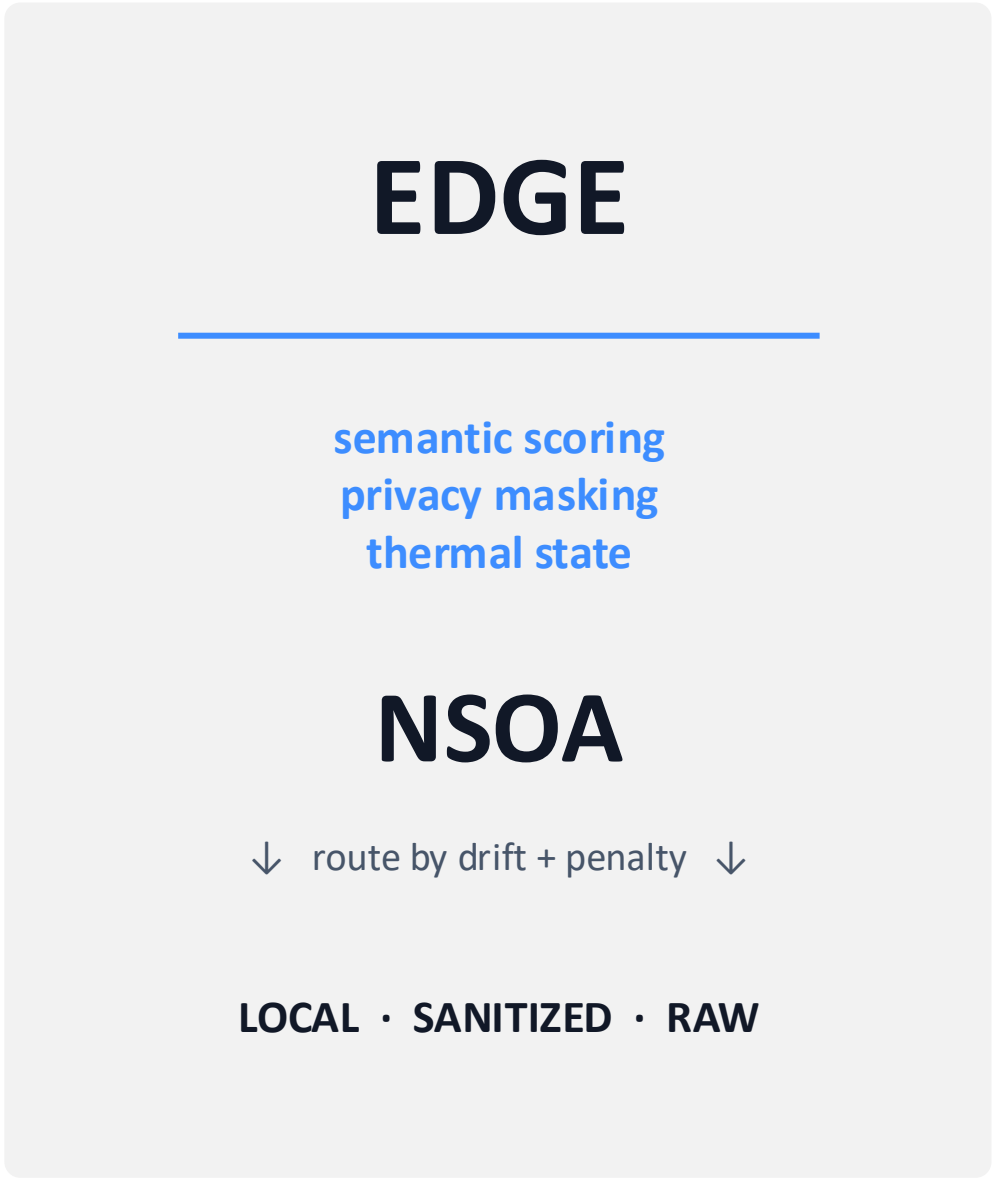
Stochastic AI Orchestration for Neural Semantic-Aware Mobile Edge Inference

Samuel Romero[1] · Jörn Altmann[2] · Bernhard Egger[1]

[1] Seoul National University [2] Lucerne University of Applied Sciences and Arts

fortin@snu.ac.kr



Cloud intelligence, edge security - pick two?

FULL CLOUD

High reasoning capacity and elastic compute

Sensitive corporate data must leave the premises.

Cybersecurity, data-integrity, and compliance exposure increases.

A central router would need to receive the prompt before routing it.

FULL LOCAL

Data stays private, but capability is constrained

Mobile devices have limited compute, battery, and thermal capacity.

Complex generative tasks can exceed local reasoning capability.

Sustained NPU inference can trigger thermal throttling.

78%
organizations use AI in at least one function

25%
enterprise GenAI apps predicted to face repeated minor incidents by 2028

60%
global firms expected to split AI stacks across sovereign zones by 2028

Context figures are cited in the 's Introduction.

RESEARCH GAP

Existing methods solve only half the problem

STREAM	WHAT IT CONTRIBUTES	WHAT REMAINS MISSING
Lyapunov offloading	Online decisions and long-term queue stability	Prompts are treated as homogeneous packets; semantics are ignored.
DNN partitioning	Splits neural layers across edge and server	Autoregressive generation would add a network trip for every output token.
Hardware-aware control	Models energy, memory, clocks, and thermal cost	No semantic privacy gate tied to the physical state.
Semantic / LLM routing	Reads task meaning and can triage workloads	No Drift-plus-Penalty proof for privacy and thermal stability.

Missing combination: on-device semantic scoring + Lyapunov routing + commercial mobile NPU

The gap becomes three testable questions

RQ1

PRIVACY

To what extent does semantic privacy scoring within a Lyapunov framework reduce the privacy violation rate?

RQ2

LATENCY

To what extent does Lyapunov routing produce statistically differentiated latency across edge and cloud execution lanes?

RQ3

THERMAL

To what extent does the virtual thermal queue correctly track and stabilize NPU temperature under sustained inference load?

Research questions reproduced from the , Section 2.

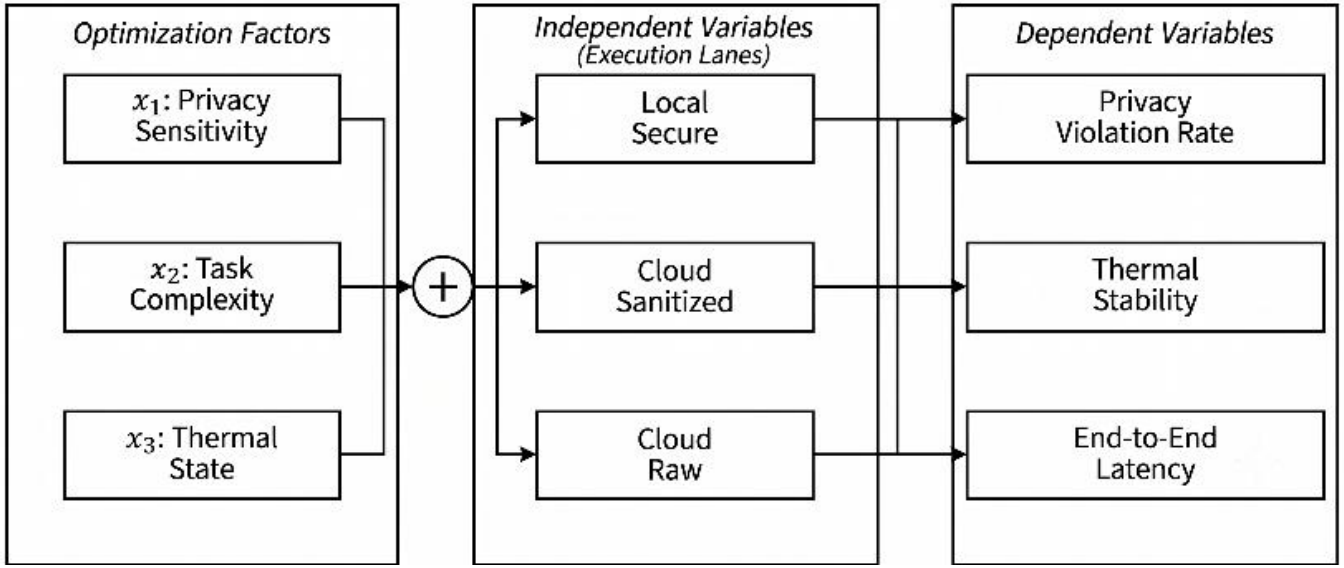
NSOA joins semantics with stable control

THREE CONTRIBUTIONS

Formal control. Prompt routing becomes a Lyapunov Drift-plus-Penalty problem with privacy as an unbounded multiplier.

Real implementation. Privacy scoring and PII masking run on a Snapdragon 8 Elite NPU, not in simulation.

Measured evidence. A 300-event evaluation quantifies routing, privacy, latency, and thermal queues.



Decision unit: one complete prompt

The request stays whole; its destination changes.

x1 privacy sensitivity · x2 task complexity · x3 thermal state

Source: Fig. 1 and Introduction contributions.

Four layers place orchestration on-device

LAYER	ROLE	EXPERIMENTAL IMPLEMENTATION
L1	**Application View**	Prompt as a TOSCA node: data class, PII flags, compute needs, and SLA constraints; Qwen3-4B x1 score plus keyword matching.
L2	**Resource View**	EdgeResource and CloudResource expose live temperature, availability, latency, and trust; both are resampled every 10 prompts.
L3	**Trusted Knowledge**	Peer DIDs and Logical Proximity combine latency, trust deficit, and energy; this layer also hosts the Ztherm virtual queue.
L4	**Orchestration Space**	The NSOA consumes L1-L3 state, applies Drift-plus-Penalty, and executes through the Qwen NPU or Groq LPU API.

Source: Table 1.

DATA PATH

Nothing leaves before privacy is assessed

Privacy classification happens before any cloud transmission.

If Cloud Sanitized is selected, PII masking is also completed on-device before offloading.

Source: Section 3.1 architecture rationale and Table 1.

Lyapunov control balances stability and cost

AT EACH TIME SLOT, MINIMIZE

$$t: \Delta(Z(t)) + V \cdot E[P(t)]$$

Δ(Z(t))

DRIFT

Change in virtual queues, including thermal buildup(x3); the controller pushes long-term instability down.
Drift = (x3{60}−42) = 0.60
30

V

CONTROL PARAMETER

Tunable weight between immediate performance and hardware protection.

E[P(t)]

PENALTY

Immediate cost of latency(x2) and privacy(x1) exposure for the selected action.

***Plocal**=latency weight{0.001}×estimated local ms{40}+quality weight{2}×example x2{0.70} =1.44*
***Psanitized**=privacy weight{4}×example x1{0.20}×remaining risk after masking{0.05}+latency weight{0.001}×(example network ms{300}+masking estimate{50}) = 0.39*
Praw=privacy weight{4}×example x1{0.20}+latency weight{0.001}×example network ms{300}−quality weight{2}×example x2{0.70} *= - 0.30*

Sensitive prompt + public cloud → P(t) approaches infinity → every other priority is overridden

Source: Equation (1), Section 3.1.

OPTIMIZATION MODEL

Three factors determine every route

x1

PRIVACY SENSITIVITY

Qwen3-4B via Genie SDK

High x1 makes Cloud Raw mathematically unreachable; the privacy penalty dominates.

x2

TASK COMPLEXITY

tiktoken cl100k_base + semantic keywords

High complexity raises the local quality penalty and favors a cloud lane when x1 permits.

x3

THERMAL STATE

Live NPU temperature → Ztherm

Rising heat increases local routing cost; cloud shedding is triggered above 78°C.

Source: , Section 3.2.

SETUP

Three lanes preserve privacy without isolation

LANE	EXECUTION PATH	PRIVACY GUARDRAIL
LOCAL SECURE	**Qwen3-4B inference on the Snapdragon 8 Elite NPU**	x1≥0.85 - forces local-only processing
CLOUD SANITIZED	**PII masking on-device, then inference through the Groq LPU API**	0.5≤x1<0.85 - used for the CONFIDENTIAL label and violation measurement
CLOUD RAW	**Direct Groq LPU inference for near-zero-sensitivity public content**	x1<0.50 - After-the-decision evaluation: Cloud Raw would be recorded as a privacy violation

The testbed uses commercial mobile hardware

EDGE DEVICE

Samsung Galaxy S25 Ultra

Snapdragon 8 Elite NPU
Qwen3-4B via Genie SDK
Scoring, masking, local inference

ORCHESTRATION HOST

Python NSOA

adb prompt delivery
/sys/class/thermal sampling
Latency and event collection

CLOUD ENDPOINT

Groq LPU API over Wi-Fi

llama-3.1-8b-instant
Round-trip latency observed
Sanitized or raw inference

3 independent sessions × 100 prompts

Ztherm reset between sessions

Peer DID trust: edge 0.95 · cloud 0.70

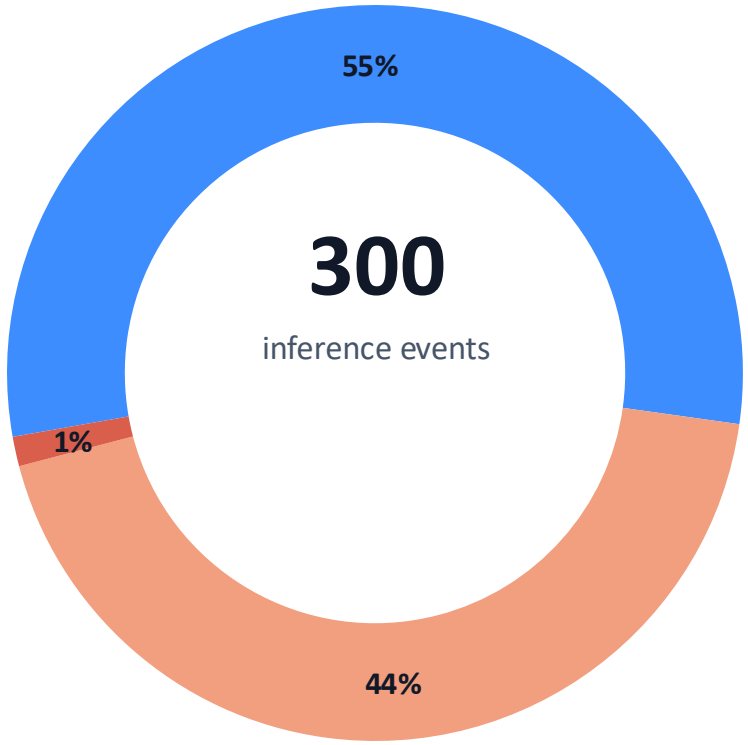
Energy: edge 0.30 · cloud 0.70

Resources resampled every 10 prompts

Live state, not a simulation

Source: Table 2 and Section 4.1.

Routing used all three lanes



LOCAL SECURE 55.0% SANITIZED 43.7% RAW 1.3%

Lane	S1	S2	S3	Total
Local Secure	62%	48%	55%	55.0%
Cloud Sanitized	38%	50%	43%	43.7%
Cloud Raw	0%	2%	2%	1.3%
Violations	0%	0%	0%	0.00%

Stable across sessions

Local Secure: 48-62% · Cloud Sanitized: 38-50%

All lanes were exercised; Cloud Raw remained deliberately rare.

Source: , Section 4.2 and Table 3.

RQ1 - To what extent does semantic privacy scoring within a Lyapunov framework reduce the privacy violation rate?

Privacy Routing Policy

0 / 300

privacy violations under the adopted policy

18 / 18

RESTRICTED prompts routed to Local Secure

84 / 84

prompts with x1 >= 0.60 excluded from Cloud Raw

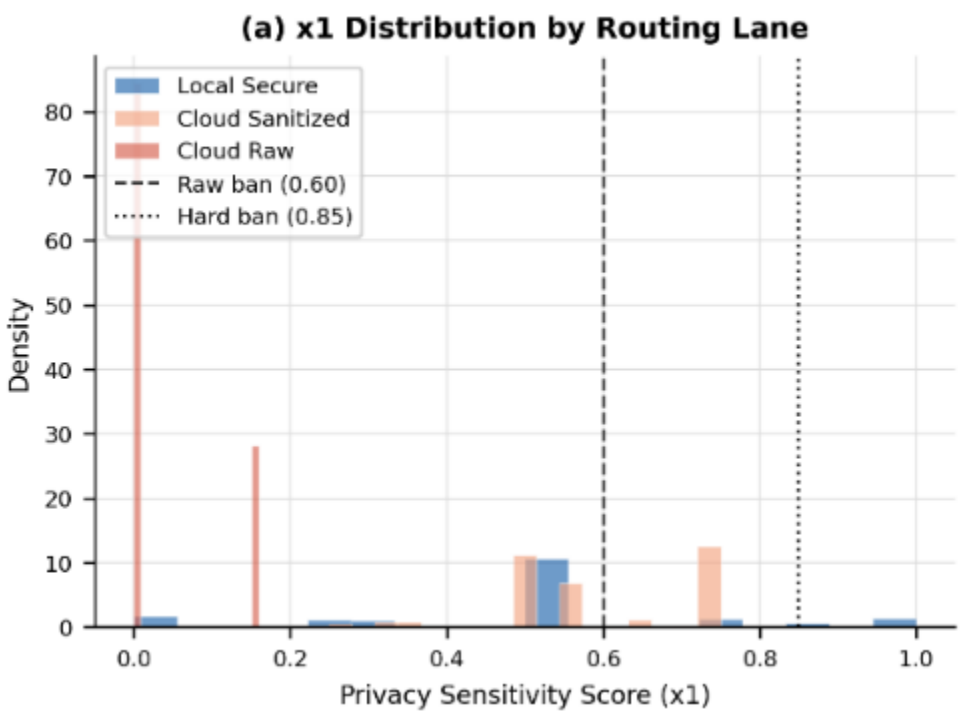
Static full-cloud counterfactual	NSOA result	Interpretation
28.0% theoretical policy violations	**0.00%**	**the result certifies that routing obeyed the x1-based policy**

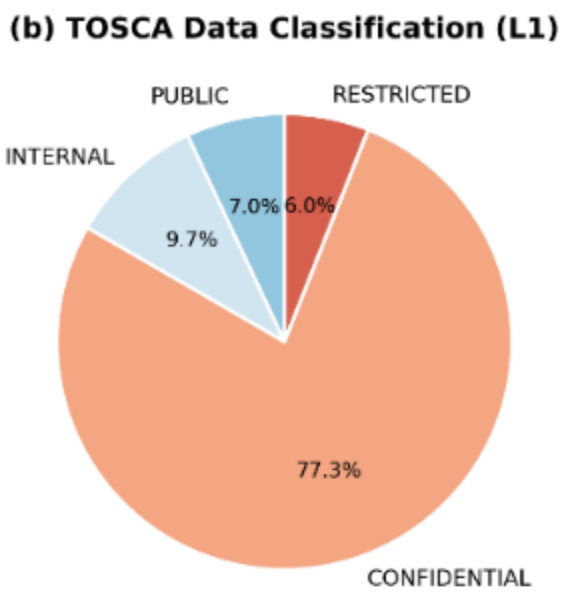
Important: this tests enforcement fidelity against x1 - not whether the classifier assigns x1 correctly.

Source: , Section 4.3 and Table 4.

RQ1

Sensitive prompts never reached Cloud Raw





WHAT THE FIGURE SHOWS

Raw lane: entirely below x1 = 0.20;

Policy boundaries: raw-cloud ban at 0.50 and hard local-only ban at 0.85.

Sanitized mean: x1 = 0.594; complex sensitive work retains cloud access after masking.

Corpus mix: 77.3% CONFIDENTIAL and 6.0% RESTRICTED.

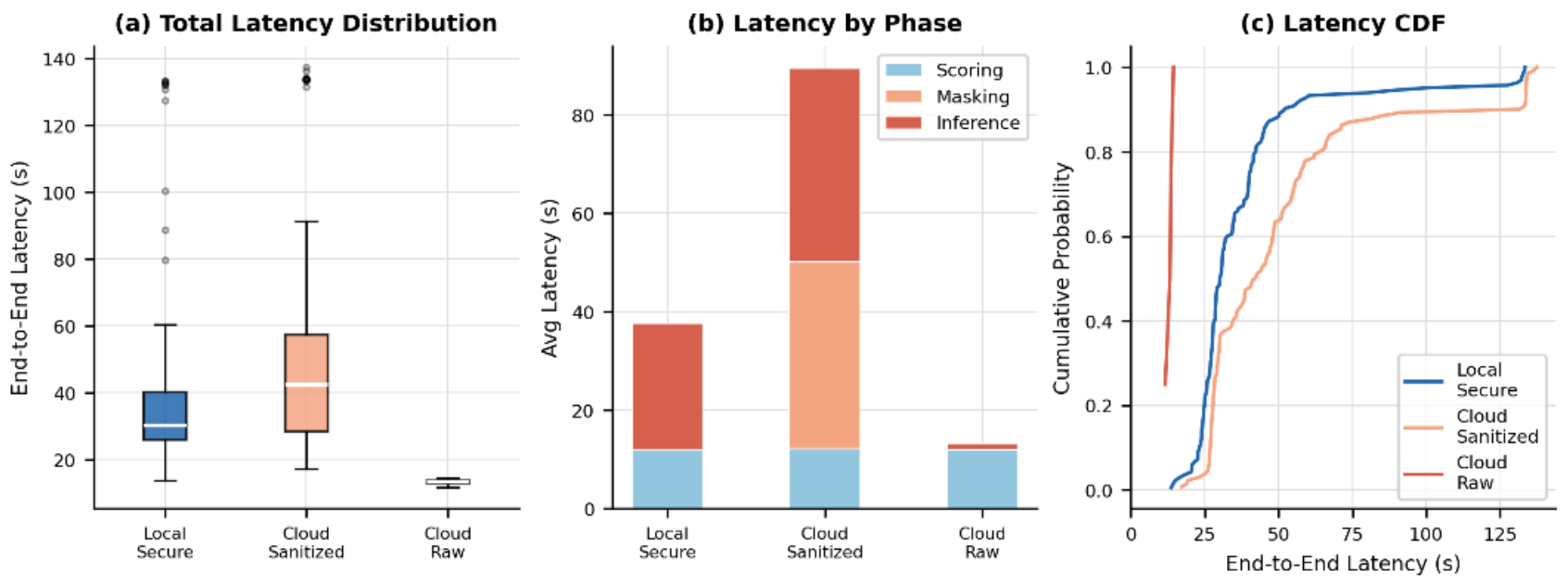
Not a classifier test

The score and policy use the same x1 signal.

Source: , Fig. 2 and Section 4.3.

RQ2 - To what extent does Lyapunov routing produce statistically differentiated latency across edge and cloud execution lanes?

Latency differs across all lanes, unlikely to be caused by random variation



13.2 s
Cloud Raw mean

38.0 s
Local Secure mean

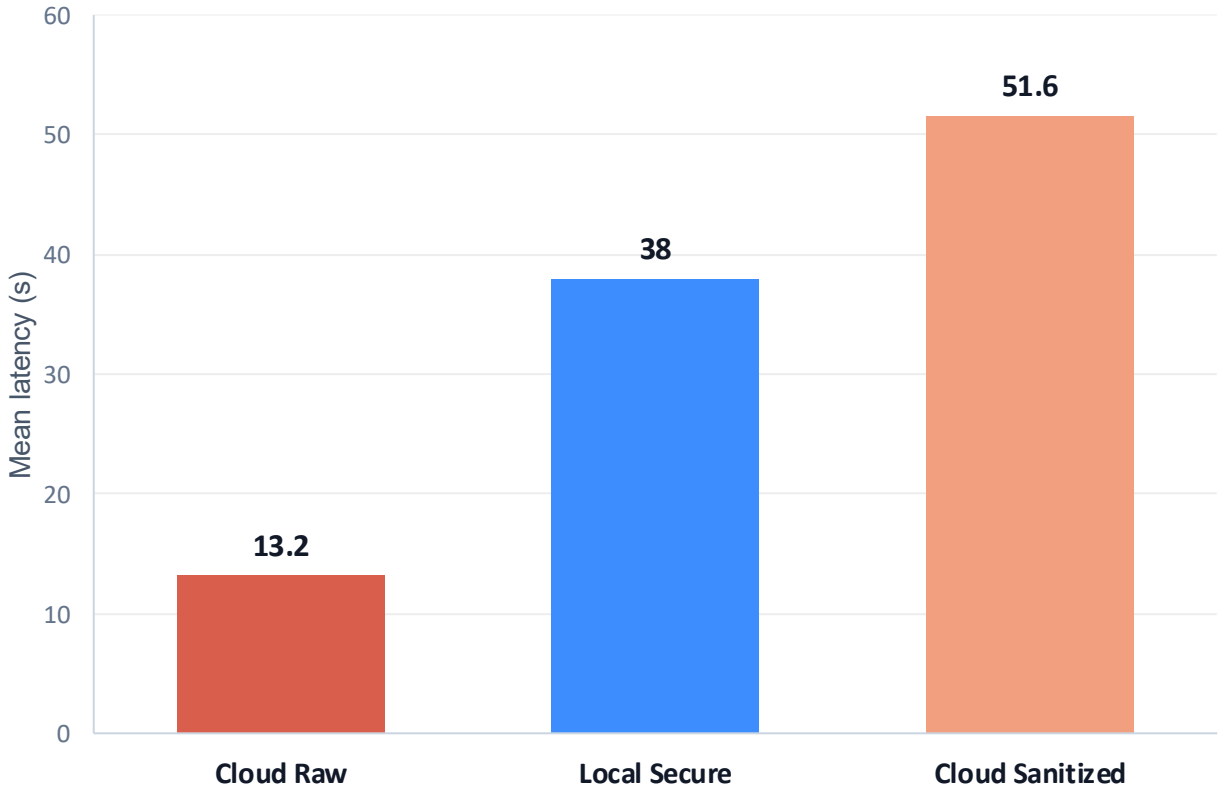
51.6 s
Cloud Sanitized mean

Kruskal-Wallis H = 38.019, p < 0.0001 · all pairwise Mann-Whitney comparisons p < 0.001

Source: Fig. 3, Table 5, and Section 4.4.

RQ2

Privacy protection explains the latency order



WHERE TIME IS SPENT

~6 s — On-device x1 scoring is the dominant fixed cost across all lanes.

~14 s — Cloud Sanitized adds on-device PII masking before the cloud round trip.

1.19 s — Groq LPU inference average for Cloud Raw.

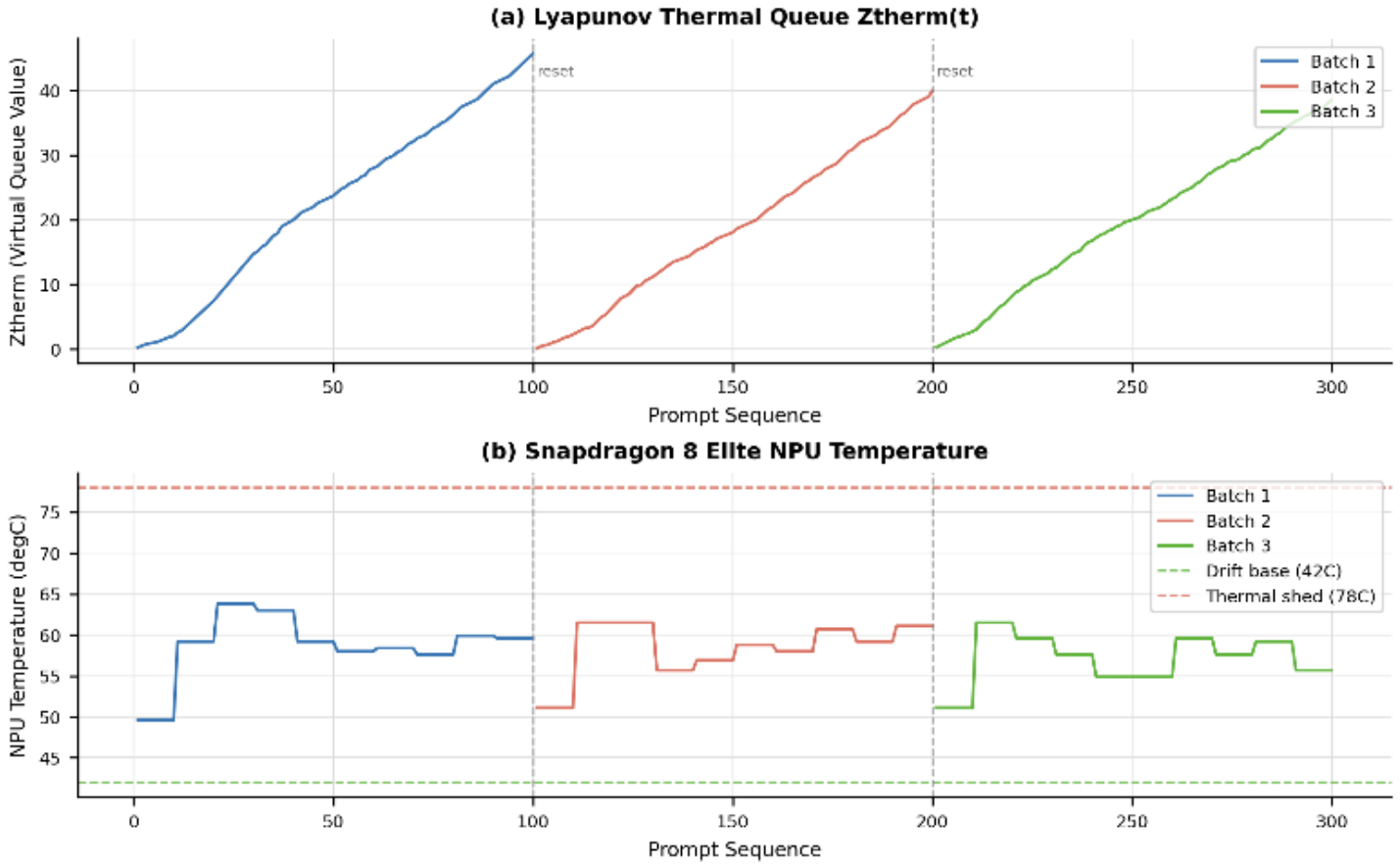
SPEARMAN CORRELATIONS

x2 vs. local inference: rho = +0.449, p < 0.001
x1 vs. total latency: rho = +0.278, p < 0.001

At 13-52 seconds mean latency, the prototype suits background batch or scheduled analysis - not interactive replies.

RQ3 - To what extent does the virtual thermal queue correctly track and stabilize NPU temperature under sustained inference load?

The thermal queue tracked real heat



49.6-63.8°C
observed NPU range

78°C
shedding threshold - never triggered. If both true x3≥78◦C and x1<0.60, local is penalized, pushing the prompt to cloud

45.710
maximum Ztherm peak in Session 1

Queue-hardware coupling

rho = +0.138, p < 0.05

Session averages: 58.8°C · 58.5°C · 57.2°C | Ztherm peaks: 45.710 · 39.985 · 38.494

BOUNDARIES

Four limitations define the next experiments

WHAT THIS STUDY DOES NOT YET ESTABLISH

Classifier accuracy. 0.00% measures enforcement fidelity against x1, not precision or recall.

Lyapunov-only benefit. The baseline is static full-cloud; regex, fixed-threshold, and greedy latency-first baselines are absent.

Thermal shedding. The 78°C threshold was never reached; the maximum was 63.8°C.

Generalization. 300 prompts, one device, one NPU architecture, and one cloud provider.

PRIORITIES FOR THE NEXT EVALUATION

Run controlled thermal stress with elevated ambient temperature and minimal inter-prompt idle time.

Replace the 4B scorer with an embedding-based classifier.

Measure labelled privacy-classification precision and recall.

Compare head-to-head with regex and threshold routers.

Extend to more devices, cloud providers, and runtime swarm discovery.

The next experiments for the journal article should separate policy enforcement, classifier quality, and thermal-control performance.

Source: , Limitations and Future work.

CONCLUSION

The evidence supports adaptive edge-cloud AI

A decentralized NSOA can enforce a privacy policy per prompt while retaining both local execution and cloud capability.

0.00%

policy violation rate across 300 events

55.0%

workloads retained on the local NPU

49.6-63.8°C

stable observed NPU temperature range

On-device semantic masking + Lyapunov control offers a practical alternative to static cloud-versus-local policies.

Source: , Abstract, Discussion, and Conclusion.

Q&A

Stochastic AI Orchestration for Neural Semantic-Aware Mobile Edge Inference

Samuel Romero · fortin@snu.ac.kr

Complete decision process

1. Read the current state: x_1, x_2, temperature x_3, network latency, and Z_{therm}.
2. Calculate P_L, P_S, and P_R.
3. Multiply each penalty by V.
4. Add the normal thermal-drift and Z_{therm} components.
5. Apply the privacy restrictions:
    1. $x_1 \geq 0.85$: allow only Local Secure.
    2. $0.50 \leq x_1 < 0.85$: exclude Cloud Raw.
6. Check the emergency thermal threshold:
    1. If $x_3 < 78°\text{C}$: continue normally.
    2. If $x_3 \geq 78°\text{C}$ and $x_1 < 0.50$: add a huge penalty to Local Secure, pushing the prompt to a cloud lane.
    3. Never bypass privacy restrictions merely because the phone is hot.
7. Choose the allowed lane with the smallest final score.
8. Execute the selected lane.
9. Update Z_{therm} according to the selected lane for the next prompt.