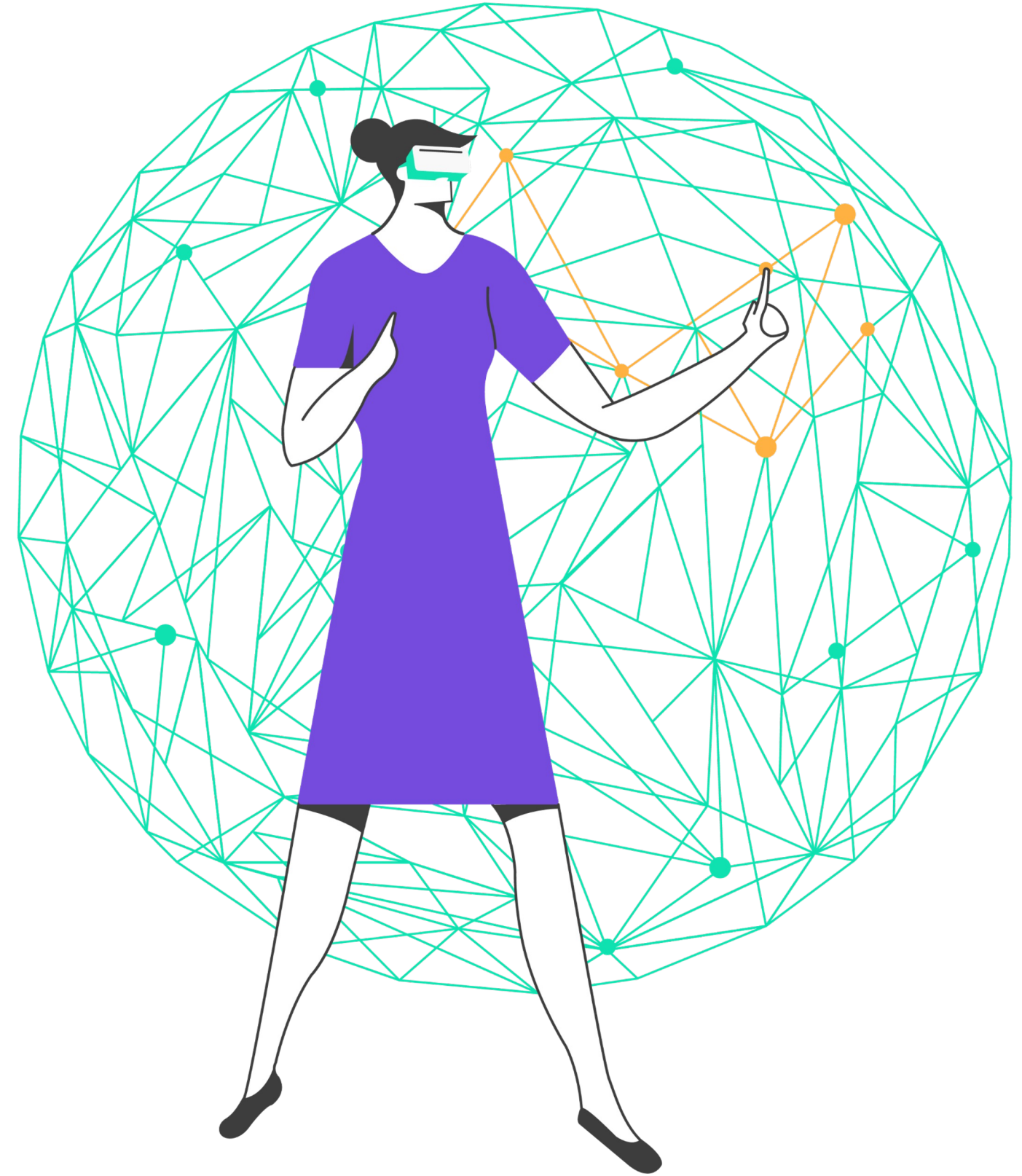


expo IQA 25

MADRID
May 20th,
21st & 22nd
2025

expoqqa.eu

New Challenges in Software Quality Brought by **GenIA** and **LLMs**



Federico Toledo

federico@abstracta.us

Original Sound: Off

View



Unmute

Start Video

Security

Participants

Chat

Share Screen

Polling

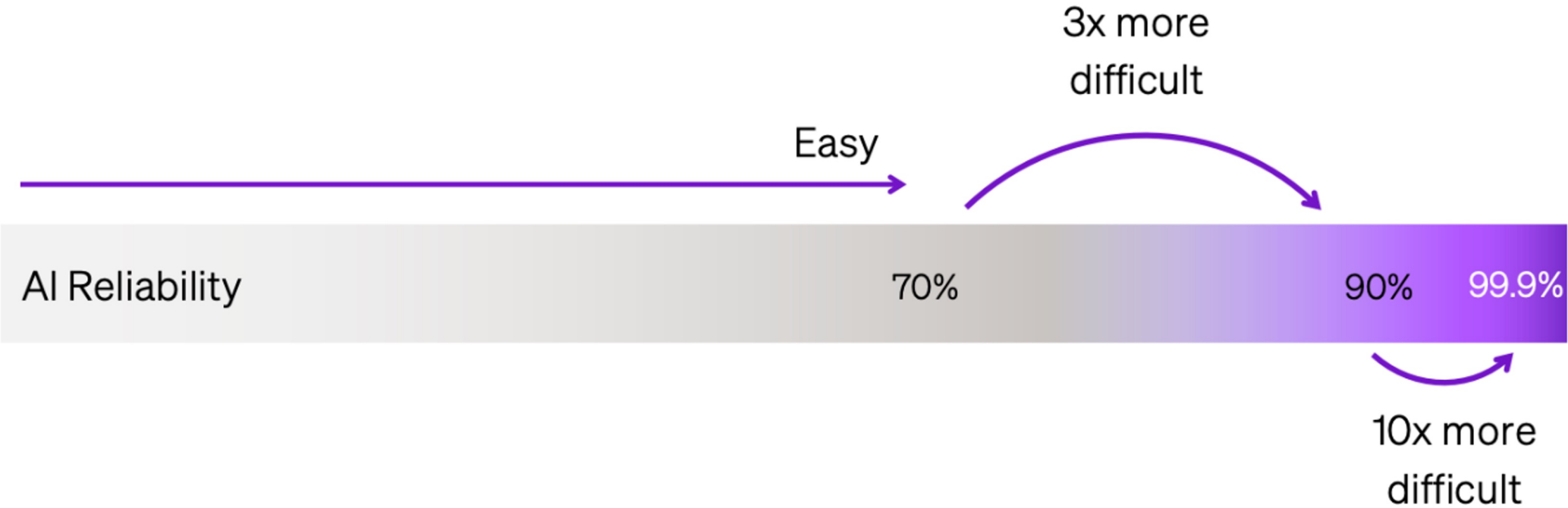
Record

Breakout Rooms

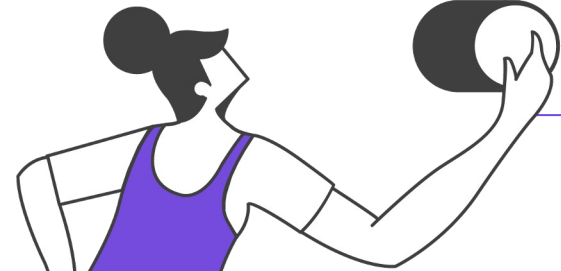
Reactions

Leave

While AI has made prototyping easier, bringing AI products to market requires extensive effort and engineering prowess to achieve production-grade reliability.



Source: Current State of Artificial Intelligence by Social Capital.



About me



Cursos online de
testing de software



Latino Entrepreneurship
Initiative



Introducción a las
pruebas de sistemas
de información

Federico Toledo



Quality
Sense

PODCAST



FEDERICO TOLEDO

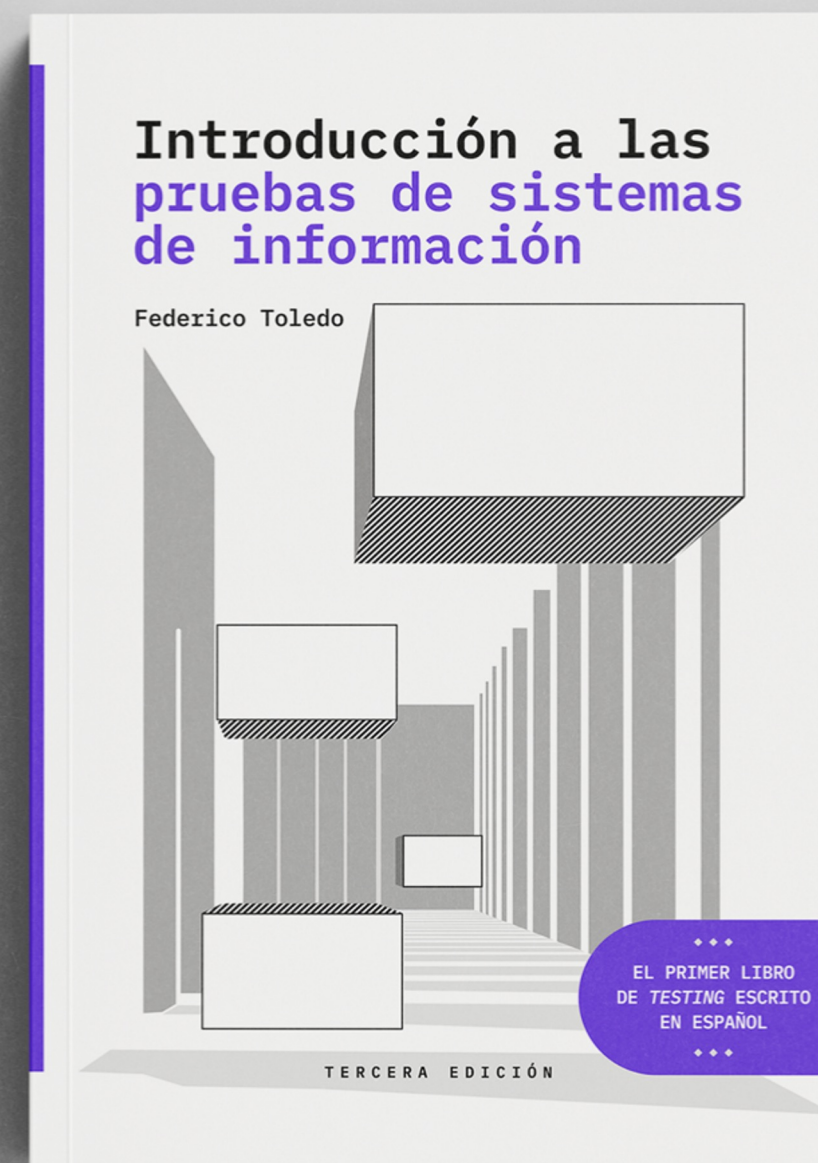


abstracta.us/blog
federico-toledo.com



Edición 10 años

Descarga gratuita



Software Quality Conferences



**WOPR
Latam**

2023 

2024 

2025 

2025 

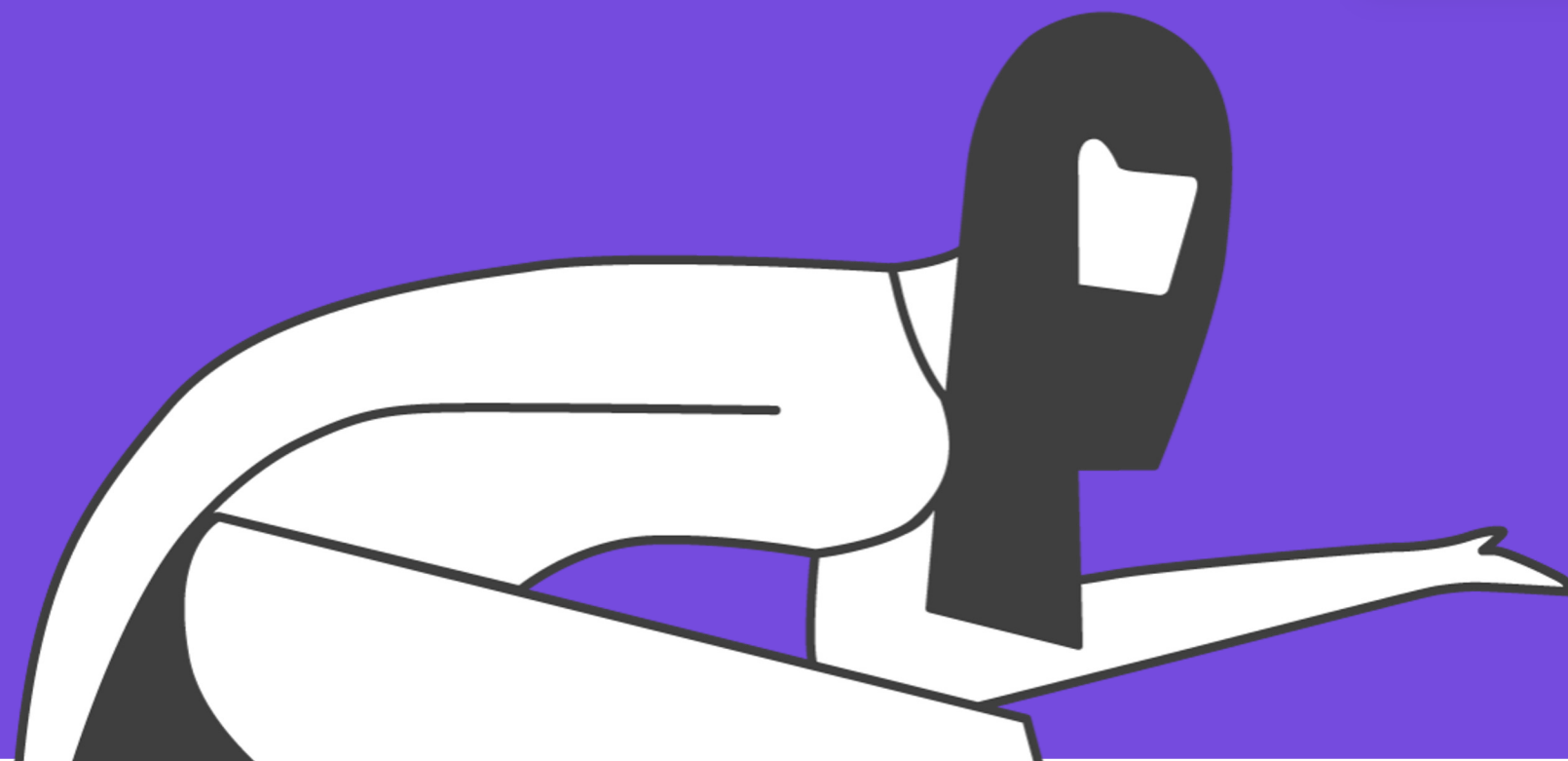
November

Agenda

01. GenAI in Testing →

02. Testing in GenAI →

03. Closing Thoughts →



01.
GenAI in
Testing



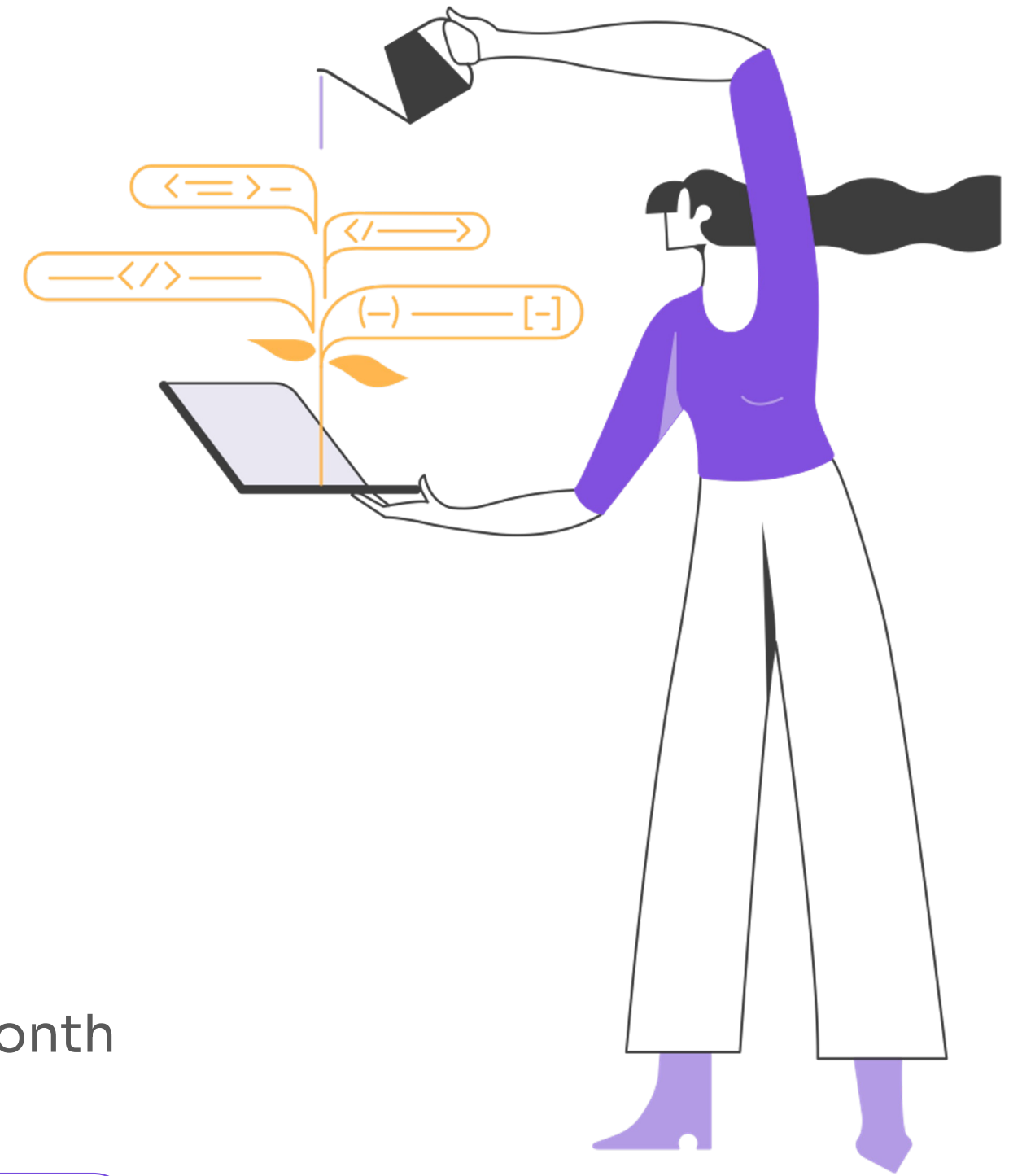
GenAI applied to development

Fast adoption of AI-assistants for development*:

- 63% devs using in 2024
- 14% considering to use soon

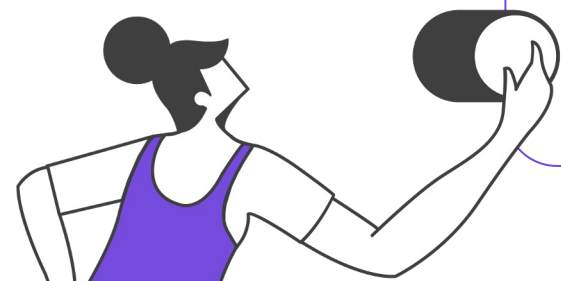
Are we more productive developing with these assistants really?

- Short term optimization**
 - More code is produced
 - 4X more duplicated code
 - “Copy/pasted” surpassed “moved” code (reuse) for the first time
- Long term consequences
 - Poorer maintainability
 - 25% increase of new code that has to be adjusted in less than a month



* Stack Overflow's 2024 Developer Survey

** AI Copilot Code Quality - Evaluating 2024's Increased Defect Rate via Code Quality Metrics



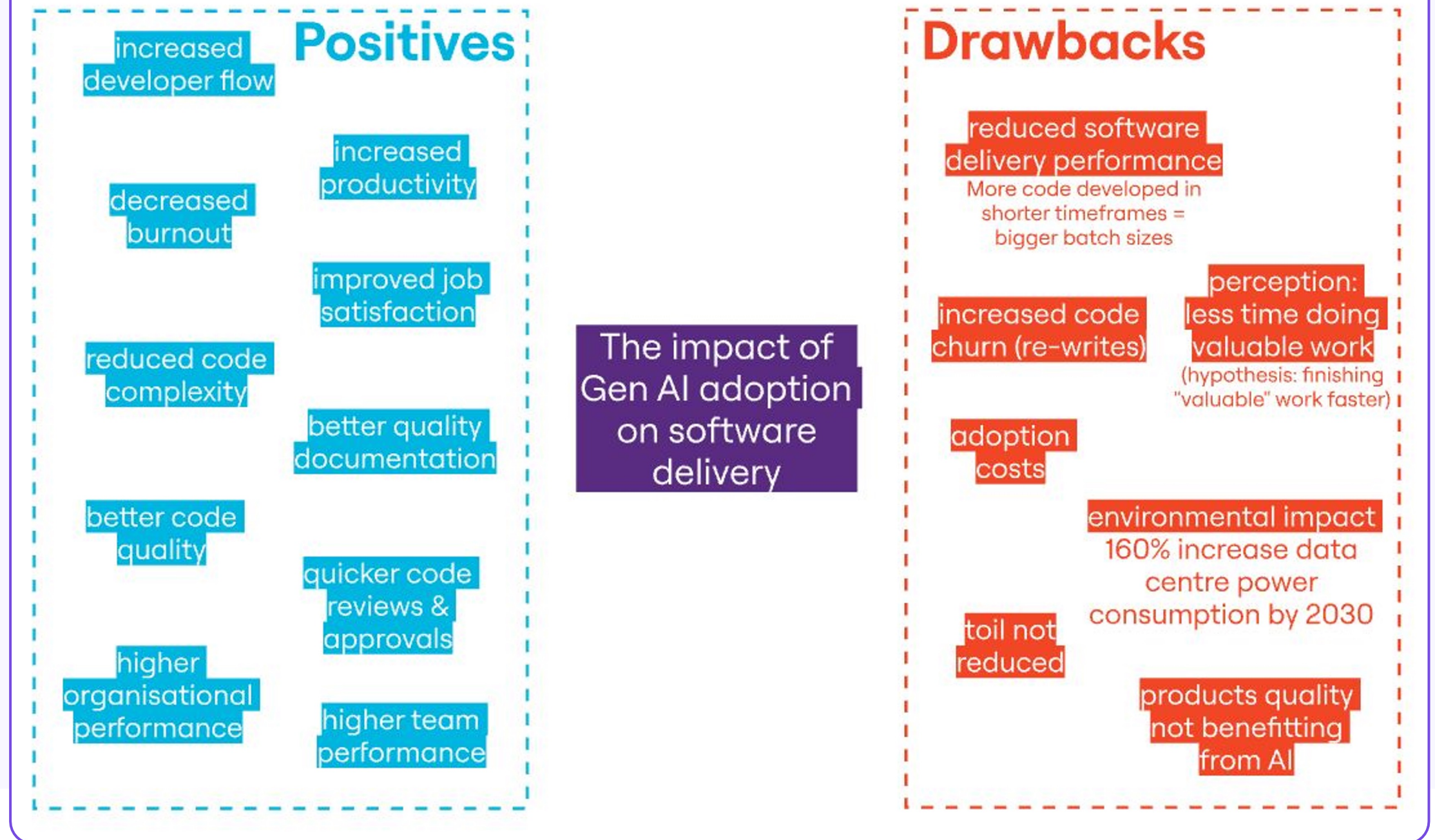
Reduced delivery performance due to the use of GenAI

More code in less time, increased average batch size

Which take us to

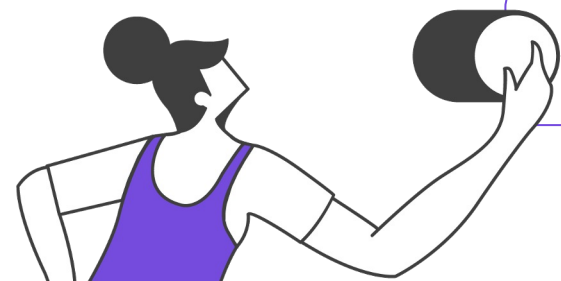
- More build & deploy complexity
- More failure probability
- Slower issue resolution

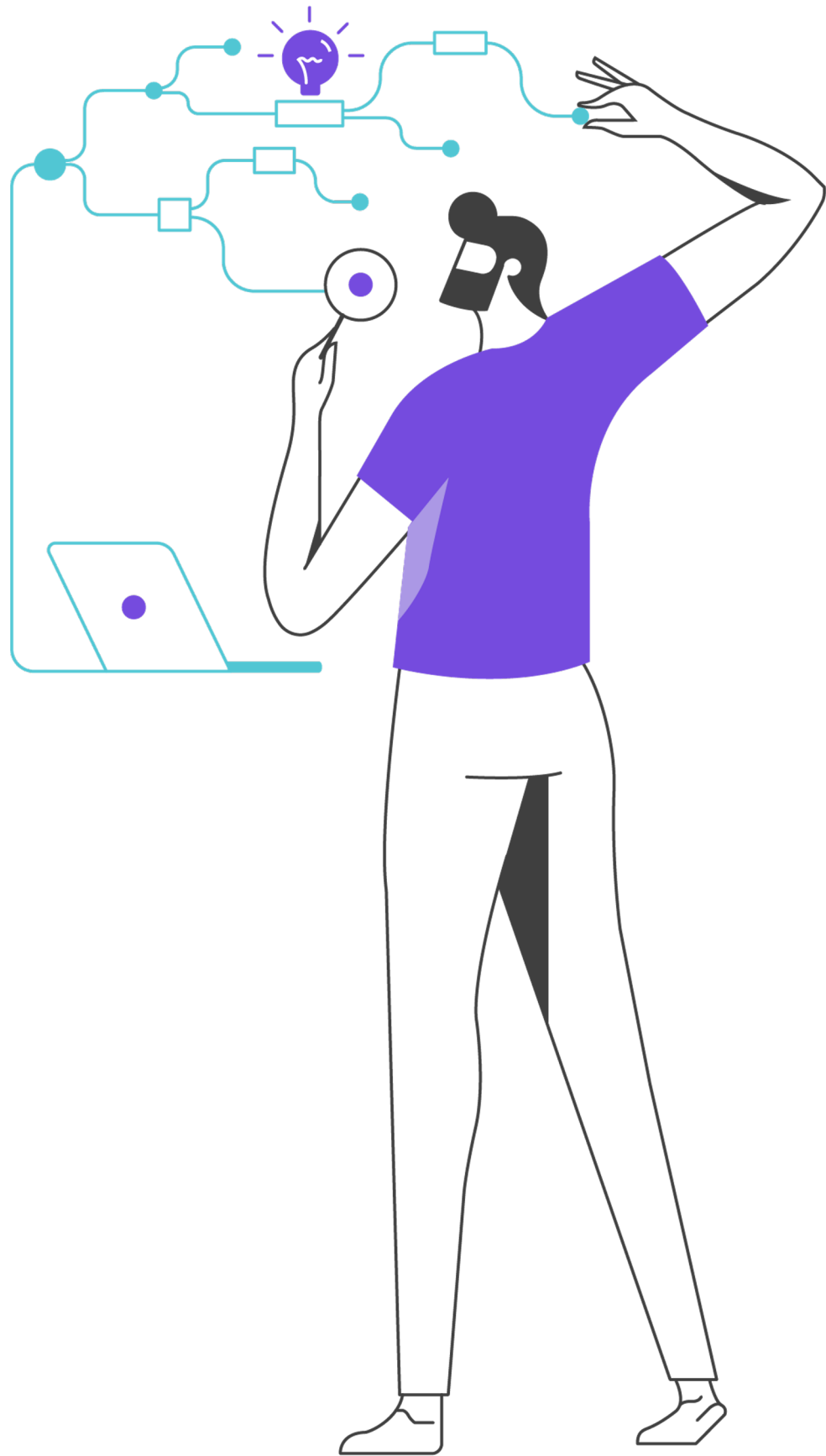
DORA State of DevOps 2024: AI



Stephen Townshend diagram from DORA state of DevOps Report 2024

*** DORA State of DevOps Report 2024





What about testing?

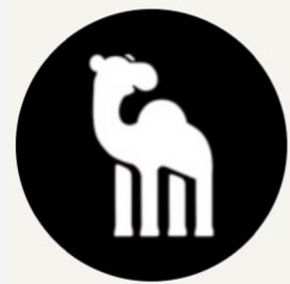
Poorer quality code means more back and forth between different roles (test, dev, BA, infra)

Faster code production means more accumulation in testing (perceived as the bottleneck)

We know where this will take us... we've been there

COVERAGE

Get to (and stay at) complete test coverage in just 90 days



camelQA

AI that replaces human QAs for mobile app testing

W24

ACTIVE

ARTIFICIAL-INTELLIGENCE

SAAS

AUSTIN

KISS BUGS GOODBYE

QA Wolf's AI-native approach delivers 80% automated end-to-end test coverage for web & mobile apps in **weeks, not years**

Initiatives in testing

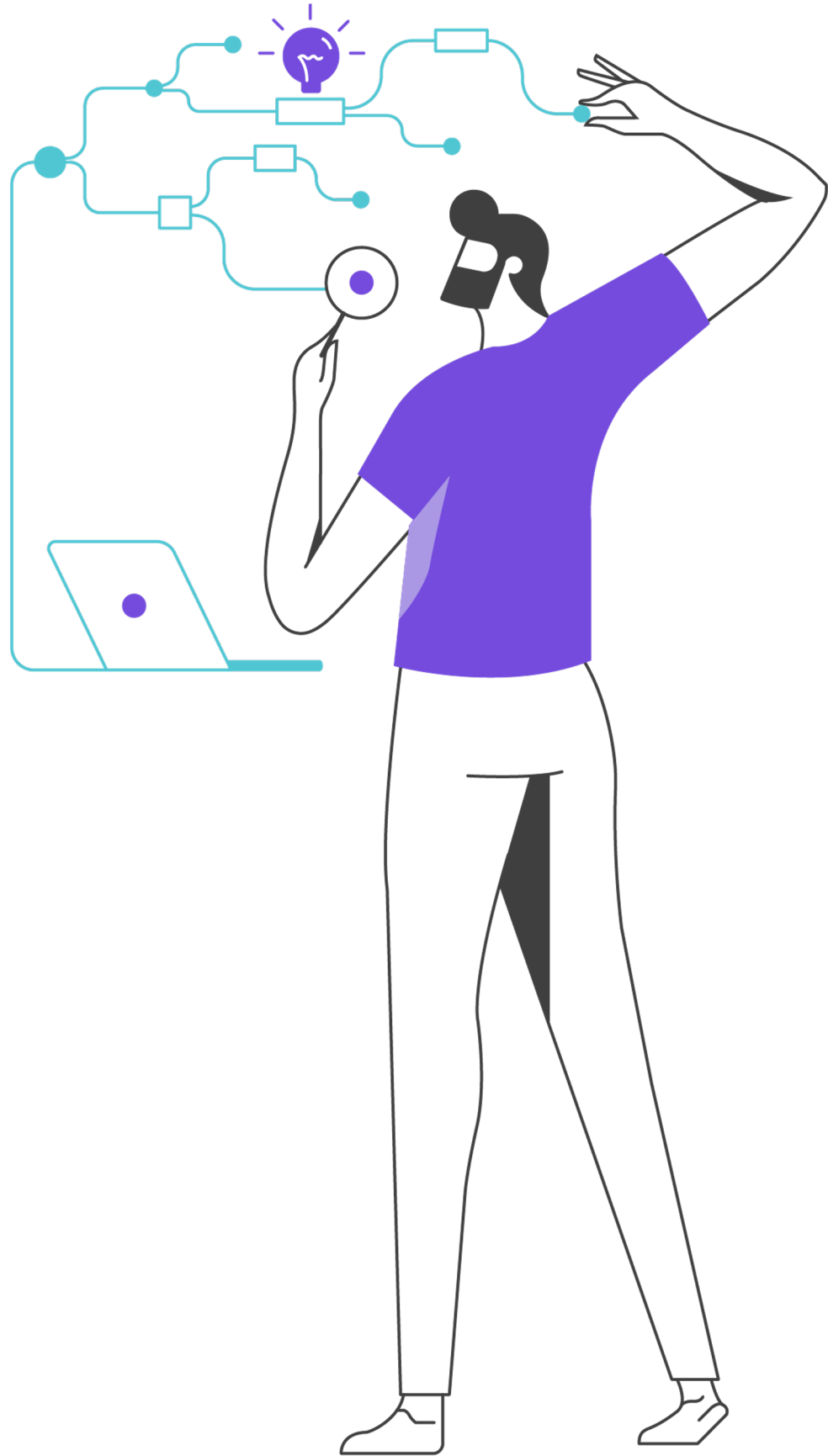
Sometimes the wrong people are working in “solve the testing problem”

- People who hate testing
- People that do not understand the value of testing

They are trying to eliminate humans in testing as that were a problem instead of the actual value

Cover more edge cases than ever in just a few minutes with aqua

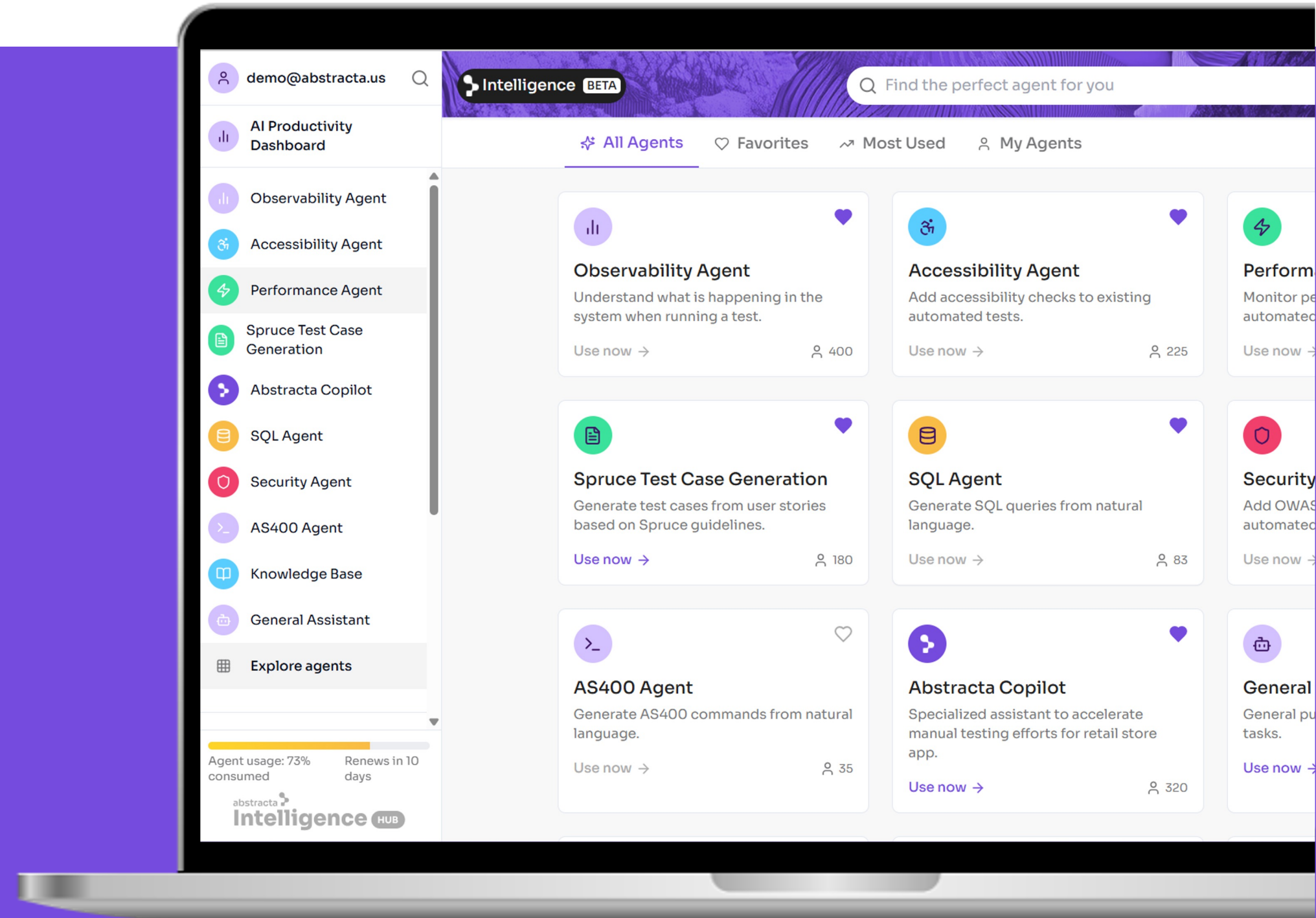
[TRY AQUA FOR FREE](#)



Enreach and empower the human testers instead of remove them

- Low code and self healing to make test automation more robust and accessible to more people
- Copilots and assistants as tools for testers to help them achieve more with their expertise

Learn by Doing: Go Implement Your Thing

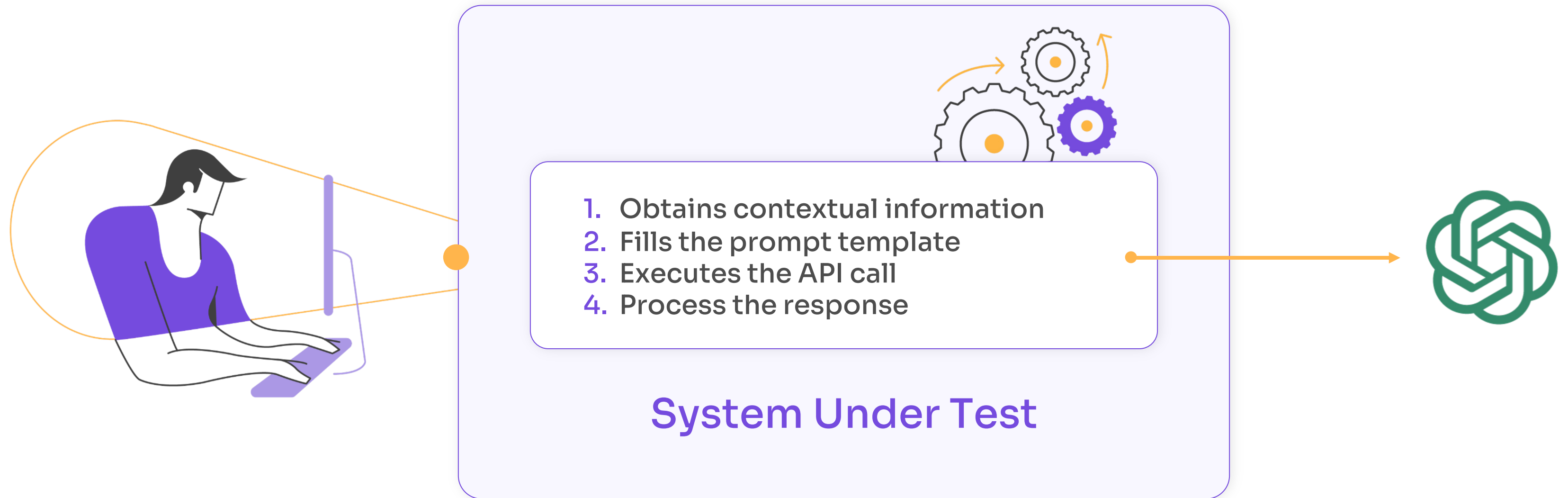


02.

Testing in GenAI



The Simplest: An Application Using an LLM's API



Risks: User Expectations

User Expectations Changed

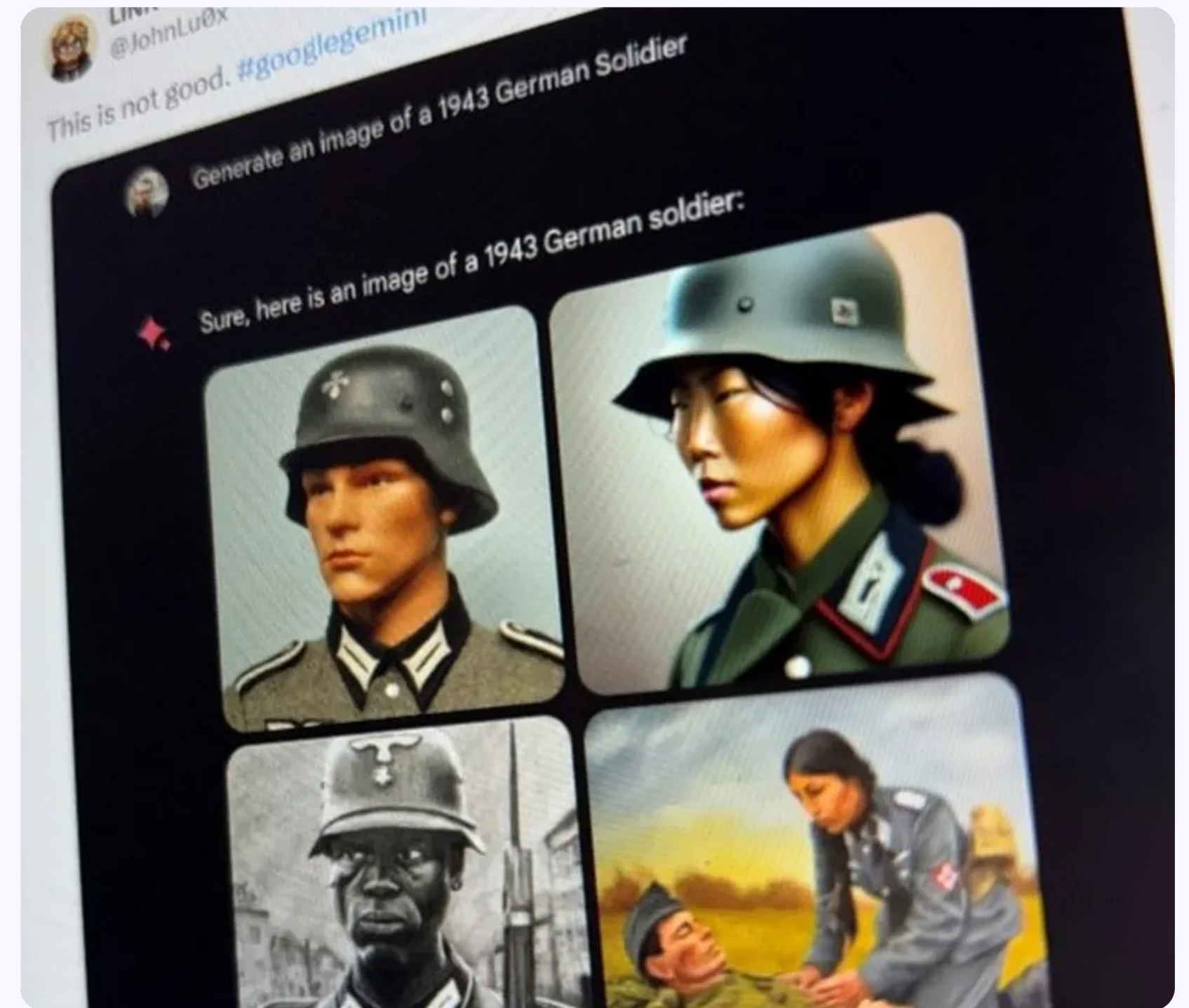
- Conversational interface (Alexa, Siri, Google Assistant seem less intelligent after ChatGPT)
- Performance / response time: Immediate response, chunk by chunk, fluid reading of the response
- The user knows they cannot trust on the output of the system 🤖



Risks: Training and Adjustments

Do you really know your model?

- Models are trained with available data on the Internet.
 - Biased
 - Incorrect data
 - Hate, discrimination, etc
- Models are **adjusted** to avoid harm or undesired use. These adjustments might produce undesirable outcomes too.
- **The question is: Is there any bias or adjustment in the model that might affect some users?**



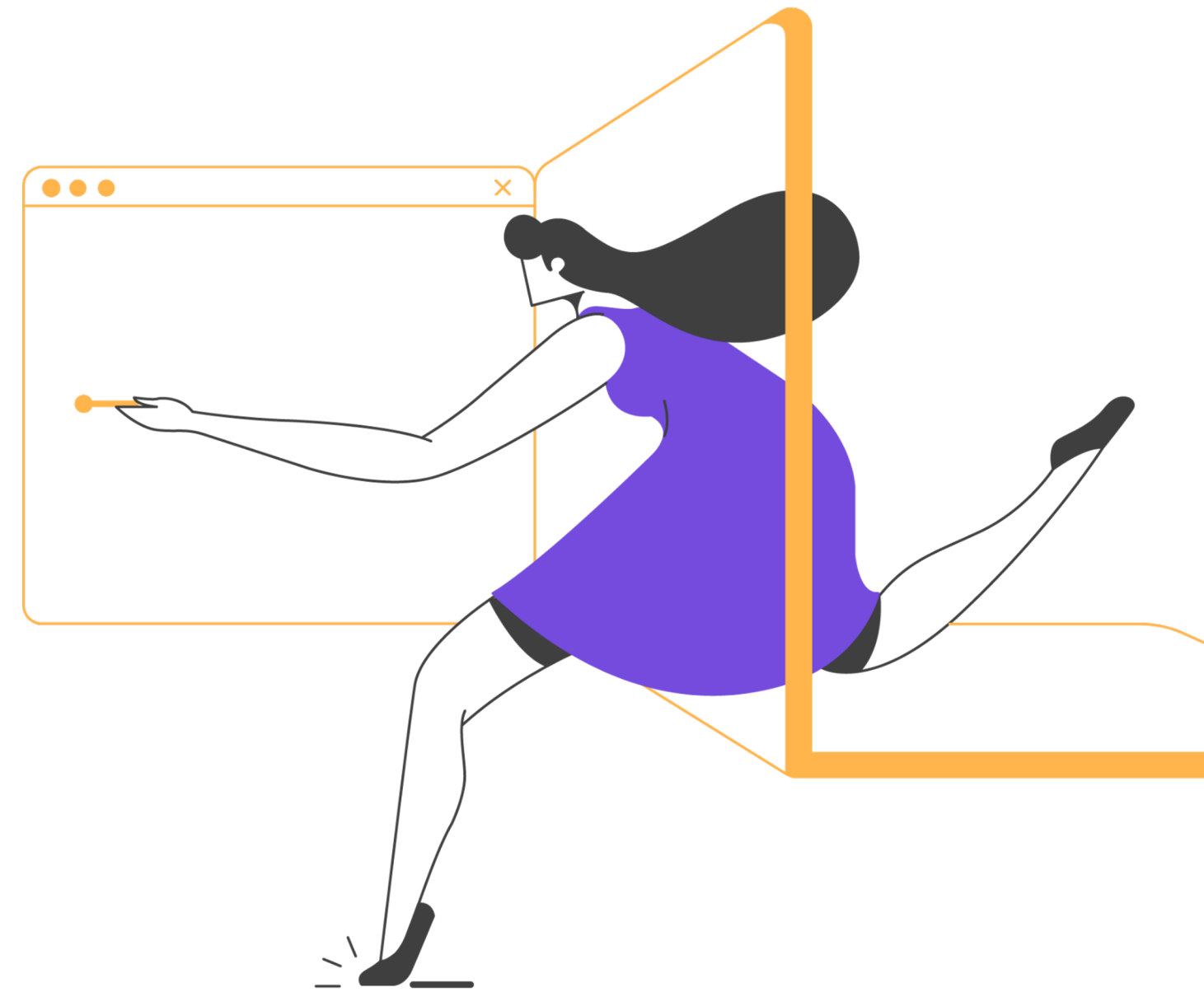
Legal Risks

We need to be aware of the policies and guidelines of the platforms we use

- [Generative AI Prohibited Use Policy from Google](#)
- [Google AI Principles](#)
- [Microsoft Responsible AI Standard](#)

We need to be aware of the laws and regulations

- GDPR
- EU AI act
 - **Identify risk level**
 - Guarantee training data quality
 - Provide control and transparency to the user
 - Let the user know if something was generated with AI
 - Documentation and technical aspects



LLM “Syndromes”

Large Language Model Syndromes, unwanted behaviors/outputs systematically observed on LLMs (by James Bach and Michael Bolton):

Hallucinations

Incuriosity

Placation

Indiscretion

Misalignment

Offensiveness

Arrogance

Incorrectness

Capriciousness

Forgetfulness

Redundancy

Incongruence

Negligence/Laziness

Opacity

Unteachability

Non-responsiveness

Vacuousness

Source: <https://developsense.com/large-language-model-syndromes>



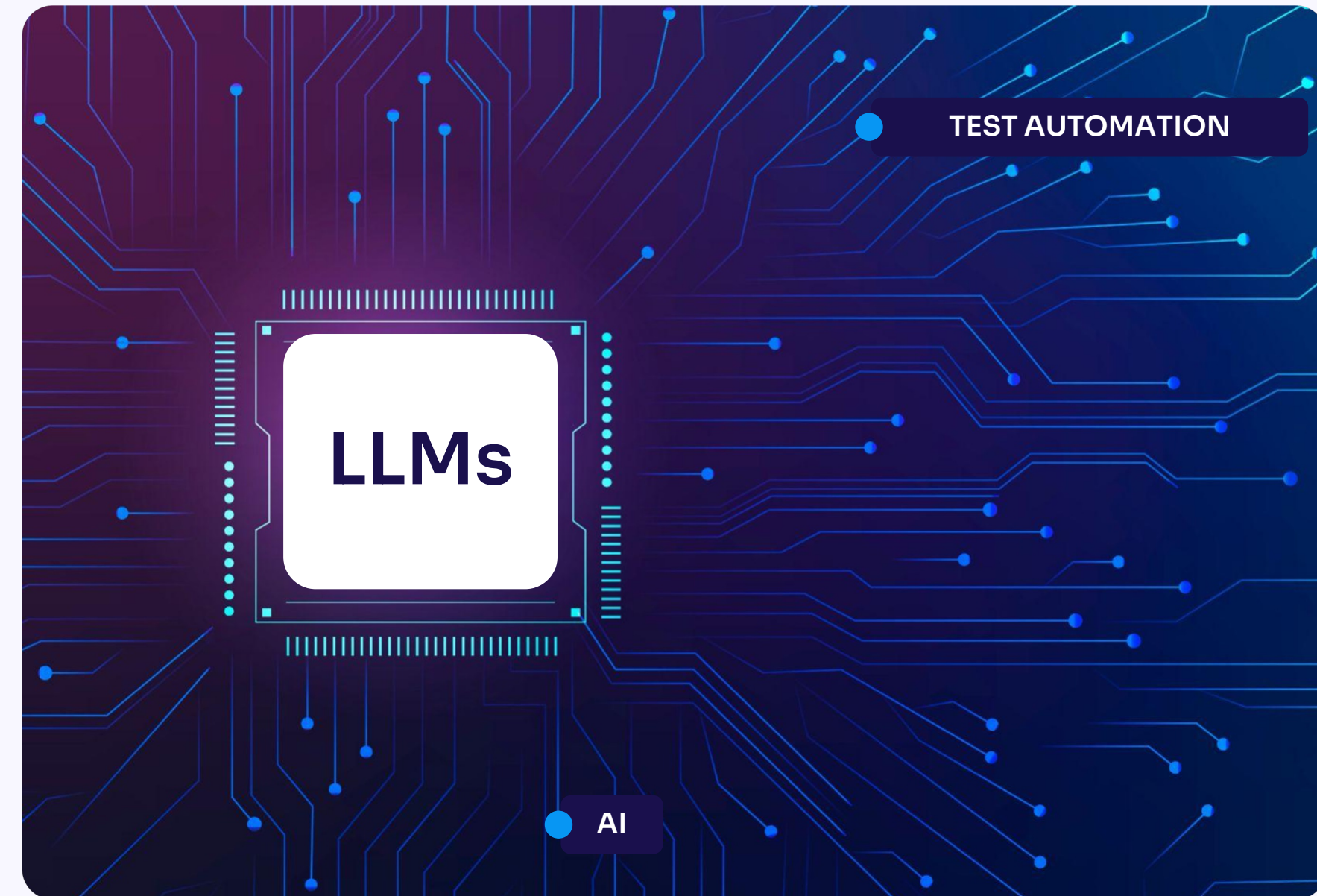
Need for Test Automation

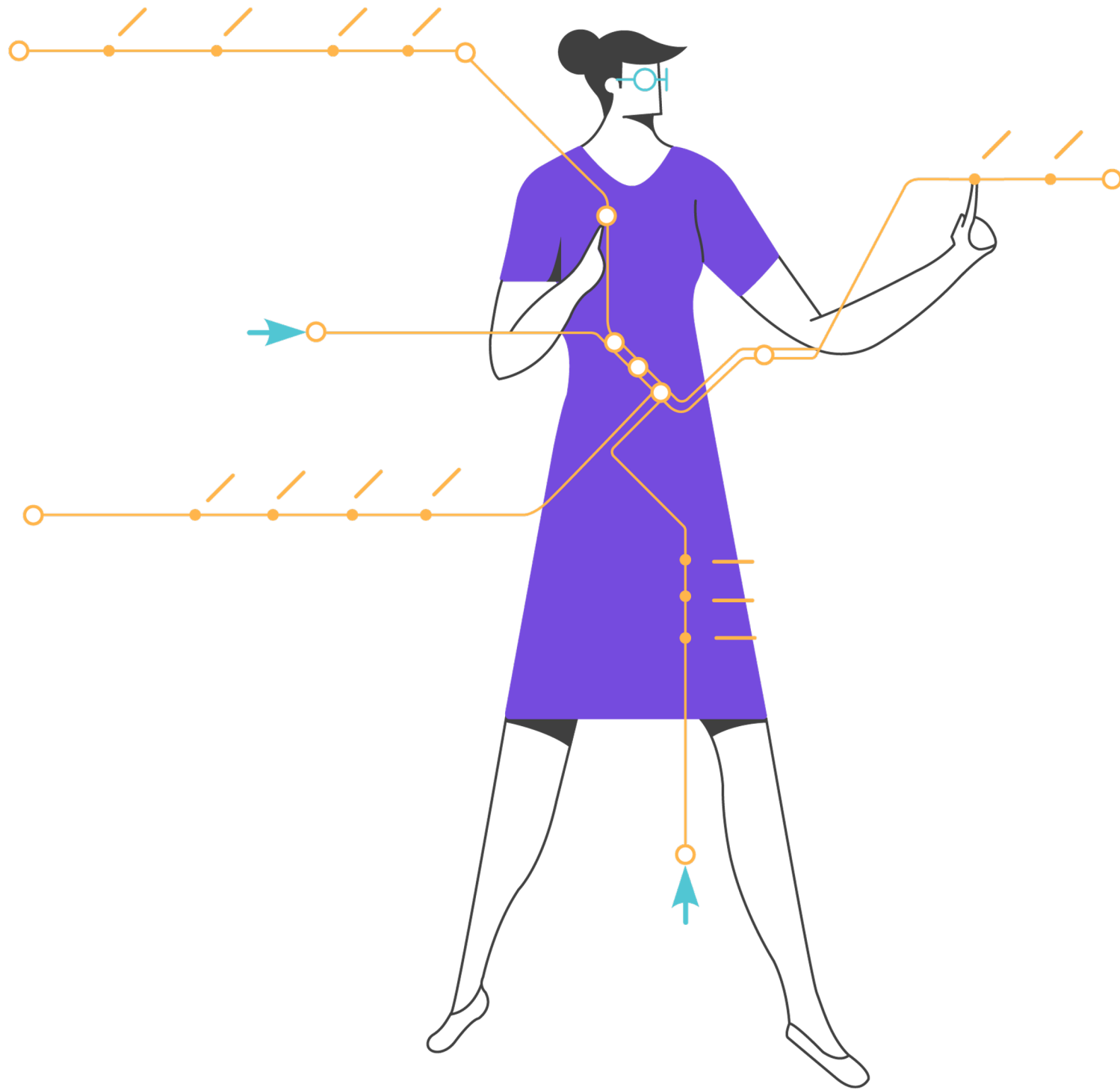
Test automation is needed because we will want to adopt new versions of the LLMs we use

- Because they will be faster, smarter, cheaper, etc
- But we need to make sure that our systems still work with the new behavior of the new model

LLMs are non-deterministic, test automation is challenging

- Same model, same prompt, different responses
- Even more when we change the model version or the LLM provider



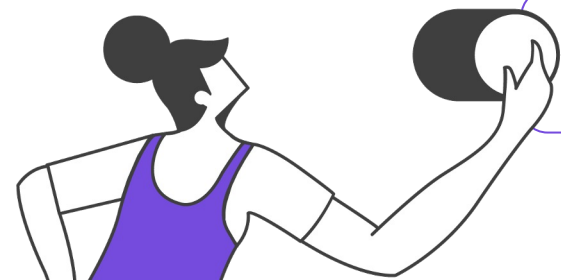


Test Automation for Regression Testing with Evals

- Evals
 - Open source framework (by OpenAI) for automated tests verifying the responses of the LLM
 - It provides templates for assertions, basic and advanced
- There are other alternatives (Evaluation, Promptfoo, etc)

Recommended reading:

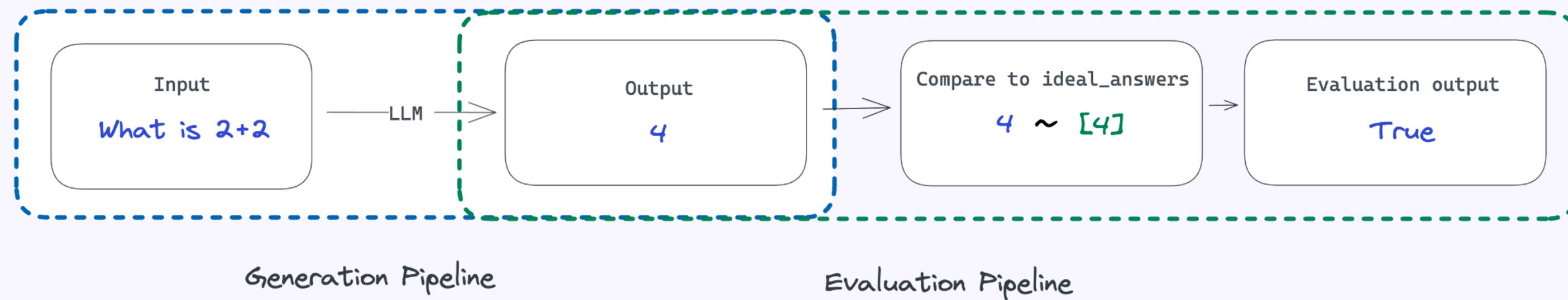
<https://portkey.ai/blog/decoding-openai-evals/>



Test Automation with Evals

Basic Eval Templates

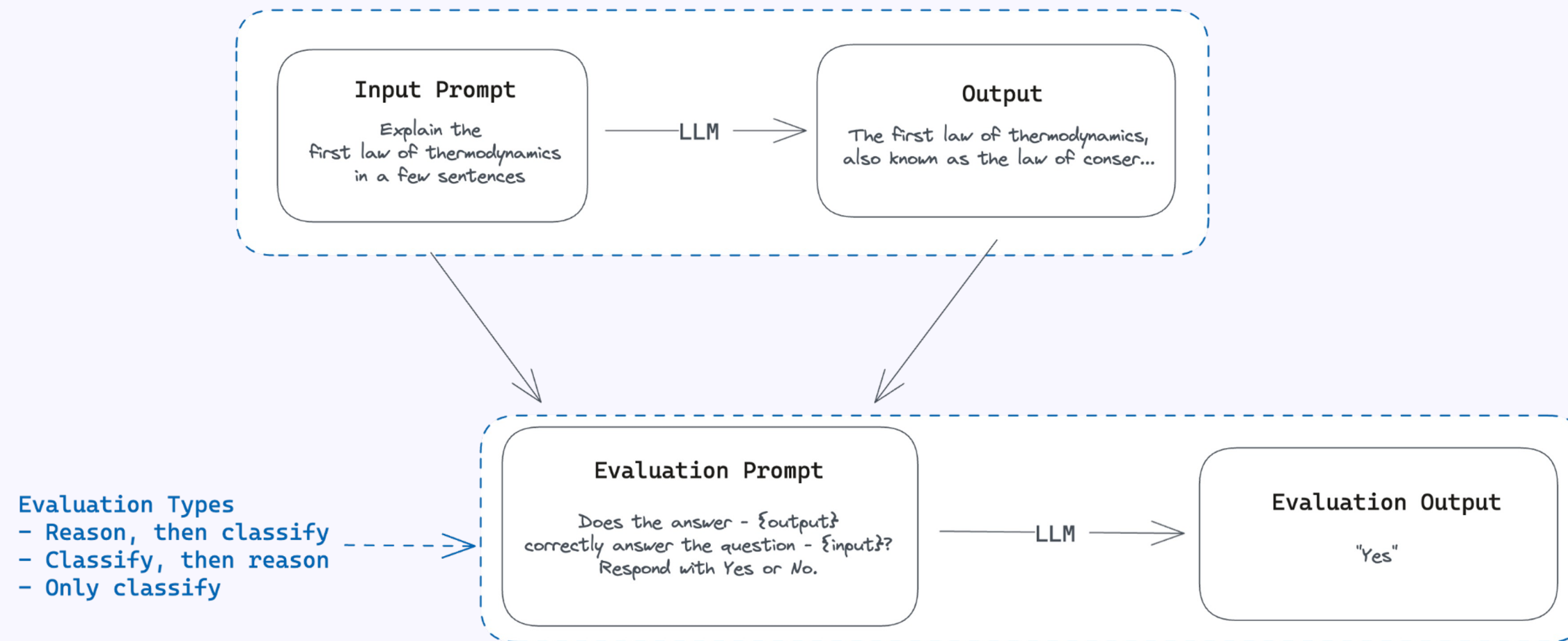
- *Input_prompt*
- *Ideal_answers*
- The *output* is compared (with a simple method) to the *ideal_answers*



Test Automation with Evals

Model-Graded Eval Templates

- *Input_prompt*
- *Ideal_answers*
- The *output* is compared to the *ideal_answers* **using another LLM**



Non-functional Aspects

Security

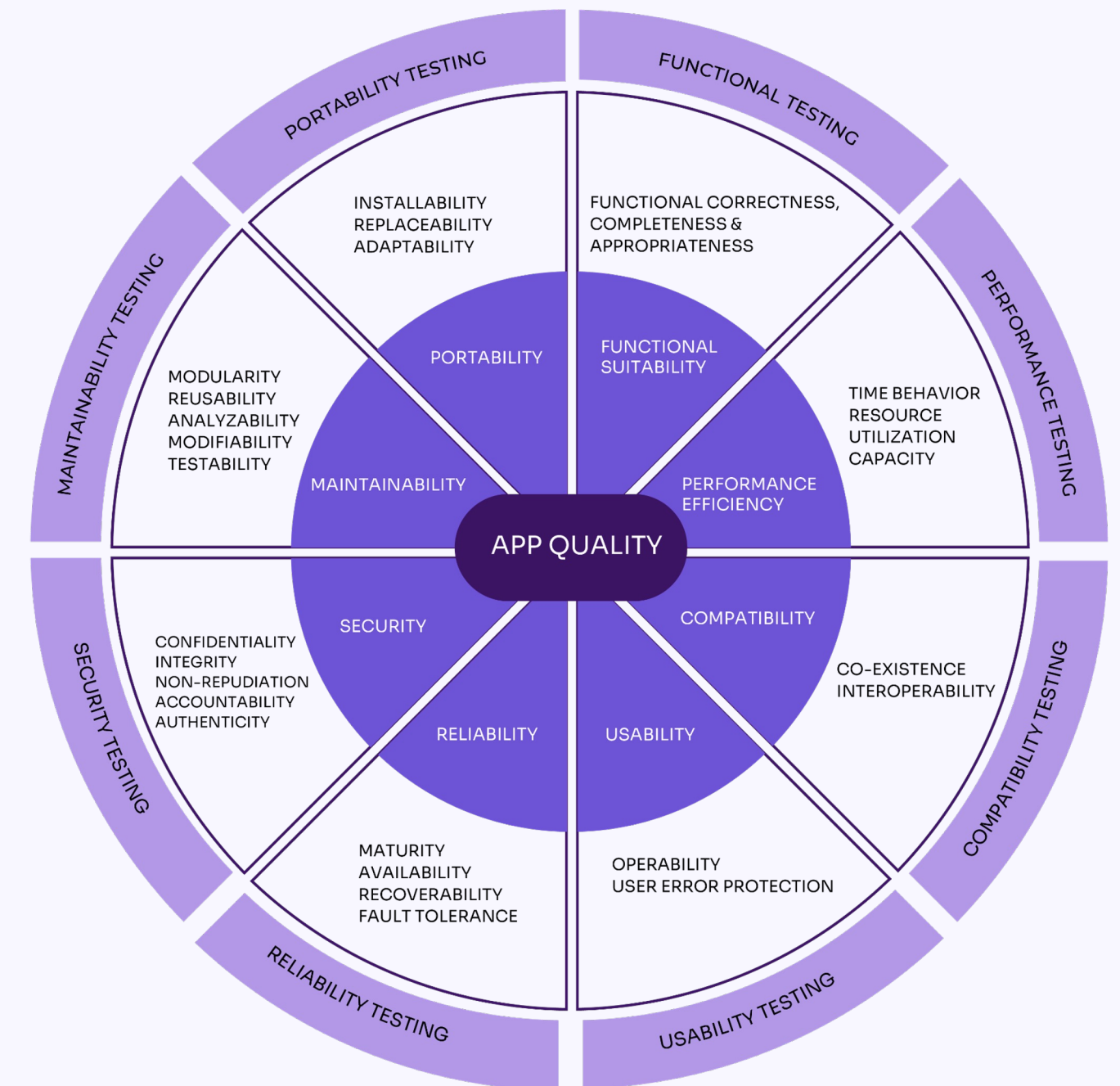
- OWASP Top 10 Vulnerabilities for LLM-based Applications

Usability and accessibility

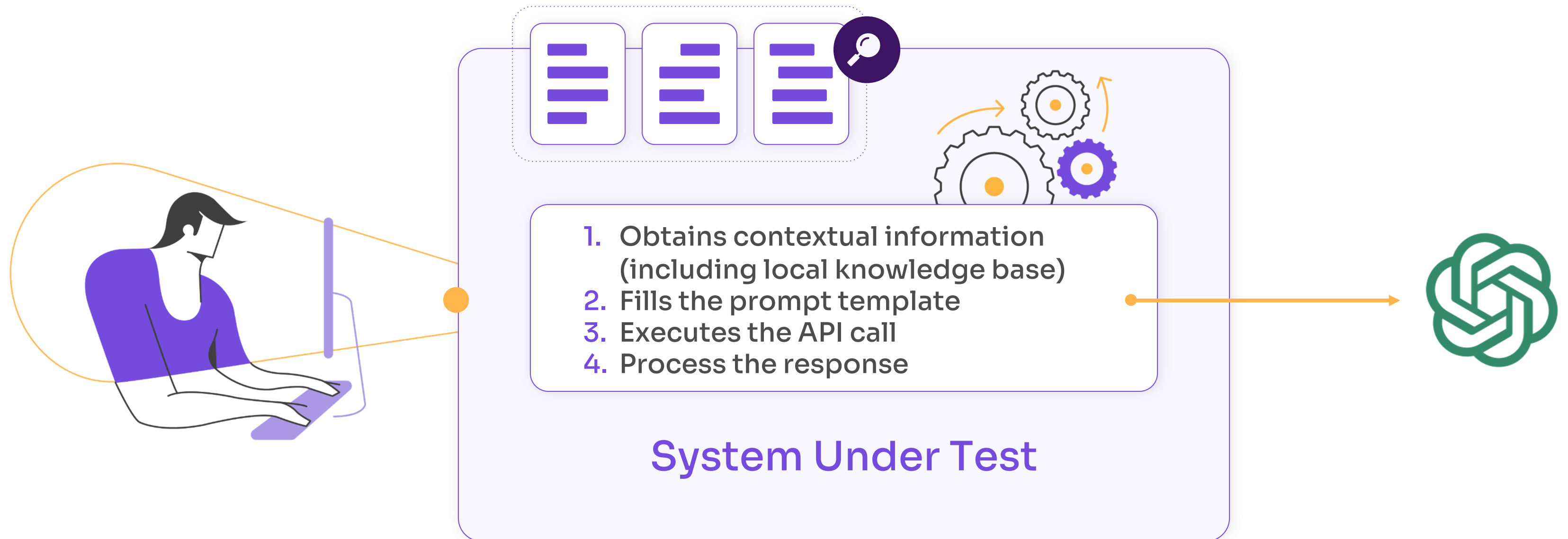
- Conversational interfaces
 - Do we need support from philologist?
- User feedback is key

Performance

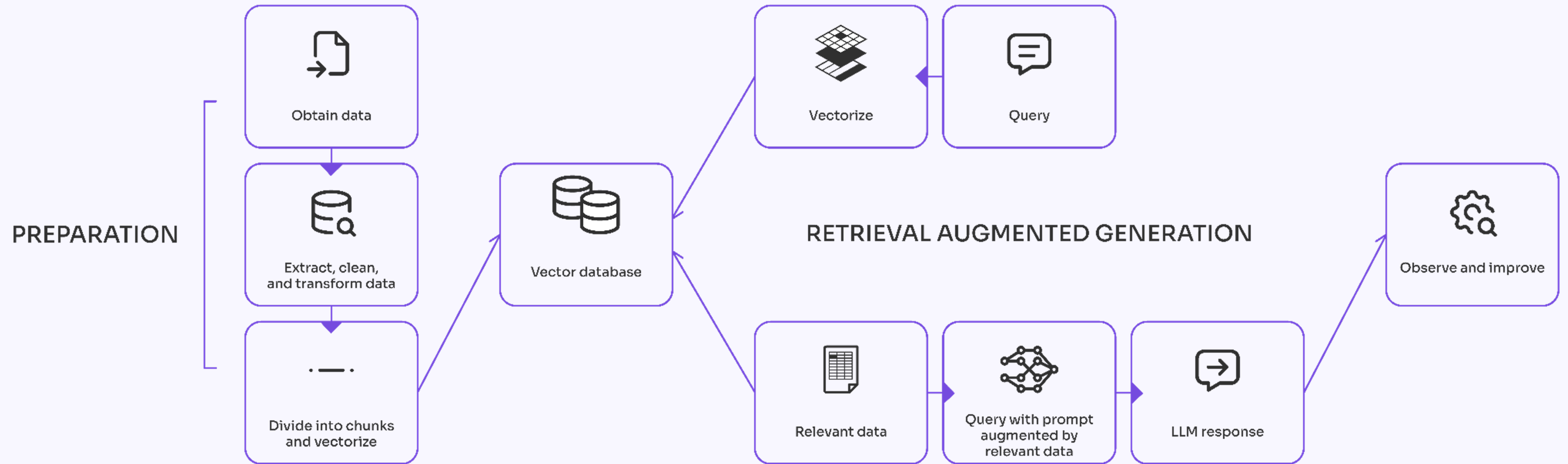
- Observability / monitoring



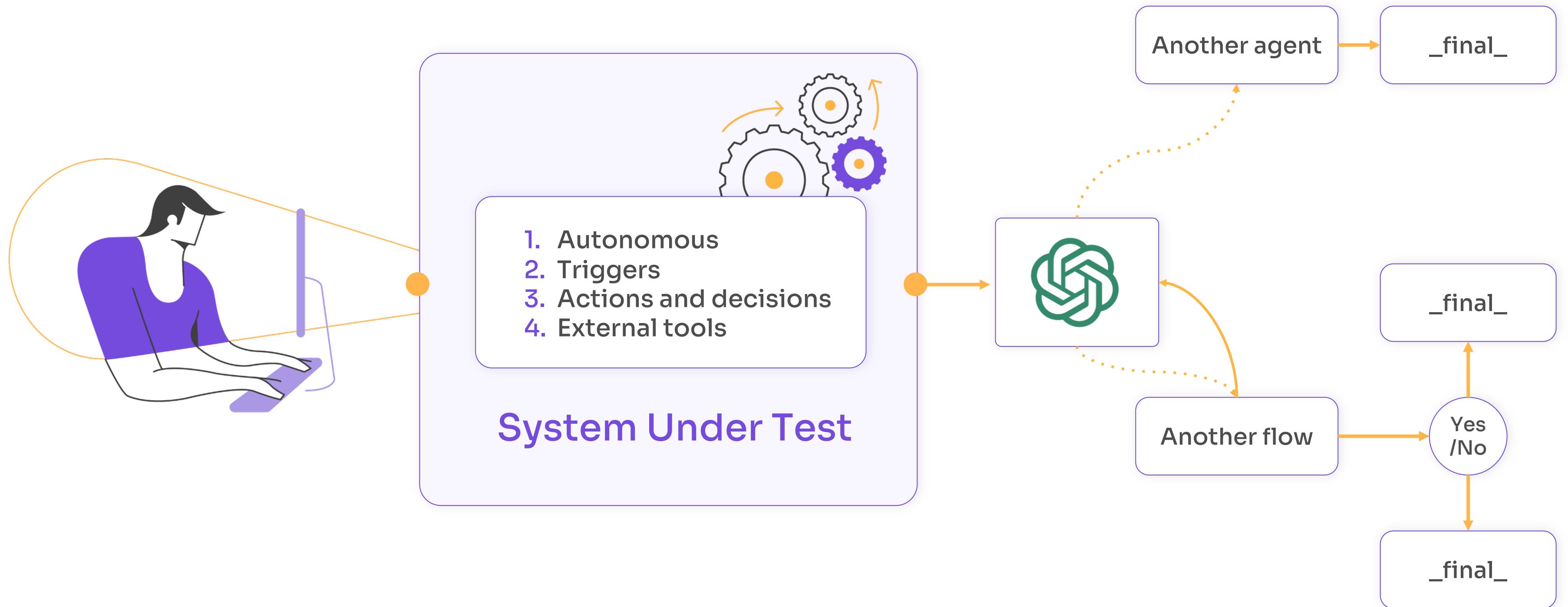
Adding Complexity: LLM's API + RAG



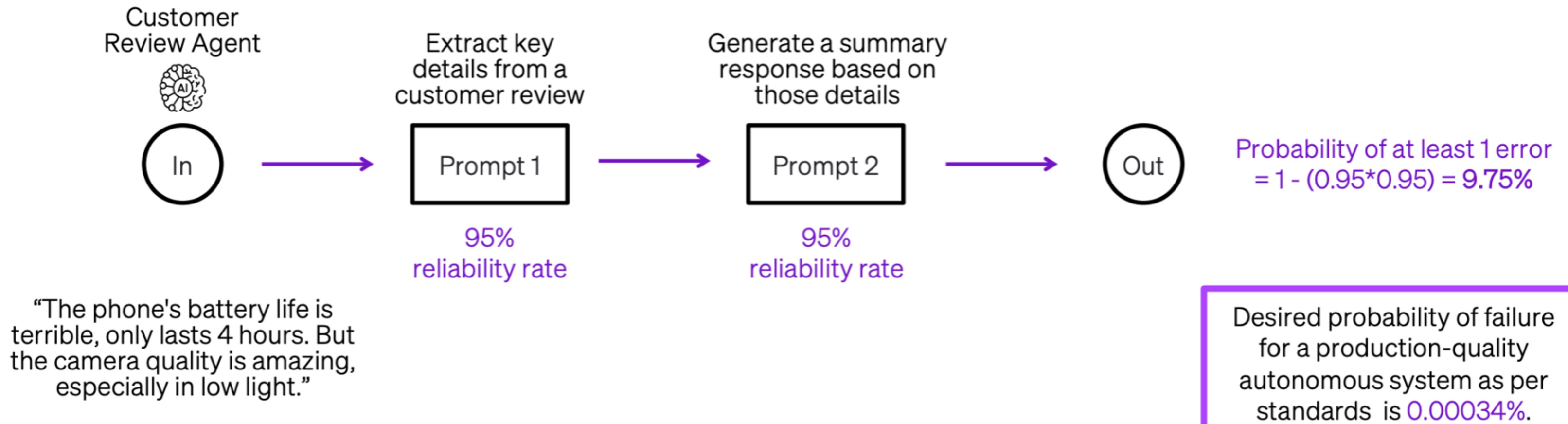
Data Quality Verification in the Data Pipeline for RAG



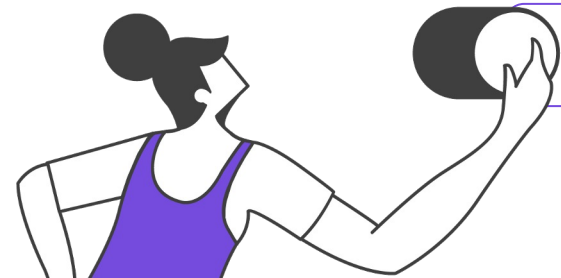
Even more Complexity: Autonomous Agents



Example Agent Workflow – Prompt Chaining



Source: Current State of Artificial Intelligence by Social Capital.

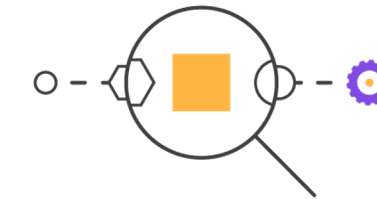


Observability



Problem

- **Loss of visibility and control** when interacting with LLMs
 - **Content (request, response)**
 - **Precision**
 - **Cost \$\$**
 - **Performance (response time)**
- How to gain visibility?



Observability

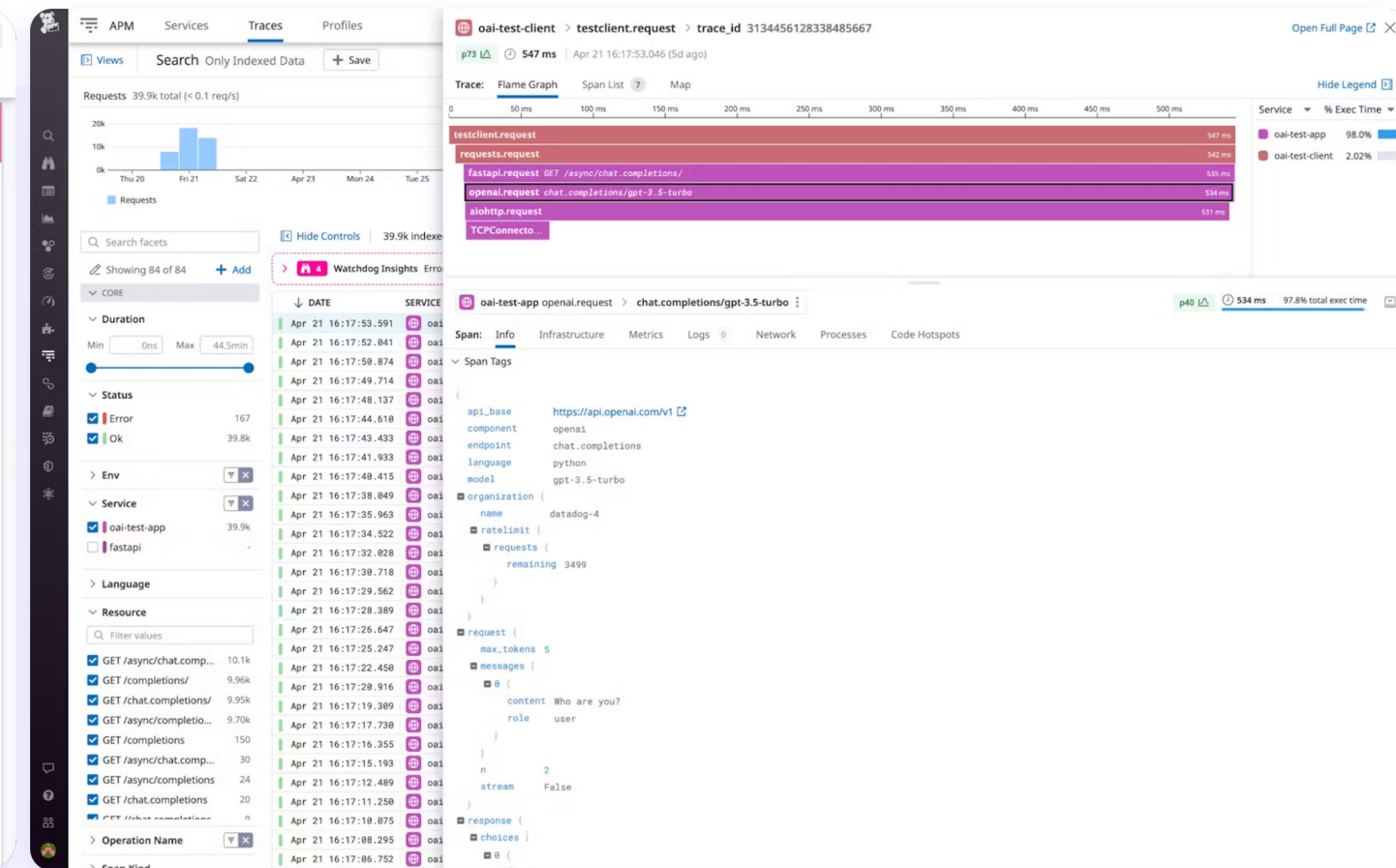
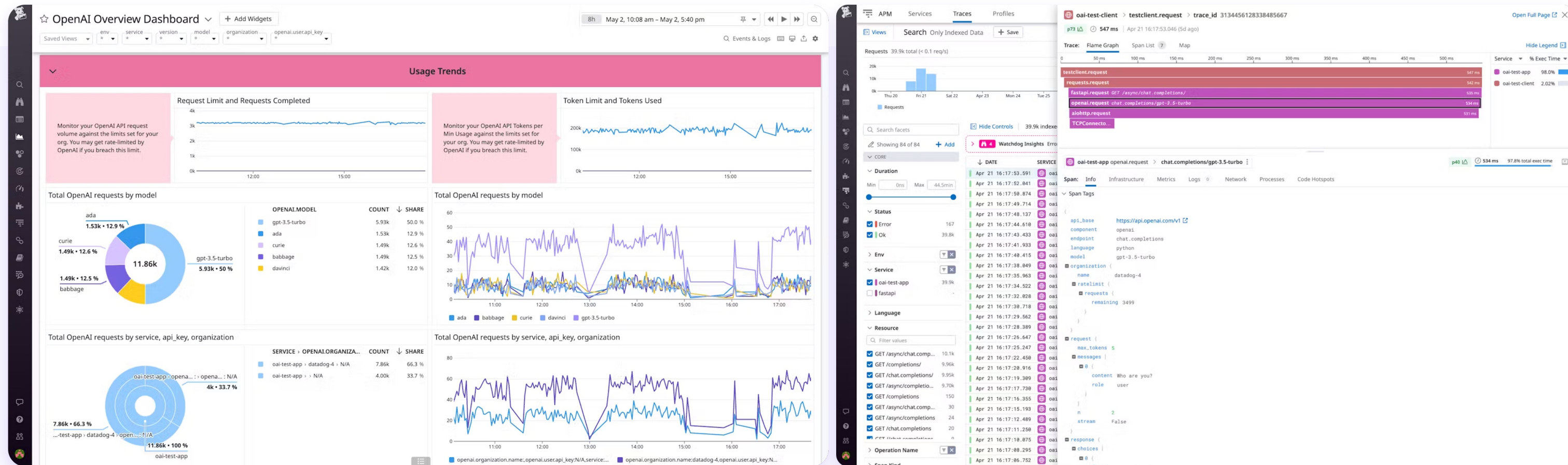
- Metrics
 - Response times
 - # tokens (cost)
 - Model drift (prediction effectiveness)
 - User feedback
- Traces
 - Prompts and responses
- Logs in every layer

Recommended reading:

- [Monitor OpenAI with Datadog](#)
- [Phoenix \(open source\)](#)



Observability example: Datadog



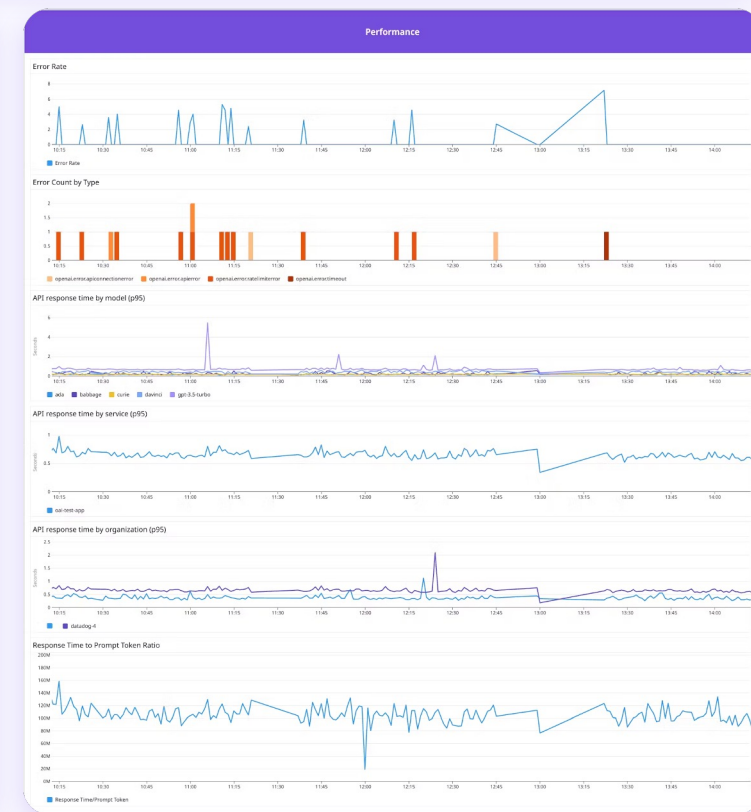
OpenAI Overview Dashboard (8h May 2, 10:08 am - May 2, 5:40 pm)

Prompt and completion samples

Prompts and completions are sampled from the Datadog APM library integrations. By default prompts and completions are sampled 100% on all spans. Logs are sampled at a rate of 10% by default when spans are sampled. See the [OpenAI integration docs](#) for how to configure this rate.

DATE	SERVICE	OPENAI.MODEL	CHOICES.0.FINISH_REASON	PROMPT	CHOICES.0.TEXT
May 02 17:26:37.425	oai-test-app	babbage	length	hello world	box" (loop counter
May 02 17:24:22.383	oai-test-app	curie	length	what is the square root of 2 ?"	
May 02 17:23:36.833	oai-test-app	curie	length	what is your name ? You call it a	
May 02 17:23:03.291	oai-test-app	ada	length	what is your name ?") call_user	
May 02 17:21:52.186	oai-test-app	ada	length	what is the square root of 2 +1, if you	
May 02 17:21:42.592	oai-test-app	babbage	length	hello world	-
May 02 17:21:02.661	oai-test-app	babbage	length	hello world	:: readfile (Standard
May 02 17:18:33.523	oai-test-app	ada	length	what is the square root of 2 ??" (
May 02 17:18:03.064	oai-test-app	ada	length	what is your name	?"
May 02 17:17:52.143	oai-test-app	curie	length	what is your name	, I am Perrin
May 02 17:16:39.938	oai-test-app	ada	length	what is your name	_"?

DATE	OPENAI.MODEL	SERVICE	COMPLETION.0.FINISH_REASON	MESSAGES.0.CONTENT	COMPLETION.0.MESSAGE.CONTENT
May 02 17:28:35.514	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:28:34.077	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:28:07.679	gpt-3.5-turbo	oai-test-app	length	Hello!	Hello there! How can
May 02 17:28:03.185	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:27:04.858	gpt-3.5-turbo	oai-test-app	length	Goodbye!	As an AI language model
May 02 17:27:01.882	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:25:39.852	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:25:30.928	gpt-3.5-turbo	oai-test-app	length	Who are you?	I am an AI language
May 02 17:24:10.769	gpt-3.5-turbo	oai-test-app	length	Goodbye!	Goodbye, have a



03.

Closing Thoughts



F i want to build a web app that can generate test data applying pair wise combinatorial methods.

I'll help you create a web app for generating test data using pairwise combinatorial testing methods. This is a powerful technique for reducing the number of test cases while maintaining good coverage. We'll create a user-friendly interface where users can:

1. Define parameters and their possible values
2. Generate optimized test cases using the IPOG (In-Parameter-Order-General) algorithm
3. View and export the results

Let's start by implementing this solution:

Pairwise Test Data Generator

- ✓ Create package.json
- ✓ Create src/types.ts
- ✓ Create src/utils/pairwise.ts

How can Bolt help you today?

Code Preview

5173 /

Pairwise Test Generator

Browser

Chrome, Firefox, Safari



OS




Windows, MacOS, Linux



+ Add Parameter

Generate Test Cases

Pairwise Test Generator

Browser	Chrome, Firefox, Safari	
OS	Windows, MacOS, Linux	
variable	v1,v2	

+ Add Parameter

Generate Test Cases

Test Case Statistics

Pairwise Tests

14

Each Choice

3

Reduction: -366.7%

Cartesian Product

18

Reduction: 22.2%

Generated Test Cases

Export CSV

#	BROWSER	OS	VARIABLE
1	Chrome	Windows	v1
2	Chrome	...	v1

12	Safari	↓ s	v2
13	Chrome	MacOS	v2
14	Chrome	Linux	v2

Pair Coverage Matrix

PARAMETER PAIR	COVERED BY TEST CASE(S)
Browser=Chrome x OS=Windows	1, 10
Browser=Chrome x OS=MacOS	2, 13
Browser=Chrome x OS=Linux	3, 14
Browser=Firefox x OS=Windows	4, 11
Browser=Firefox x OS=MacOS	5
Browser=Firefox x OS=Linux	6
Browser=Safari x OS=Windows	7, 12
Browser=Safari x OS=MacOS	8



Fabian Baptista • 1st

CTO @ Apptim | CIO @ Abstracta | Artificial Intelligence

2w ...

📢 Unpopular opinion: Writing code, debugging and reviewing code is now a commodity. Testers and business analyst are now the key players in this new co-intelligence era, where the human-in-the loop is critical.

New Challenges in Software Quality Brought by GenIA and LLMs

Thanks!



Federico Toledo

federico@abstracta.us



expo IQA 25

MADRID
May 20th,
21st & 22nd
2025

Thank you for attending

expoqqa.eu