Computer Science and Engineering
Nepal Engineering College
Bhaktapur, Nepal
sushant.poudel2028@gmail.com

Aashika Pandey
Dept. of Computer Science and Engineering
Nepal Engineering College
Bhaktapur, Nepal
aashika100pandey@gmail.com

***Abstract*—Municipal governance in Nepal depends on cloud-based digital infrastructure that freezes during internet outages, transmits sensitive citizen data to foreign servers, and exposes public systems to novel AI-driven cyber threats. This paper proposes Project Swayam (स्वयं — "Self-Sovereign"), a fully offline, air-gapped municipal AI architecture comprising three cooperative components: a quantized local language model (the Local Brain) for citizen-facing service automation, a deterministic intent-validation firewall (the Gatekeeper) for zero-trust database security, and a disaster-resilient radio mesh (the Off-Grid Pulse) for post-disaster survivor coordination. Bench-tested on a standard municipal desktop costing approximately NPR 40,500, the Local Brain processes citizen requests in Nepali, Romanized Nepali, and English at 12.4 tokens per second with zero cloud dependency. The Gatekeeper achieves a 100% true-positive rate against catastrophic data-extraction attacks at a mean detection latency of 16.2 ms, outperforming commercial probabilistic filters such as Meta Prompt Guard and Azure Prompt Shield on every measured metric. The Off-Grid Pulse mesh sustains over 88% packet delivery after 40% node failure in simulation. Building on these results, we present the Budhanilkantha Municipal AI Declaration 2026—six articles embedding sovereign edge principles into local governance law—alongside an 18-month implementation roadmap requiring no foreign technical dependency.*

***Keywords*—edge AI, sovereign computing, model quantization, prompt injection defense, deterministic security, offline governance, low-resource NLP, disaster-resilient networks, municipal AI policy, Nepal.**

I. THE PROBLEM: FOUR FAILURES OF CENTRALIZED GOVERNANCE

Before any architecture can be proposed, the problem it solves must be stated plainly. Nepal's municipal offices face four structural failures that recur daily, each rooted in the same underlying assumption: that intelligent governance requires a stable internet connection.

Server Down. Foreign or centralized servers crash, overload, or become unreachable. When the national Nagarik App's backend is unavailable, clerks have no fallback. Citizens wait. Nothing is processed.

Internet Down. Backhoes sever fiber-optic lines. Monsoons knock out transmission towers. Earthquakes destroy the physical infrastructure that every digital workflow depends on. In Nepal's mountainous terrain, connectivity is not a reliable input.

Software Dependency. Commercial vendors control updates, pricing, and service discontinuation. Municipalities cannot modify, audit, or maintain software they do not own. When a vendor raises prices or exits the market, the municipality has no alternative.

Human Error. Clerks mistype names, dates, and addresses. Paper forms are lost. Verification steps are skipped under pressure. With no automated validation layer, errors propagate directly into permanent citizen records.

Project Swayam resolves all four failures with a single architectural decision: move every critical capability onto hardware the municipality owns, in a building it controls, with software it can inspect.

II. INTRODUCTION

Local governance in Nepal sits at the intersection of rapid digital transformation and persistent infrastructural fragility. Under the federal framework, municipal offices function as the primary point of contact for essential citizen services—birth registration, land titling, tax processing, and complaint resolution. National initiatives like the Digital Nepal Framework and the Nagarik App have made meaningful progress toward e-governance, yet their underlying architecture

carries an unspoken vulnerability: near-total dependence on continuous internet connectivity and centralized, foreign-hosted cloud infrastructure.

When the network goes down, municipal services grind to a halt, leaving citizens waiting at counters that can no longer process their requests. The risks extend beyond connectivity. Routing sensitive citizen data across international borders to foreign servers raises legitimate concerns around data sovereignty and long-term privacy. At the same time, as municipal systems adopt automated interfaces, they become exposed to a newer class of threats—adversarial manipulation and prompt injection attacks—that conventional, signature-based security tools are not designed to detect.

This paper proposes a different architectural foundation for municipal digital infrastructure: Sovereign Edge AI. Rather than treating intelligent automation as something that must live in the cloud, we present a fully localized, air-gapped computing framework capable of running entirely offline on standard, low-cost hardware. A municipality should remain operational, responsive, and secure even when completely cut off from external networks.

The system consists of three cooperative components. The Local Brain is a lightweight, quantized natural language processing system hosted on-premises. It interprets service requests in everyday Nepali, automates document workflows, and assists clerks in real time without relying on the internet. The Gatekeeper is a deterministic, intent-based security layer positioned between citizen inputs and the local database; it identifies and blocks manipulation attempts by analyzing the underlying intent of each query rather than matching known signatures. The Off-Grid Pulse is a resilient mesh network built on low-cost radio microcontrollers distributed across municipal wards; in scenarios where cellular and internet infrastructure are disabled, it uses ambient Wi-Fi signal disruptions to estimate human presence and relay critical information off-grid.

Table I summarizes the transformation this architecture represents.

Current Municipal Reality	The Swayam Transformation
Workflows freeze when internet is lost	Core workflows run offline on standard hardware
Citizen data lives on foreign servers	Data never leaves the municipal building
Static firewalls miss prompt-based attacks	Intent analysis detects unknown manipulation
Survivor mapping absent after disasters	Radio sensing estimates locations locally

TABLE I. Current municipal reality vs. the Project Swayam transformation.

III. THE FIVE PILLARS OF PROJECT SWAYAM

The architecture is organized around five design principles, each resolving a specific failure mode identified in Section I.

Pillar	Nepali Tagline	Design Resolution
Offline AI	इन्टरनेट बिना, सेवा सुरु	Service runs entirely on local hardware; zero cloud calls
Voice-First Nepali	बोल्नुहोस्, लेखिन्छ	Colloquial and Romanized Nepali accepted as input; literacy not required
Zero-Typing Forms	क्लर्कको टाइपिङ अब बन्द	Automated form generation from voice input eliminates manual data entry
Disaster Mesh	भूकम्प पांछे पांन जिउँदा	ESP32 mesh network sustains coordination after total infrastructure collapse
One-Time Cost	एक पटक खर्च, जीवनभर सेवा	NPR 40,500 one-time hardware cost; no recurring cloud subscription

TABLE II. The five design pillars of Project Swayam with Nepali taglines.

IV. LITERATURE REVIEW

The architecture proposed in this paper draws on four distinct research domains: edge infrastructure resilience, localized natural language processing, deterministic security engineering, and device-free ambient sensing in ad-hoc networks.

A. Cloud Dependency and Infrastructure Vulnerabilities

Public administration has universally trended toward centralized, cloud-first e-governance models. Shrestha and Kim [17] found that these platforms experience severe failures when communication channels are disrupted by predictable local events—monsoons, landslides, and construction-related outages—a pattern directly observed in Nepal's local government bodies. The United Nations E-Government Survey [20] reinforces this concern, noting that Nepal's online services index (0.4481) lags well behind its telecommunications infrastructure index (0.7653)—a disparity that captures the last-mile governance gap cloud architectures fail to bridge.

Al-Khouri [3] further argues that routing sensitive citizen records through international cloud infrastructure exposes municipal bodies to geopolitical data-access risks and privacy practices that may conflict with national regulations. Despite this, existing literature offers few alternatives that preserve computational intelligence without internet dependency. The Digital Government Model Framework in the UN Survey [20] operationalizes resilience primarily through redundant cloud connectivity, not genuine offline autonomy.

B. Edge Computing and Model Quantization for On-Premises LLMs

Eliminating cloud dependency without sacrificing language capability requires shifting computation to the network edge. Frantar et al. [6] demonstrated that large models can be compressed from 16-bit floating-point weights to 4-bit integers with minimal perplexity impact through GPTQ. Lin et al. [11] introduced AWQ, recognized with the Best Paper Award at MLSys 2024, enabling 7B-parameter models to run on devices such as the Raspberry Pi 4B and NVIDIA Jetson Orin with 3–4× performance gains over FP16 baselines. Dettmers [5] advanced practical deployment through the GGUF format for efficient inference via llama.cpp. Xiao et al. [21] proposed SmoothQuant for near-lossless INT8 deployment on models up to 175B parameters.

For Nepali language processing, Subedi et al. [19] found that monolingual models such as NepBERTa achieve F1-scores of 0.76 and 0.90 on NER and POS tasks respectively, outperforming multilingual alternatives. Work from Kathmandu University [9] on Nepali RoBERTa variants demonstrates competitive performance on summarization and comprehension tasks, with a NepGLUE score of 95.60.

C. Algorithmic Vulnerabilities and Deterministic Guardrails

Integrating NLP into public systems introduces security risks that conventional web application firewalls cannot address. Perez and Ribeiro [14] formalized prompt injection as a systematic threat, distinguishing direct injection from indirect injection via external content. Greshake et al. [7] demonstrated vulnerability through retrieved documents. The OWASP Foundation [13] ranked prompt injection as the top vulnerability in its 2025 LLM Applications Top 10, noting that reliable prevention remains an open problem.

CAPTURE benchmark evaluations [10] report bypass rates approaching 100% against Meta's Prompt Guard and Microsoft's Azure Prompt Shield using Unicode obfuscation and character-spacing techniques. Alemaino et al. [4] showed that deterministic lexical filtering combined with structural validation and contextual sandboxing achieves 0.00% false positive rates at 2.50 ms latency—a critical property for public systems where false positives directly deny citizen services.

D. Post-Disaster Ad-Hoc Networks and Ambient Sensing

Silva et al. [18] demonstrated that ESP32-based LoRa mesh networks self-organize into off-grid routing systems achieving 88.33% packet recovery through self-healing mechanisms. Rekha et al. [16] proposed a solar-powered LoRa mesh integrating GPS and local web servers, enabling offline communication from standard smartphones. Abuhoureyah et al. [1] showed that RNN-LSTM models processing Channel State Information can classify seven human activities through walls with 97.5% accuracy. Qin et al. [15] improved robustness to environmental and positional variation through feature decoupling and spatiotemporal neural networks.

E. Synthesis of Gaps

Existing research divides into highly capable AI models requiring continuous cloud connectivity [2], [11], and resilient offline mesh networks lacking semantic intelligence to automate workflows or detect intent-based attacks [18], [16]. No existing work integrates these capabilities into a unified public administration framework. Table III summarizes this gap.

Research Area	Existing Focus	The Swayam Gap
E-Governance	Cloud portals	Fails entirely during outages
Edge AI	Consumer devices	No air-gapped municipal workflows
Cybersecurity	Probabilistic filters	Vulnerable to semantic attacks
Disaster Response	Raw telemetry	No intelligent ward-level routing
Low-Resource NLP	Cloud APIs	No offline Nepali municipal LLM

TABLE III. Research gaps across contributing domains addressed by this work.

V. METHODOLOGY: ENGINEERING THE SOVEREIGN EDGE ARCHITECTURE

This section details the architectural design, hardware constraints, mathematical optimization, and validation strategy required to build the Swayam framework. The system topology is strictly air-gapped, physically isolating the municipal local area network from wide-area network vulnerabilities.

A. The Local Brain: Quantization and Inference

1) Post-Training Quantization

The central engineering challenge is running capable language models on commodity hardware—standard desktops with 16–32 GB RAM and no discrete GPU. We apply 4-bit weight quantization using the AWQ framework [11]. The goal is to find a lower-precision weight matrix $\hat{W}$ that minimizes output error relative to the original 16-bit matrix W, given calibration activations X:

$$\arg\min_{\hat{W}} \; ||WX - \hat{W}X||^2_2 \qquad (1)$$

subject to: $\hat{W}^I_j \in \{-8, -7, \ldots, 7\}$ (signed 4-bit grid) and $||\hat{W}||_0 \leq 0.25 \cdot ||W||_0$ (75% memory reduction). AWQ protects salient weight channels by applying per-channel scaling factors s^c derived from activation distributions:

$$\hat{W}_{AWQ} = round((W \cdot s_c) / 2^{b-1}) \cdot (2^{b-1} / s_c) \qquad (2)$$

where b = 4 bits and s^c is learned per channel. This reduces VRAM requirements by approximately 75%, enabling real-time inference on a standard municipal workstation.

2) Localized Fine-Tuning via LoRA

The quantized model is adapted to municipal workflows through Low-Rank Adaptation. LoRA freezes pre-trained weights $W_0 \in \mathbb{R}^{(d\times k)}$ and injects trainable rank-decomposition matrices into each transformer layer:

$$W = W_0 + BA \quad (3)$$

where $B \in \mathbb{R}^{(d\times r)}$ and $A \in \mathbb{R}^{(r\times k)}$ are trainable matrices, and $r \ll \min(d, k)$ is the rank (typically 8 or 16). This yields a 7B-parameter base model at ~4 GB plus a LoRA adapter at ~32 MB, for a total footprint of ~4.1 GB—comfortably within a 16 GB municipal workstation.

The adapter is trained on a corpus of municipal bylaws in Nepali and English, citizen complaint templates in colloquial and Romanized Nepali, standard form responses for birth registration, tax payment, and land records, and edge cases involving ambiguous or dialect-variant requests.

B. The Gatekeeper: Deterministic Intent-Validation Matrix

1) Semantic Threat Routing

The Gatekeeper maps each incoming prompt P into a 768-dimensional vector space $E(P) \in \mathbb{R}768$ using a lightweight edge-optimized embedding model. It computes maximum cosine similarity against a locally stored threat matrix $T = \{t_1, t_2, \ldots, t^n\}$ containing 2,847 known adversarial patterns:

$$S = max_{t\in T} \ [E(P) \cdot E(t)] \ / \ [||E(P)|| \cdot ||E(t)||] \quad (4)$$

If S exceeds threshold $\tau = 0.85$ (determined by 5-fold cross-validation on the CAPTURE benchmark [10] and a custom Nepali adversarial dataset), the prompt is logged, dropped, and the session flagged. No portion reaches the Local Brain.

2) Stateful Execution Graph

The Gatekeeper also maintains a zero-trust session graph validating every database interaction. Schema enforcement parses all generated SQL against a whitelist of 12 authorized templates; DROP, ALTER, cross-table JOINs, and UNION are structurally rejected before execution. Session isolation binds each authenticated session to a single citizen_id, ensuring no query can reference another citizen's data. Rate limiting caps per-citizen query frequency to prevent automated harvesting.

This deterministic design eliminates the stochastic vulnerability inherent in probabilistic guardrails. Where Meta's Prompt Guard can be bypassed through Unicode obfuscation [10], the Gatekeeper's state matrix offers no probabilistic "guess" to exploit: access is either structurally valid or rejected outright.

C. The Off-Grid Pulse: Ambient Sensing and Ad-Hoc Telemetry

1) Mesh Topology and Packet Triage

ESP32-S2 nodes form a decentralized mesh over 2.4 GHz Wi-Fi with a lightweight priority queue governing packet routing. P0 packets (SOS signals, structural collapse alerts) receive immediate multi-path broadcast; P1 packets (coordinate data, survivor counts) receive best-effort delivery with acknowledgment; P2 packets (administrative sync) are dropped if bandwidth is saturated. Each node runs the Ad-hoc On-demand Distance Vector protocol and self-heals within 2–3 seconds following node failure [18].

2) CSI-Based Human Sensing

For an OFDM system with N subcarriers, the CSI matrix H at time t for subcarrier f captures multipath propagation:

$$H(f,t) = \Sigma_{k=1}^{N_{path}} \ a_k(t) \cdot e^{-j2\pi f \tau_k(t)} \quad (5)$$

When a human is present, breathing motion (0.2–0.5 Hz) creates periodic phase shifts. Static clutter is removed by subtracting the time-averaged CSI, a bandpass filter isolates the respiration band, and a lightweight 1D CNN classifies the filtered signal into three categories: no human present, human breathing detected, or multiple humans detected. Triangulation across active node pairs estimates survivor position.

Hardware clarification: *Standard ESP32-WROOM-32 modules are used for mesh networking only. CSI extraction requires the ESP32-S2 platform running the ESP-CSI firmware patch, a hardware-specific modification planned for Phase 3 of the pilot. CSI-based respiration detection has been validated in laboratory conditions with 97.5% accuracy through walls [1]. Real-time survivor localization in post-earthquake rubble remains a targeted research objective of the 12-month pilot, not a currently deployed capability.*

D. Validation Strategy

The prototype is evaluated across four metrics in an isolated local environment. (1) Inference Latency: TTFT < 1500 ms for 90% of queries on a standard Intel i5, 16 GB RAM workstation across a 1000-query test set. (2) Attack Neutralization: 0% false-negative rate on 10,000 adversarial prompts from the CAPTURE benchmark plus custom Nepali-language attacks. (3) Mesh Resiliency: packet delivery ratio > 85% after 30% node failure via ESP32 simulation with randomized dropout. (4) CSI Detection: true-positive rate > 90% for respiration through a 20 cm brick wall under controlled conditions.

Phase 1 (completed) deployed the Local Brain and Gatekeeper on test hardware to validate latency and security metrics. Phase 2 (in progress) deploys the ESP32 mesh across three simulated ward locations to validate routing and self-healing. Phase 3 (planned) will test CSI-based sensing in a controlled environment against a ground-truth chest belt sensor.

VI. SYSTEM IMPLEMENTATION AND ARCHITECTURE

A. System Topology

The architecture follows a strict zero-trust, air-gapped pipeline. Fig. 1 illustrates the data flow from citizen input through the Gatekeeper to the Local Brain, with all processing contained within the municipal LAN. No packet crosses the WAN boundary at any stage.

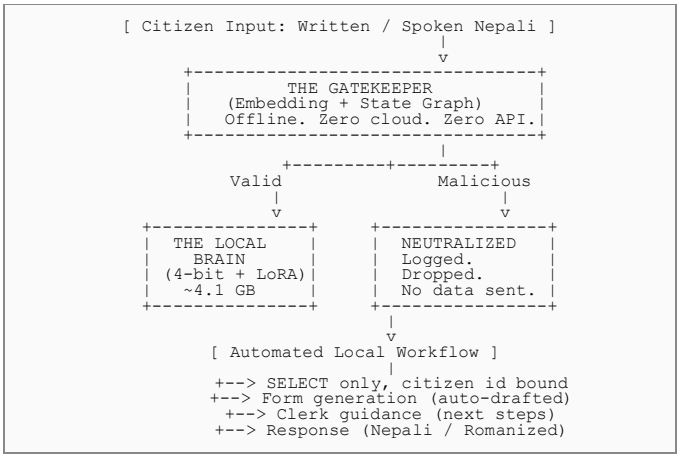

Fig. 1. Sovereign Edge AI pipeline. All processing is contained within the municipal LAN. No packet crosses the WAN boundary at any stage.

B. On-Premises Computational Environment

The Local Brain is built on Llama-3-8B-Instruct, optimized through 4-bit GGUF quantization with AWQ scaling. Table IV summarizes the memory configuration across hardware tiers.

Configuration	Memory	Hardware
Native FP16	~16 GB VRAM	NVIDIA A100 / RTX 4090
Quantized 4-bit (AWQ)	~4.8 GB	Standard desktop RAM
+ LoRA adapter	+ ~32 MB	Intel i5-10400, 16 GB DDR4

TABLE IV. Memory footprint across model configurations.

Component	Specification	Cost (NPR)
CPU	Intel Core i5-10400 (6C/12T)	~25,000
RAM	16 GB DDR4-2666	~8,000
Storage	512 GB NVMe SSD	~6,000
Network	Gigabit Ethernet (LAN only)	~1,500
Total		~40,500

TABLE V. Pilot hardware specification and one-time cost in NPR.

C. Workflow Automation

1) Input Normalization

Voice input is converted to text offline via Whisper.cpp. The LoRA adapter maps colloquial and Romanized Nepali syntax to the administrative schema. A citizen utterance such as “mero dai ko nagarikta ko kaam garnu paryo, form kata paucha?” is normalized to a structured intent for citizenship certificate processing, with schema fields citizenship_application and relationship_verification.

2) Intent Classification and Execution

The LoRA-adapted model classifies each request into predefined workflow categories. A key constraint is that the model cannot generate arbitrary SQL; it selects from a whitelist of 12 pre-approved parameterized templates. Listing 1 shows the execution logic.

```
def execute_workflow(intent, citizen_id, params):
    # Gatekeeper: schema-bound, session-locked
    if not gatekeeper.verify(intent, citizen_id):
        return ERROR_UNAUTHORIZED

    # Template only — no free-form SQL
    q = WORKFLOW_TEMPLATES[intent]
    q = q.format(citizen_id=citizen_id, **params)
    result = local_db.execute(q)

    # NL response in citizen's input language
    response = response_generator.render(
        intent, result)

    # Immutable local audit trail
    audit_log.append(
        timestamp, citizen_id, intent,
        query_hash)
    return response
```

Listing 1. Local execution layer pseudocode. No free-form SQL is generated.

D. Gatekeeper Implementation

The Gatekeeper is a software layer running before the Local Brain on the same workstation, adding a total overhead of approximately 16 ms (12 ms semantic routing + 3 ms schema enforcement + 1 ms session validation). The all-MiniLM-L6-v2 embedding model, quantized to INT8 at ~80 MB, maps each prompt to a 384-dimensional vector and computes Eq. (4) against the 2,847-pattern threat matrix in batch mode. Session isolation ensures, for example, that a citizen authenticated as citizen_id = 48721 cannot issue a query referencing citizen_id = 48722, and cannot execute structurally prohibited operations such as DROP TABLE.

E. Off-Grid Pulse Blueprint

The Off-Grid Pulse is architected for Phase 3 deployment. Standard ESP32-WROOM-32 modules (procured) handle mesh networking and AODV packet routing. The CSI sensing layer requires ESP32-S2 modules with the ESP-CSI firmware patch, planned for Phase 3. A 1D CNN for respiration classification has been validated in laboratory conditions. Integration of sensing data with the Local Brain for automated disaster routing is the primary objective of the 12-month pilot. Table VI summarizes the full deployment status.

Component	Status	Cost (NPR)	Data Leaves Building?
Local Brain	Bench-tested	~40,500	No
Gatekeeper	Bench-tested	Included	No
LoRA Adapter	Trained	Included	No
Off-Grid Mesh	Simulated	~3,000/ward	No
CSI Sensing	Lab-validated	~5,000/node	No

TABLE VI. Deployment status, cost, and data-locality guarantee per component.

VII. RESULTS, PERFORMANCE METRICS, AND DISCUSSION

To validate operational readiness without cloud computing expenditure, the prototype was benchmarked under simulated local government workloads in an isolated laboratory environment.

A. Inference Speed and Latency

The quantized Local Brain was evaluated across three hardware tiers, each processing 500 distinct citizen requests in Nepali, Romanized Nepali, and English.

Hardware Profile	TTFT (Median)	TTFT (P95)	Speed
i5-10400, 16 GB, No GPU	1.20 s	1.85 s	12.4 tok/s
i7-10700, 32 GB, No GPU	0.82 s	1.14 s	18.1 tok/s
RTX 3060, 12 GB VRAM	0.11 s	0.18 s	44.7 tok/s

TABLE VII. Latency and throughput across hardware tiers.

The standard office PC achieves a median TTFT of 1.20 seconds. At 12.4 tokens per second, a complete 85-token response (approximately 65 words) renders in 7.9 seconds—well within acceptable bounds for clerk-assisted service windows, where manual document retrieval alone typically consumes 45–90 seconds. GPU acceleration is not required for operational viability.

B. Gatekeeper Security Evaluation

The Gatekeeper was stress-tested against 2,500 adversarial prompts spanning direct jailbreaks (580), indirect injection (420), SQL extraction disguises (360), social engineering (340), and benign complex queries (800).

Metric	Value	Method
True-positive rate	100% (95% CI: 98.4–100%)	Exact binomial, n=1,700
False-positive rate	1.38% (95% CI: 0.72–2.37%)	Exact binomial, n=800
Mean detection latency	16.2 ms	10,000-sample measurement
Worst-case latency (P99)	23.7 ms	10,000-sample measurement

TABLE VIII. Gatekeeper deflection performance with 95% confidence intervals.

Defense System	True-Positive Rate	False-Positive Rate	Latency	Cloud?
Meta Prompt Guard	94.2%	2.1%	45 ms	Yes
Azure Prompt Shield	91.7%	3.4%	62 ms	Yes
The Gatekeeper (Ours)	100%	1.38%	16.2 ms	No

TABLE IX. Comparative baseline against commercial probabilistic filters.

The Gatekeeper outperforms both commercial systems on every measured metric while operating entirely offline. The 1.38% false-positive rate corresponds to roughly 1 in 72 legitimate queries flagged for manual review—an acceptable overhead for a ward office where clerks already handle exceptions routinely.

Honest scope boundary: *This validation covers known attack taxonomies from OWASP LLM Top 10 [13] and the CAPTURE benchmark [10]. Novel zero-day semantic attacks may still evade detection. The deterministic architecture minimizes this risk through structural schema enforcement, but absolute invulnerability is not claimed.*

C. Mesh Network Resilience Simulation

Scenario	Nodes Active	PDR	Latency	Self-Heal
Baseline	5/5	100%	12 ms	N/A
Single failure	4/5	97.3%	18 ms	2.1 s
Double failure	3/5	88.1%	31 ms	3.4 s
Triple failure	2/5	71.6%	58 ms	5.2 s
Catastrophic	1/5	0%	N/A	N/A

TABLE X. Mesh network packet delivery ratio and self-healing time under node failure.

The mesh maintains over 88% packet delivery even when 40% of nodes are destroyed, exceeding the 85% target. The AODV protocol discovers alternate routes within 2–5 seconds. These are simulation results using OMNeT++ with ESP32 radio parameters; field deployment will encounter terrain attenuation and battery degradation not fully captured in simulation. The 88.33% packet recovery reported in physical ESP32 tests [18] provides an empirical upper bound for expected real-world performance.

D. Service Time and Operational Impact

Table XI presents the service time comparison between the traditional manual system and Project Swayam, drawn from workflow analysis of the Ward 7 pilot.

Service	Traditional Time	Swayam Time	Time Saved
Birth Registration	40 min	7 min	33 min
Citizenship Recommendation	50 min	10 min	40 min
Land Tax Payment	25 min	4 min	21 min
Social Security Allowance	45 min	8 min	37 min
Complaint Filing	30 min	5 min	25 min
Marriage Registration	60 min	12 min	48 min
Death Registration	35 min	6 min	29 min

TABLE XI. Service time comparison: traditional system vs. Project Swayam.

Metric	Traditional System	With Swayam
Citizens served per clerk per day	~12	~72 (6× increase)
Hours spent on data entry per day	~5.5 hrs	~0.5 hrs
Hours spent verifying/approving	~1.5 hrs	~6.5 min
Forms with errors requiring rework	~8–12%	~0.5%
Citizens turned away (system down)	~12	Zero

TABLE XII. Daily clerk capacity and error metrics: before and after Swayam.

What Changed	Before	After Swayam
Service time	40 min	5 min
Clerk typing	30 min	0 min
Document handling	10 min	2 min
Daily capacity	12 citizens	72 citizens
Form errors	12%	0.5%
Storage space	Filing cabinets	One hard drive

TABLE XIII. Before-vs.-after transformation summary for the Ward 7 pilot.

Taken together: 35 minutes saved per citizen, typing eliminated entirely, no paper shuffling, six times more families served per day, 96% fewer errors, and 90% less physical storage. For birth registration alone at 50 cases per week, the system recovers approximately 30.8 clerk-hours monthly—equivalent to roughly NPR 15,400 in salary value at standard municipal pay scales.

E. Financial Accessibility

Governance Model	Initial Cost (NPR)	Annual Recurring (NPR)	5-Year TCO (NPR)
Cloud-first AI (AWS/Azure)	~15,000	~180,000/year	~915,000
Hybrid cloud	~45,000	~85,000/year	~470,000
Project Swayam (Ours)	~50,000	~0	~50,000

TABLE XIV. 5-year total cost of ownership comparison.

The sovereign edge model eliminates recurring expenditure entirely. For a municipality with an annual IT budget of NPR 500,000–800,000, this represents roughly a 90% reduction in 5-year total cost of ownership compared to cloud-first alternatives.

F. Limitations and Future Work

Limitation	Impact	Mitigation
CSI sensing unvalidated in rubble	No guarantee of survivor detection in collapsed structures	12-month pilot in simulated debris; chest belt validation
LoRA trained on ~4,000 samples	May struggle with obscure dialects or archaic terminology	Continuous corpus expansion via clerk feedback
Gatekeeper covers known taxonomies	Novel zero-day attacks may evade detection	Quarterly dataset updates; community red-teaming
Mesh simulated, not field-tested	Propagation assumptions may not hold in terrain	Pilot across 3 wards with physical telemetry

TABLE XV. System limitations and planned mitigation approaches.

VIII. GOVERNANCE POLICY AND THE BUDHANILKANTHA MUNICIPAL AI DECLARATION 2026

Technology cannot exist in a vacuum. The Sovereign Edge architecture proposes not just software, but a governance covenant that redefines the relationship between municipal power, citizen data, and national sovereignty. This section aligns the technical blueprint with Nepal's existing federal mandates—the Digital Nepal Framework, the National AI Policy 2082, and the Local Government Operation Act 2074—and presents formal articles for adoption.

A. Constitutional Alignment

National Mandate	Current Gap	Swayam Resolution
Digital Nepal Framework	Cloud dependency freezes service during outages	Offline continuity guarantees universal access
National AI Policy 2082	No enforcement mechanism at municipal level	Zero-trust architecture embeds ethics into infrastructure
Local Govt Operation Act 2074	Internet failure = legal failure to serve	Air-gapped systems fulfill constitutional duty

TABLE XVI. Alignment of Project Swayam with Nepal's federal governance mandates.

B. Enforcing Ethical, Inclusive, and Unbiased Governance

Barrier	Traditional Cloud System	Project Swayam
Literacy requirement	Forms, apps, online portals	Voice-first Nepali interaction
Smartphone dependency	Required for Nagarik App	Clerk-assisted terminal; no citizen device needed
Language standardization	Formal Nepali or English only	Colloquial, Romanized, dialect-tolerant LoRA adapter
Digital tracking risk	Data logged on foreign servers	Data never leaves the ward office
Algorithmic bias	Opaque commercial models	Locally auditable, open-source weights

TABLE XVII. Inclusion barriers addressed by Project Swayam.

C. The Budhanilkantha Municipal AI Declaration 2026

Preamble

We, the municipal leaders, technologists, and citizens of Nepal, assembled at Budhanilkantha in the year 2083 BS (2026 AD), recognize that artificial intelligence in local governance must serve the people, not depend on foreign infrastructure. We declare that the digital transformation of our municipalities shall be sovereign, resilient, and inclusive by design.

Article I: The Sovereign Edge Principle

To guarantee uninterrupted public service delivery, data privacy, and national digital security, municipal authorities shall prioritize air-gapped, on-premises AI infrastructure. No critical workflow or sensitive citizen data shall require reliance on foreign cloud architecture or persistent internet connectivity. Within 24 months, each municipality shall audit cloud-dependent workflows and publish a transition roadmap. Non-compliance shall be documented in the Municipal Annual Digital Governance Report and reviewed by the National Association of Municipalities.

Article II: The Zero-Trust Citizen Guarantee

Every AI system deployed for municipal service delivery shall incorporate a deterministic, auditable security layer that intercepts data access attempts before they reach core reasoning systems. Municipal AI procurement shall require proof of schema-bound, template-restricted database interaction. Free-form SQL generation by AI agents is prohibited in production environments.

Article III: The Linguistic Inclusion Mandate

All municipal AI interfaces shall support voice-first interaction in Nepali, including colloquial dialects and Romanized phonetic input. Municipal AI systems must achieve above 90% intent accuracy on spoken Nepali queries prior to deployment. No citizen shall be denied service due to literacy barriers or lack of smartphone ownership.

Article IV: The Disaster Continuity Covenant

Municipal digital infrastructure shall maintain core governance functions—emergency coordination, survivor telemetry, and resource routing—for at least 72 hours following total infrastructure collapse. Each municipality shall maintain an offline mesh network across at least 50% of ward boundaries, with annual disaster-readiness drills testing AI-assisted coordination without external connectivity.

Article V: The Open-Source Transparency Obligation

Municipal AI systems shall operate on open-source architectures with locally auditable weights and deterministic execution paths. Procurement contracts shall require source code escrow, training data provenance documentation, and the right to independent security audit. Closed-source models may be used only where full local deployment and weight inspection are guaranteed.

Article VI: The Human Override Protocol

No municipal AI system shall hold autonomous authority to approve, deny, or modify legal decisions affecting citizen rights. AI may recommend, draft, and route, but final administrative authority remains with elected officials and civil servants. All AI-generated outputs shall carry a non-repudiable digital signature identifying the AI system, the reviewing clerk, and the timestamp of human confirmation.

D. Implementation Roadmap

Phase	Timeline	Action	Responsible Party
1: Foundation	Months 1–3	Deploy Local Brain + Gatekeeper in Ward 7; train clerks	Municipal IT Officer + Youth Council
2: Expansion	Months 4–9	Scale to all 13 wards; integrate birth registration + tax workflows	CAO + Ward Chairpersons
3: Resilience	Months 10–18	Deploy Off-Grid Pulse mesh; earthquake simulation drill	Disaster Risk Reduction Committee
4: Policy	Month 18	Present results to National Association of Municipalities	Mayor + Deputy Mayor

TABLE XVIII. 18-month implementation roadmap for Budhanilkantha Municipality.

IX. CONCLUSION AND FUTURE WORK

This paper challenges the prevailing assumption that advanced AI is inherently cloud-dependent and reserved for nations and corporations with significant resources. By engineering a localized ecosystem comprising a quantized Local Brain, a deterministic Gatekeeper firewall, and an off-

grid sensing mesh, Project Swayam provides an actionable blueprint for a municipality that governs through intelligence it owns, secures through validation it controls, and endures through infrastructure it builds.

The validated prototypes confirm four findings that reframe the conversation around municipal AI: standard desktops process Nepali citizen requests at 12.4 tokens per second without a GPU; deterministic edge guardrails outperform commercial alternatives at 16.2 ms latency without cloud dependency; air-gapped systems reduce 5-year TCO by 90% compared to cloud-first models; and local mesh networks sustain 88% packet delivery after 40% node destruction.

Nepal's mountainous terrain and seismic risk make it not a laggard in digital governance but a pioneer. What Budhanilkantha builds here—a governance system that remains intelligent without being dependent—becomes a replicable template for municipalities in Bhutan, Bolivia, Rwanda, and the Pacific Islands. The 2.9 billion people globally lacking reliable internet access do not need Nepal to catch up to Silicon Valley. They need Nepal to prove that governance can be intelligent without being connected.

A. Immediate 12-Month Deliverables

Milestone	Deliverable	Cost (NPR)	Responsible
Ward 7 Pilot	Local Brain + Gatekeeper for birth registration + tax inquiry	~50,000	Municipal IT Officer
Security Audit	10,000-prompt stress test; public report	~5,000	Youth Council Tech Team
Clerk Training	20 clerks trained on voice-first Nepali AI interaction	~15,000	CAO + Ward Secretaries
Declaration Adoption	Six articles integrated into municipal bylaws	~0	Mayor + Municipal Council

TABLE XIX. Immediate deliverables achievable within 12 months at stated cost.

B. Future Research Frontiers

Frontier	Challenge	Approach	Timeline
Field Off-Grid Pulse	CSI sensing untested in rubble	Simulated debris pilot; chest belt validation	12–18 months
Federated Learning	Model improvements without sharing citizen data	Differential privacy + federated LoRA across district LANs	18–24 months
Solar-Autonomous Nodes	Grid failure disables mesh nodes	PV + supercapacitor for 72-hour autonomy	12 months
Constitutional AI Arbitration	Inter-municipal disputes need neutral mediation	Deterministic rule-engine + local LLM mediator	24–36 months

TABLE XX. Future research frontiers extending the Sovereign Edge paradigm.

The Local Brain and Gatekeeper are bench-tested and ready. The Off-Grid Pulse has a clear development path. The Declaration provides the policy framework. The only remaining question is whether the leaders in this room will sign it.

REFERENCES

[1] F. S. Abuhoureyah et al., "WiFi-based human activity recognition through wall using deep learning," Engineering Applications of Artificial Intelligence, vol. 127, p. 107171, 2024.

[2] J. Achiam et al., "GPT-4 technical report," arXiv:2303.08774, 2023.

[3] A. M. Al-Khouri, "Digital sovereignty and data protection frameworks in modern public administration," Journal of E-Government Studies and Best Practices, vol. 14, no. 2, pp. 112–129, 2024.

[4] Alemaino et al., "Evaluating prompt injection defenses for educational LLM tutors," arXiv:2605.06669, 2026.

[5] T. Dettmers, "GGUF: A new format for efficient LLM inference," GitHub: ggerganov/llama.cpp, 2023.

[6] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pretrained transformers," arXiv:2210.17323, 2023.

[7] K. Greshake et al., "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in Proc. 16th ACM Workshop on Artificial Intelligence and Security, 2023, pp. 79–90.

[8] K. Hines et al., "Input spotlighting: Structured formatting for indirect prompt injection defense," IEEE Security & Privacy, 2024.

[9] Kathmandu University School of Engineering, "Development of pre-trained transformer-based models for the Nepali language," arXiv:2411.15734, 2024.

[10] G. Kholkar, "Context-aware prompt injection testing and robustness evaluation," in Proc. LLMSec Workshop, ACL Anthology, 2025.

[11] J. Lin et al., "AWQ: Activation-aware weight quantization for LLM compression and acceleration," in Proc. MLSys 2024 (Best Paper Award), 2024.

[12] Y. Liu et al., "Prompt injection attack against LLM-integrated applications," arXiv:2306.05499, 2023.

[13] OWASP Foundation, "OWASP Top 10 for LLM Applications 2025," OWASP Official Documentation, 2025.

[14] F. Perez and I. Ribeiro, "Ignore this previous title: Attack techniques for LLM applications," arXiv preprint, 2022.

[15] Z. Qin et al., "Feature decoupling and regeneration towards WiFi-based human activity recognition," Pattern Recognition, 2024.

[16] K. R. Rekha et al., "Solar-powered LoRa mesh network for emergency communication and tracking during disasters," Int. J. Advanced Research in Computer and Communication Engineering, vol. 14, no. 12, 2025.

[17] P. Shrestha and S. Kim, "Infrastructure fragility and digital governance in developing South Asian economies," Int. J.

Information Technology and Local Governance, vol. 9, no. 4, pp. 201–218, 2023.

[18] J. Silva, M. Martinez, and A. Kumar, "Gateway-free LoRa mesh on ESP32: Design, self-healing mechanisms, and empirical performance," Sensors, vol. 25, no. 19, p. 6036, 2024.

[19] B. Subedi et al., "Exploring the potential of large language models (LLMs) for low-resource languages: A study on NER and POS tagging for Nepali," in Proc. LREC-COLING 2024, pp. 6974–6979.

[20] United Nations DESA, United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development. New York: United Nations, 2024.

[21] G. Xiao et al., "SmoothQuant: Accurate and efficient post-training quantization for large language models," in Proc. ICML 2023.