[Sample Preview] Engineering Trustworthy AI

Subtitle: Why Accuracy is the Greatest Lie in Modern Engineering

Author: Olaitan Almaroof

1. The 99% Failure

Imagine a fraud detection system with **99.2% accuracy**. In most boardrooms, that's a standing ovation. In a high-stakes production environment, that 0.8% error rate isn't just a margin of error—it's a catastrophic vulnerability.

If that 0.8% represents your highest-value clients or a specific geographic region (like the Lagos example in Chapter 1), your "accurate" model is actually a business-ending liability.

The Accuracy Paradox

Traditional metrics like the **Confusion Matrix** (Precision, Recall, F1-Score) are static snapshots. They tell you how the model performed on *yesterday's* data. They do not tell you how the model will behave when:

1. **The world changes** (Data Drift).
2. **An attacker notices the pattern** (Adversarial Decay).
3. **The infrastructure fails** (Systemic Noise).

2. The Trust Stack: A New Engineering Requirement

To build AI that survives the "wild," we must move beyond the model and look at the **Trust Stack**. A trustworthy system isn't just "smart"; it is architected across three layers:

Layer	Focus	Engineering Guardrail
Safety	Reliability & Harm	**Circuit Breakers:** If $Confidence < 0.7$, revert to heuristics.
Security	Integrity & Defense	**Adversarial Hardening:** Testing against prompt injection and poisoning.
Sovereignty	Compliance & Ethics	**Data Residency:** Ensuring logic respects GDPR/NDPA boundaries.

3. Case Study: The "Ghosting" Borrower

In Kenya, a microlending AI saw a sudden drop in smartphone pings and flagged thousands of users as "intentional defaulters."

The Reality: A regional telecom outage. **The Engineering Failure:** The model lacked "Contextual Awareness." **The Fix:**Implementing **Multi-modal Fallbacks**. When digital footprint data (Signal A) becomes erratic, the system's "Fairness Circuit Breaker" automatically shifts weight to historical repayment (Signal B).

The Lesson: Trustworthy AI doesn't just make a guess; it knows when it *can't* make a guess and seeks a safer alternative.

4. What’s Inside the Full Book?

This sample is only the beginning. The full edition of *Engineering Trustworthy AI* provides the technical blueprints for:

- **Defending the Stochastic Engine:** Hardening LLMs against prompt injection.
- **Explainability as Security:** Using SHAP and LIME to audit "Black Box" decisions.
- **The Red-Button Architecture:** When and how to kill an autonomous agent.
- **Governance at Scale:** Integrating NIST AI RMF into your CI/CD pipeline.

5. Early Praise for Engineering Trustworthy AI

"Finally, a book that moves past the AI hype and addresses the gritty reality of production-grade systems. A must-read for any Principal Engineer."