GLOBAL •I•ASSURANCE PILOT



Testing Real World GenAl Systems

Technical testing of GenAl applications across industries

Using GenAI in real-life situations - e.g., in banks, insurers, hospitals or airports – raises the bar on quality and confidence. Current AI **Safety** efforts tend to focus on the underlying **foundation Models**. This pilot focuses on the **Reliability** of the end-to-end **Applications** into which such models are embedded.

Targeted outcomes



Use case statistics

How are LLMs used in the application?

(of use cases: top 5 archetypes only)

Summarisation (15)

Retrieval augmented generation (13)

Data extraction from unstructured source (12)

Multi-turn chatbot (10)

Classification or recommendation (08)

Who is the Is a human in the loop? application for? 10 Internal – specialists 12 Yes 🔵 5 No

2 Internal - all staff

5 Public / Customers

Top 5 Risks Tested



echniques Test	ing A	pproaches
Use-case specific historical or synthetic test data	Ev	aluators
Red-teaming (adversarial)	Ø	Human expert
	Ĵţ	LLM as a judge
Simulation testing (non-adversarial)	60	Non-LLM model
Off-the-shelf benchmarks	jag	Rule-based logic

Lessons Learnt

Test what matters

Your context will determine what risks you should (and shouldn't!) care about. Spend time upfront to design effective tests for those

Don't expect test data to be fit for purpose

No one has the "right" test dataset to hand. Human and AI effort is needed to generate realistic, adversarial and edge case test data

Look under the hood

Testing just the outputs may not be enough. Interim touchpoints in the application pipeline can help with debugging and increase confidence

0₁0

Use LLMs as judges, but with skill and caution

Semantic or

Similarity metrics

Human-only evals don't scale. LLMs-as- judges are often necessary, but need careful design and human calibration. Cheaper, faster alternatives exist in some situations

The role of the human expert is still paramount for effective testing!

B

盎

E





Scan to read the full case studies.

