Integrated Guardrails for Unbiased and Adaptive Neural Network Architectures

Decoupling Safety Telemetry from Core Inference

Pradeeban Kathiravelu[1] **Tihana Galinac Grbac**[2]

[1]University of Alaska Anchorage, Anchorage, Alaska, USA
[2]Juraj Dobrila University of Pula, Pula, Croatia

12th IEEE International Symposium on Systems Engineering (IEEE ISSE 2026)
Dubrovnik, Croatia
September 22-24, 2026

Introduction & Motivation

- **Rapid AI Adoption:** Neural Network (NN) architectures are increasingly deployed in high-stakes fields like *healthcare* and *financial services*.
- **Generative AI (GenAI) Risks:** Autoregressive models (e.g., LLMs) can produce biased, toxic, or unaligned outputs.
- **Role of Safety Guardrails:** Intercept and filter unsafe behavior.

Traditional Guardrail Architecture

State-of-the-art guardrails (e.g., OpenAI Moderation API, Llama Guard, Nvidia NeMo) operate as **external, monolithic supervisors** executing *post-hoc, synchronous checks*.

Limitations of Current Guardrails

1. Selective Refusal Bias

- Keyword-based static filters trigger false-positive refusal responses.
- Disproportionately blocks benign requests related to minority dialects or sensitive medical queries.
- Inadvertently creates secondary marginalization/bias.

2. Context Blindness

- Rigidity to cultural nuances.
- Inability to dynamically adjust safety thresholds based on conversation history or user trust (e.g., authenticated clinician vs. layperson).
- Vulnerability to zero-day "jailbreak" exploits.

Performance Overhead: Synchronous blocking introduces severe generational latency.

Research Questions & Contributions

Research Questions

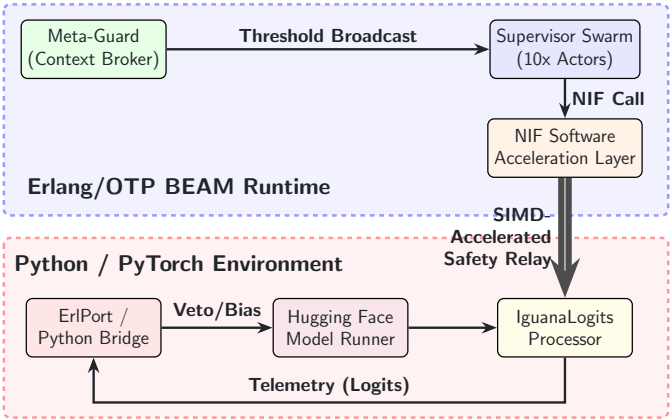
- **RQ1:** Can we decouple NN guardrails from the inference engine to dynamically adjust context-specific strictness without incurring significant generational latency?
- **RQ2:** Can such an adaptive approach mitigate the bias observed in the current guardrails?

Core Contributions of IGUANA

- **C1 (Decoupled Architecture):** Erlang-based asynchronous decentralized guardrail framework using actor concurrency.
- **C2 (Adaptive Alignment):** Dynamic thresholding via real-time Shannon entropy evaluation.
- **C3 (Performance Layer):** Hardware-level SIMD-accelerated Native Implemented Functions (NIFs).

The IGUANA Architecture

- **Decoupled Termination:** Intercepts probabilities at decoding stage without modifying model weights.
- **BEAM VM:** Concurrently processes evaluations.
- **Out-of-band Loop:** No Python GIL bottlenecks or thread priority inversion.



Out-of-Band Concurrency & Fault Tolerance

Asynchronous Messaging Loop

- Hugging Face `LogitsProcessor` extracts Top-K ($K = 100$) probabilities.
- Serializes token info across the ErlPort boundary asynchronously via non-blocking `gen_cast`.
- Swarm processes telemetry in parallel with the model's next GPU-bound forward pass.

Fault Tolerance & Scalability

- **"Let It Crash" Philosophy:** Configured with a `one_for_one` restart strategy for the supervisor tree. Failure isolation prevents Python inference engine crashes.
- **Scale-Out Safety:** Mesh-connected cluster via Erlang's `net_kernel`. Dynamic worker tracking with process groups (pg) to route telemetry to least-loaded nodes.
- **Security Boundary:** Node communication encrypted via TLS; cookie-based cluster authentication.

Mathematical Formulation & Soft Correction

- **Entropy Telemetry:** The supervisor swarm evaluates the Shannon entropy $H(P)$ of the token distribution:

$$H(P) = -\sum_{i \in V} P(x_t^i) \log_2 P(x_t^i) \tag{1}$$

- **Dynamic Thresholds:** Safety veto (τ_v) and Bias injector (τ_b) are scaled based on the initialized user `trust_score`.

Scenario A: $H(P) < \tau_v$ (Critical Risk)

Triggers an immediate `veto_token` sequence interrupt, zeroing logits to force an End-Of-Sequence (EOS) yield.

Scenario B: $H(P) > \tau_b$ (Bias/Uncertainty)

Computes SkewPNN soft bias vector:

$$B_t = A_2(C) \times \Phi\left(\frac{x - \xi}{\omega}\right) \tag{2}$$

Steers logits pre-softmax without weight mutation.

Deployment Modes & Latency Mitigation

Handling Asynchronous Token Propagation Delay

Safety evaluation of token t runs in parallel with the forward pass of token $t+1$. A veto event might experience a 1-token propagation lag. IGUANA supports two solutions:

1. **Zero-Leak Compliance (Synchronous Mode):** Python bridge blocks synchronously on Erlang evaluation. Sacrifices throughput for strict prevention.
2. **Streaming UI Buffer Mode:** Client-side UI buffers $N = 1$ or 2 tokens. Purges objectionable tokens before they are rendered, preserving full asynchronous throughput.

Orthogonality to Speculative Decoding

IGUANA is highly synergistic with speculative decoding. Candidate sequences are evaluated asynchronously by the Erlang VM, avoiding GPU execution stalls during validation.

Preliminary Assessments: Latency & Throughput

Workload Simulation (LLaMA-7B)

- **Baseline (Synchronous):**
  - Latency: **69.09 ms**/token
  - Throughput: **14.47 tokens/s**
- **IGUANA (Asynchronous):**
  - Latency: **26.03 ms**/token (BEAM VM)
  - Throughput: **38.40 tokens/s**
  - Speedup: **2.65x** throughput increase
- **IPC Overhead:** Averages just **1.8 ms**.
- **Top-K Optimization:** Reduced telemetry footprint from $\approx$ 250 KB to $\approx$ 0.8 KB (300x IPC reduction).
- **NIF Speedup:** SIMD compiler autovectorization yields **1.27x Speedup**.

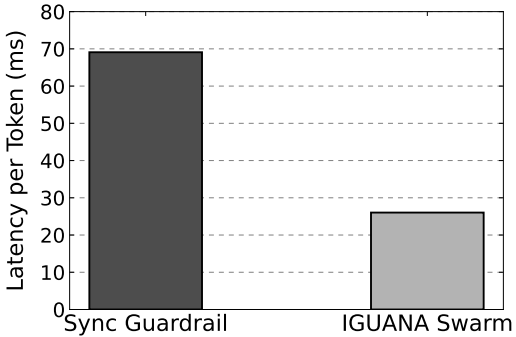


Figure: End-to-end token latency comparison.

Preliminary Assessments: Verification & Bias

Correctness Verification

Verified via 10-case EUnit / Common Test correctness suite.

- Validated entropy calculations.
- Confirmed process lifecycles.
- Verified cluster swarm handshake, dynamic scalability, and sync mode.

Demographic Bias Mitigation (VAL1)

Soft-correction rebalancing via SkewPNN + A2 reduces bias manifestation in generated corpora by **38.64%** without inducing selective refusal.

ID	Component	Objective	Res.
TC1	Entropy Formula	Asserts uniform distribution entropy	✓
TC2	Entropy	Asserts deterministic entropy $= 0.0$	✓
TC3	Threshold Logic	Asserts veto when $H(P) < \tau_v$	✓
TC4	Threshold Logic	Asserts accept when $\tau_v \leq H(P) \leq \tau_b$	✓
TC5	Lifecycle	Asserts gen_server init & queries	✓
TC6	State Mutation	Asserts trust_score updates τ_b	✓
TC7	Math Fidelity	Asserts Skew-Normal purity	✓
TC8	Cluster Swarm	Asserts cluster node handshakes	✓
TC9	Adaptiveness	Asserts dynamic A2 parameter sync	✓
TC10	Sync Mode	Asserts sync block mode evaluation	✓
VAL1	Bias Efficacy	Verifies 38.64% reduction in bias	✓

Table: EUnit & Common Test Results

Multi-Domain Use Cases: Healthcare & Water-Food

Clinical Healthcare & Mental Health

- Mitigates diagnostic biases across patient demographics.
- Distinguishes layperson queries from professional clinical expert queries.
- Avoids blocking help-seeking wellness requests by steering away from clinical overreach rather than categorical blocking.

Water-Food Circular Economy

- Dynamic "water-on-demand" decision support.
- Reshapes generation to reflect multi-stakeholder needs (utilities, agronomy, regulation).
- Evaluates nutrient benefits and potential contaminant risks (CECs, POPs) dynamically without applying static reuse rules.

Multi-Domain Use Cases: Finance & Legal

Financial Services & Risk Assessment

- Low-latency automated risk disclosure and investment advising.
- Avoids demographic/geographic credit biases in small-business reviews.
- Parallel evaluations handle high concurrent-session demands without request queuing.

Legal & Governmental Support

- Mitigates historical socio-economic and demographic skews in draft sentencing/parole recommendations.
- Rebalances output probability without rigid, blocking blocklists, retaining access to legal research.

Discussion & Future Work

Discussion

- **Functional Concurrency:** BEAM VM processes safety telemetry out-of-band, preserving highly optimized, GPU-bound Python ecosystems.
- **Soft Logit-Level Correction:** Confining safety intervention to pre-softmax logit scores avoids the complexities of attention-weight mutations or subnetwork pruning.
- **Attestation Layering:** Trust score determination is handled at the application layer to keep the guardrail core domain-agnostic.

Future Outlook

- Integrating the SIMD-accelerated NIF kernel with high-density GPU clusters to explore scaling limits on petascale models.
- Incorporating structured session onboarding to map declared user roles to normalized trust scores $[0.0, 1.0]$.

Conclusion

- **IGUANA:** A decoupled, asynchronous, and adaptive safety guardrail framework for unbiased neural network architectures.
- **Performance:** Eliminates sequential execution bottlenecks, reducing per-token delay to **26.03 ms** and increasing throughput **2.65x** to **38.40 tokens/s**.
- **Adaptation:** Leverages entropy-guided SkewPNN soft corrections to reduce bias manifestation by **38.64%** while mitigating selective refusal.
- **Fault Tolerance:** Natively implements the BEAM actor model for isolated recovery and resilient distributed mesh operations.

Acknowledgements

This work is funded by the EU NextGeneration under the Juraj Dobrila University of Pula institutional research project number IIP_UNIPU_010162.

Project webpage: `https://tgalinacgrbac.github.io/equisys/`