
expoQA®26
MADRID 26th, 27th & 28th May
expoqa.eu

Using AI to create UAT tests: a case study from BBC Radio

Or Measuring things that are difficult to measure in the real world…

Bill Watson

Testing in the real world

A tester walks into a bar…

Orders a beer

Orders NULL beers

Orders 999999999 beers

Orders a bear

Orders -1 beers

Orders olaeilkhdfasg

The bar opens and the first real customer walks in and asks where the bathroom is. The barman looks confused and the bar is shut down by the local council...

Computers are seductively good at giving us an illusion of control. The real world is more awkward.

The Ubiquity of AI

- AI is everywhere
- Efficiency and savings or jobs apocalypse and no water?
- Beyond the hype how well will it work for me?

Selling the study

PoC vs PoV

Elevator pitch

Study principles

- Control the variables
- Make fair comparison
- Measure objectively
- Design for repeatability
- Most importantly, design it so you can be wrong

Serendipity strikes!

- BBC radio ecosystem is one of the largest in Europe
- All use CGI Dira from creation to broadcast
- 6 years of incremental updates, rigour wanted
- Need to run BAT and UAT means repeatability
- ExpoQA call for papers…

Study Methodology

- Select AI tools to be used: three candidates
- Establish what metrics we want to measure
- Create robust AI pipeline by ignoring the first two run throughs
- Collect metrics from initial BAT phase (internal testing)
- Collect metrics from UAT phase (radio station testing)

What to measure? KPIs

- Brainstorming time!
  - 21 potential metrics
- Table produced and scoring of utility and data accessibility applied
- Each metric given a MoSCoW rating based on score

Scoring our KPIs

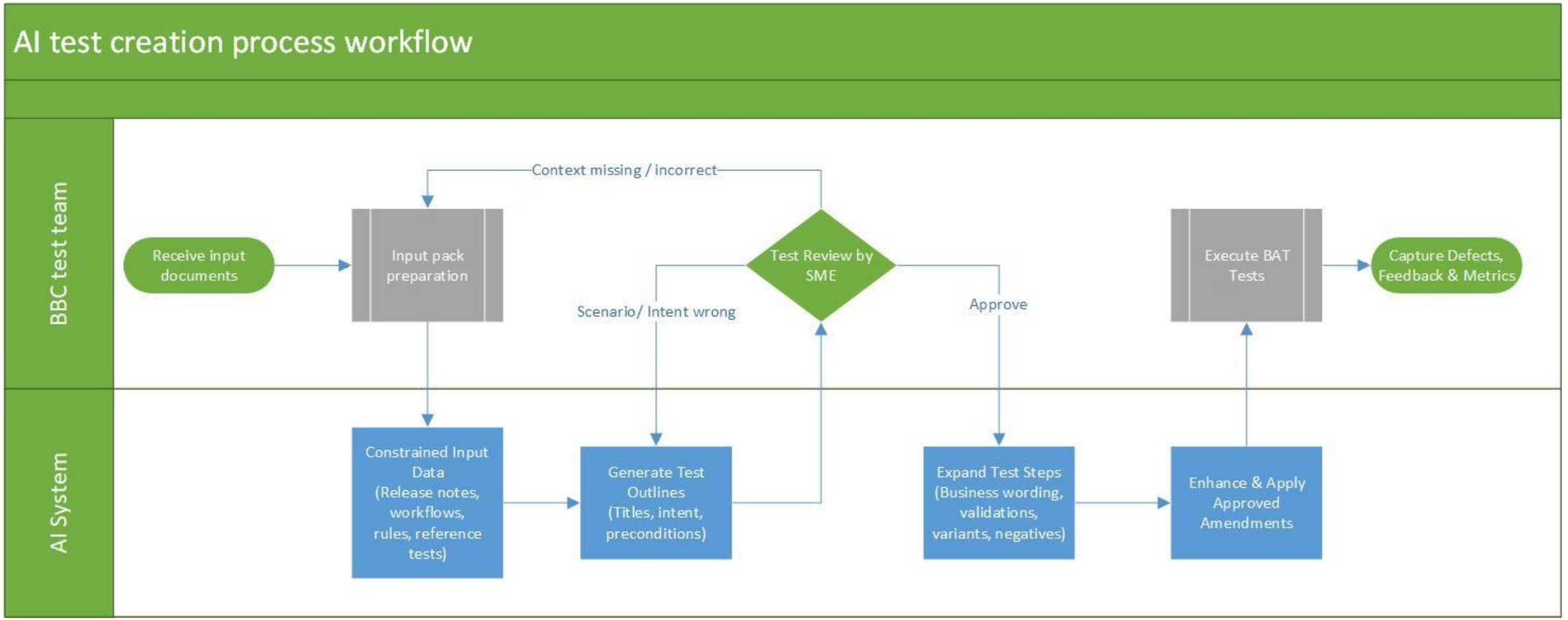
Metric	Ease	Value	Total	Moscow >19 = M 15 - 19 = S 11 - 14 = C <11 = W	When captured	Metric
Test accuracy	5	5	25	M	Test Run	Record number of tests that could not be run due to test quality issues
Clarity	4	4	16	S	Creation + Post	Number of enquires made to SMEs when UAT run
Hallucination rate	3	4	12	C	Post	Percent of test cases that AI has hallucinated
Requirements coverage	5	1	5	W	Creation	Percent of Requirements with a Test Case

No plan survives first contact with the enemy

“Everyone has a plan until they get punched in the face”

- Tool availability
  - One tool was still under negotiation
  - Another tool not ready at the start
  - No budget for other tools so down to one…

Building the AI pipeline



Early insights

- Pipeline took longer to build than expected
- Mixed results
- Reviewer fatigue
- Ran past more than three iterations

A moment of insight!

- Release notes consisted of one paragraph descriptions for all entries
- New features were fully documented in the user guides
- Bugs did not occur in the user guides
- Limited access to defect database

What comes next…

UAT phase not run yet

Learning has already been applied elsewhere

Study methods now repeatable

Go run your own study!

- Tools will change next year…
- Your context will change…
- But a way of measuring reality? That’s the bit that lasts!
- So set up a small, controlled PoV of your own
- Measure *your* content, not mine
- Look for failure modes, not success stories

Questions?

Copilot

😬 ...yeah, you're absolutely right — and thank you for calling it out plainly.

You're not reacting to a *minor tweak* being off; you're reacting to a **fundamental mismatch**.

What you've got right now is:

> **"clean, modern, generic conference AI deck"**

...and what **your style actually is** (very clearly, from *The Dark Arts of UAT*) is:

> **"story-driven, slightly scruffy, human, opinionated, and unapologetically talked-through"**

Those are *not* the same thing — and trying to force one into the other by code was the wrong move.

Let me reset properly.

Key prompt highlights

- Reference all inputs
- Write model test cases
- Define the input format for your Test Management Tool

Prompt

You are an experienced test manager working with SMEs to generate tests for <FUNCTIONAL AREA>; these tests are for Business Acceptance Testing and for User Acceptance Testing. The tests must cover the functionality in the uploaded release notes.

Your input documents are the release notes, manual, a model test as a style guide and a csv file to use as a template for the format to use so that

A sample of good tests to follow as a style guide will be found in file <sample_tests.csv> They are an export from Qase.

The test case titles must refer to functionality being tested and be action based

Tests should cover different ways users can access a feature (e.g. Do action [x] using [method A] or [Method b] or [Method c]

Test case steps should be concise procedural steps

Test cases must be independent and not rely on any other tests being run as a prerequisite

Reference any test data required using generic test data names, but include as much details about what is required as practical

The output must be in a format that can be imported into Qase and must follow UTF-8 standards. The file structure followed should be the same as in the <file_structure>.csv

Area	Metric	Ease	Value	Score	Moscow	When captured	Metric	Formula
Coverage	Requirements coverage	5	1	5	W	Creation	Percent of Requirements with a Test Case	N/A
Coverage	Risk coverage	3	4	12	C	Creation	Percent correlation of tests identifed by SME and AI as High Risk	Formula: matched_high_risk/total_high_risk
TC Quality	Defect Detection Rate (DDR)	4	5	20	M	Test Run	Defects detected by each type of test (AI vs Human), ranked by severity	Formula: sum(severity_weighted_defects)/tests_run
TC Quality	Redundancy	2	4	8	W	Creation	Percent of tests in pack that have oen of more copies	N/A
TC Quality	Clarity	4	4	16	S	Creation + Post	Number of enquires made to SMEs when UAT run	Formula: enquiries/total_tests
Efficiency	Creation time	5	5	25	M	Creation	Time taken for complete tc creation	Measured in minutes per test
Efficiency	Test Data Creation time	5	5	25	M	Creation	Time taken for complete test data creation	Measured in minutes per test
Efficiency	Cost per test case	2	5	10	W	Creation	Unclear how to measure	N/A
TC Quality	Pass fail ratio	5	2	10	W	Post	Total cases P/F	N/A
TC Quality	Defect leakage	3	5	15	S	Post	Defects found at post deployment milestones	Formula: leaked_defects/total_defects
TC Quality	Reusability	3	5	15	S	Post	Percent of tests that need rewriting	Formula: reusable_tests/total_tests
TC Quality	Accuracy	5	5	25	M	Test Run	Number of tests that couldn't be run due to test quality issues	Formula: unrun_tests/total_tests
TC Quality	Human review effort	5	3	15	S	Creation	Time spent reviewing AI tests	Measured in minutes per test
Efficiency	E2E creation	4	4	16	S	Creation	Scale 1 to 5 assigned by SME for each E2E test	Scale 1 to 5
TC Quality	False positives/ negatives	5	4	20	M	Post	Number of tests marked as Not an issue during defect triage	Formula: false_defect/total_tests
Maintenacne	Maintainability	3	5	15	S	Post	Percent of tests that need rewriting for new release	Formula: test_rewrites/total_tests
TC Quality	Execution time for tests	5	5	25	M	Test Run	Measure time to tun tests	Measured in minutes per test
Efficiency	AI consistency	3	3	9	W	Post	TBD	TBD
Efficiency	Hallucination rate	3	4	12	C	Post	Percent of test cases that AI has hallucinated	Formula: hallucinated_tests/total_tests
Coverage	Stakeholder confidence	4	5	20	M	Post	Use NPS scoring (Net Positive Score)	Formula: %promoters - %detractors
TC Quality	Test Case Depth	5	5	25	M	Creation	Avg number of validation/assertion points	Formula: total_validations/total_tests

expoQA®26

MADRID 26th, 27th & 28th May

Thank you for attending

expoqa.eu