



LLM

Building Quality into LLM-powered applications



What is an LLM-powered application



CODE WRITTEN
LARGELY BY AI

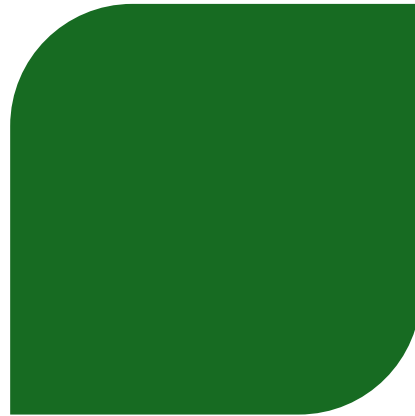


APPLICATION THAT
USES AI WIDELY FOR
OUTPUT

Future Proofing Your Build



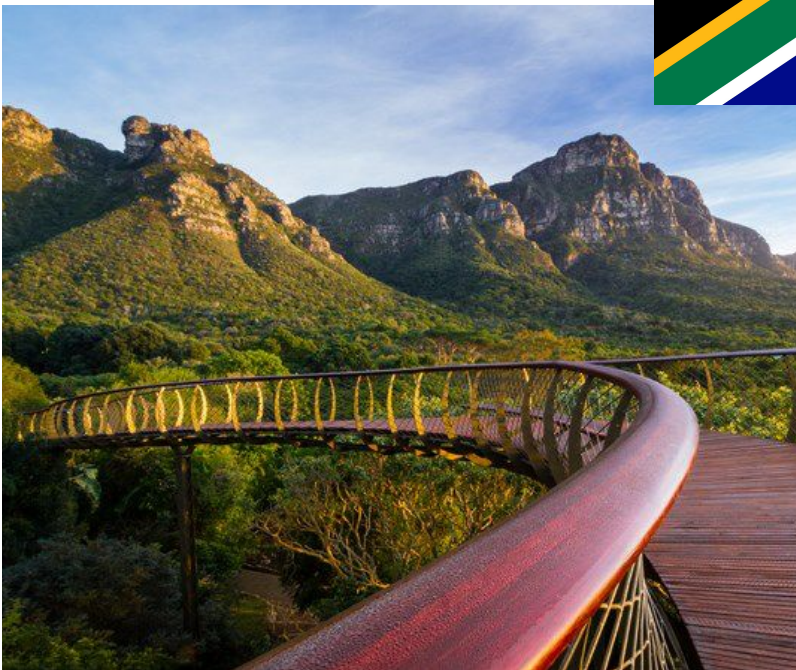
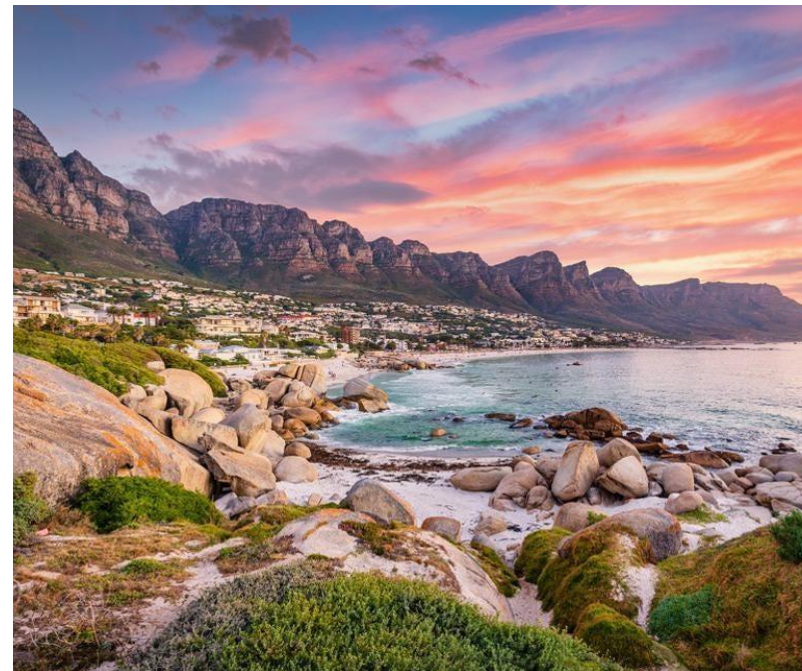
UNDERSTANDING SOFTWARE
DEVELOPMENT IS CHANGING

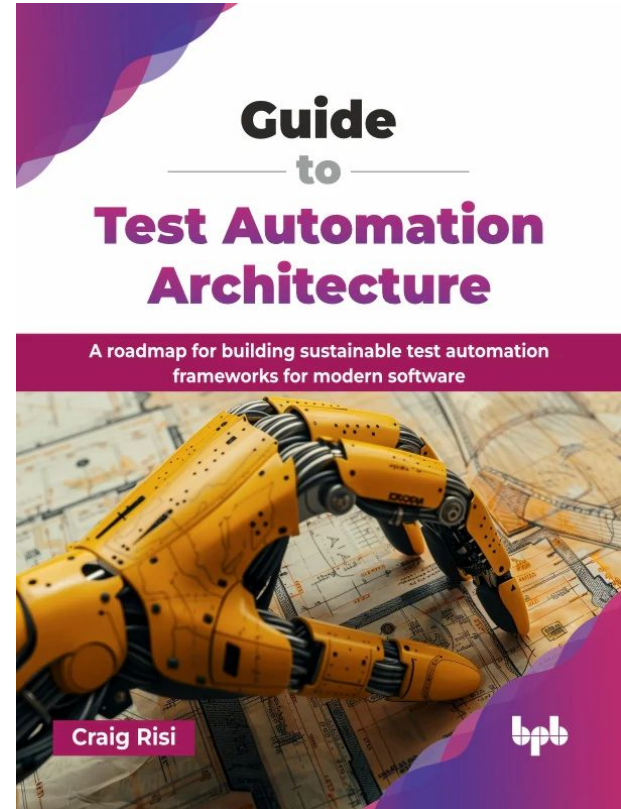
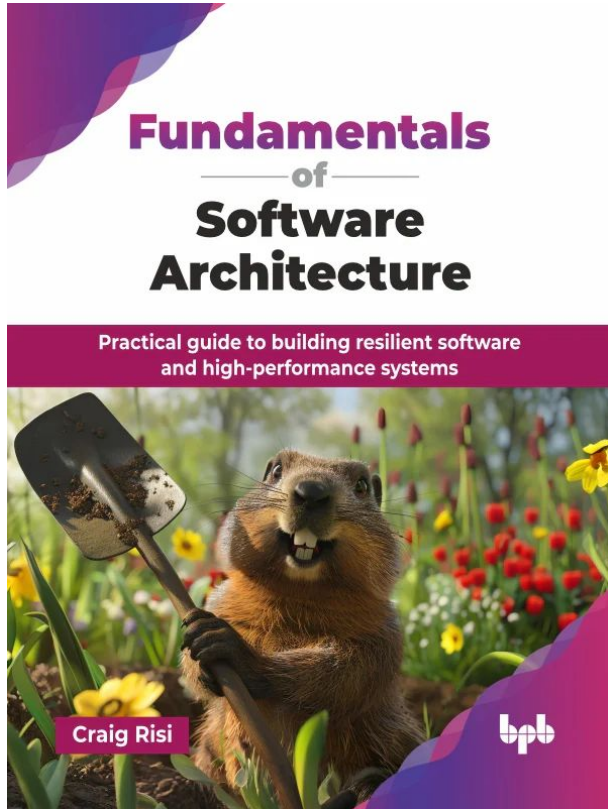


CHANGING THE LENS ON
DEFINING QUALITY



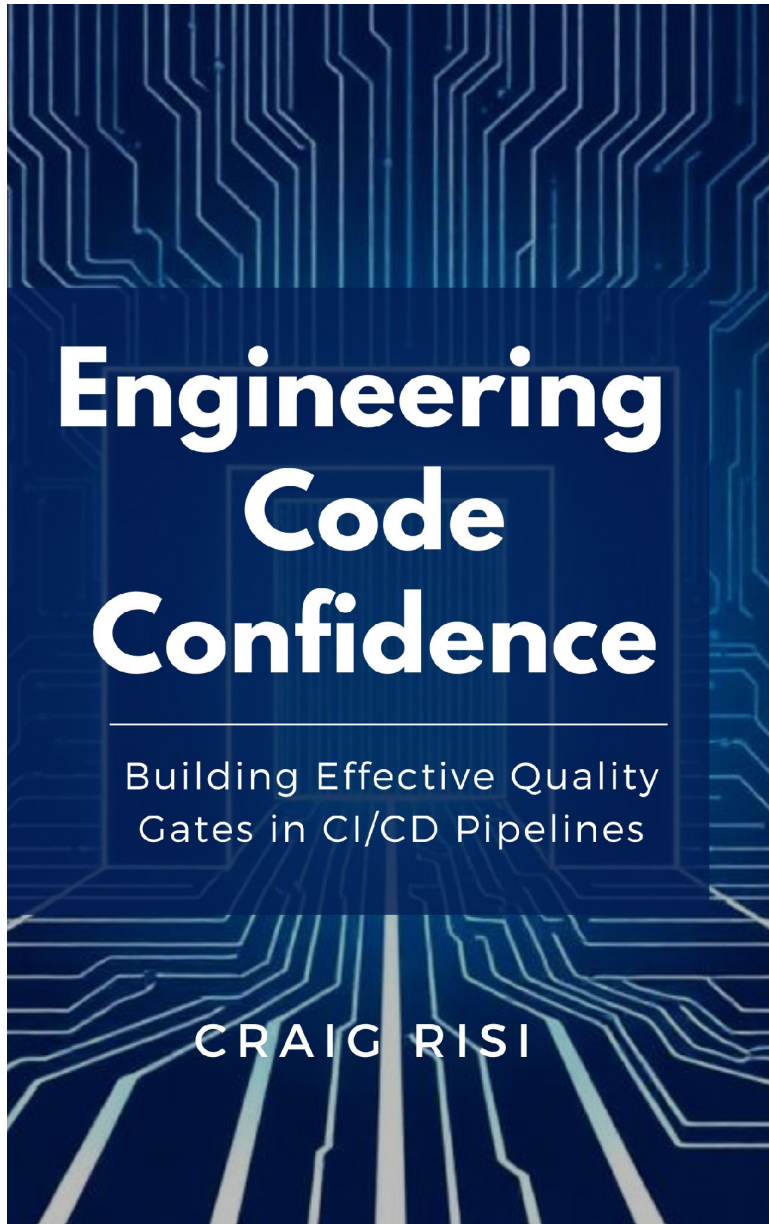
STRATEGIES FOR BUILDING
QUALITY AND TESTING INTO
AI-DRIVEN DEVELOPMENT





www.craigrisi.com





www.craigrisi.com

PAST







FUTURE

NOW



AI Projects

Software Development Projects

 Nature of Work	Exploration, discovery, probabilistic outcomes.	Building deterministic, functional systems
 Uncertainty	High (depends on data quality, model feasibility)	Low (requirements defined upfront)
 Iteration	Frequent revisiting of data, models, and evaluation	Iteration exists (e.g., Agile), but less exploratory.
 Output	Insights, predictions, trained models	Functional software (e.g., applications).
 Success Metrics	Accuracy, precision, recall; subjective evaluation	Binary: software works per specifications or not.
 Data Dependency	Central (requires preprocessing/exploration)	Secondary (data as input/output).



Challenges working with LLMs

Unique Challenges of modern LLM-powered Applications

Non-deterministic outputs

Context management

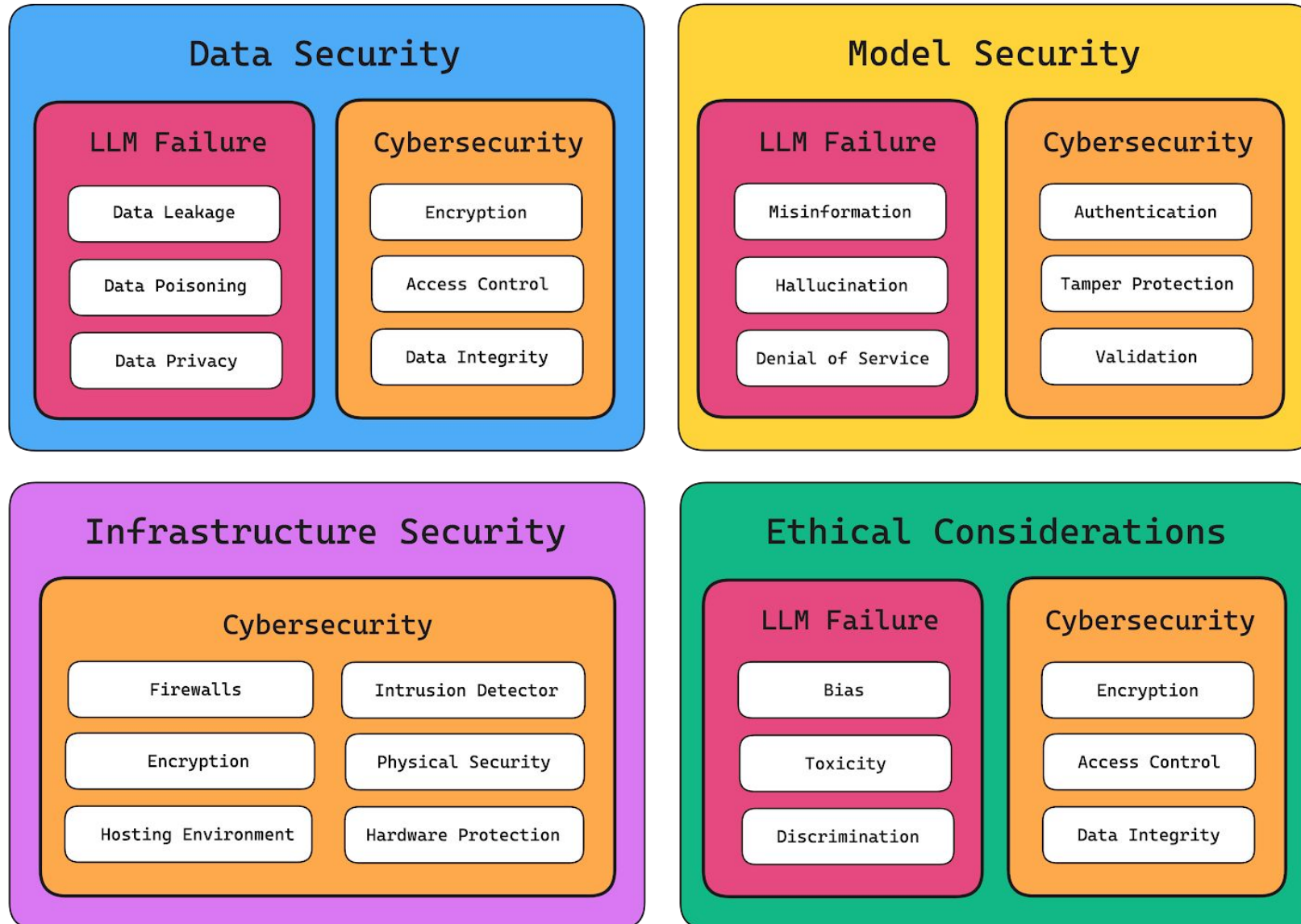
Data quality

Evaluation difficulty

Integration complexity

Cost Optimization

4 Pillars of LLM Security





Redefining Quality

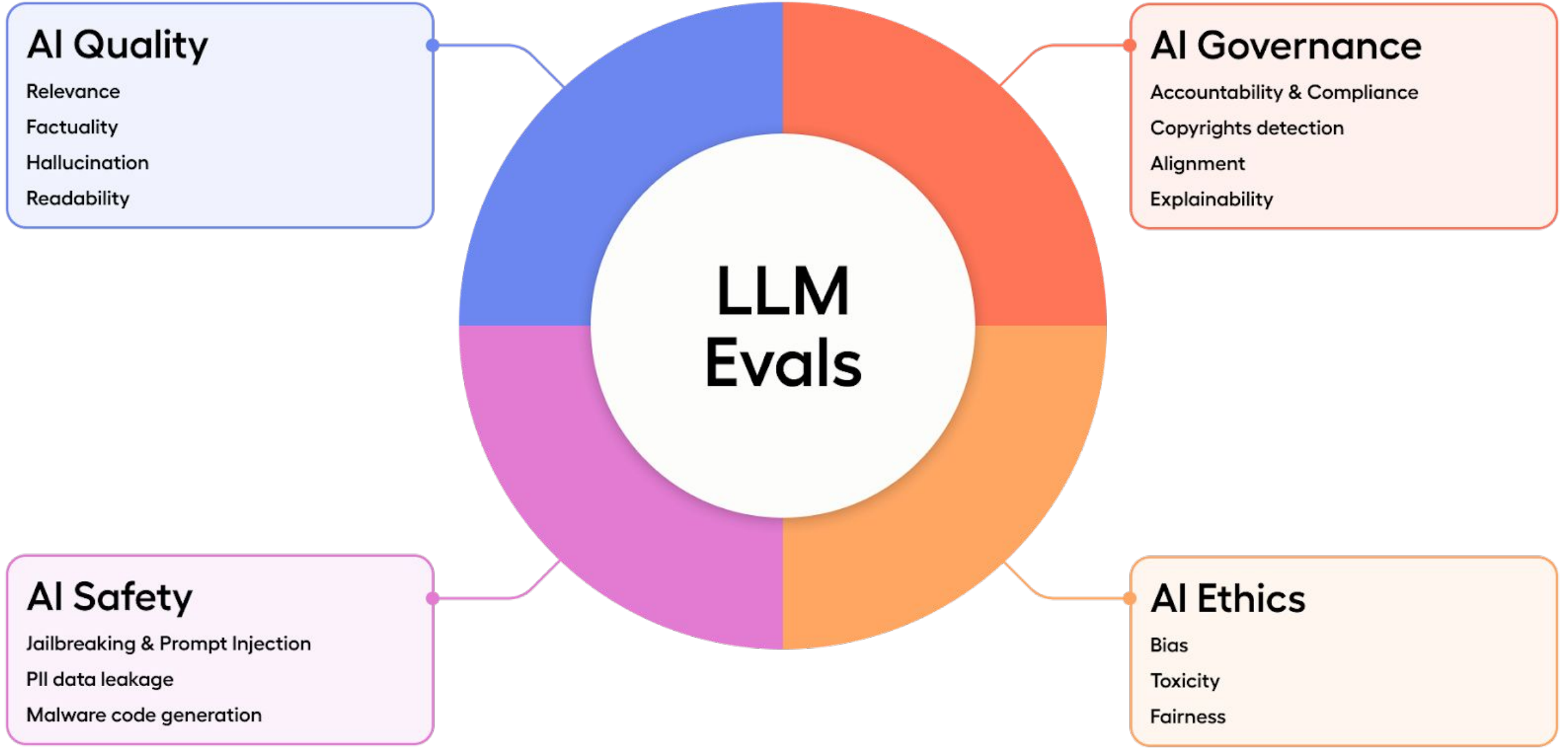
Changing the Dimensions of Quality

Accuracy: Does the model produce correct or factual results?

Reliability: Consistent behaviour across inputs and contexts.

Safety & Ethics: Avoiding harmful, biased, or unsafe outputs.

Performance: Speed, scalability, and cost efficiency.



A futuristic robot with a white and blue metallic body stands in a high-tech environment. The robot has a rounded head with two blue eyes and is positioned in the center. The background is filled with various data visualizations and glowing blue lines, including molecular structures, graphs, and network diagrams. The overall scene is illuminated with a strong blue light, creating a futuristic and technological atmosphere.

Unlocking LLM Potential

Strategies for Building Quality



**ENGINEERING
PRINCIPLES**



**TESTING &
EVALUATION
FRAMEWORKS**



**SAFETY &
ETHICS CHECKS**



**OBSERVABILITY
& FEEDBACK
LOOPS**

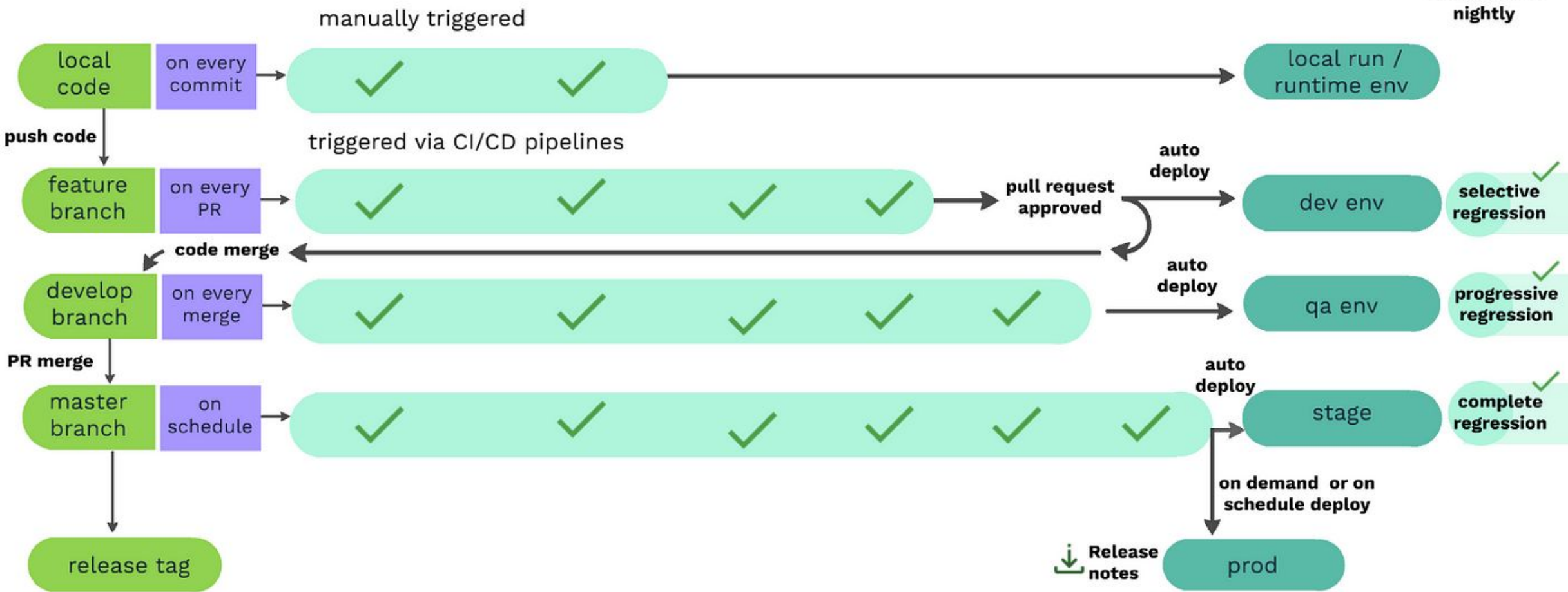


**GUARDRAILS &
FAIL-SAFES**





on schedule nightly



Testing LLM-powered applications



DATA QUALITY
TESTING



TEST-DRIVEN
DEVELOPMENT
(STRICT UNIT
TESTING)



WORK WITH
GOLDEN DATASETS



LINTING AND
QUALITY CONTROL




SECURITY AND
PERFORMANCE
TESTING




EARLY MONITORING
AND OBSERVABILITY

AI Testing Strategies & Methodologies



01 Simulate Real-World Scenarios

Test how the AI performs in messy, unpredictable environments—typos, slang, contradictions, and real-world input noise.




02 Test for Bias and Fairness

Identify & fix unfair outcomes across demographic groups by auditing for bias & applying mitigation strategies.




03 Data Validation & Quality Testing

Ensure training and input data is clean, complete, consistent, and representative to prevent downstream model errors.




04 Red Teaming for AI

Actively simulate adversarial threats like prompt injection, model inversion, and data poisoning.




05 TDD for AI

Shift AI modeling from reactive fixes to goal-driven design with test-first thinking.




06 Scenario-Based Testing

Design specific interaction flows to evaluate logic, consistency, and system behavior under complex rules.




07 Edge Case Testing

Challenge the model with rare, distorted, or unusual inputs to test reliability and identify weak spots in prediction confidence.



08 Model Drift Testing

Continuously compare model performance over time. Detect when evolving data causes accuracy or fairness to decline, & take action to retrain or adjust.



09 Human-in-the-Loop (HITL) Testing

Use human testers to validate subjective outputs esp in areas like moderation or diagnostics where automation falls short.



10 Domain Expert Involvement

Involve domain experts to ensure compliance, ethical decisions, and context-aware testing in high-risk sectors.



Questions