Contribution ID: **18** Type: **Poster**

From Cloud to Clarity: Architecting Resilient AI Systems in Enterprise Environments

Monday 13 October 2025 19:10 (20 minutes)

Delivering AI systems in enterprise settings requires more than model optimization —it demands infrastructure clarity, cross-functional orchestration, and the ability to navigate complexity over time. This poster highlights real-world lessons from deploying intelligent systems within a cloud-native architecture, combining applied technical experience with strategic delivery in organizational environments undergoing transformation.

The work draws from two significant experiences: the completion of the **UC Berkeley Professional Certificate in Machine Learning and AI**, which provided a strong foundation in applied AI practices; and active participation in the **2025 NVIDIA GTC Hackathon**, where real-time problem-solving and innovation under pressure shaped the architecture of resilient, scalable ML pipelines.

The poster explores the intersection of infrastructure, AI delivery, and organizational complexity through a practitioner lens. Topics include:

1. **Infrastructure design** using GCP, Kubernetes, and hardened microservices
2. Techniques for building AI systems that can scale across teams and environments
3. Patterns for **security**, **observability**, and operational clarity at deployment
4. Navigating delivery friction in systems where multiple priorities compete

Many teams attempt machine learning projects, but few sustain them to completion. This poster examines why some AI efforts quietly stall, even with strong talent and intent, and what's required to close that gap.

Drawing from the personal experience of completing a rigorous internal nanodegree, this work surfaces the often-invisible factors that determine whether AI systems succeed —including follow-through, detachment from noise, and **infrastructure maturity**.

In addition to modeling static user baselines, the system leverages **sequential modeling** techniques to capture time-ordered patterns in user activity. By understanding not only what a user does but when and in what sequence, the model can detect context-aware anomalies—such as unusual login behavior or compromised session flow—that would go unnoticed with traditional static rules. This approach supports long-term adaptability and improves detection of subtle shifts in behavior across evolving enterprise systems.

This poster will be of interest to engineers, data scientists, and research leaders who seek to operationalize AI in large organizations —particularly those undergoing cloud migration or seeking to create lasting research-ready infrastructure. It connects theory to delivery and highlights the critical infrastructure layer where sustainable AI impact begins.

Primary author: CHEN, Weiying

Presenter: CHEN, Weiying

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: Infrastructures to Support Data-Intensive Research - Local to Global