



How to test AI; using AI

www.testzeus.com

Hello, I am Robin



Robin Gupta
Co-founder - **CEO**



- 15+ years leading Product & Engineering teams of up to **45** members
- Published author on Testing with **4.8** ★ on Amazon
- **Open source contributor** to community projects like Selenium.

accenture

athenahealth



Models

OpenAI GPT 4.5, Deepseek R1 etc.



Wrapper

RAG applications, Chatbots



Agents

TestZeus, Bolt.new, HarveyAI



Changing needs for Software testing

Criterion	Traditional Software	LLM based apps
<i>Behaviour</i>	Predefined Rules	Probability+Prediction
<i>Output</i>	Deterministic (1 input=1 output)	Non-deterministic (1 input = Many possible outputs)
<i>Testing Strategy</i>	Evaluate as right or wrong	Evaluate on: Accuracy Quality Bias Consistency Toxicity

Automated evals: what and when

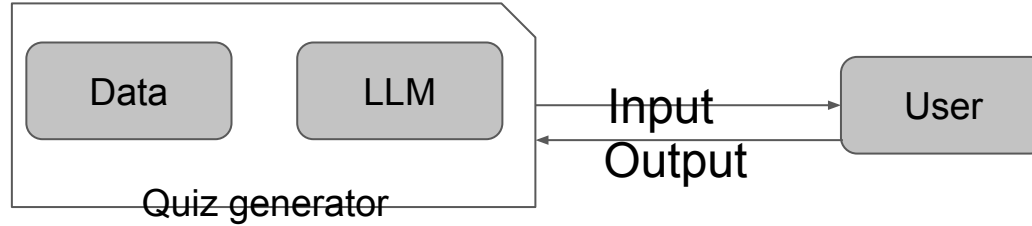
What should you evaluate?

- Context adherence
- Context relevance
- Correctness
- Bias and Toxicity

When should you evaluate?

- After every change (bug fixes, feature updates, data changes)
- Pre-deployment (merges to prod branch, end of sprint, prior to shipping hotfix)
- Post-deployment (on demand business needs)

Demo Application under test: Quiz generator

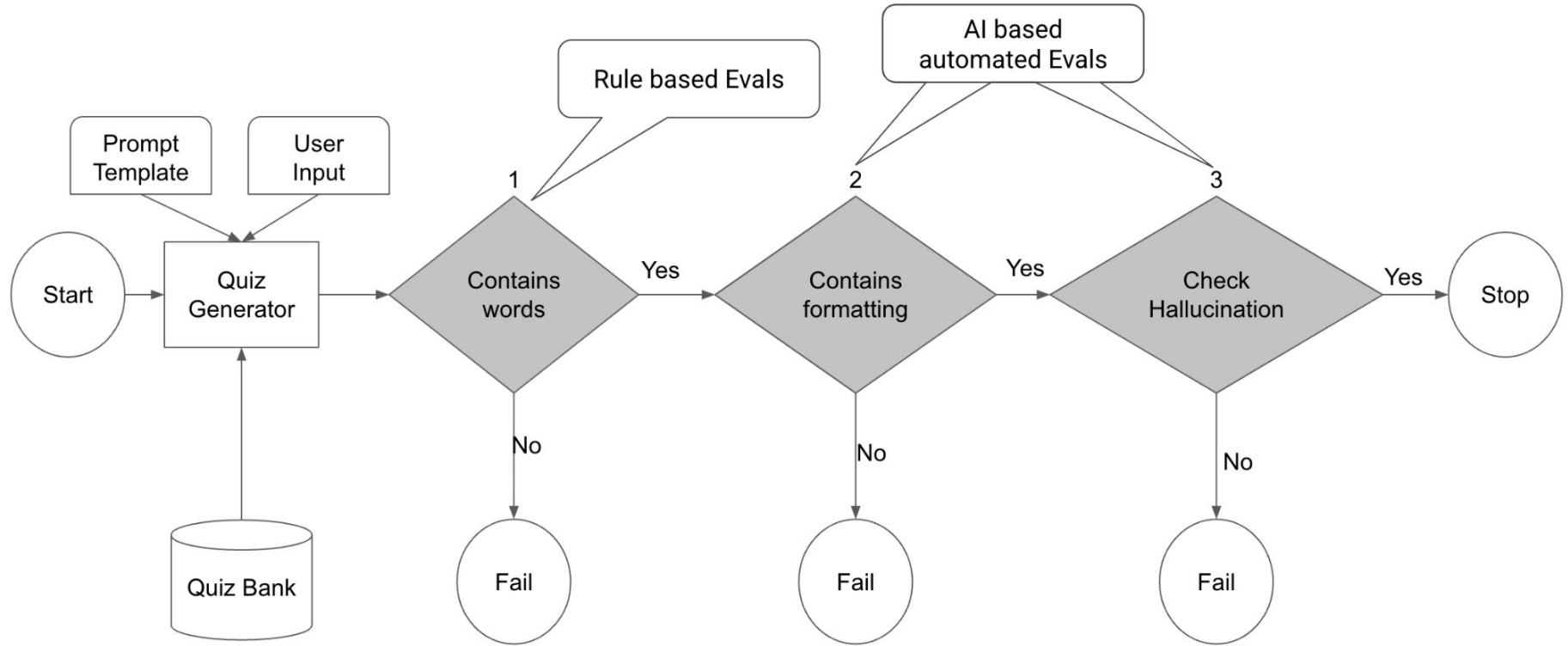


Input: Write a quiz about science

Output: Sure ! Here are three science questions for you:

1. True or False: Water slows down the speed of light.
2. What did Marie and Pierre Curie discover in Paris?
3. Where were the first refracting telescopes invented?

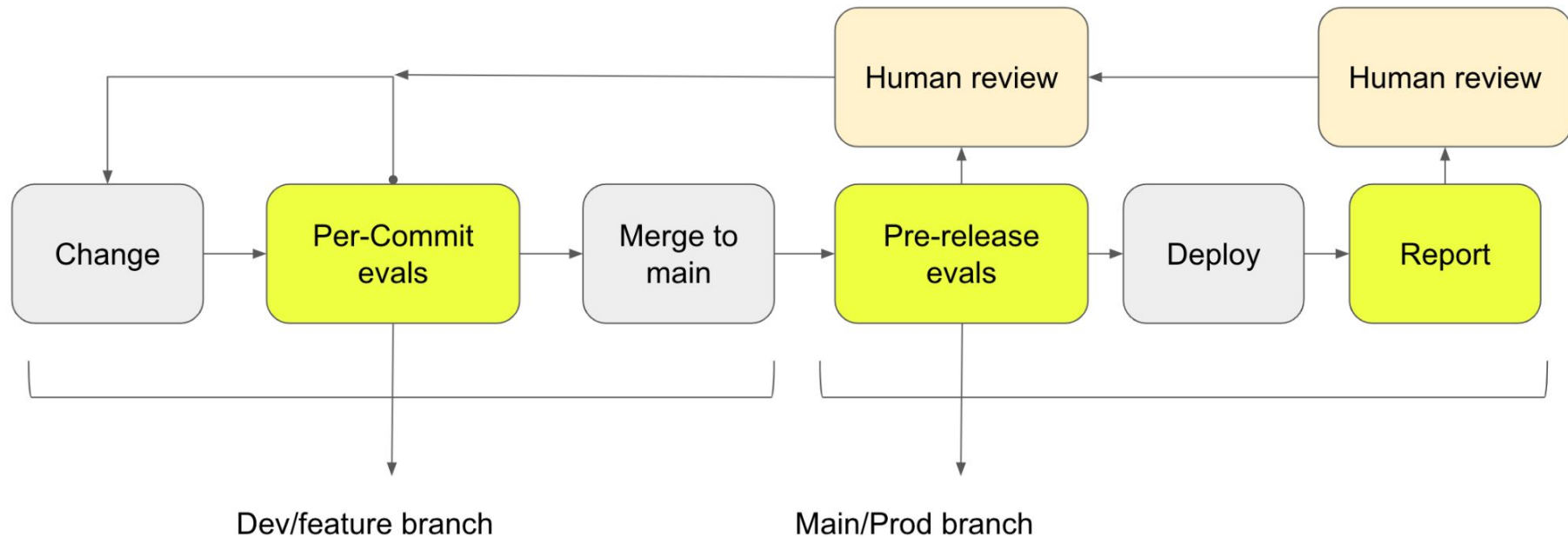
Demo Application under test: Quiz generator





Time to roll up your sleeves

CI/CT pipeline for AI apps



1. Notebooks from the demo:
 - a. <https://colab.research.google.com/drive/1ciWohOo4f1pdt6rOlxyypsLDj44k96O0?usp=sharing>
 - b. <https://colab.research.google.com/drive/14-2Et2j8T55JwJE8U5pOJDIFYMPbacHV?usp=sharing>
 - c. <https://colab.research.google.com/drive/1IsGy0RIfBLEt29d1Y3Xd0HFjLc4NAoxa?usp=sharing>
2. <https://github.com/Infosys/Infosys-Responsible-AI-Toolkit/tree/Release-2.1.0>
3. <https://www.promptfoo.dev/docs/red-team/>
4. [GitHub - openai/evals: Evals is a framework for evaluating LLMs and LLM systems, and an open-source registry of benchmarks.](#)
5. [\[2307.03109\] A Survey on Evaluation of Large Language Models](#)

Thanks and Questions



<https://github.com/test-zeus-ai/testzeus-hercules>



Software quality at godly speed

www.testzeus.com