Hamiltonian Research

High-Performance, Sovereign Machine Learning Systems

Engineering next-generation non-autoregressive and state-space architectures optimized from the hardware-level up to deliver highly efficient, secure, and fully autonomous local model deployments.

THE PROBLEM

The Enterprise AI Bottleneck

- **Quadratic KV-Cache Scaling**
  Autoregressive Transformers scale quadratically ($O(N^2)$) with context length, creating massive memory bottlenecks and driving up operational compute costs at scale.
- **Sequential Decoding Wall**
  The token-by-token generation bottleneck forces memory-bound execution, severely limiting GPU hardware utilization and introducing high latency.
- **Sovereign and Data Risk**
  Reliance on external cloud APIs or closed source models introduces unacceptable latency pipelines, data residency vulnerabilities, and high dependency risks for regional enterprise infrastructure.

THE SOLUTION

Next-Generation Inference Architectures

- **Non-Autoregressive (NAR) Design**
  Bypassing the sequential decoding bottleneck by generating entire sequences or structures in parallel, reducing the execution loop from O(N) to O(1) or small, highly optimized iterative steps.
- **Linear-Time State Space Models**
  Leveraging state-of-the-art SSMs (like Mamba-2) that bypass attention mechanisms entirely, maintaining a constant-size recurrent state and ensuring linear scaling (O(N)) for long-context windows.
- **Hardware-Aware Co-Design**
  Custom CUDA kernel engineering and highly performance-oriented local backends designed specifically to maximize physical hardware efficiency and throughput on local GPU clusters.

THE SYSTEM ADVANTAGE

Unmatched Throughput and Local Efficiency

Low-Latency Parallel Pipelines

Parallel sequence generation eliminates traditional autoregressive decoding overhead, yielding up to 10x improvement in generation throughput.

Computational Resource Optimization

Linear memory complexity enables local deployments of complex models on standard local enterprise hardware, cutting cloud hosting and infrastructure requirements by up to 80%.

Native Sovereign Footprint

System runs fully locally on physical regional hardware, bypassing cloud API dependencies, ensuring air-gapped security, and securing regional data autonomy.

TRACTION AND MILESTONES

Grounded Technical Validation

01 **Open-Source Deployment**

Released model weights for hr-diffuse-1-nano (288M parameters) on Hugging Face, validating the core architectural viability of non-autoregressive discrete diffusion models.

02 **Systems Research**

Completed and published technical preprints detailing custom CUDA kernel performance and hardware-aware optimizations for non-autoregressive modeling.

03 **Ecosystem Recognition**

Invited guest speaker at the "Make It in the Emirates" forum, presenting edge AI solutions and high-performance sovereign computing to UAE government and industrial leaders.

THE TEAM

Faris Allafi

Solo Systems Architect

Independent machine learning systems researcher specializing in high-throughput architectures, custom CUDA kernel design, and low-latency systems engineering. Proficient in high-performance Rust systems development and hardware-level optimizations. Youngest speaker in "Make It in the Emirates" forum history.

MISSION

Abu Dhabi Enterprise Sandboxes

- **Localized Inference Pilots**
  Partnering with regional enterprise and government institutions to deploy highly secure, energy-efficient local inference sandboxes.
- **Accelerating UAE Deep Tech**
  Securing absolute data sovereignty by scaling model deployments natively on local, physical hardware, completely bypassing foreign cloud bottlenecks.