AI – Current State

Prepared by:

Mei Reyes-Tsai, TTC General Manager – Technology and Delivery

Contributors:

Bernard Roux, Zespri Head of Quality Assurance and Release Management
Ian Kalmakoff, Otago University Group Leader Test and Performance
Matthew James, Otago University Analyst Test and Performance
Rodney Colvin, Tauranga City Council Senior Test Lead

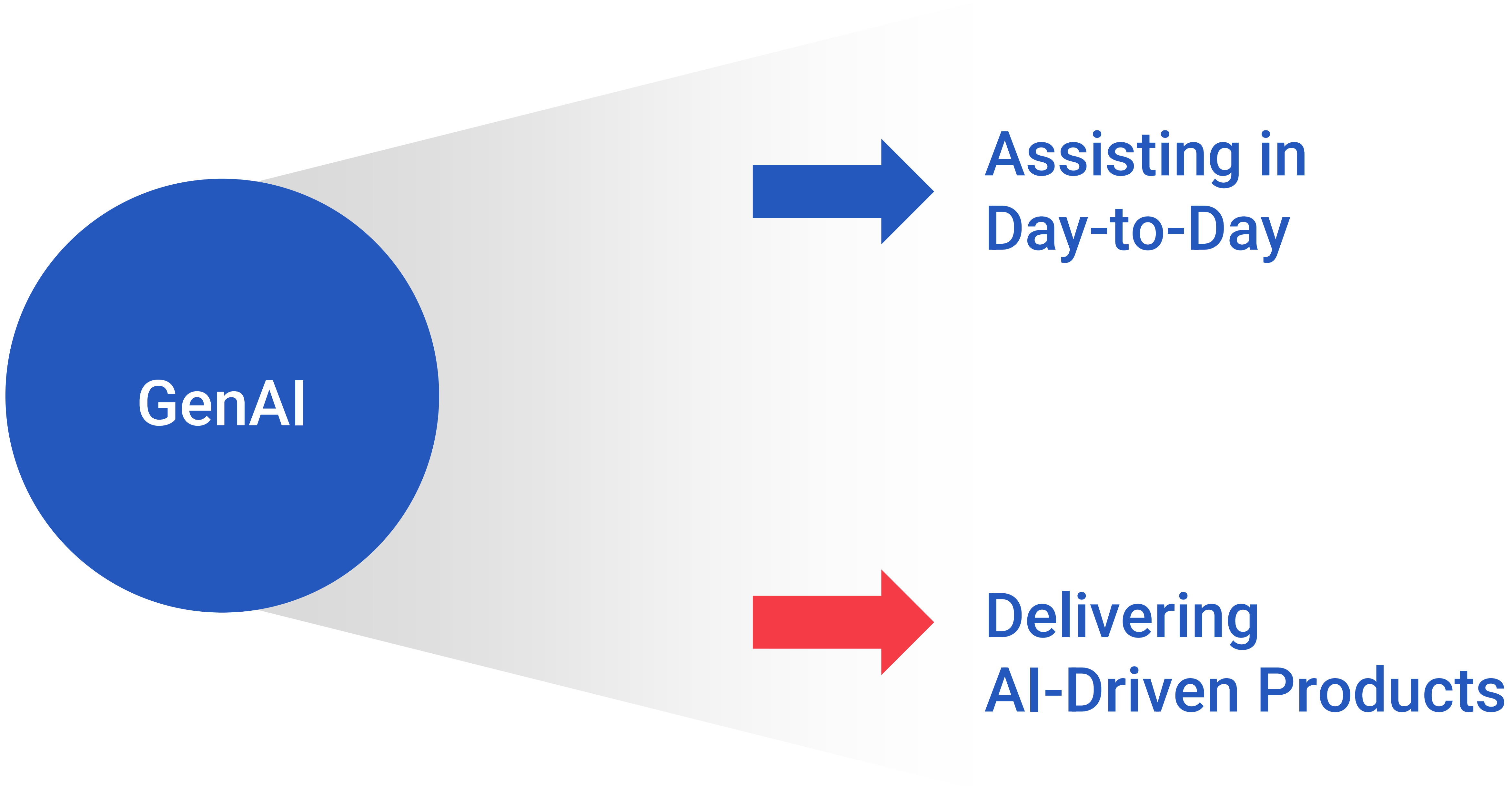
19th May 2025

TTC Confidential – Do Not Distribute

AGENDA

1. GenAI – Current State
2. Using GenAI on Day-to-Day
3. Testing AI

AI Impact to the Software Development Industry



Assisting in Day-to-Day

Delivering AI-Driven Products

Huge potential to:

Increase

- Productivity
- Quality
- Profits

Reduce

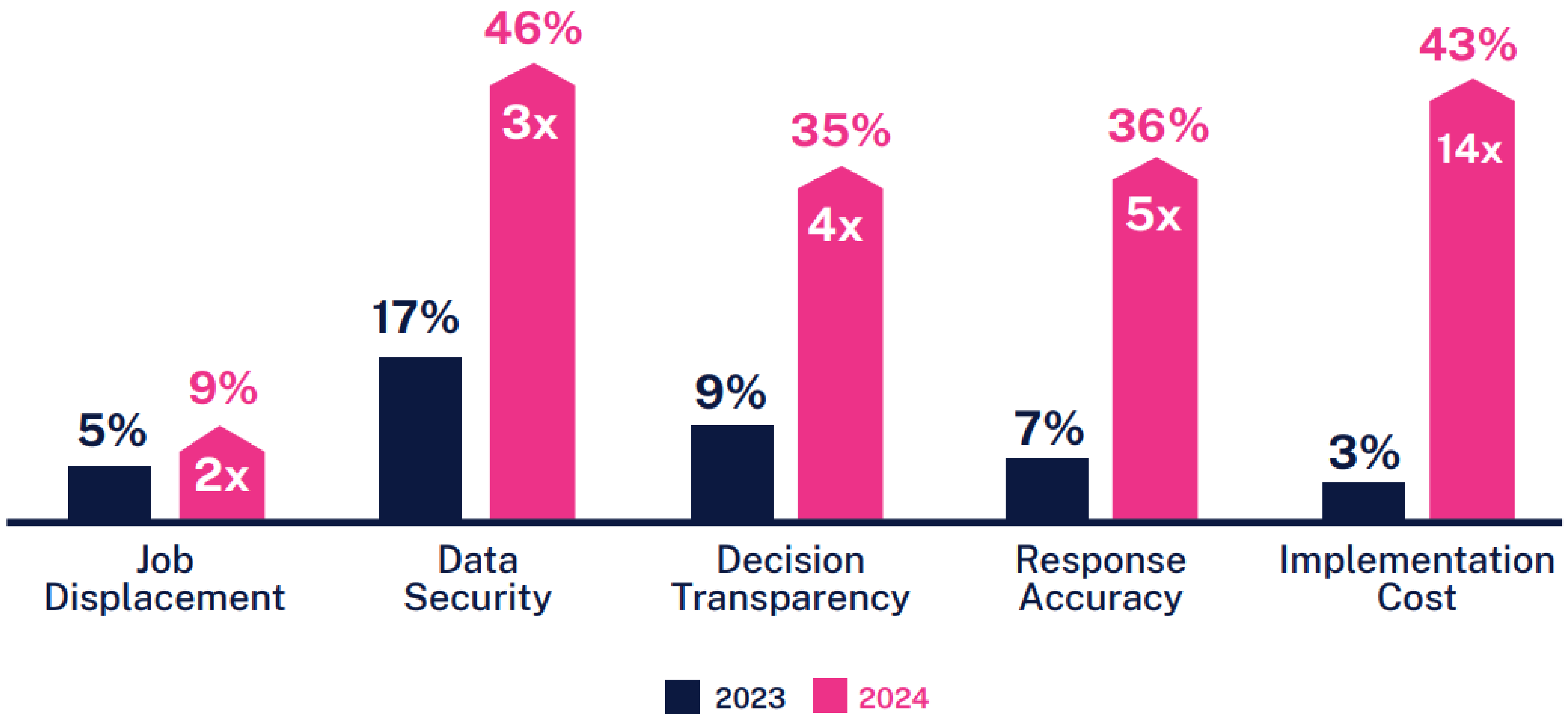
- Time to market
- Costs
- Risks

...if done right

GenAI Adoption Studies

Significantly Increasing Concerns

Top Gen AI Concerns 2023 v. 2024



GOVERNANCE

Companies understand the critical need for responsibility around data privacy, transparency, and fairness as they adopt new generative AI practices.

Most Successfully Deployed Governance AI Initiatives:

- Standard Gen AI tools and models defined to ensure alignment
- Restricted access to Gen AI tools and data based on role
- Gen AI guidelines defined and distributed to minimize risk

GENERAL & ADMINISTRATIVE COST REDUCTION

Today, with concerns around implementation costs skyrocketing, the need to balance innovation with costs is top of mind for business leaders.

Most Successfully Deployed G&A Cost Reduction AI Initiatives:

- Gen AI for QA testing and debugging code
- Provide employees with help and FAQs
- Gen AI generates first draft of new code

GenAI Adoption Studies – Testing

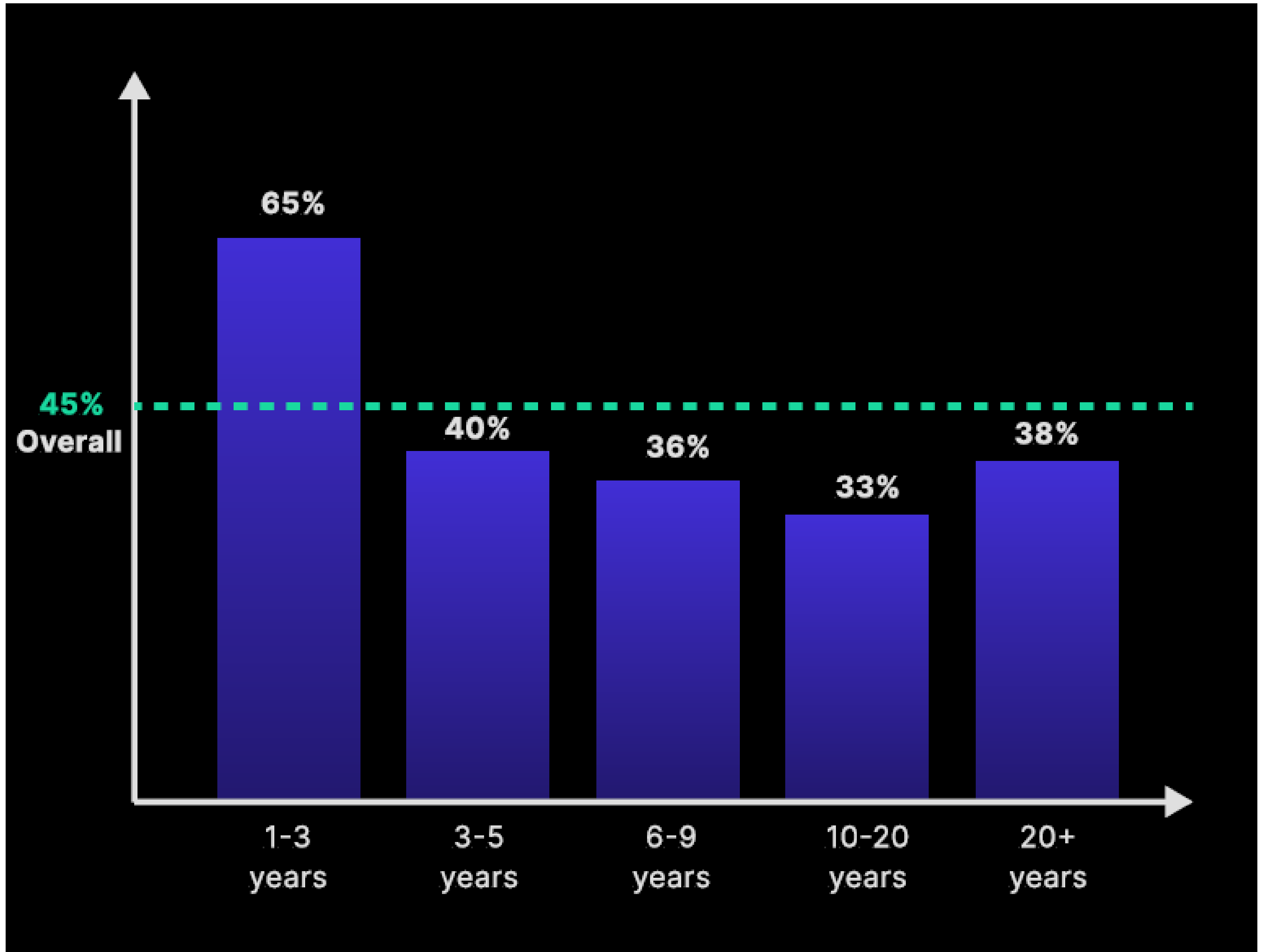
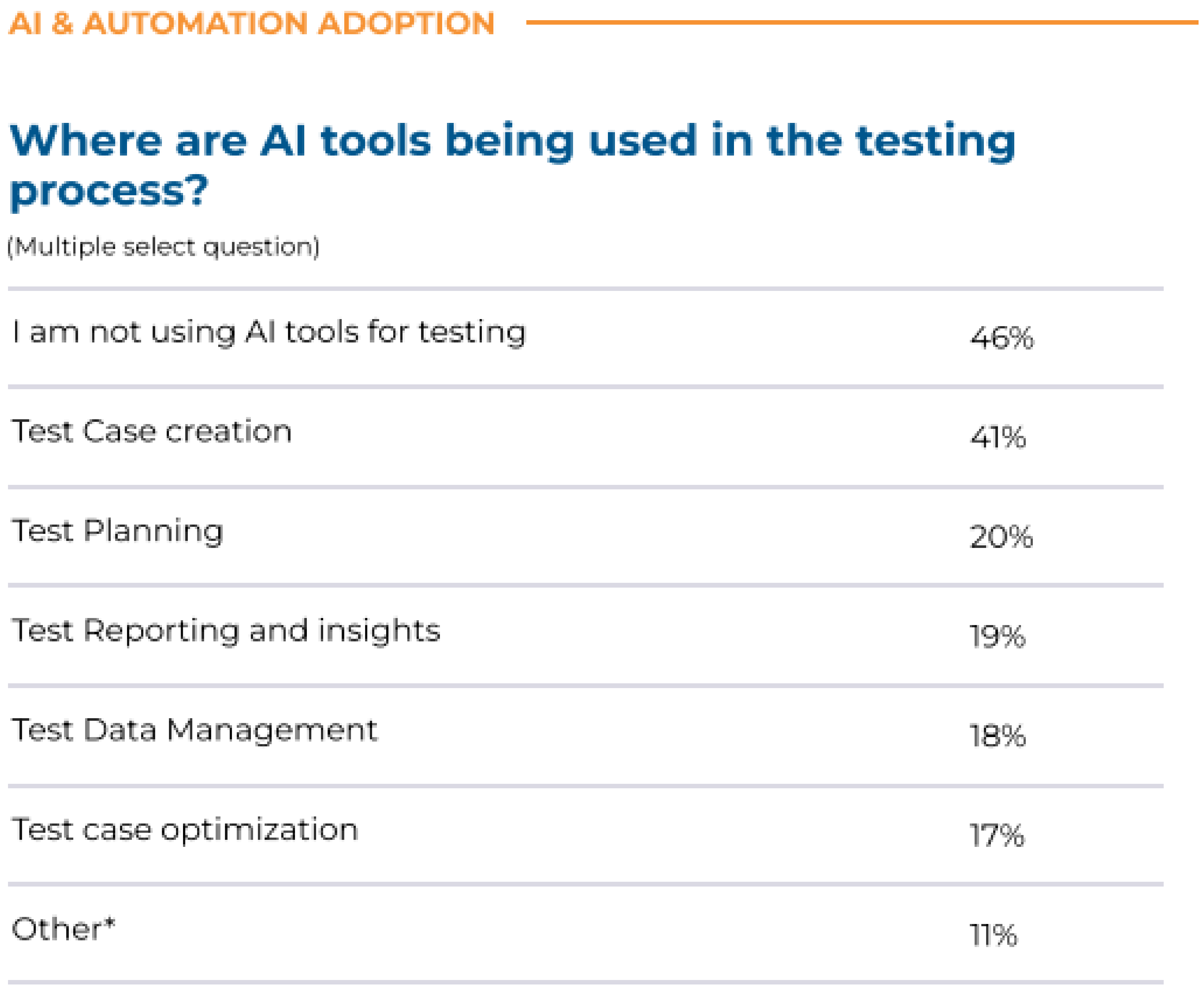


Figure 14. Rates of AI adoption for QE activities



AI & AUTOMATION ADOPTION

Where are AI tools being used in the testing process?

(Multiple select question)

I am not using AI tools for testing	46%
Test Case creation	41%
Test Planning	20%
Test Reporting and insights	19%
Test Data Management	18%
Test case optimization	17%
Other*	11%

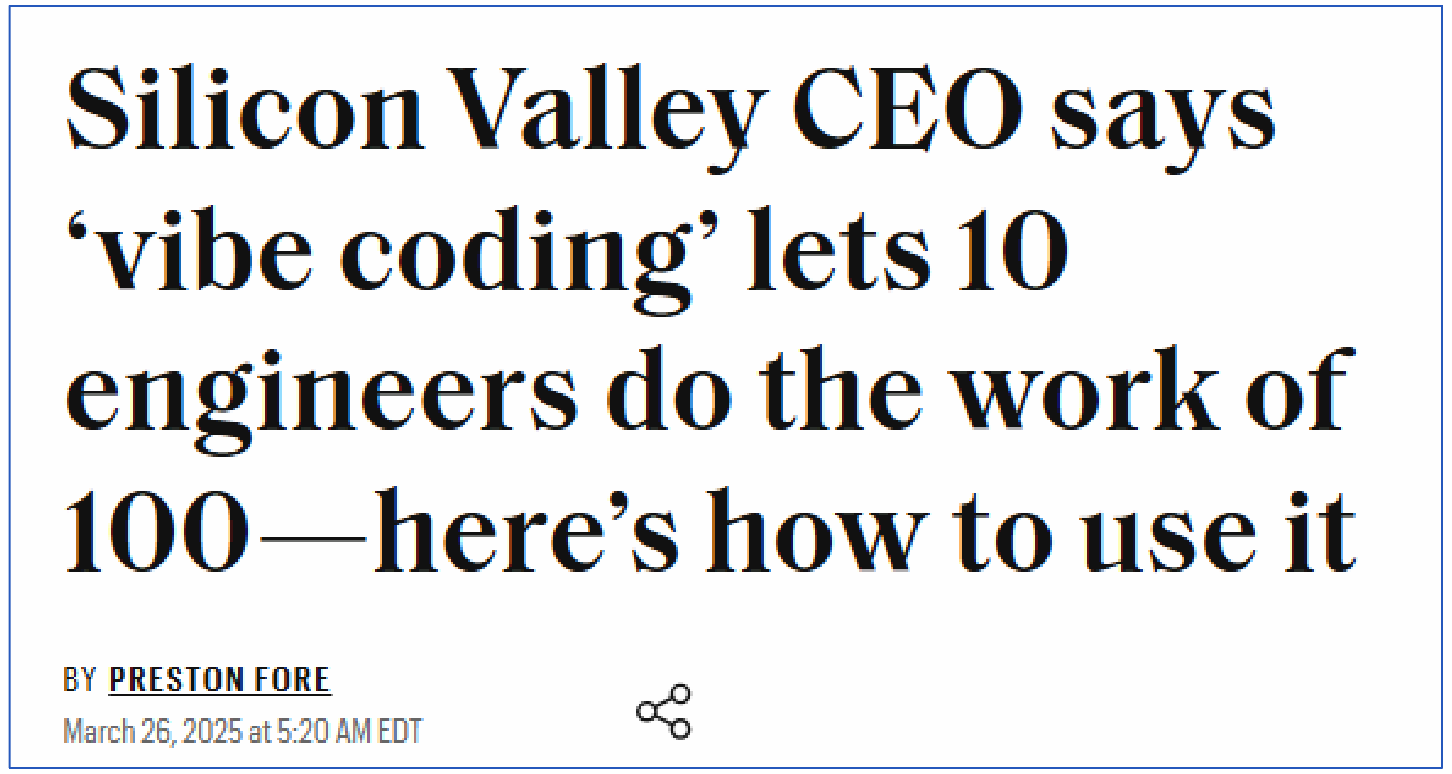
USING AI ON DAY-TO-DAY

Impact on Day-to-Day - The Promise

CLAIMS with a GenAI-Enhanced Structure

- **Fewer but more versatile roles**:
  - Full-stack developers (rather than specialised juniors)
  - QA engineers (focus on test strategy rather than manual testing)
  - AI specialist (new role to optimize GenAI tools and workflows)
  - DevOps engineer (manages CI/CD with integrated AI testing)
- **Cross-functional collaboration** enabled by GenAI tools that bridge skill gaps
- **Flatter organisation** where technology leadership focuses on strategy

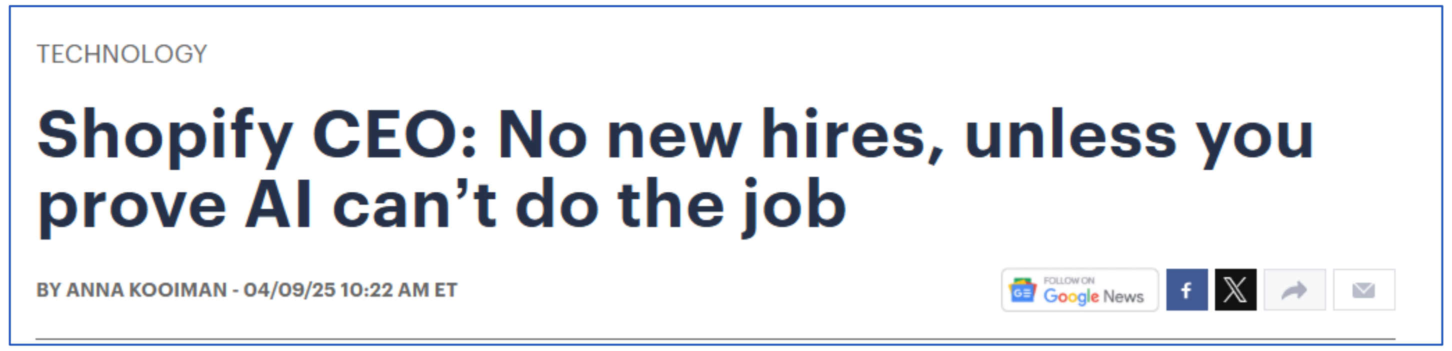
> *Developers using AI Code Assistants are 55% more productive.*
>
> – Github Copilot Analysis



Silicon Valley CEO says 'vibe coding' lets 10 engineers do the work of 100—here's how to use it

BY PRESTON FORE

March 26, 2025 at 5:20 AM EDT



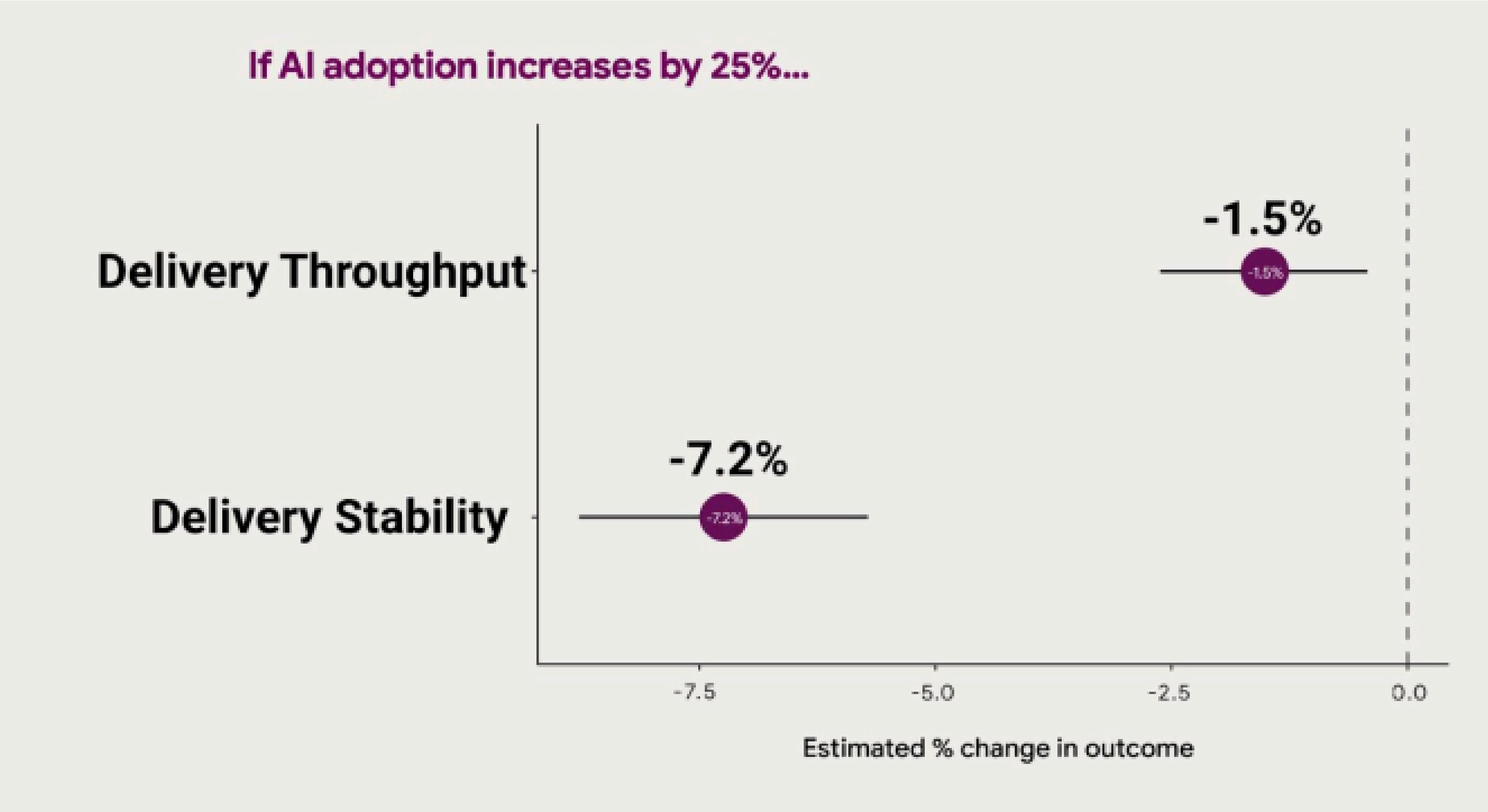
TECHNOLOGY

Shopify CEO: No new hires, unless you prove AI can't do the job

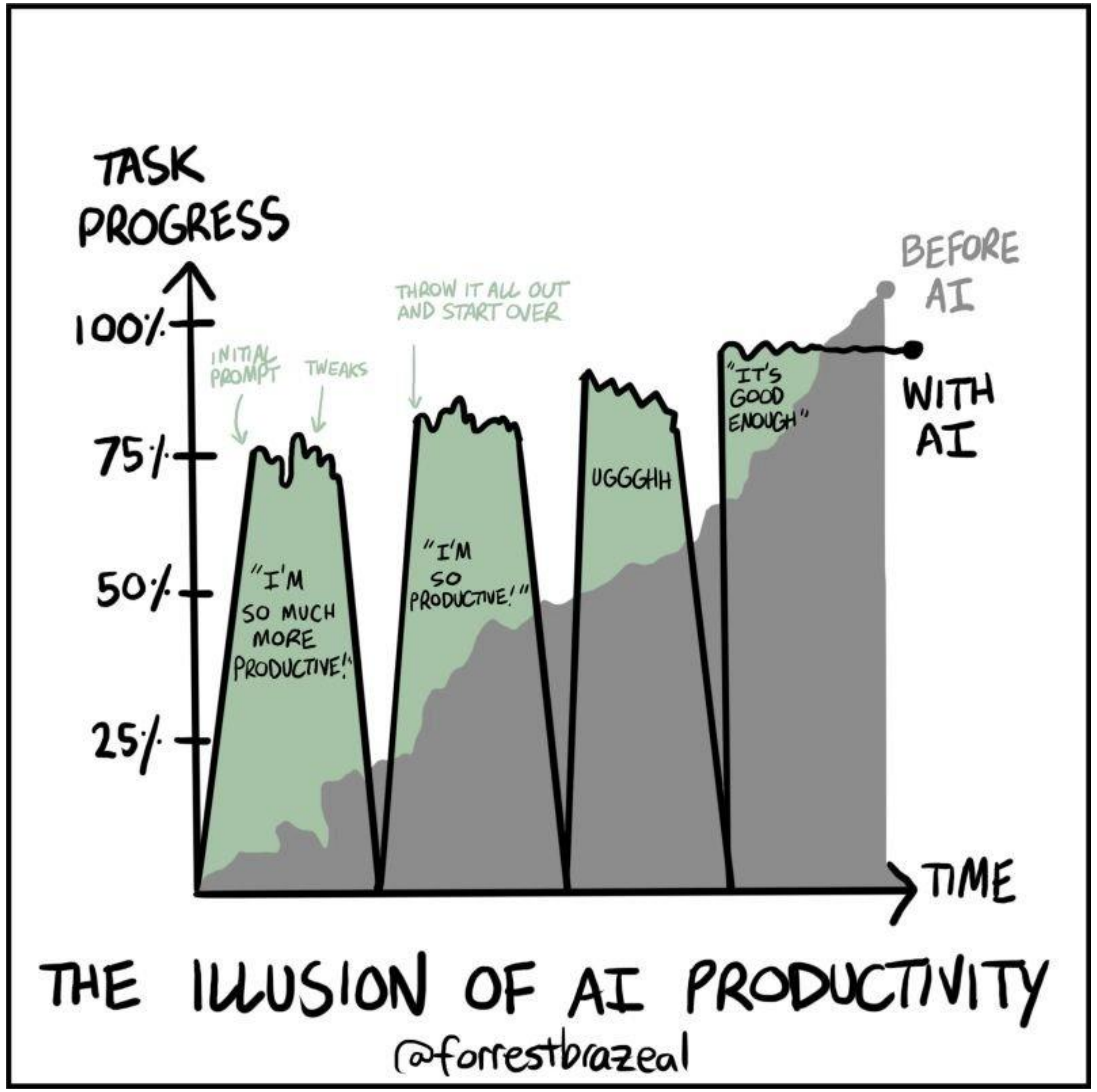
BY ANNA KOOIMAN - 04/09/25 10:22 AM ET

The Reality…

Development Delivery with AI

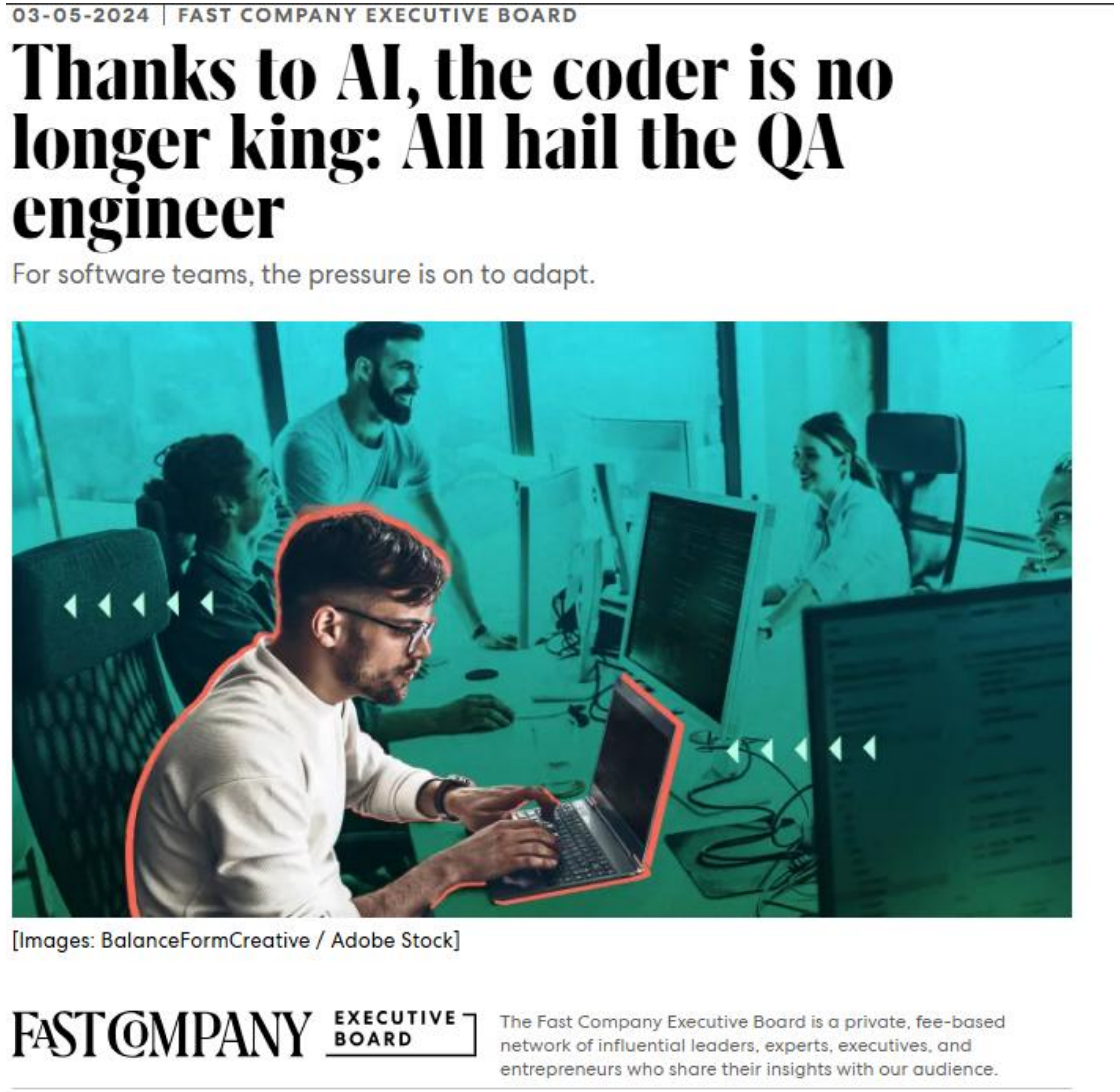


Google DORA 2024's extrapolated change in delivery stability per 25% increase in AI



Impact to Testing

Developers using AI Code Assistants are 55% more productive



03-05-2024 | FAST COMPANY EXECUTIVE BOARD

Thanks to AI, the coder is no longer king: All hail the QA engineer

For software teams, the pressure is on to adapt.

[Images: BalanceFormCreative / Adobe Stock]

FAST COMPANY EXECUTIVE BOARD

The Fast Company Executive Board is a private, fee-based network of influential leaders, experts, executives, and entrepreneurs who share their insights with our audience.

The Two Questions

HOW CAN WE LEVERAGE ADVANCEMENTS IN AI TO TEST SOFTWARE BETTER?

HOW CAN WE TEST SOFTWARE THAT LEVERAGES AI?

AI Use Cases in Testing

Test Prioritisation
Use Machine Learning to predict an optimal set of tests based on risk of code or functional change.

Self Healing
Leverage Artificial Intelligence to repair automated test cases in real-time and find the most likely replacement candidate.

Test Data Generation
Generate meaningful & realistic synthetic test data for your test environments.

Automated Test Script Generation
Use Generative AI to automatically generate meaningful automation from written test cases.

IDE Code Assistants
Use LLMs to sit beside the user and help out

Visual Functional Automation
Leverage AI to identify elements on screen and use OCR to translate text. Allows for automation over Citrix / RDP connections.

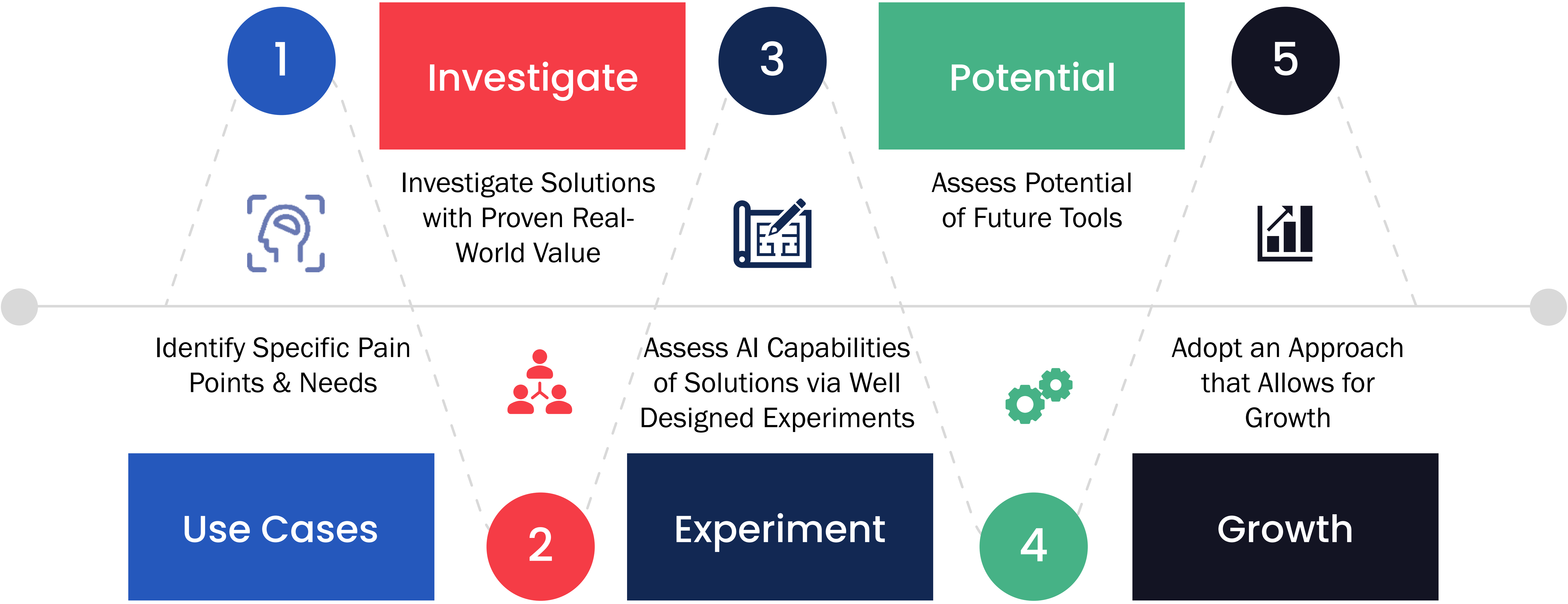
Manual Test Case Generation
Use Generative AI to automatically generate meaningful and understandable manual Test Cases.

Visual Testing
Use Machine Learning to identified which changes in rendered screen are important to the users.

API/Contract Testing
Use Machine Learning to analyze API Specs and Build Tests

Autonomous Testing
Point it at an application / logs it returns a report.

AI Test Tool Selection & Analysis Framework



1

Use Cases

Identify Specific Pain Points & Needs

2

Investigate

Investigate Solutions with Proven Real-World Value

3

Experiment

Assess AI Capabilities of Solutions via Well Designed Experiments

4

Potential

Assess Potential of Future Tools

5

Growth

Adopt an Approach that Allows for Growth

Manual Test Case Generation

Enterprise Readiness

3 of 5

How AI may help

Use Generative AI to automatically generate meaningful and understandable manual Test Cases from the requirements or user stories in the system.

Potential Benefits

- Generate comprehensive test ideas faster and with less effort.
- Increase coverage with depth of testing ideas.

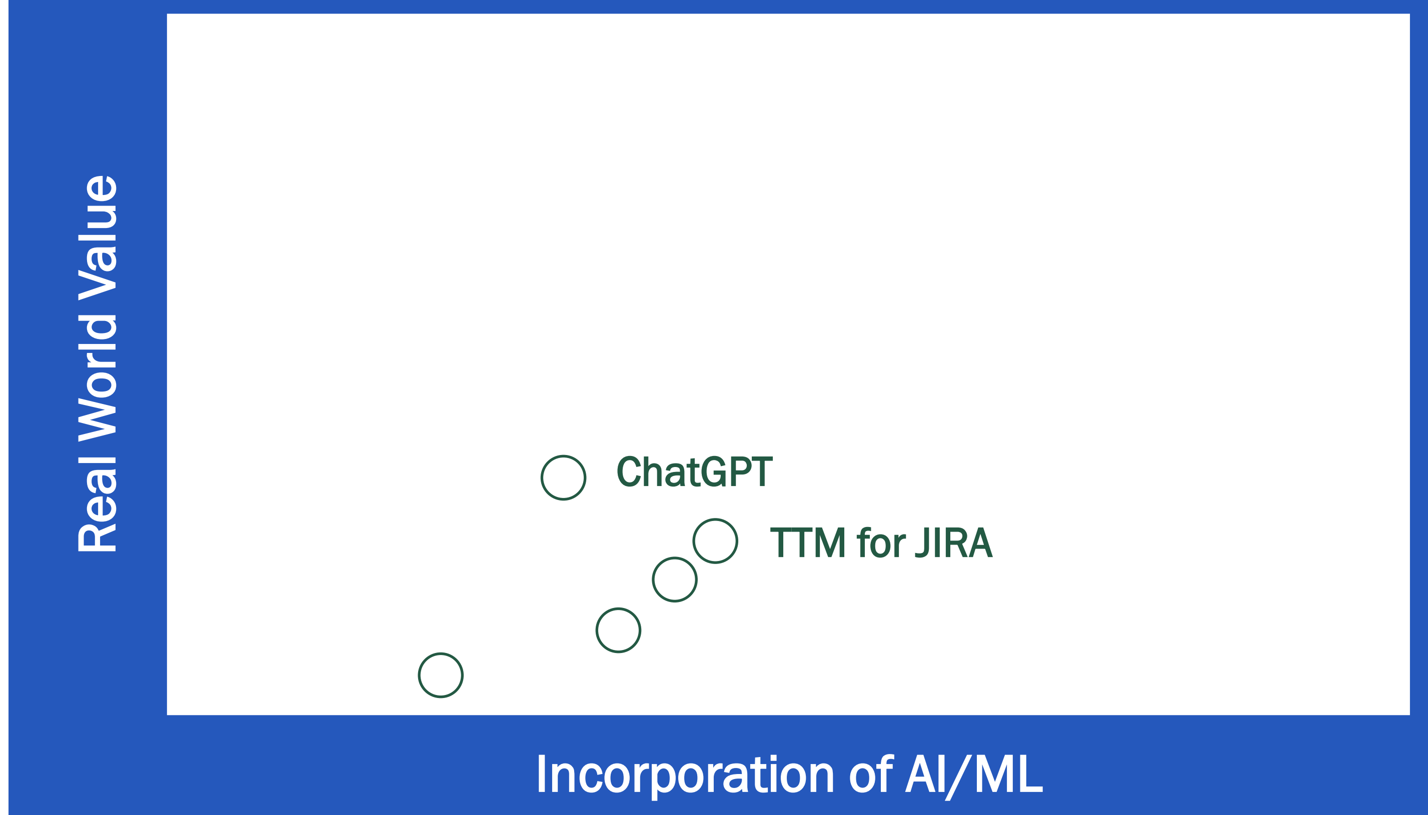
Inherent Risks

- Does not generate tests for important requirements. Leaving teams with unknown gaps.
- Generates tests that are nonsensical.
- Does not look at existing test suite for test case generation.

Current TTC Recommendation

We recommend significant human oversight – specifically around test coverage. Key features of early adopters would be lower risk, lower data complexity, more generic application flows, and mature requirements processes.

What is TTC seeing in the market?



Skillful crafting of test cases is mostly down to prompt engineering. AI-Powered Manual Test Case Generation Tools ship with custom prompts that we don't see – but that are tuned to be better than our first experiments.

Tools like ChatGPT and other general purpose LLMs allow more control over prompting and allow us to add additional context which may be critical to getting good coverage of important risks.

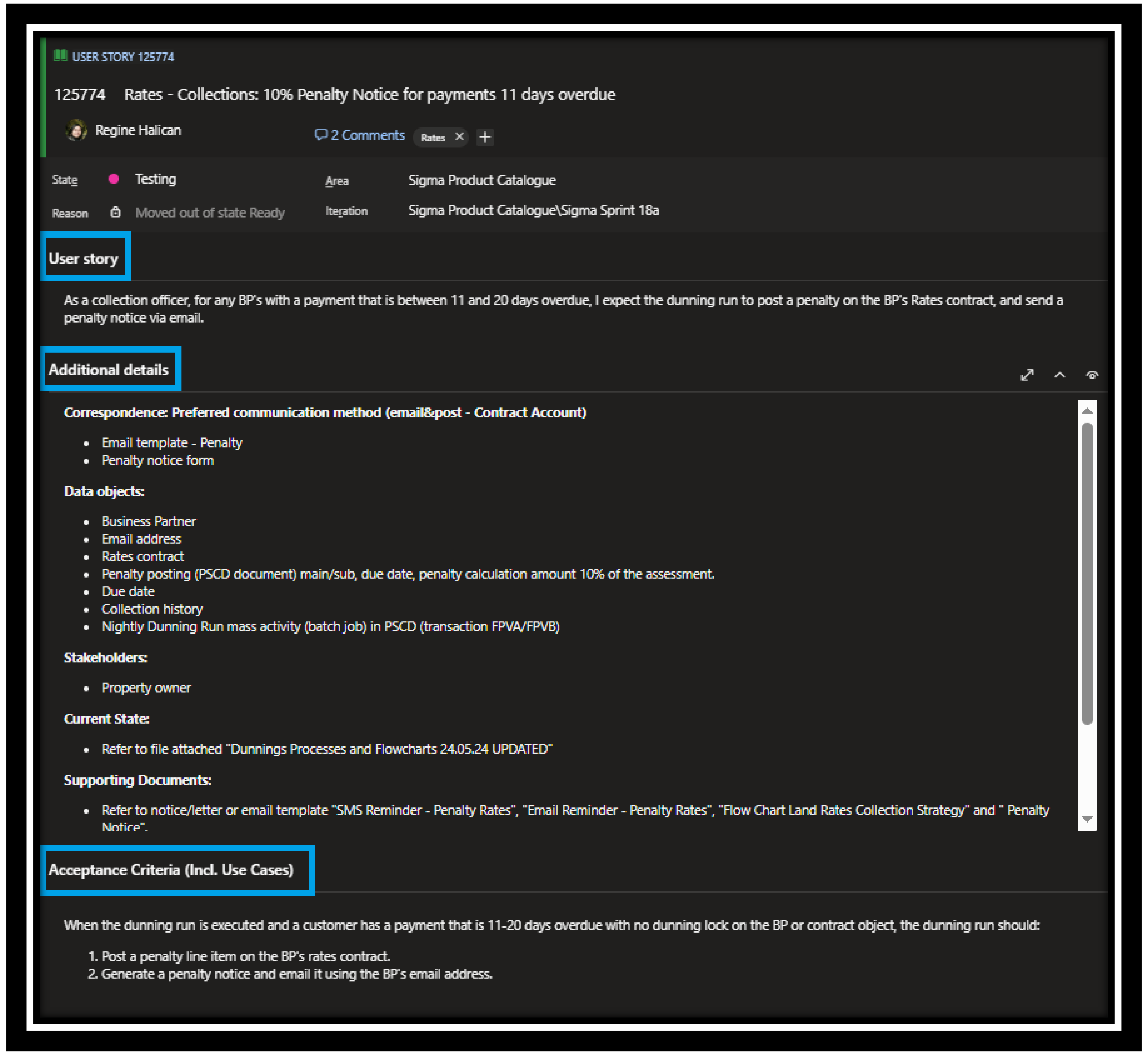
We expect the use of AI for test case generation to continue and become standard in the market.

AI-Powered Test Script Development in Azure DevOps

Tauranga City

Why AI in Test case Development?

- Deriving Test Scenarios from User Story's can be time-consuming depending on complexity and information available.
- Creation of test cases is often time-consuming
- Improve checking of requirement coverage / consider negative testing scenarios
- Lack of automation in linking test cases to user stories for traceability



125774 Rates - Collections: 10% Penalty Notice for payments 11 days overdue

Regine Halican

State: Testing
Reason: Moved out of state Ready
Area: Sigma Product Catalogue
Iteration: Sigma Product Catalogue\Sigma Sprint 18a

User story

As a collection officer, for any BP's with a payment that is between 11 and 20 days overdue, I expect the dunning run to post a penalty on the BP's Rates contract, and send a penalty notice via email.

Additional details

Correspondence: Preferred communication method (email&post - Contract Account)

- Email template - Penalty
- Penalty notice form

Data objects:

- Business Partner
- Email address
- Rates contract
- Penalty posting (PSCD document) main/sub, due date, penalty calculation amount 10% of the assessment.
- Due date
- Collection history
- Nightly Dunning Run mass activity (batch job) in PSCD (transaction FPVA/FPVB)

Stakeholders:

- Property owner

Current State:

- Refer to file attached "Dunnings Processes and Flowcharts 24.05.24 UPDATED"

Supporting Documents:

- Refer to notice/letter or email template "SMS Reminder - Penalty Rates", "Email Reminder - Penalty Rates", "Flow Chart Land Rates Collection Strategy" and " Penalty Notice".

Acceptance Criteria (Incl. Use Cases)

When the dunning run is executed and a customer has a payment that is 11-20 days overdue with no dunning lock on the BP or contract object, the dunning run should:

1. Post a penalty line item on the BP's rates contract.
2. Generate a penalty notice and email it using the BP's email address.

The AI-Driven Solution

01

Easy to use solution for users, One-click test generation from User Stories information

02

Uses Azure OpenAI API service to generate structured test cases

03

Automatically adds test cases to a test plan In Azure DevOps

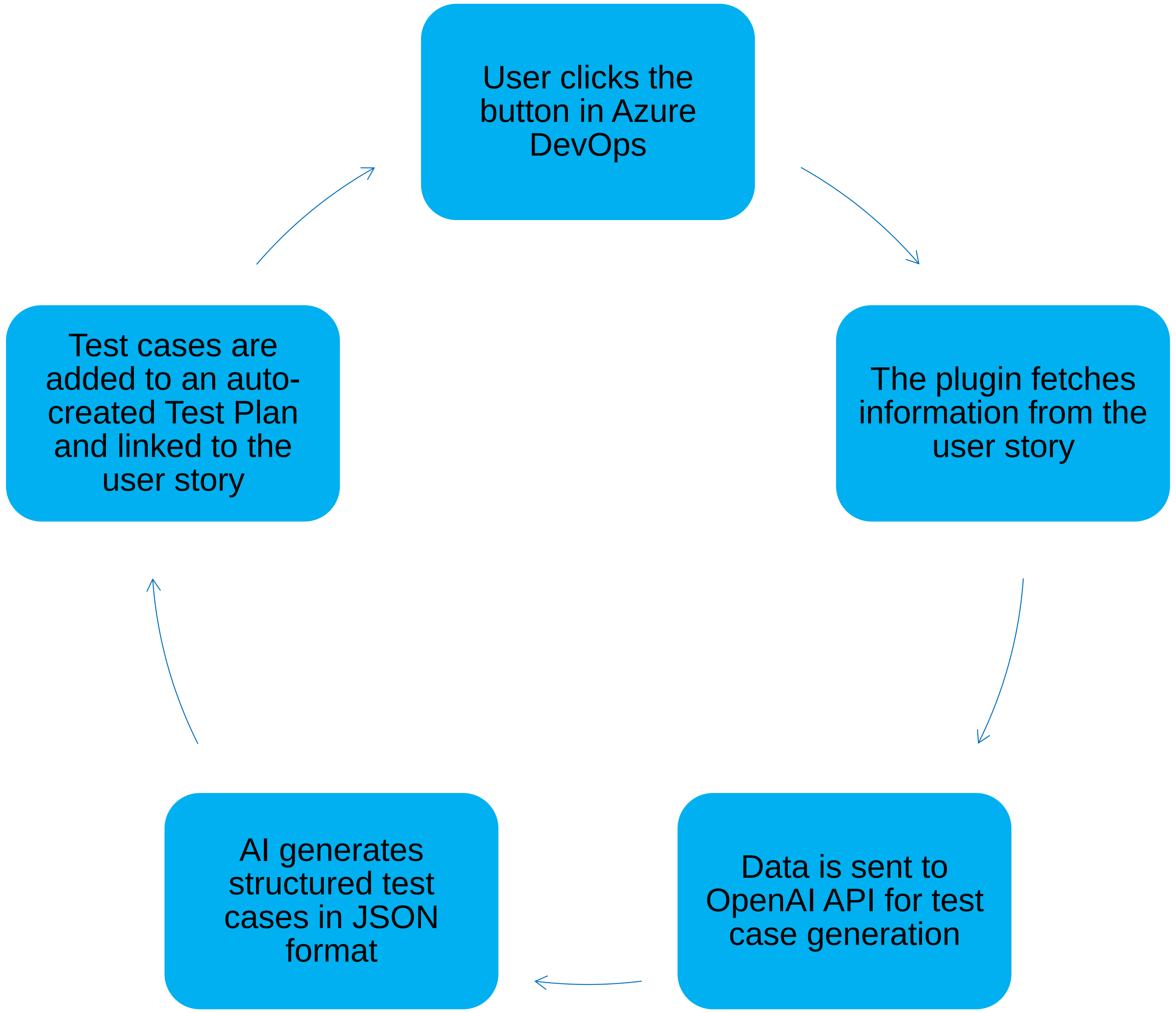
04

Testcases should be linked to the User Story

05

The solution should be secure, and Costs should be within reason

How It Works - High-Level Workflow



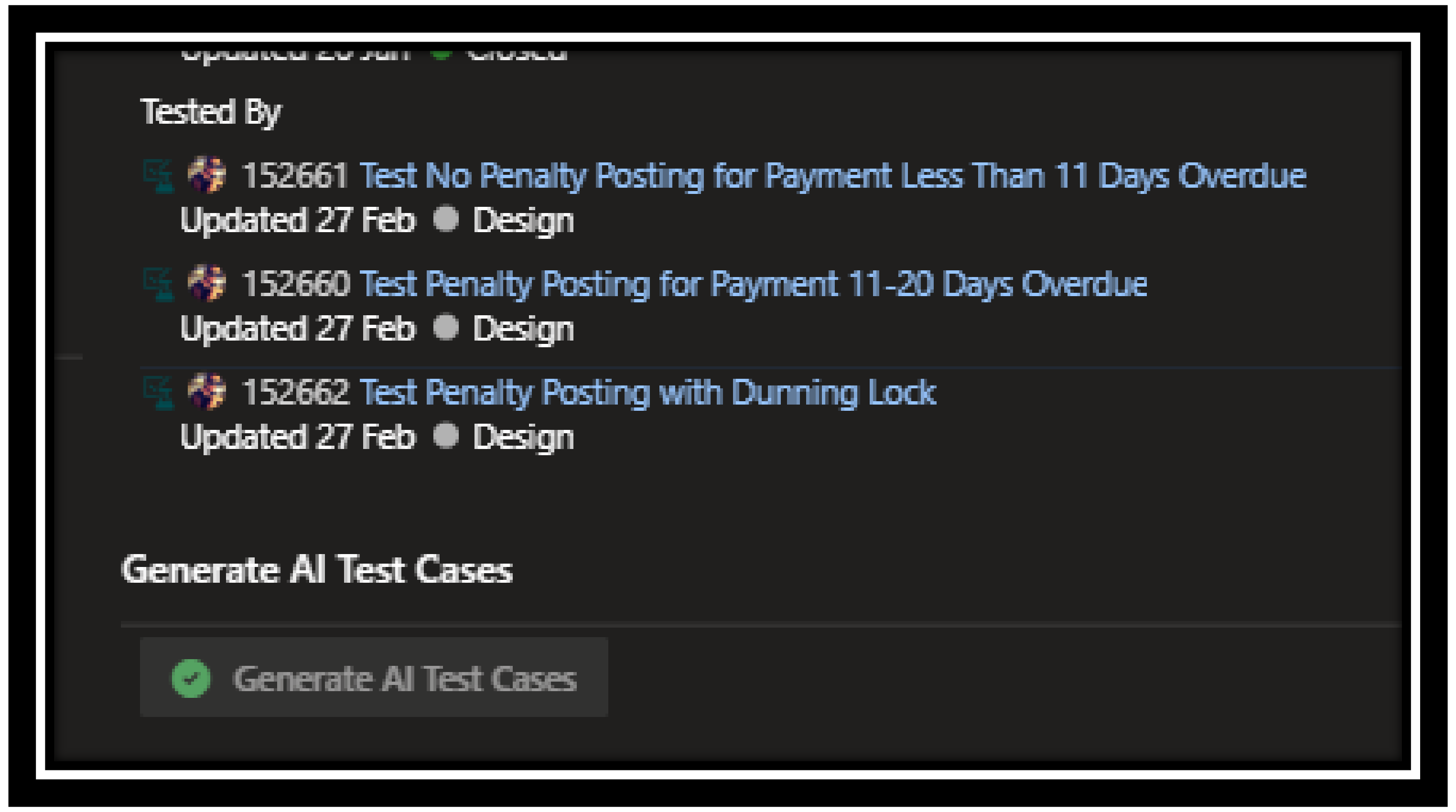
User clicks the button in Azure DevOps

The plugin fetches information from the user story

Data is sent to OpenAI API for test case generation

AI generates structured test cases in JSON format

Test cases are added to an auto-created Test Plan and linked to the user story



Tested By

152661 Test No Penalty Posting for Payment Less Than 11 Days Overdue
Updated 27 Feb ● Design

152660 Test Penalty Posting for Payment 11-20 Days Overdue
Updated 27 Feb ● Design

152662 Test Penalty Posting with Dunning Lock
Updated 27 Feb ● Design

Generate AI Test Cases

✓ Generate AI Test Cases

Lessons Learned & Challenges

01

Testcases generated are only as good as the information passed into the prompt

02

For complex systems and custom code context is required

03

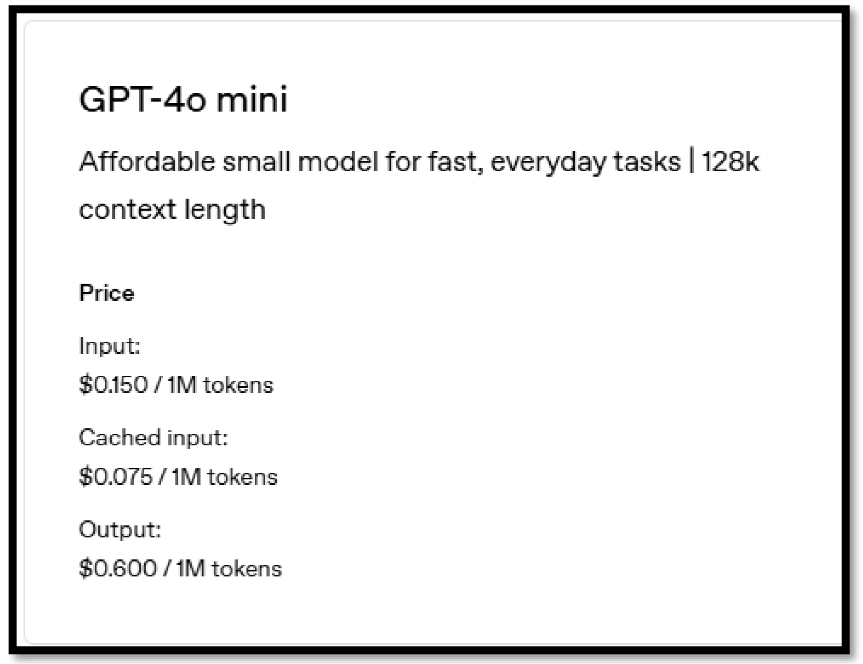
Cost effective $0.000672 NZD for the example shown used 1300 total tokens

04

Another example for a complex testcase with 45 detailed steps containing 14000 words cost $0.017911 NZD

05

Effective from Clicking the button to the testcases being in DevOps is around 25 seconds



GPT-4o mini

Affordable small model for fast, everyday tasks | 128k context length

Price

Input:
$0.150 / 1M tokens

Cached input:
$0.075 / 1M tokens

Output:
$0.600 / 1M tokens

Future Enhancements

Fine-tuning AI responses for better accuracy using RAG (Retrieval-Augmented Generation)

Standardize User Story Formats across sprint teams for alignment Connextra / MoSCoW

Experimenting with different models

Look for other opportunities e.g Video transcriptions conversion to User Stories/ Testcases

AI Use Cases in Testing

Test Prioritisation
Use Machine Learning to predict an optimal set of tests based on risk of code or functional change.

Self Healing
Leverage Artificial Intelligence to repair automated test cases in real-time and find the most likelv replacement can

utomatically ngful automation from written test cases.

IDE Code Assistants
Use LLMs to sit beside the user and help out

Mutation/Fuzz Testing
Implement mutations to your test cases to increas defect detection. Lever nprove fuzzing.

e Generation
automatically gful and al Test Cases.

al Testing
achine Learning to identified which changes in rendered screen are important to the users.

API/Contract Testing
Use Machine Learning to analyze API Specs and Build Tests

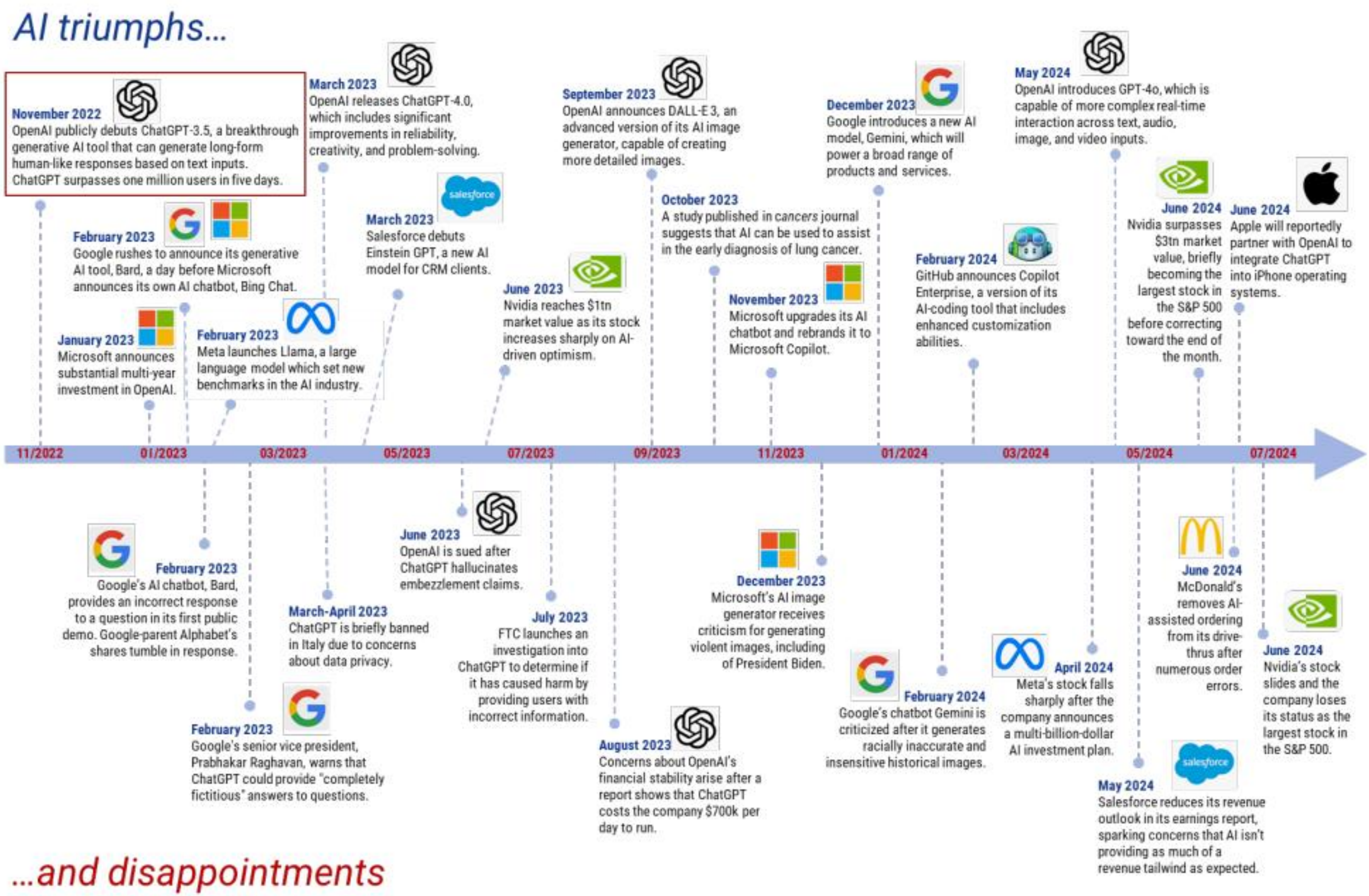
Autonomous Testing
Point it at an application / logs it returns a report.

You are responsible for the behaviour of the output of an AI agent you chose to use.

TEST IT!

Use your critical thinking skills.

TESTING AI



AI triumphs...

November 2022
OpenAI publicly debuts ChatGPT-3.5, a breakthrough generative AI tool that can generate long-form human-like responses based on text inputs. ChatGPT surpasses one million users in five days.

January 2023
Microsoft announces substantial multi-year investment in OpenAI.

February 2023
Google rushes to announce its generative AI tool, Bard, a day before Microsoft announces its own AI chatbot, Bing Chat.

February 2023
Meta launches Llama, a large language model which set new benchmarks in the AI industry.

March 2023
OpenAI releases ChatGPT-4.0, which includes significant improvements in reliability, creativity, and problem-solving.

March 2023
Salesforce debuts Einstein GPT, a new AI model for CRM clients.

June 2023
Nvidia reaches $1tn market value as its stock increases sharply on AI-driven optimism.

September 2023
OpenAI announces DALL-E 3, an advanced version of its AI image generator, capable of creating more detailed images.

October 2023
A study published in *cancers* journal suggests that AI can be used to assist in the early diagnosis of lung cancer.

November 2023
Microsoft upgrades its AI chatbot and rebrands it to Microsoft Copilot.

December 2023
Google introduces a new AI model, Gemini, which will power a broad range of products and services.

February 2024
GitHub announces Copilot Enterprise, a version of its AI-coding tool that includes enhanced customization abilities.

May 2024
OpenAI introduces GPT-4o, which is capable of more complex real-time interaction across text, audio, image, and video inputs.

June 2024
Nvidia surpasses $3tn market value, briefly becoming the largest stock in the S&P 500 before correcting toward the end of the month.

June 2024
Apple will reportedly partner with OpenAI to integrate ChatGPT into iPhone operating systems.

February 2023
Google's AI chatbot, Bard, provides an incorrect response to a question in its first public demo. Google-parent Alphabet's shares tumble in response.

February 2023
Google's senior vice president, Prabhakar Raghavan, warns that ChatGPT could provide "completely fictitious" answers to questions.

March-April 2023
ChatGPT is briefly banned in Italy due to concerns about data privacy.

June 2023
OpenAI is sued after ChatGPT hallucinates embezzlement claims.

July 2023
FTC launches an investigation into ChatGPT to determine if it has caused harm by providing users with incorrect information.

August 2023
Concerns about OpenAI's financial stability arise after a report shows that ChatGPT costs the company $700k per day to run.

December 2023
Microsoft's AI image generator receives criticism for generating violent images, including of President Biden.

February 2024
Google's chatbot Gemini is criticized after it generates racially inaccurate and insensitive historical images.

April 2024
Meta's stock falls sharply after the company announces a multi-billion-dollar AI investment plan.

May 2024
Salesforce reduces its revenue outlook in its earnings report, sparking concerns that AI isn't providing as much of a revenue tailwind as expected.

June 2024
McDonald's removes AI-assisted ordering from its drive-thrus after numerous order errors.

June 2024
Nvidia's stock slides and the company loses its status as the largest stock in the S&P 500.

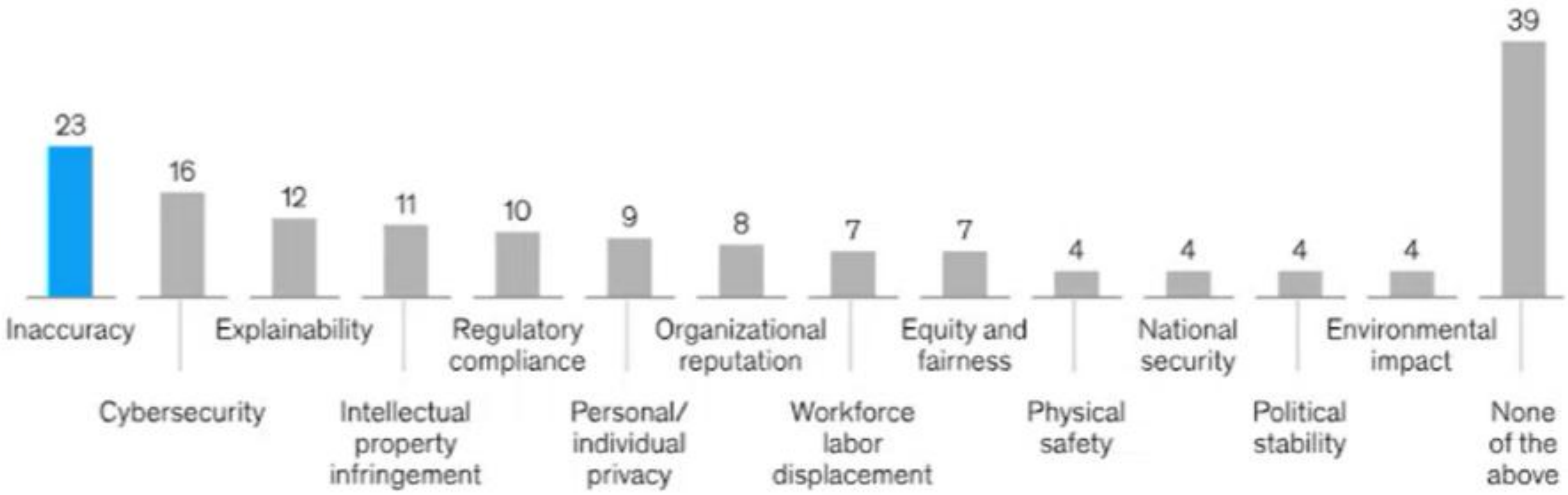
...and disappointments

Note: This does not constitute an exhaustive list of all AI-related developments.
Source: BBC, cancers, OpenAI, tech.co, Google, various news sources, compiled by Goldman Sachs GIR.

AI Adoption Studies

Nearly one-quarter of respondents say their organizations have experienced negative consequences from generative AI's inaccuracy.

Generative-AI-related risks that caused negative consequences for organizations,[1] % of respondents

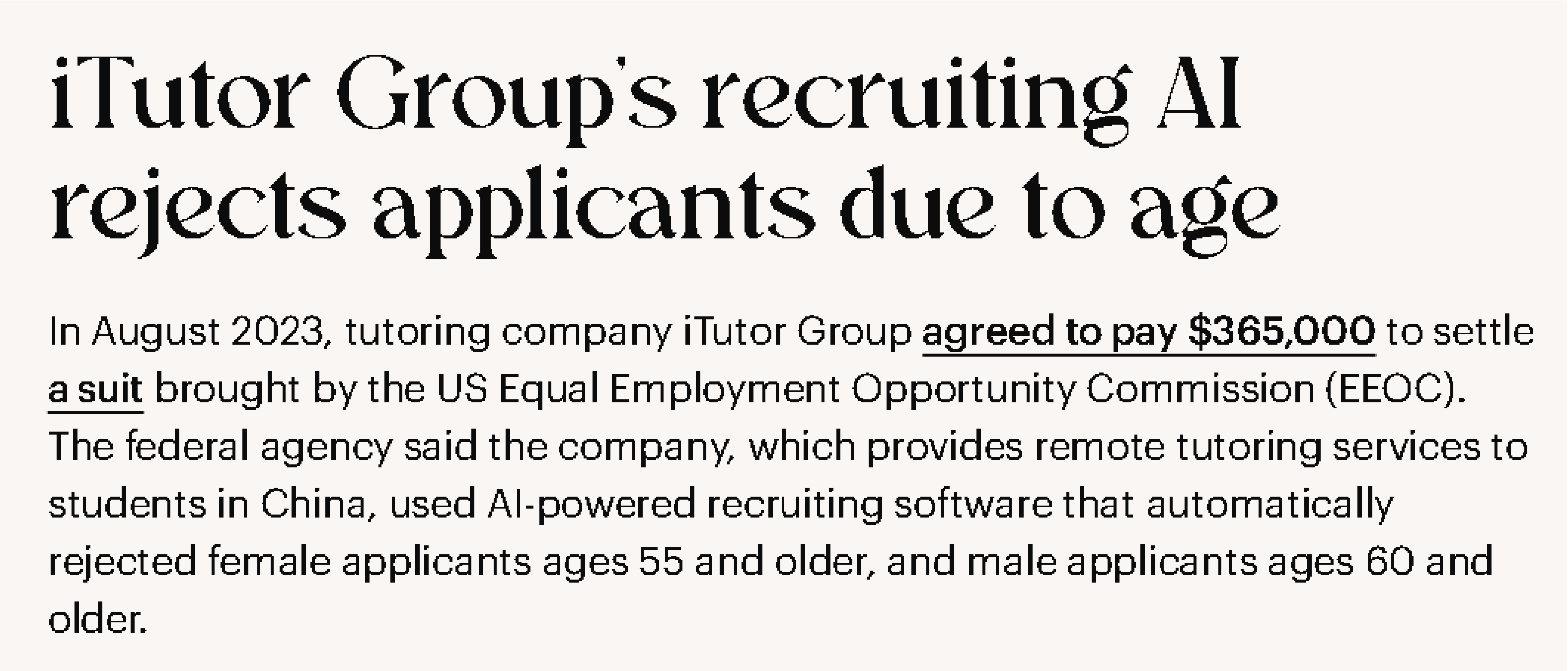


[1] Question was asked only of respondents whose organizations have adopted generative AI in at least 1 function, n = 876. The 17 percent of respondents who said "don't know/not applicable" are not shown.

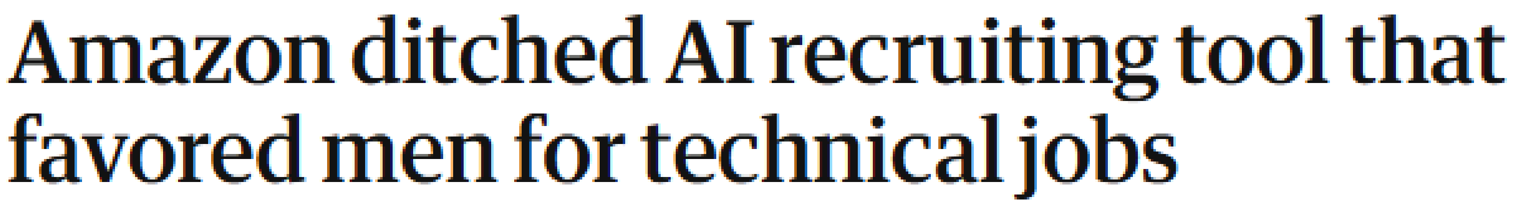
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

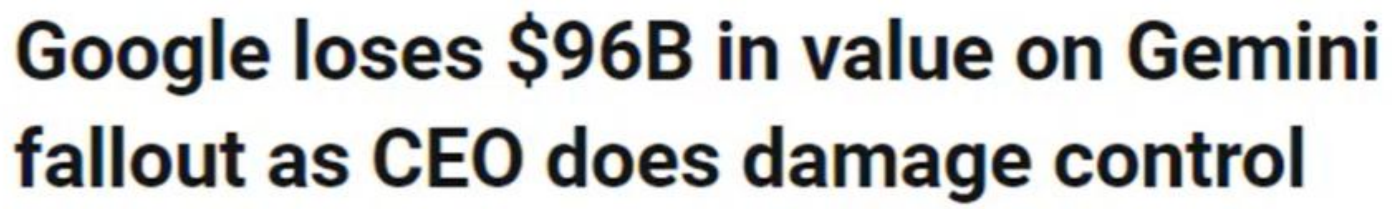
McKinsey & Company

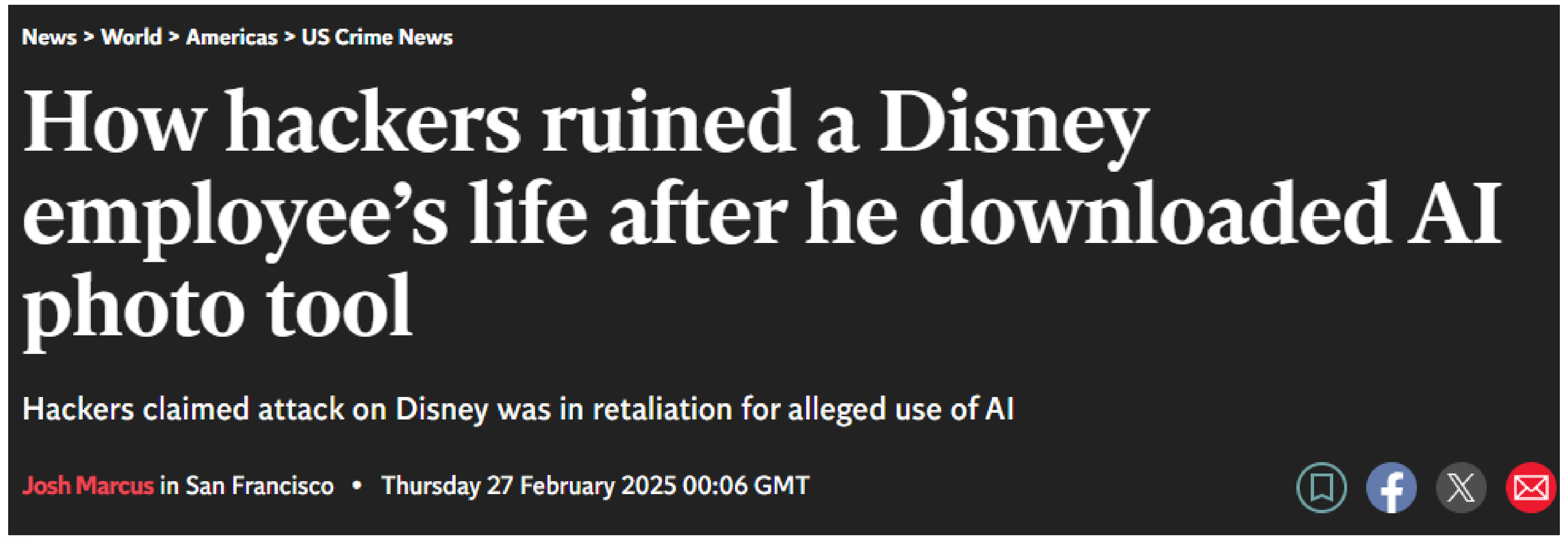
AI in Production – Issues


iTutor Group's recruiting AI rejects applicants due to age

In August 2023, tutoring company iTutor Group agreed to pay $365,000 to settle a suit brought by the US Equal Employment Opportunity Commission (EEOC). The federal agency said the company, which provides remote tutoring services to students in China, used AI-powered recruiting software that automatically rejected female applicants ages 55 and older, and male applicants ages 60 and older.


Amazon ditched AI recruiting tool that favored men for technical jobs

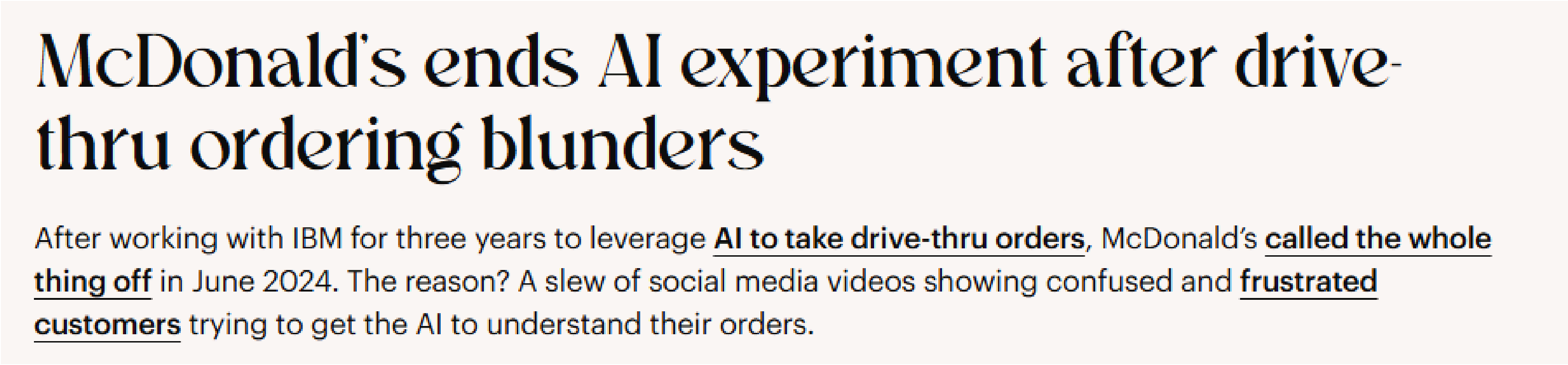

Google loses $96B in value on Gemini fallout as CEO does damage control


News > World > Americas > US Crime News

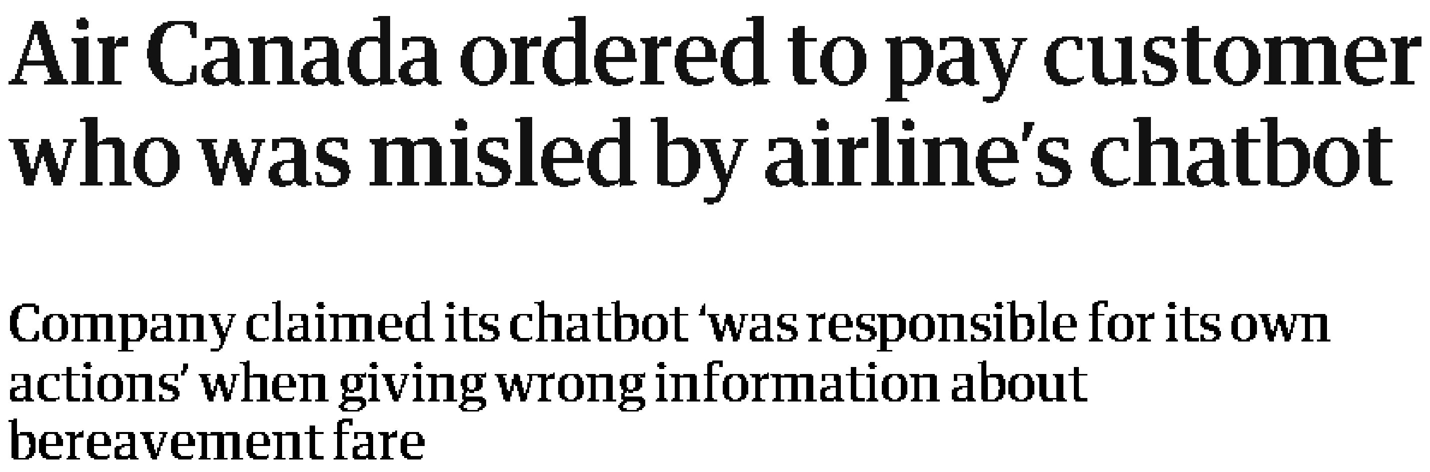
How hackers ruined a Disney employee's life after he downloaded AI photo tool

Hackers claimed attack on Disney was in retaliation for alleged use of AI

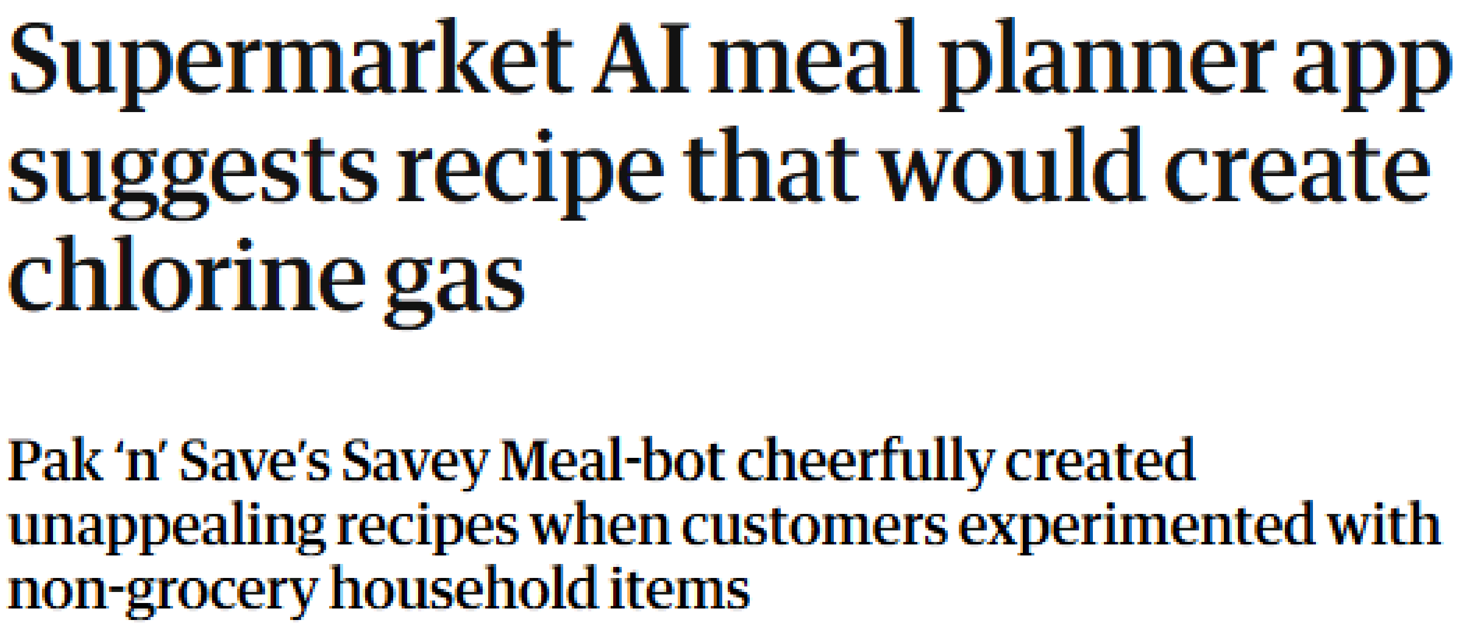
Josh Marcus in San Francisco • Thursday 27 February 2025 00:06 GMT


McDonald's ends AI experiment after drive-thru ordering blunders

After working with IBM for three years to leverage **AI to take drive-thru orders**, McDonald's called the whole thing off in June 2024. The reason? A slew of social media videos showing confused and frustrated customers trying to get the AI to understand their orders.


Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

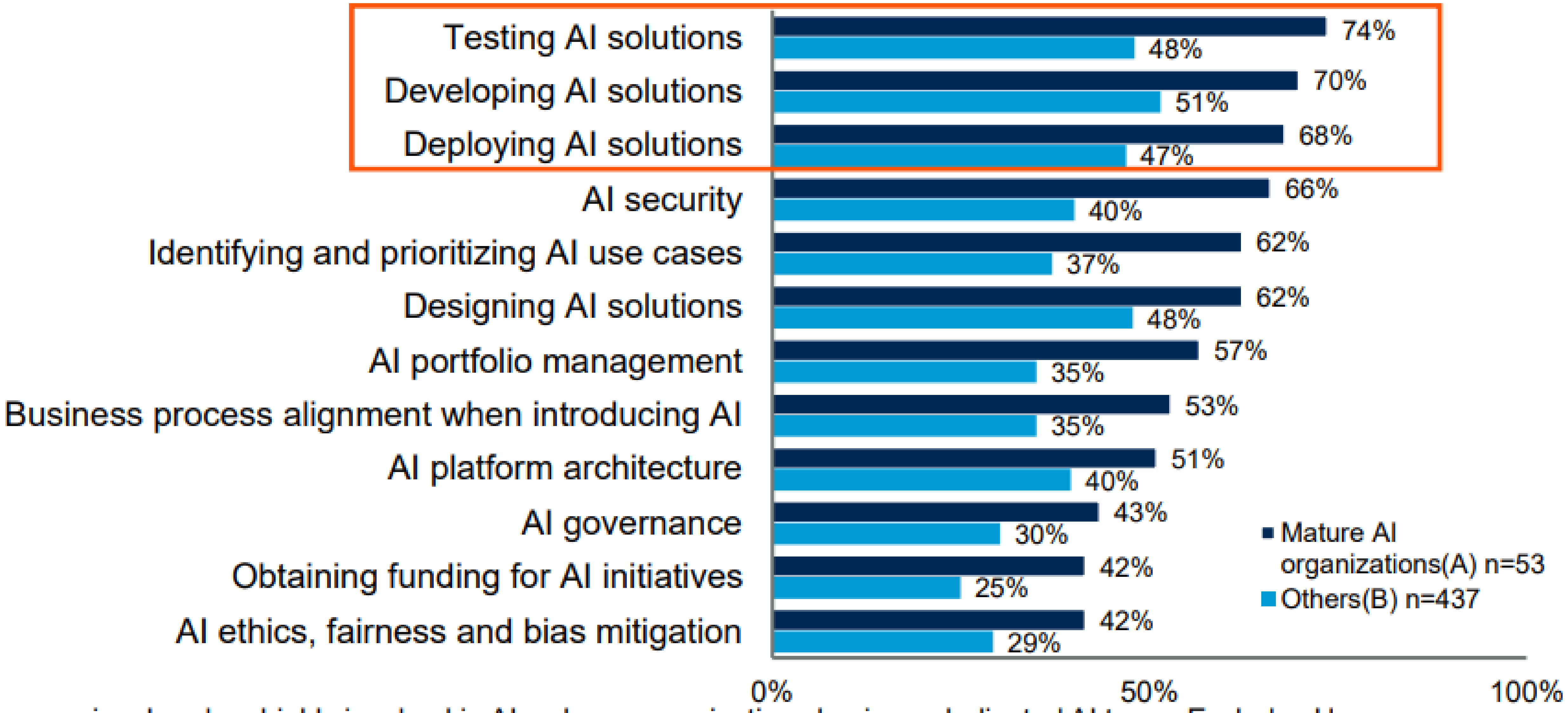

Supermarket AI meal planner app suggests recipe that would create chlorine gas

Pak 'n' Save's Savey Meal-bot cheerfully created unappealing recipes when customers experimented with non-grocery household items

Testing AI

Mature Organizations Double Down on AI engineering

Tasks dedicated to AI team by AI maturity
Multiple responses



n varies, Leaders highly involved in AI , whose organizations having a dedicated AI team; Excludes Unsure

Q02: What tasks is the dedicated AI team accountable for?

Source: 2023 Gartner AI in the Enterprise Survey

What Makes Testing AI Different/Challenging

Emergent Behaviour

Large Machine Learning Models exhibit “Emergent Behaviour”. That is behaviour that the model was not explicitly designed for and is not easily understandable.

These leads to challenges in testing models including small changes having a large impact, difficulty in isolating the impact of a change, lack of transparency/visibility, and unintended negative impacts of changes.

Non-Determinism

AI systems also often exhibit non-determinism either intentionally or unintentionally.

This makes typical testing approaches difficult or impossible to implement. For example, there will not always be a simple pass/fail result, tests may need to be repeated to see the variability within a system, and it also increases the risk of an important defect being missed.

Qualitative Assessment

As AI systems can generate human-like results, they often require a combination of both qualitative and quantitative methods for evaluation.

Did the AI communicate clearly? Is the code generated efficient, maintainable and readable, not just effective. Evaluations need to consider the impact of changes to the model across multiple prompts/contexts and prioritize importance. This is very different to traditional functional testing.

Tools For Automated LLM Testing

DeepEval

An Open Source framework written in python for evaluation and benchmarking of LLMs.

PromptFoo

An Open Source framework written in JavaScript using Node.js to evaluate and benchmark prompt variation

ML Flow

An Open Source framework to manage LLM lifecycle management written by DataBricks.

TruLens

An open source framework written in python for LLM evaluation and benchmarking.

Giskard

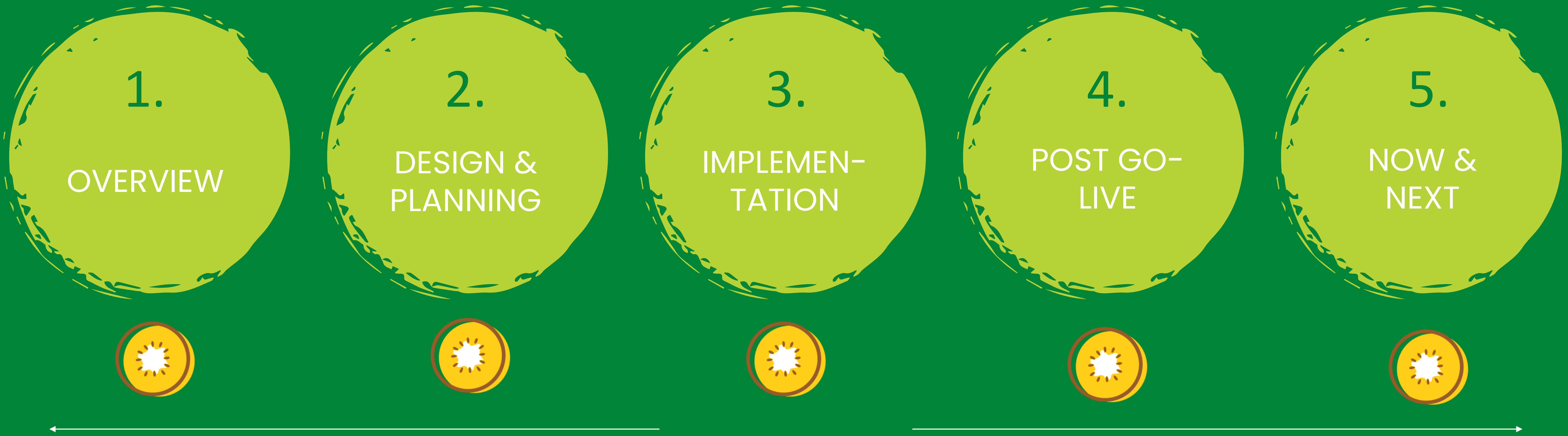
An evaluation framework in python with enterprise reporting dashboards. Available in an open source base version and an commercial enterprise management platform.

Patronous.ai

A commercial platform with custom evaluation data sets and benchmarks. Includes evaluations particularly tuned for financial analysis, copyright detection, and other critical functions.

Zespri's AI Journey

Zespri's AI Journey



1. OVERVIEW

- Background
- Complexity
- Vast amounts of information to stay compliant and productive in the Kiwifruit industry

2. DESIGN & PLANNING

- Introduced LLM (Large Language Models)
- Tight integration within the Canopy portal

3. IMPLEMEN-TATION

- SIT
- UAT
- Security

4. POST GO-LIVE

- Rolled out in stages
- Positive feedback from growers

5. NOW & NEXT

- Microsoft Copilot for productivity gains
- Test case generation using AI for productivity gains

Study Match

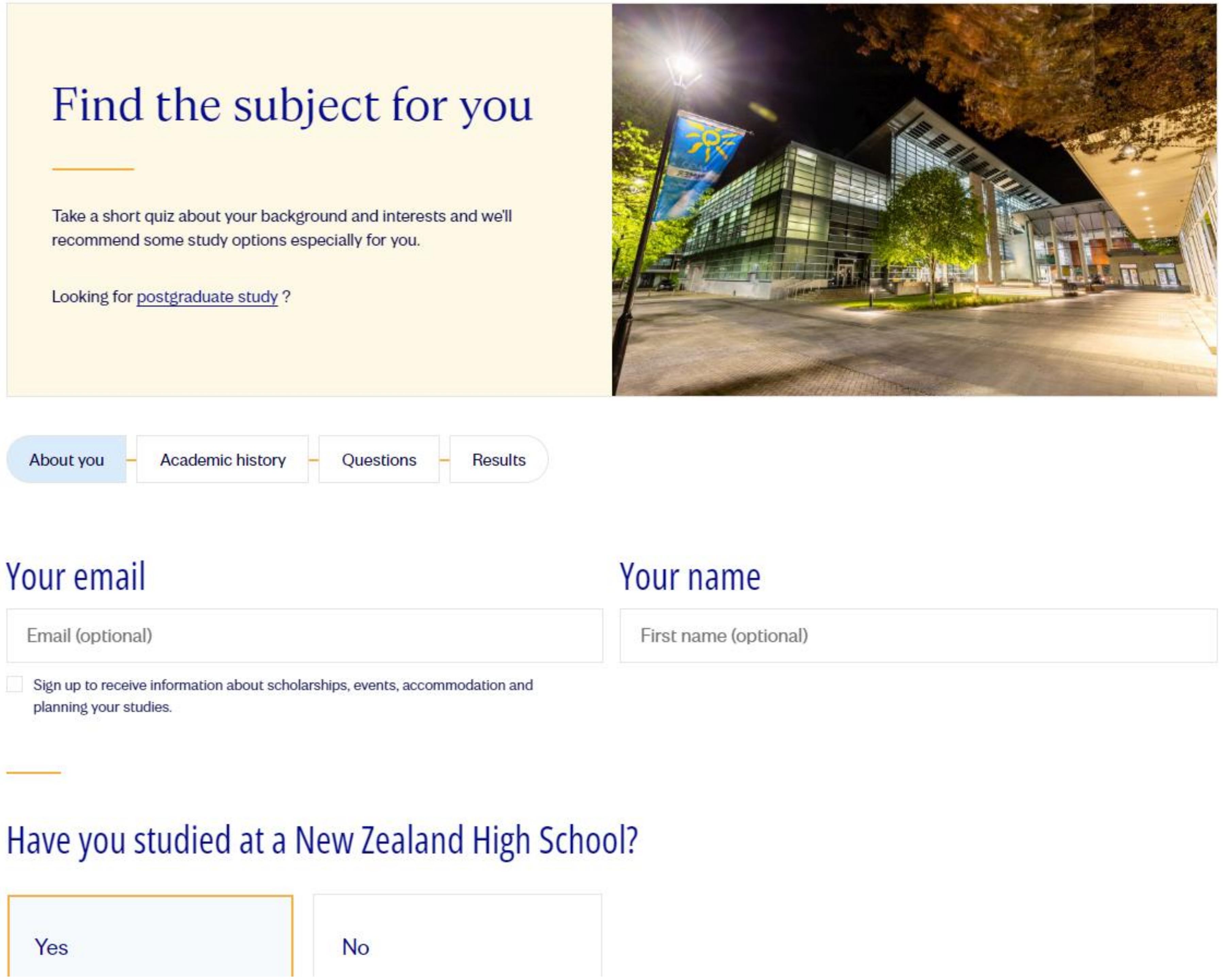
A Course Finder tool using AI

University of Otago

ŌTĀKOU WHAKAIHU WAKA

Introducing Study Match

- It is a tool to assist early-journey future students who want help exploring tertiary study options.
- With the large number of subjects available, maintaining a predetermined decision tree was impractical, particularly with changes to subjects that occur over time.
- AI is used to generate the results instead. The AI receives anonymised student responses, generates search keywords, and these keywords are then sent to a search package to return 15 results based on the subject pages.
- Subjects and subject content are then sent back to the AI to write a rationale.



Find the subject for you

Take a short quiz about your background and interests and we'll recommend some study options especially for you.

Looking for postgraduate study ?

Your email

Email (optional)

□ Sign up to receive information about scholarships, events, accommodation and planning your studies.

Your name

First name (optional)

Have you studied at a New Zealand High School?

Yes

No

Test Approach

Automated testing – 19+ rounds of 1,000 inputs

- Randomised and real data
- All subjects returned
- Distribution is reasonable

User acceptance testing

- stakeholders testing functionality

Usability testing

- in-person testing with students to assess usability and satisfaction

Beta testing

- distributing tool to students with survey for general feedback

Experience and lessons from testing an AI-driven tool

Non-Deterministic Behaviour impact

Automated testing and Exploratory Testing are absolutely critical

Guardrails are important

Testing AI - Key Take Aways

1

Testing AI is Different

Some of our traditional expectations will change. New techniques will be needed.

2

Testing AI is Exciting

New challenges, new tools to learn, new ways of thinking.

Might include moving past the test case paradigm.

3

You can Test AI

Your critical thinking skills, understanding of risk, and abilities to communicate what you discover are still going to be useful. This is not impossible for you to take up.

Questions & Discussion

✓ttc | Contact Us

New Zealand

Shed 19/Level 1 Princes Wharf
137 Quay St, City Centre,
Auckland 1010, New Zealand
+64 9 948 2225
info@ttcglobal.com

United States

25211 Grogans Mill Rd #450
The Woodlands,
Texas 77380
(832) 813-8063
sales.us@ttcglobal.com

Europe | UK

10 John Street
London WC1N 2EB
United Kingdom
+44 7384 719098
uk@ttcglobal.com

Singapore

Hong Leong Building
6 Raffles Quay, #33-03
Singapore 048581
+65 9822 6679
singapore@ttcglobal.com

United Arab Emirates

14th Floor, Al Khatem Tower
Wework Hub 71 Abu Dhabi
Global Market Square, Al
Maryah Island Abu Dhabi, UAE
+971 58 5233912
UAE@ttcglobal.com

India

6 Floor Westport S.No.
32/1A/1/30 to 38 & 54 Pan
Card Club Rd, Baner, Pune,
Maharashtra 411045
india@ttcglobal.com

Australia

Level 4, 50 Miller St
North Sydney
NSW 2060
+61 2 8999 1965
australia@ttcglobal.com