Trajectory Guard

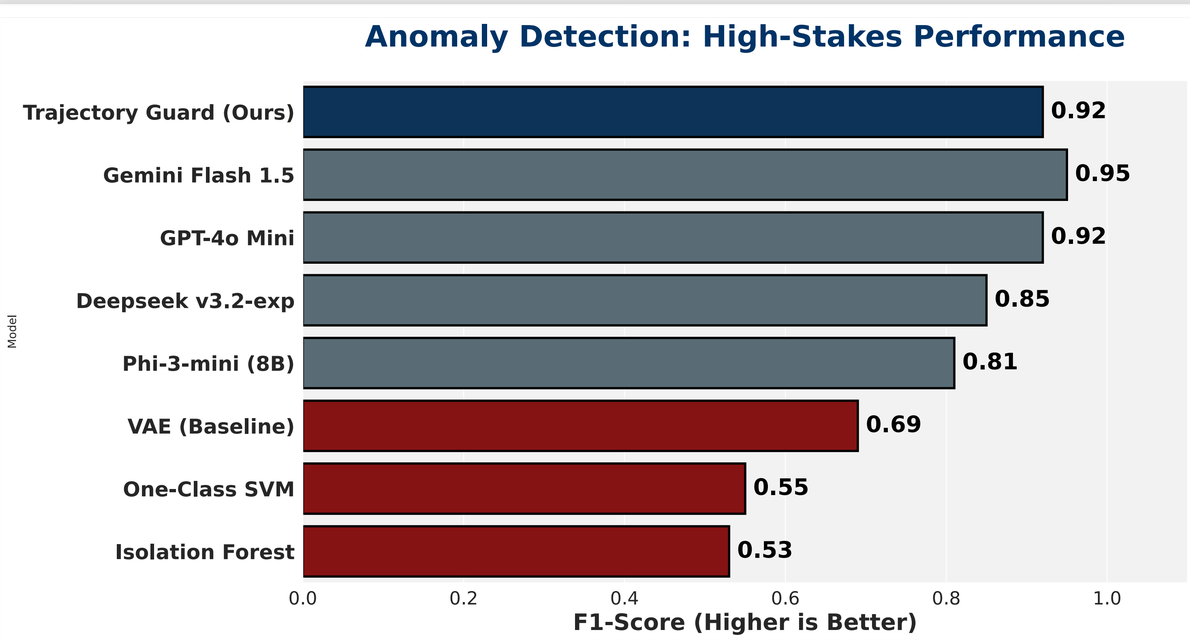
A Lightweight, Sequence-Aware Model for Real-Time Anomaly Detection in Agentic AI

Laksh Advani - laksh.advani@colorado.edu

University of Colorado **Boulder**

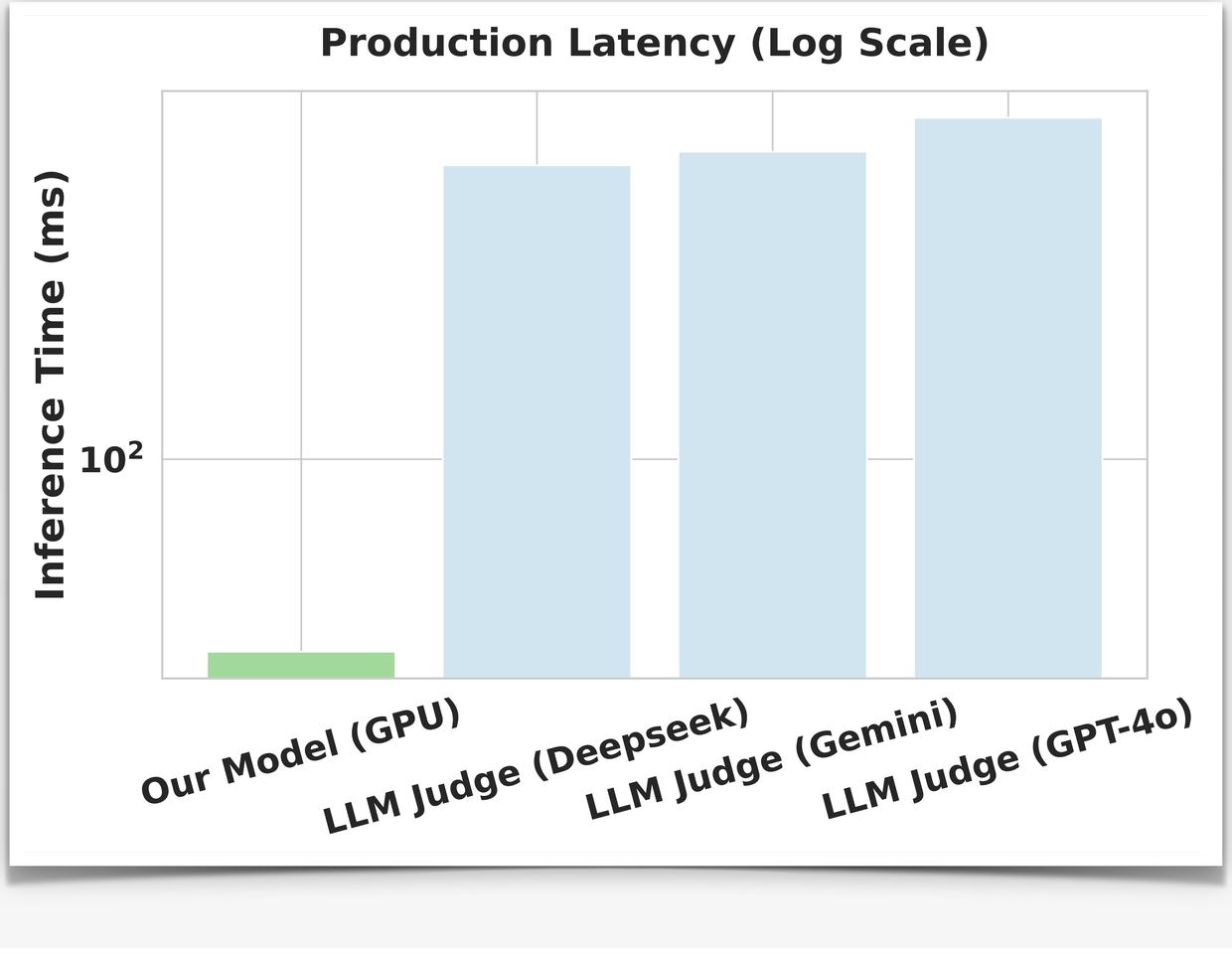
Trajectory Guard

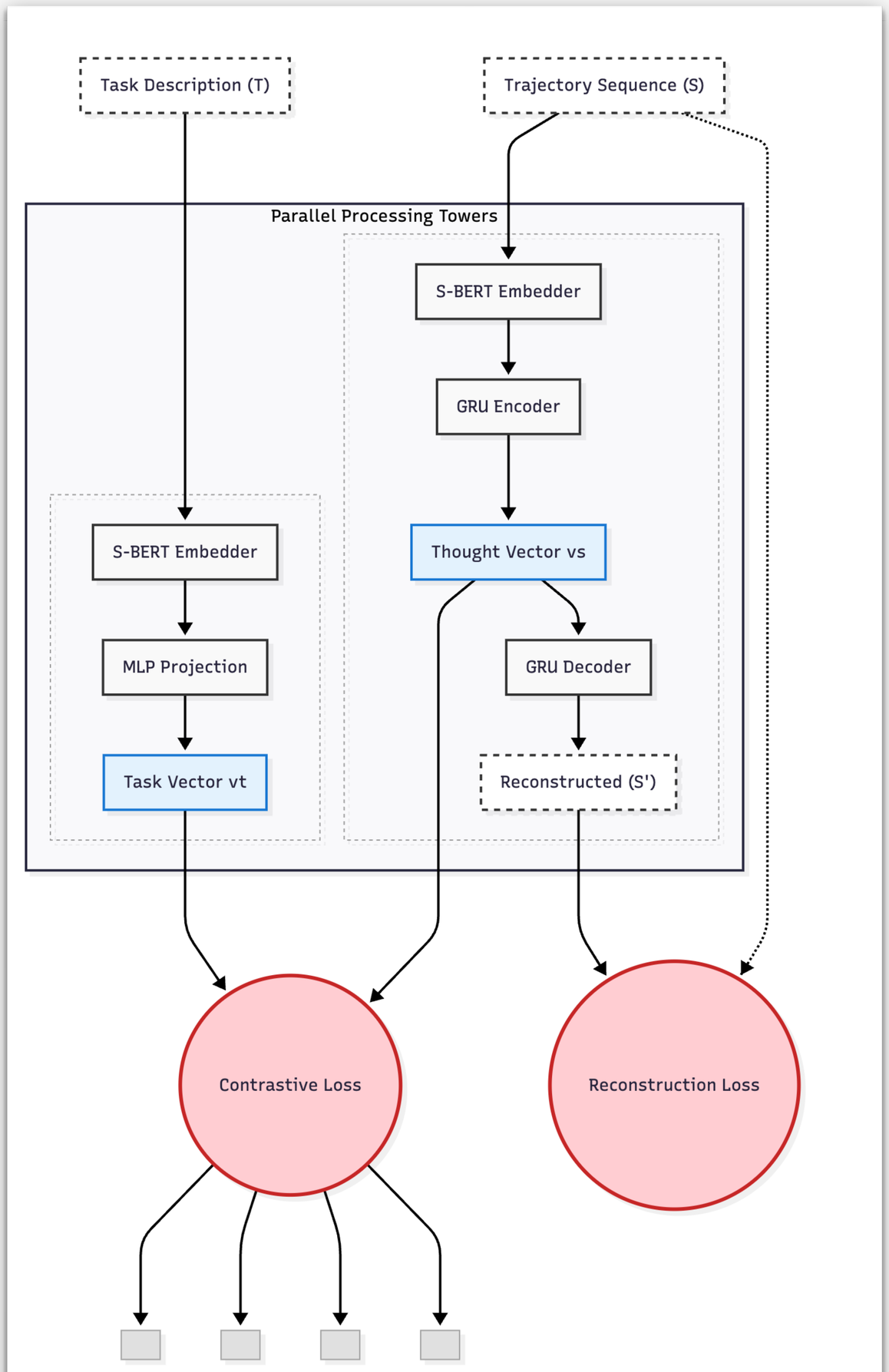
- **Core Problem:** Autonomous LLM agents generate multi-step "trajectories" to solve tasks, but these often fail due to semantic misalignment or structural incoherence.
- **The Problem with Baselines:** Standard unsupervised methods (VAEs, Isolation Forests) fail to capture the semantic and sequential nuances of valid plans, with F1-scores limited to 0.69.
- **The Latency Barrier:** Heavyweight "LLM Judges" provide high accuracy but introduce 556–735 ms of latency, making them unsuitable for real-time production safety guards.
- **The Trajectory Guard Solution:** We introduce a lightweight Siamese Recurrent Autoencoder that validates agent plans in 32 ms—enabling real-time "Trust Alerts" before execution.



Architecture

- **Design Rationale:** A robust safety guard must distinguish between contextual anomalies (mismatch between task and plan) and structural anomalies (illogical or incoherent steps).
- **Siamese Recurrent Autoencoder:** The model employs a dual-tower design to process the task and the action sequence in parallel:
- **Task Tower:** An MLP projection that maps task embeddings into a 128-dimensional latent vector.
- **Trajectory Tower:** A GRU encoder compresses the sequence into a "thought vector," while a GRU decoder reconstructs the sequence to verify its structural "grammar."
- **Dual-Objective Loss Innovation:** The model is trained using a Hybrid Objective L_{total} that combines two distinct safety signals: Contextual Alignment (via the Siamese contrastive loss) and Structural Validity (via reconstruction error).





Conclusion

- **Real-Time Latency:** Trajectory Guard achieves a 32.48 ms inference latency, representing a 17.1x–22.6x speedup over traditional LLM Judge baselines.
- **Infrastructure Efficiency:** Our architecture maintains high throughput on commodity hardware (NVIDIA T4), outperforming significantly larger models (e.g., Phi-3-mini) hosted on A100-class compute.
- **High-Stakes Reliability:** The model demonstrates high sensitivity, maintaining Recall between 0.86 and 0.92 on real-world failure logs—minimizing the probability of Type II errors (false negatives) in safety-critical deployments.
- **Generalization Power:** The system exhibits zero-shot transferability, successfully validating diverse benchmarks (RAS-Eval for Security; Who&When for Multi-agent logs) without the need for domain-specific fine-tuning.

Verification at Scale

In the era of autonomous agents, unverified accuracy is merely a coincidence: Trajectory Guard enables deterministic safety in real-time production environments.