AI Testing and Assurance

Bryan Jones

QAT Architect
2i

We test.
You impress.

HUSTEF

How do we test AI?

Challenges

- The Oracle Problem
  - No Expected Results
  - Non-Deterministic
  - Probabalistic Answers
- The code doesn’t represent the algorithm
- Data Complexity & Volume
- Self-Optimising
- System Complexity
- Attempting to Mimic Human Abilities
- Bias
- Hallucinations/Confabulation
- Test Coverage?

HUSTEF

2i We test. You impress.

Test Levels

Familiar

- Unit or Component testing,
- Integration testing
- End-to-End testing
- UAT

New

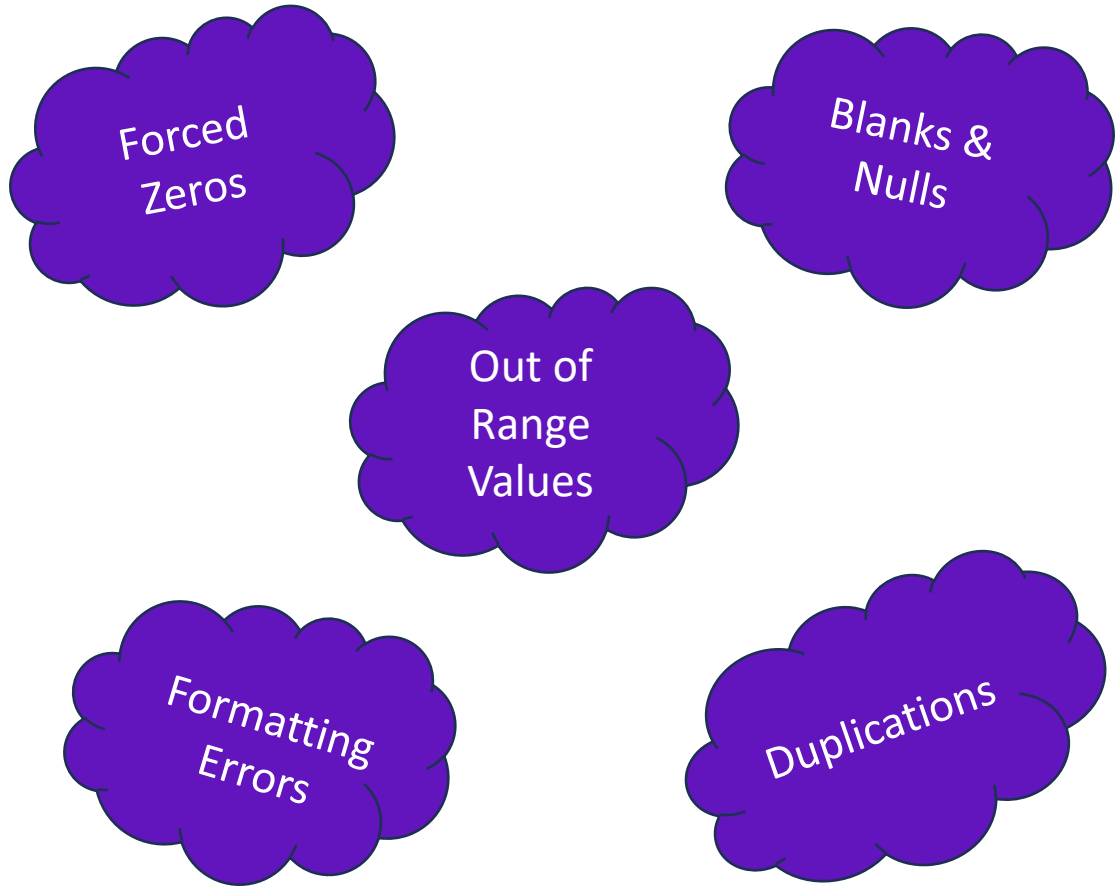
- Model Testing
- Data Testing
- Monitor in live

HUSTEF

We test.
You impress.

Component Test

“Automation without Requirements”



HUSTEF

We test.
You impress.

Data Test

"The code is not the algorithm,
The data defines the behaviour"

- Bias & Fairness
  - Don't forget Proxies
- Data Diversity and Representativeness
- Data Labelling and Annotation
- Data Splitting Problems
- Statistical Analysis

Data Scientists Are Our Friends

HUSTEF

2i We test. You impress.

Integration Test

Familiar but …

Probabilistic!

Chained Probabilities multiply

35% cat, 55% dog, 15% blueberry muffin

HUSTEF

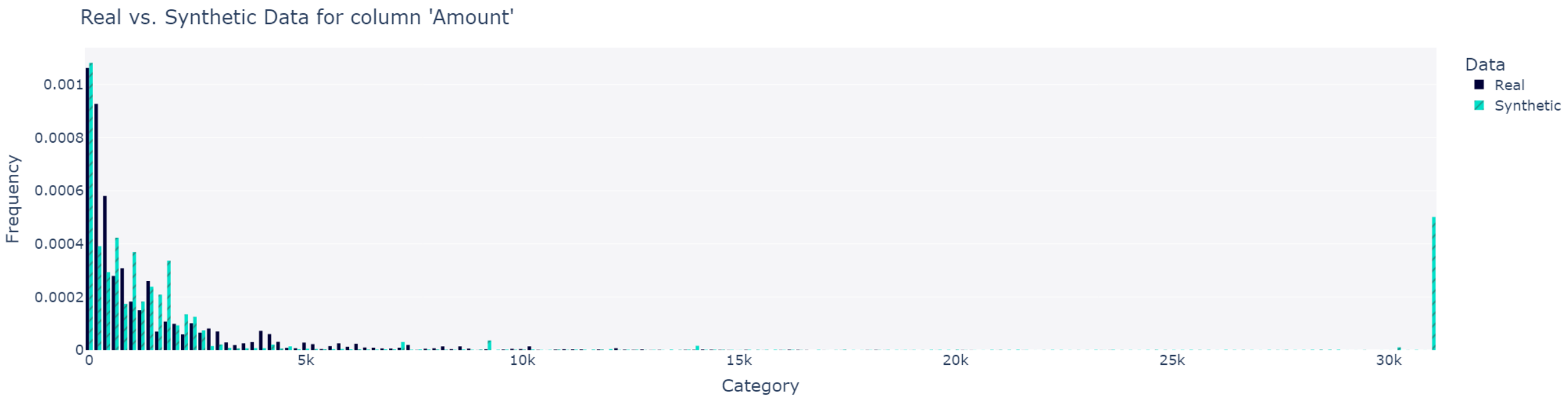
We test.
You impress.

Model Testing

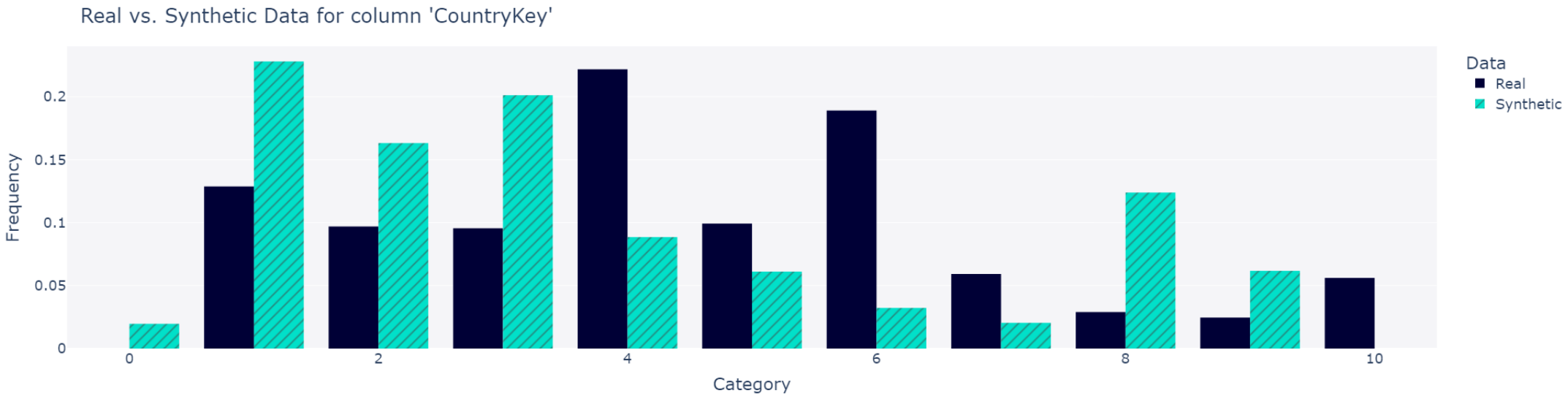
- Statistical in nature
- Large data volumes
- Complex to interpret
- Data Scientists are our friends

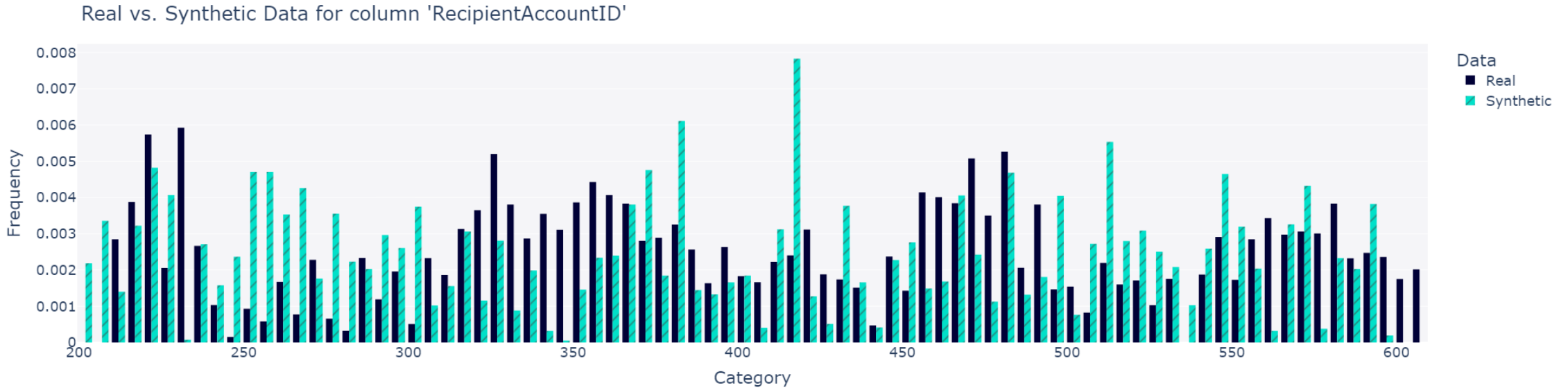
Still benefits from a Tester's curiosity, critical thinking, system thinking, and Question Asking!

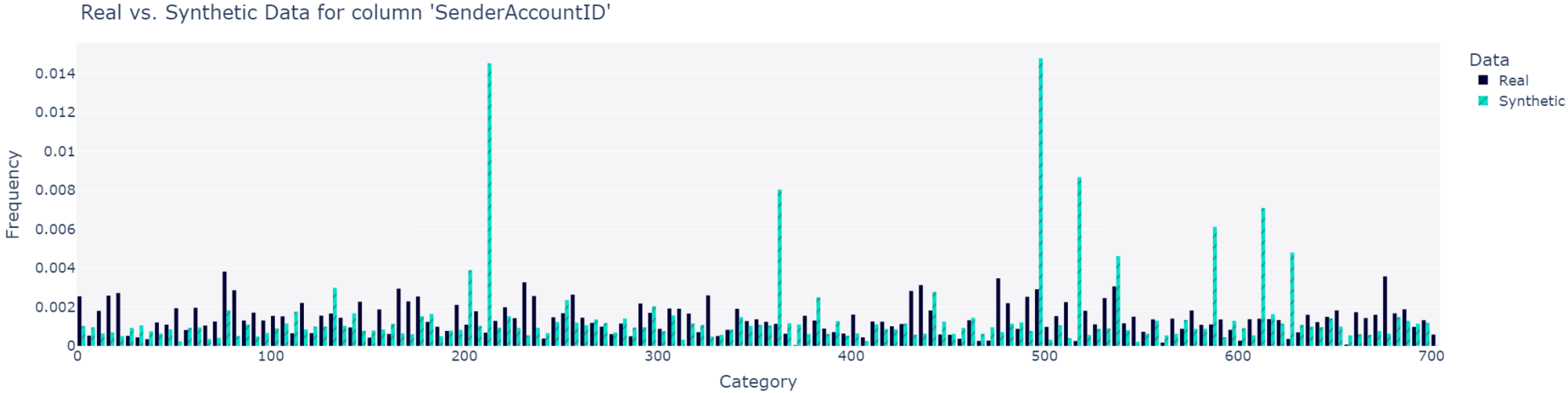
Evaluating the Synthetic Data



We test.
You impress.







End-to-End

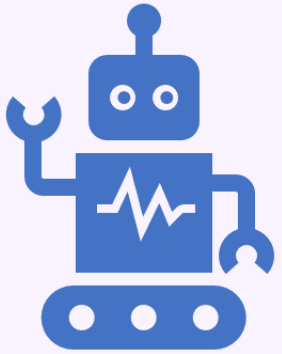
- Very Familiar
- Beware of complexity!
- Involve Experts to help decide what is a “Right” answer

HUSTEF

2i We test. You impress.

UAT

Pretty much the same



But watch out for Automation Bias!

Monitor in Live

- Performance
- Drift
  - Context, Concept or Model Drift
  - Data Drift

All the other usual stuff

- A11y
- Usability
- Security – Check out Adversarial Testing
- Performance – but …

e.g. *Accuracy, Precision, Recall, F1-Score, Confusion Matrix, ROI (or value add), Programmability, Energy/Power, Throughput/latency, GOPS, frame rate, delay, cost, footprint

*Accuracy measures how well your AI model performs on new or unseen data. Precision indicates the relevance of results to your target audience or problem. Recall shows how comprehensive the results are. F1-score is a measure of the balance between precision and recall.

Test Techniques

- Explainability
- A/B Testing
- Parallel Testing
- Statistical Analysis
- Exploratory Testing
- Use Experts
- Pairwise/Orthogonal Testing

More Test Techniques

- Metamorphic Testing
- Adversarial Testing
- Model Backtesting
- Dual Coding/Algorithm Ensemble
- Coverage Data
- Cross Validation
- Affordances Modelling

The Holy Trinity

The 4 Factors of the AssureAi Quality Score

Accuracy

Ensuring AI systems produce accurate, reliable outputs that meet user requirements.

- Data Integrity
- Model Validation
- Continuous Monitoring
- Feedback Loops
- Compliance and Standards

Robustness

Ability of AI systems to maintain performance under varying conditions and against adversarial attacks.

- Adversarial Testing
- Data Diversity
- Fault Tolerance
- Scalability
- Security Measures

Explainability

Making AI decisions understandable to humans, including how and why decisions are made.

- Transparent Algorithms
- Feature Importance
- User-Centric Design
- Audit Trails
- Documentation

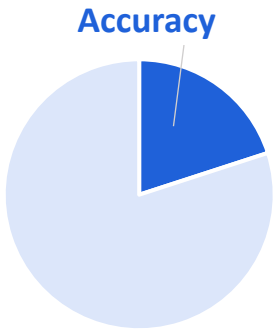
Performance

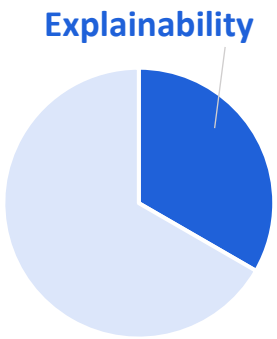
The speed of AI systems in executing tasks.

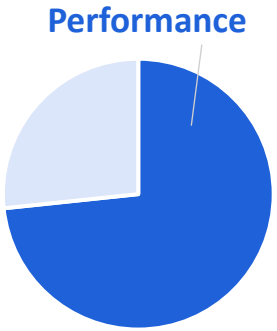
- Benchmarking
- Optimization
- Scalability Testing
- Latency Reduction
- Resource Management

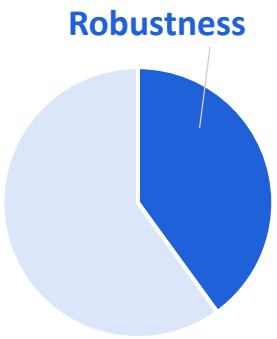
We test.
You impress.

AssureAi Scores









Parameters	Score
Data Integrity	Silver
Model Validation	NPR
Continuous Monitoring	NPR
Feedback Loops	NPR
Compliance and Standards	Gold

Parameters	Score
Transparent Algorithms	Bronze
Feature Importance	Silver
User-Centric Design	NPR
Audit Trails	NPR
Documentation	Silver

Parameters	Score
Benchmarking	Bronze
Optimisation	Gold
Scalability Testing	Gold
Latency Reduction	Silver
Resource Management	Silver

Parameters	Score
Adversarial Testing	Bronze
Data Diversity	NPR
Fault Tolerance	Bronze
Scalability	Bronze
Security Measures	Silver

NPR = Not production ready

We test.
You impress.

Useful Resources & Tools

- HuggingFace
- Fairlearn
- AI Fairness 360
- What-If-Tool
- Google Vertex AI
- Amazon SageMaker
- MLflow
- Neptune.ai
- MonkeyLearn
- AWS Comprehend
- Scikit Learn
- SciPy
- Tensorflow Extended (TFX)
- PyTorch

Contact Us

Bryan Jones
Bryan.Jones@2itesting.com

LinkedIn:
/bryan-jones-mbcs-96953/

www.2itesting.com

We test.
You impress.