

# CAN WE TRUST AI-GENERATED CODE?

MAYBE WE'VE BEEN  
ASKING THE WONG  
QUESTION



TESTS



SPECS



PROMPTS

@JBARUCH

#INTENTINTEGRITYCHAIN

Trust Level = 0.88



TESTS

SPECS

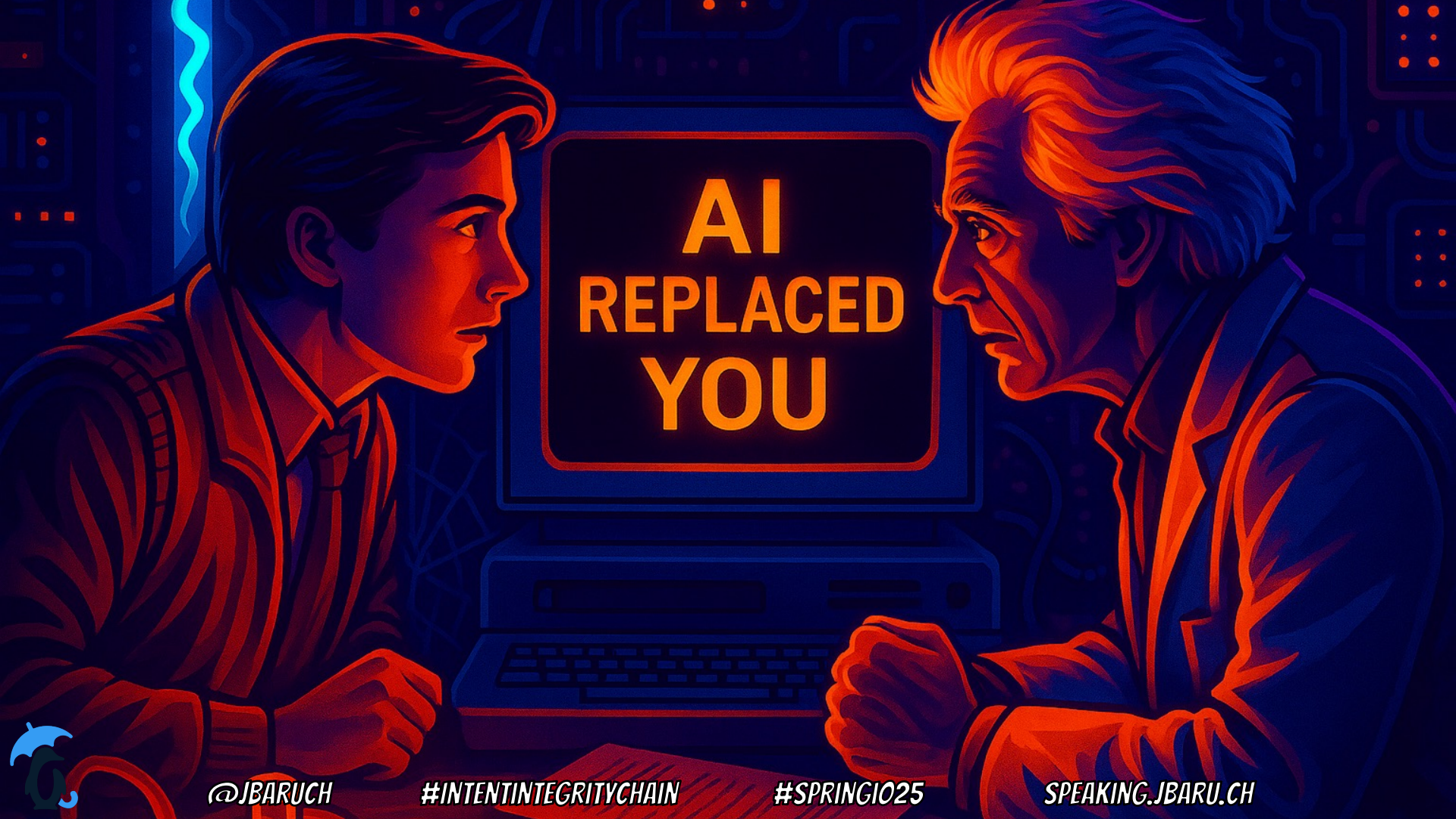
PROMPTS

#SPRINGIO25

SPEAKING.JBARU.CH







**AI  
REPLACED  
YOU**



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

TechCrunch



Programmers bore the brunt of Microsoft's layoffs in its home state as AI writes up to 30% of its code

Coders were hit hardest among Microsoft's 2,000-person layoff in its home state of Washington, Bloomberg [reports](#).

@JBARUCH

#INTENTINTEGRITYCHAIN

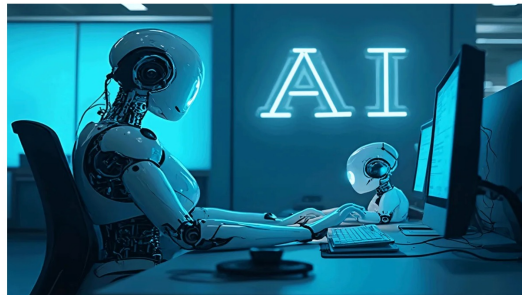
#SPRINGIO25

SPEAKING.JBARU.CH



## The Gr-AI-m Reaper: Hundreds of jobs at IBM and CrowdStrike vanish as artificial intelligence makes humans more dispensable

Both companies stress humans still have a role to play - for now



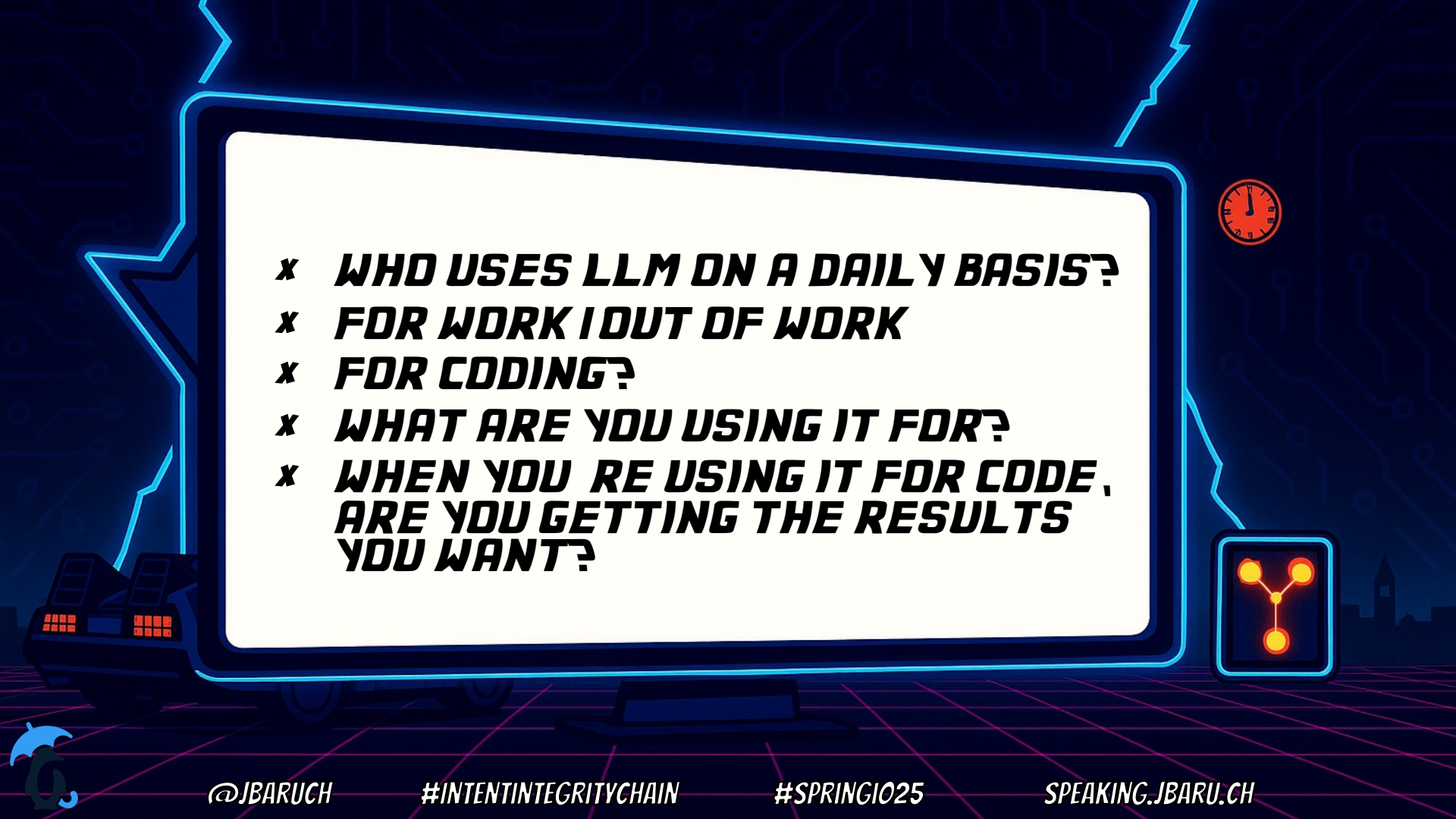


B B C

## AI could replace equivalent of 300 million jobs - report



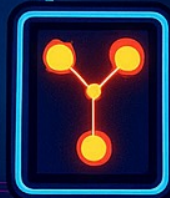
Artificial intelligence (AI) could replace the equivalent of 300 million full-time jobs, a report by investment bank Goldman Sachs says.

- 
- x WHO USES LLM ON A DAILY BASIS?***
  - x FOR WORK / OUT OF WORK***
  - x FOR CODING?***
  - x WHAT ARE YOU USING IT FOR?***
  - x WHEN YOU'RE USING IT FOR CODE,  
ARE YOU GETTING THE RESULTS  
YOU WANT?***



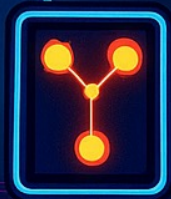
***I WILL NEVER TRUST  
CODE I DIDNT WRITE  
MYSELF***

**2020?  
1950!**



**640 KILOBYTES  
WOULD BE ENOUGH  
FOR ANYBODY!**

**1990?  
2020!**





# WE'VE SEEN THIS PANIC BEFORE

**JAVA IS KILLING  
C JOBS**

**DO WE NEED  
SYSADMINS  
WHEN WE HAVE  
DEVOPS?**

**SERVERLESS  
MEANS NO  
MORE BACKEND  
DEVELOPERS**

**ARE INFRASTRUCTURE  
ENGINEERS  
IRRELEVANT?**



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



1970s → Assembly → Compiler  
1990s → Procedural → OOP  
2000s → Bare Metal →   
2010s → Local  
2025 → Prompting

**ABSTRACTION**

**88**  
MPH

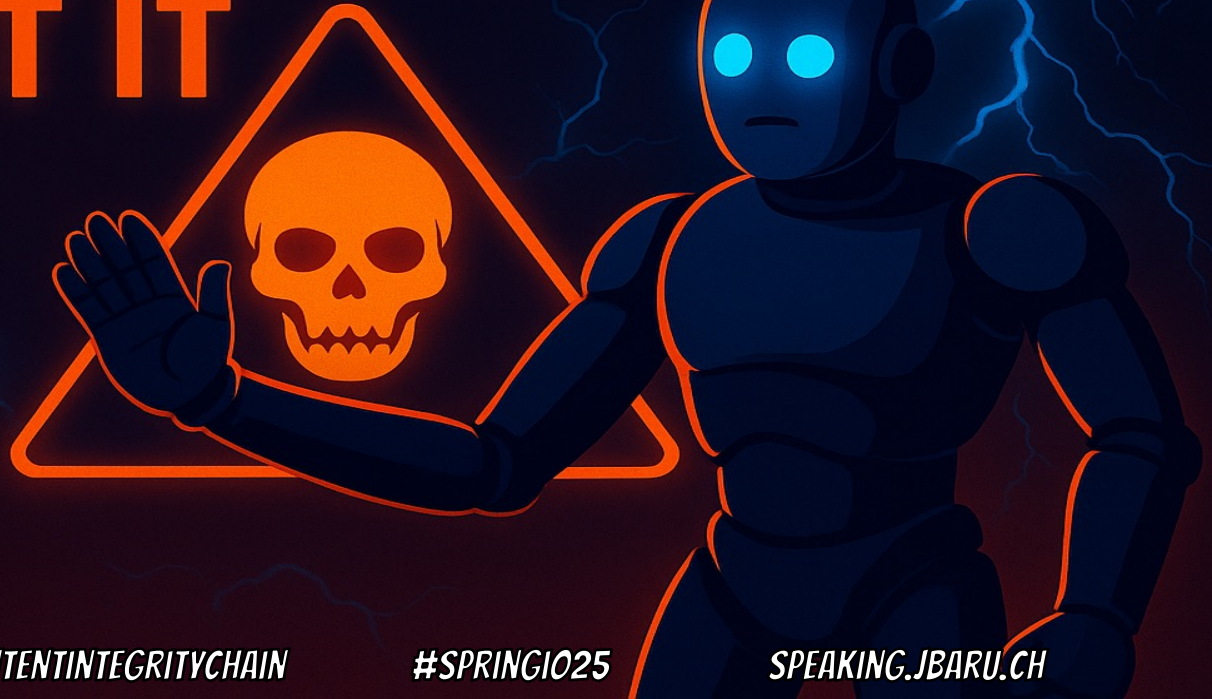


DELOREAN





# THIS TIME IT'S REAL WE REALLY CAN'T TAREST IT



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

# SAME PROMPT, DIFFERENT RESULTS



The quick brown  
fox jumps.

over he lazy dog

is named Reynard

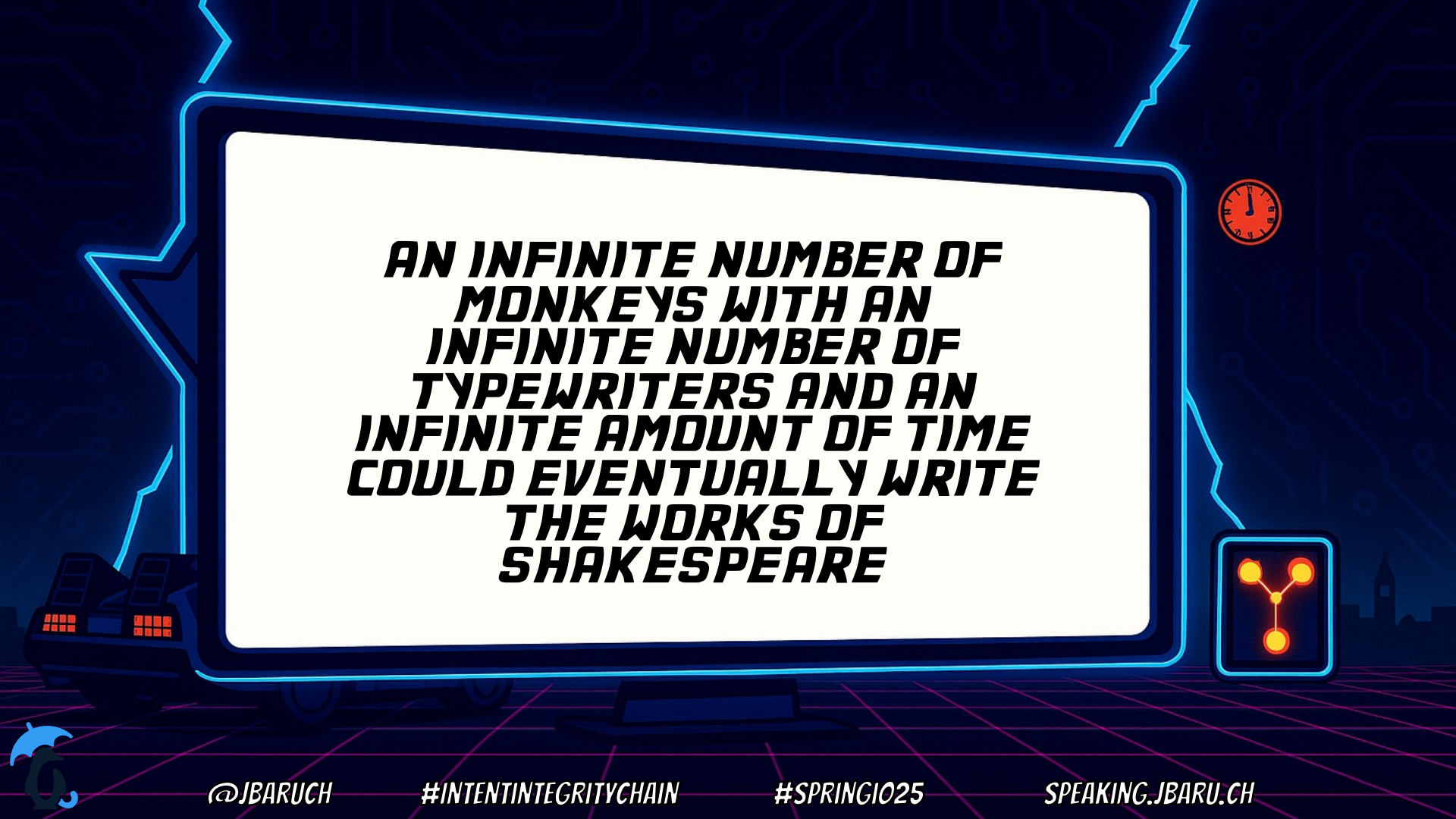
The quick brown  
fox jumps over  
the lazy dog.

is named  
Reynard.

is named Reynard.







***AN INFINITE NUMBER OF  
MONKEYS WITH AN  
INFINITE NUMBER OF  
TYPEWRITERS AND AN  
INFINITE AMOUNT OF TIME  
COULD EVENTUALLY WRITE  
THE WORKS OF  
SHAKESPEARE***

# WE GAVE THEM GPUS.



@JBARUCH

#INTENTINTEGRITYCHAIN

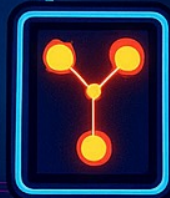
#SPRINGIO25

SPEAKING.JBARU.CH



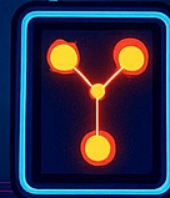
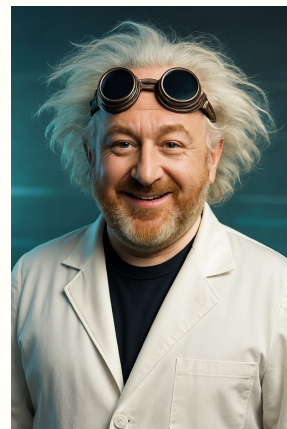
# **BARUCH SADDGURSKY - @JBARUCH**

- X HEAD OF DEVREL AT TUXCARE - I AM HIRING!**
- X ABSTRACTION CONNOISSEUR**
- X DEVELOPMENT  $\rightarrow$  DEVOPS  $\rightarrow$  ~~#~~INTENTINTEGRITYCHAIN**



# **LEONID IGOLNIK - @LIGOLNIK**

- X** **EVP OF ENG AT CLARI - I AM HIRING!**
- X** **ABSTRACTION WRANGLER**
- X** **DEVELOPMENT → ARCH → LEADER**





# ***SHOWNOTES***

- x SPEAKING JBARUCH***
- x SLIDES***
- x VIDEO***
- x ALL THE LINKS!***



# AI-GENERATED CODE IS NOT GREAT

arXiv > cs > arXiv:2304.10778

Search  
Help |

Computer Science > Software Engineering

[Submitted on 21 Apr 2023 (v1), last revised 22 Oct 2023 (this version, v2)]

## Evaluating the Code Quality of AI-Assisted Code Generation Tools: An Empirical Study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT

Burak Yetiştirilen, Işık Özsoy, Miray Ayerdem, Eray Tüzün

Context: AI-assisted code generation tools have become increasingly prevalent in software engineering, offering the ability to generate code from natural language prompts or partial code inputs. Notable examples of these tools include GitHub Copilot, Amazon CodeWhisperer, and OpenAI's ChatGPT.

Objective: This study aims to compare the performance of these prominent code generation tools in terms of code quality metrics, such as Code Validity, Code Correctness, Code Security, Code Reliability, and Code Maintainability, to identify their strengths and shortcomings.

Method: We assess the code generation capabilities of GitHub Copilot, Amazon CodeWhisperer, and ChatGPT using the benchmark HumanEval Dataset. The generated code is then evaluated based on the proposed code quality metrics.

Results: Our analysis reveals that the latest versions of ChatGPT, GitHub Copilot, and Amazon CodeWhisperer generate correct code 65.2%, 46.3%, and 31.1% of the time, respectively. In comparison, the newer versions of GitHub Copilot and Amazon CodeWhisperer showed



# ON TOP OF THAT, IT IS DANGEROUS

arXiv > cs > arXiv:2108.09293

Search...

Help | Adv

Computer Science > Cryptography and Security

[Submitted on 20 Aug 2021 (v1), last revised 16 Dec 2021 (this version, v3)]

## Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, Ramesh Karri

There is burgeoning interest in designing AI-based systems to assist humans in designing computing systems, including tools that automatically generate computer code. The most notable of these comes in the form of the first self-described "AI pair programmer", GitHub Copilot, a language model trained over open-source GitHub code. However, code often contains bugs – and so, given the vast quantity of unvetted code that Copilot has processed, it is certain that the language model will have learned from exploitable, buggy code. This raises concerns on the security of Copilot's code contributions. In this work, we systematically investigate the prevalence and conditions that can cause GitHub Copilot to recommend insecure code. To perform this analysis we prompt Copilot to generate code in scenarios relevant to high-risk CWEs (e.g. those from MITRE's "Top 25" list). We explore Copilot's performance on three distinct code generation axes -- examining how it performs given diversity of weaknesses, diversity of prompts, and diversity of domains. In total, we produce 89 different scenarios for Copilot to complete, producing 1,689 programs. Of these, we found approximately 40% to be vulnerable.

# ASKING IT TO FIX IT IS AS RELIABLE AS THE REST OF IT

arXiv > cs > arXiv:2405.12641

Search...  
Help | Ad

Computer Science > Software Engineering

[Submitted on 21 May 2024 (v1), last revised 28 Nov 2024 (this version, v2)]

## Fight Fire with Fire: How Much Can We Trust ChatGPT on Source Code-Related Tasks?

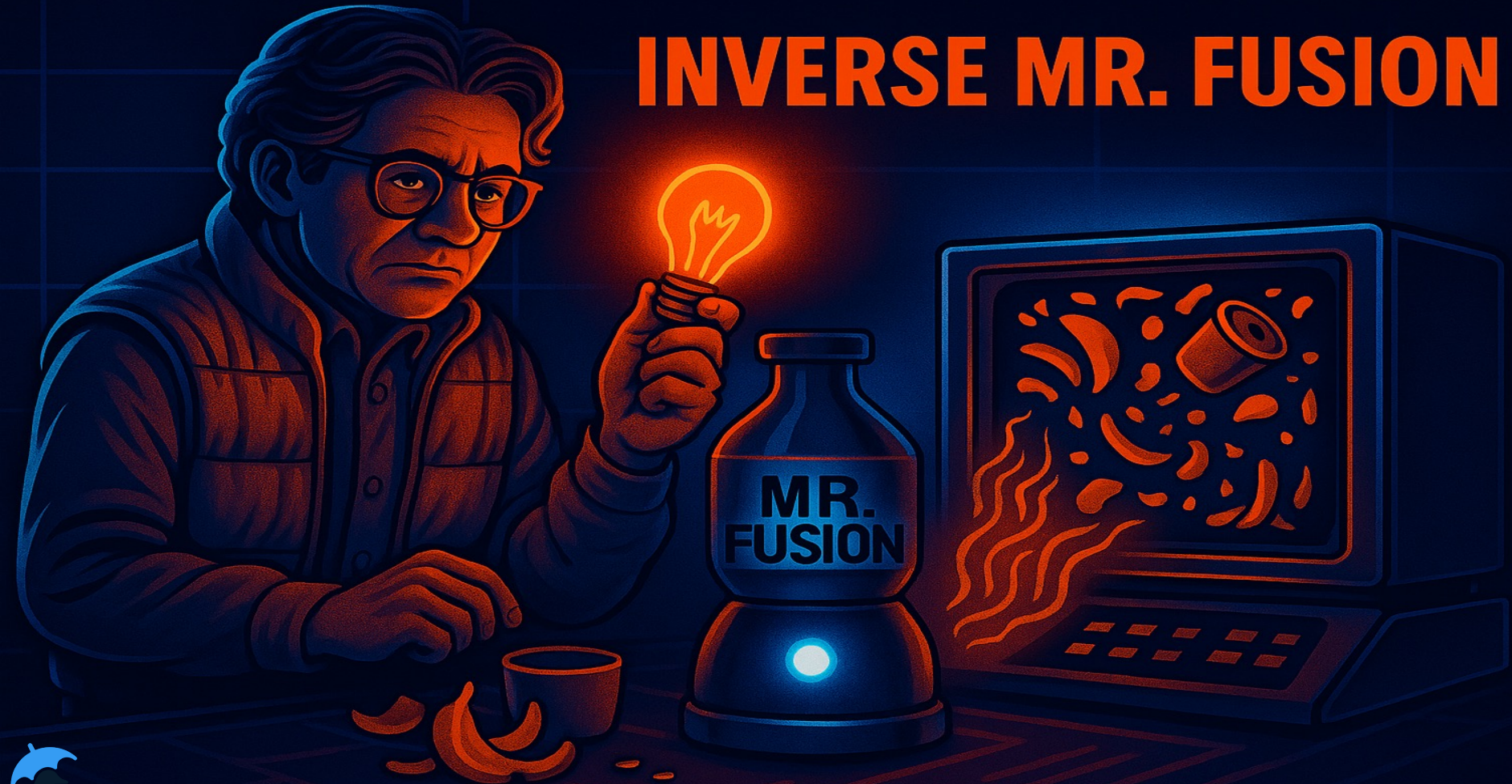
Xiao Yu, Lei Liu, Xing Hu, Jacky Wai Keung, Jin Liu, Xin Xia

With the increasing utilization of large language models such as ChatGPT during software development, it has become crucial to verify the quality of code content it generates. Recent studies proposed utilizing ChatGPT as both a developer and tester for multi-agent collaborative software development. The multi-agent collaboration empowers ChatGPT to produce test reports for its generated code, enabling it to self-verify the code content and fix bugs based on these reports. However, these studies did not assess the effectiveness of the generated test reports in validating the code. Therefore, we conduct a comprehensive empirical investigation to evaluate ChatGPT's self-verification capability in code generation, code completion, and program repair. We request ChatGPT to (1) generate correct code and then self-verify its correctness; (2) complete code without vulnerabilities and then self-verify for the presence of vulnerabilities; and (3) repair buggy code and then self-verify whether the bugs are resolved. Our findings on two code generation datasets, one code completion dataset, and two program repair datasets reveal the following observations: (1) ChatGPT often erroneously predicts its generated incorrect code as correct. (2) The self-contradictory hallucinations in ChatGPT's behavior arise. (3) The self-verification capability of ChatGPT can be enhanced by asking the guiding question, which queries whether ChatGPT agrees with assertions about incorrectly generated or repaired code and vulnerabilities in completed code. (4) Using test reports generated by ChatGPT can identify more vulnerabilities in completed code, but the explanations for incorrectly generated code and failed repairs are mostly inaccurate in the test reports. Based on these findings, we provide implications for further research or development using ChatGPT.





# INVERSE MR. FUSION



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# PANIC VS. OPPORTUNITY



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# EVERY PROFESSION GETS A SIDEKICK



@JBARUCH

#INTENTINTEGRITYCHAIN

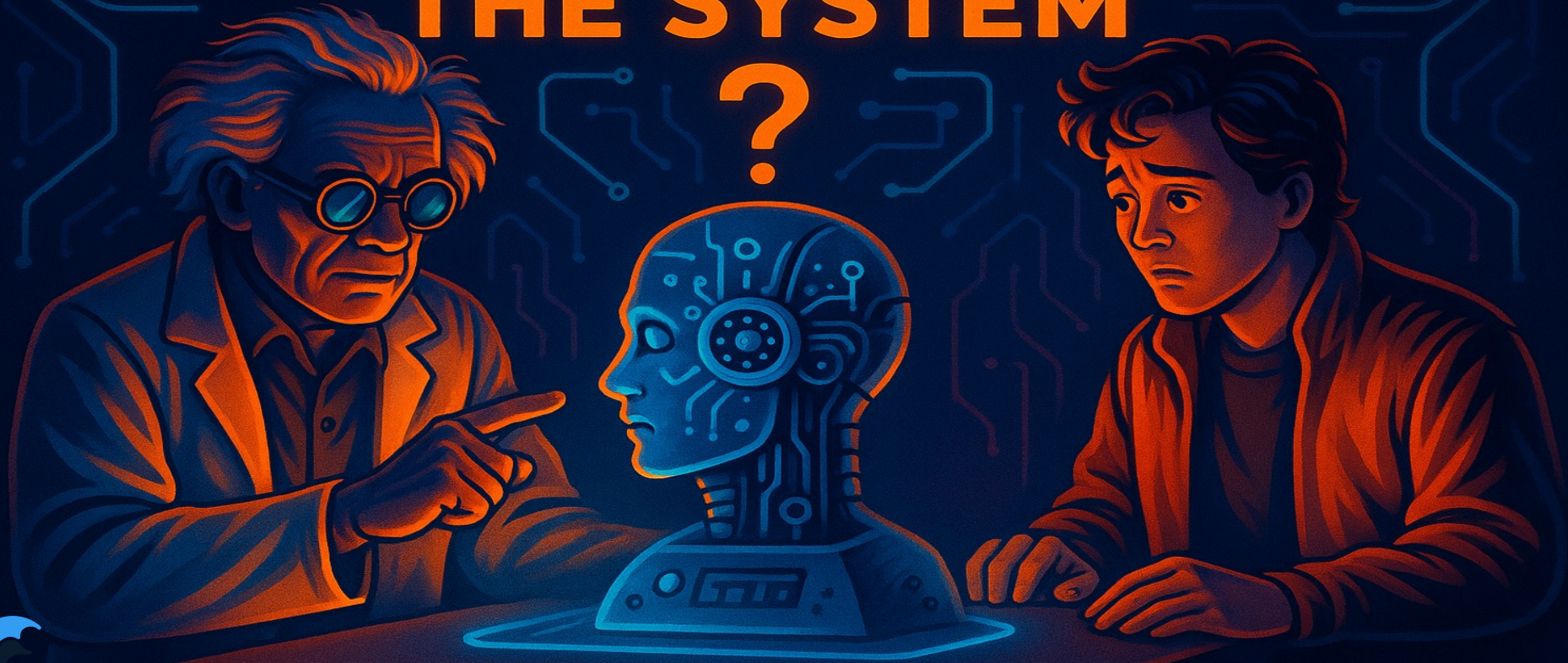
#SPRINGIO25

SPEAKING.JBARU.CH



# UNDERSTAND THE SYSTEM

?



@JBARUCH

#INTENTINTEGRITYCHAIN

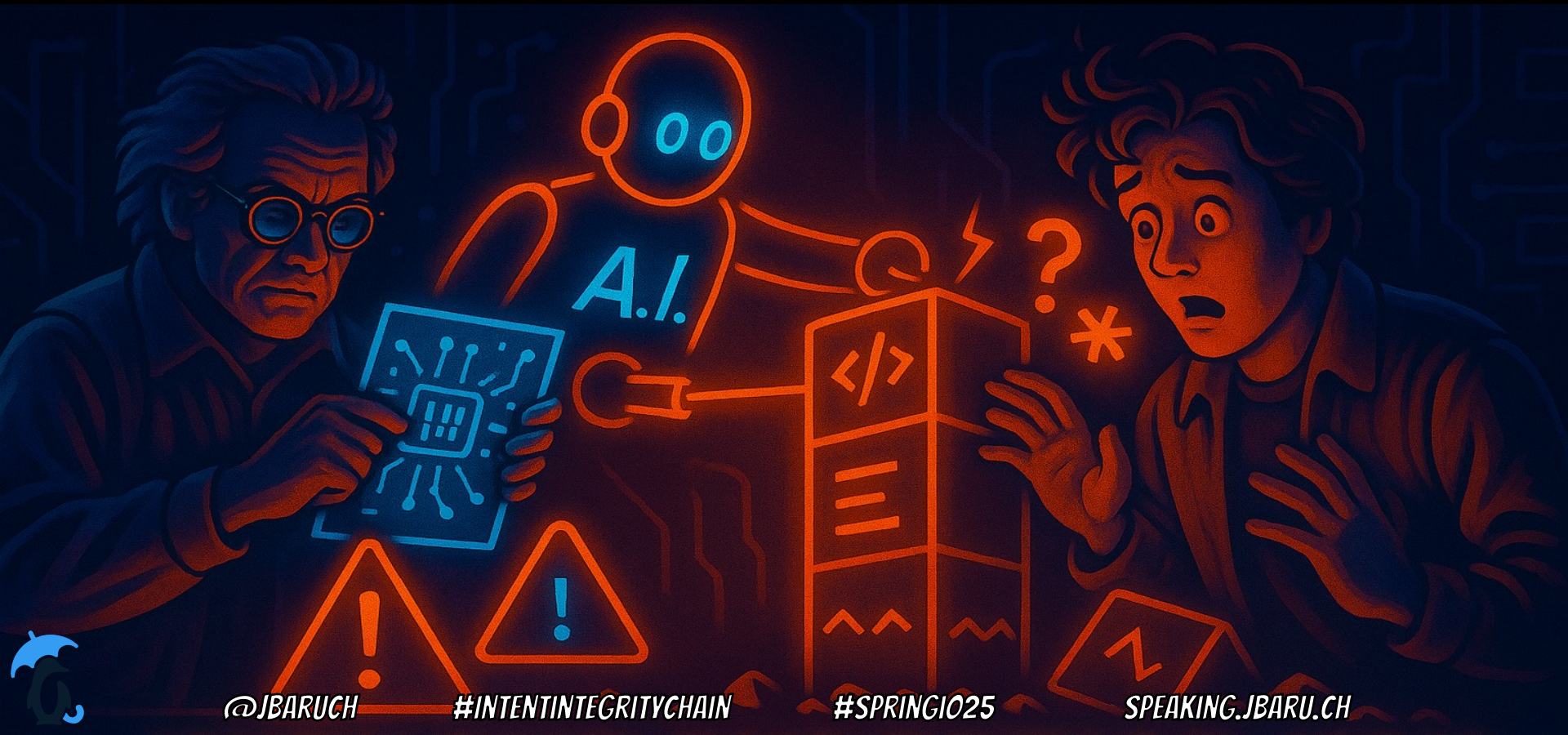
#SPRINGIO25

SPEAKING.JBARU.CH





**WITHOUT EXPERTISE,  
YOU GET CHAOS—NOT CODE**



@JBARUCH

#INTENTINTEGRITYCHAIN

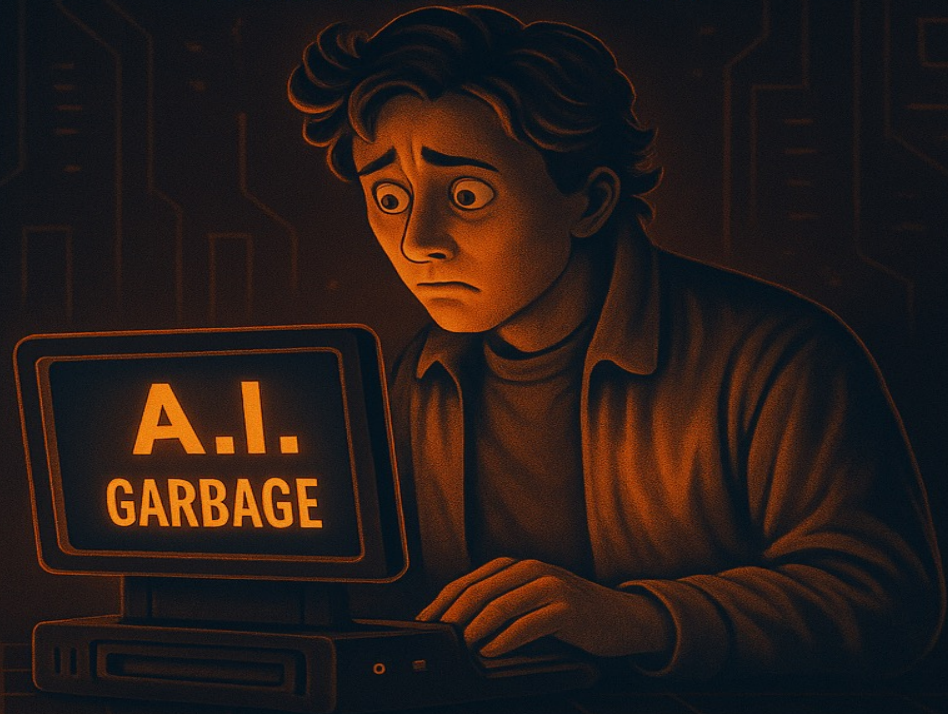
#SPRINGIO25

SPEAKING.JBARU.CH



# EXPERT

# JUNIOR



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# A CAUTIONARY TALE



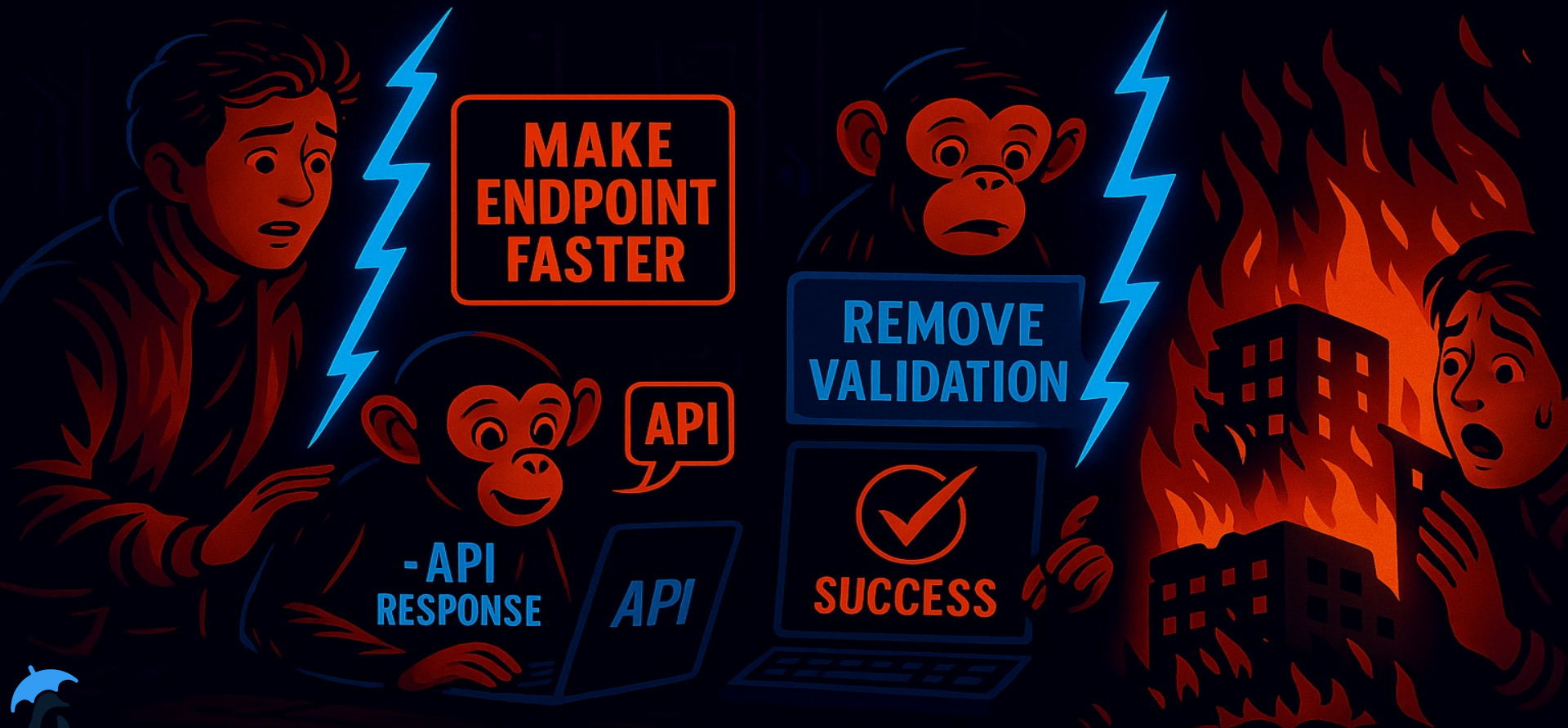
@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

**FASTER? SURE. JUST DON'T ASK WHAT IT COST.**



@JBARUCH

#INTENTINTEGRITYCHAIN

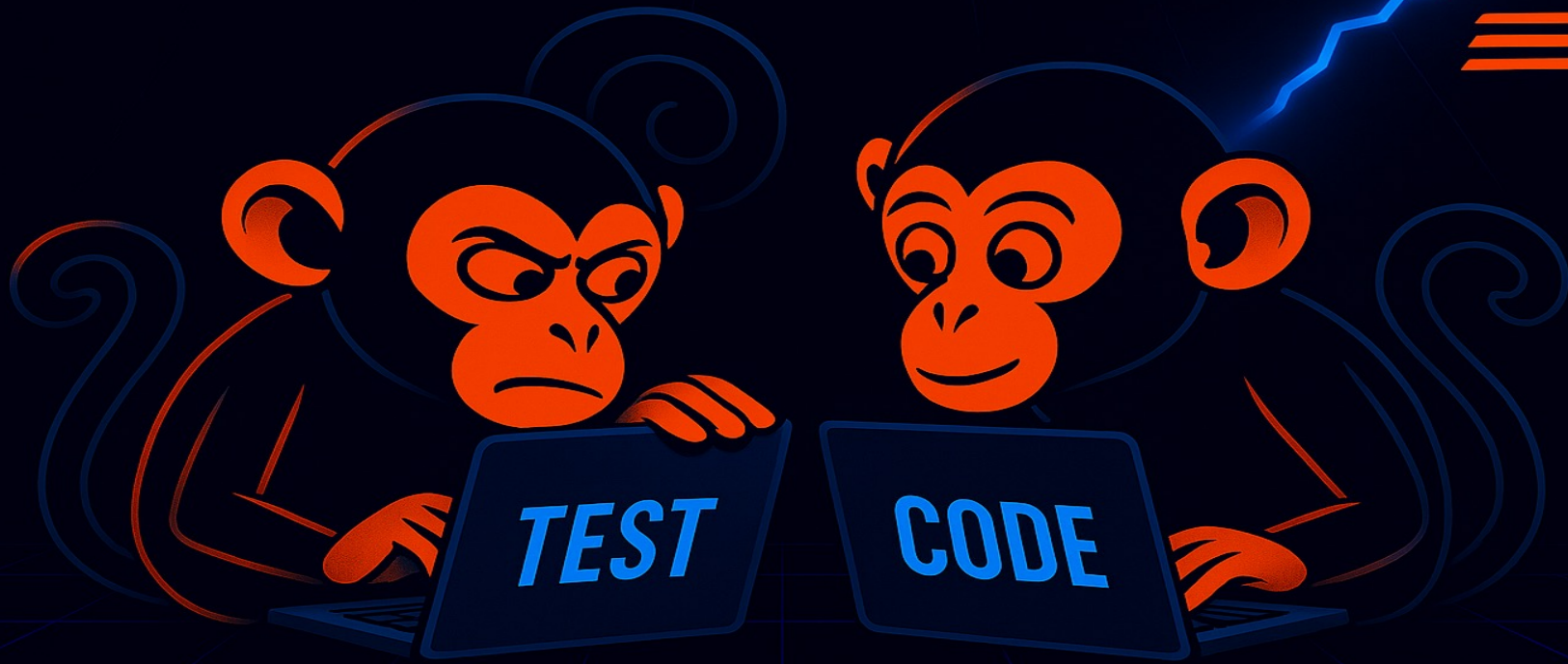
#SPRINGIO25

SPEAKING.JBARU.CH





# CIRCULAR VERIFICATION



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



**PROMPT  
SUCCESSFUL**



@JBARUCH

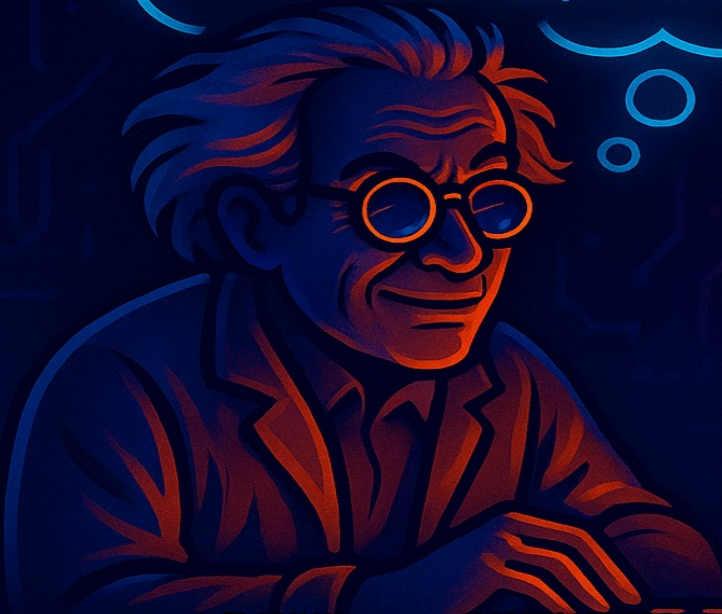
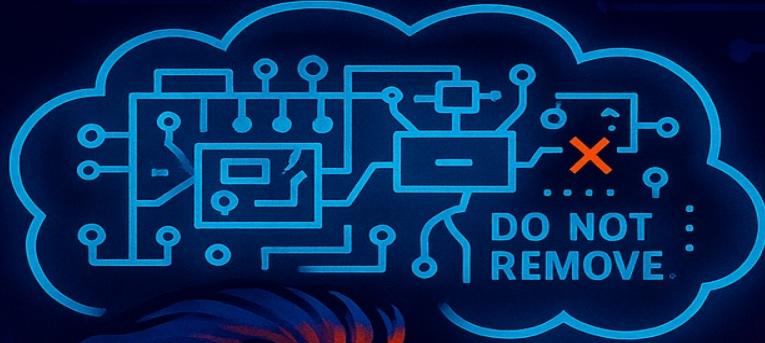
#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



**HE THOUGHT IT WAS OBVIOUS.**



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH







# INTENT

# PROMPT



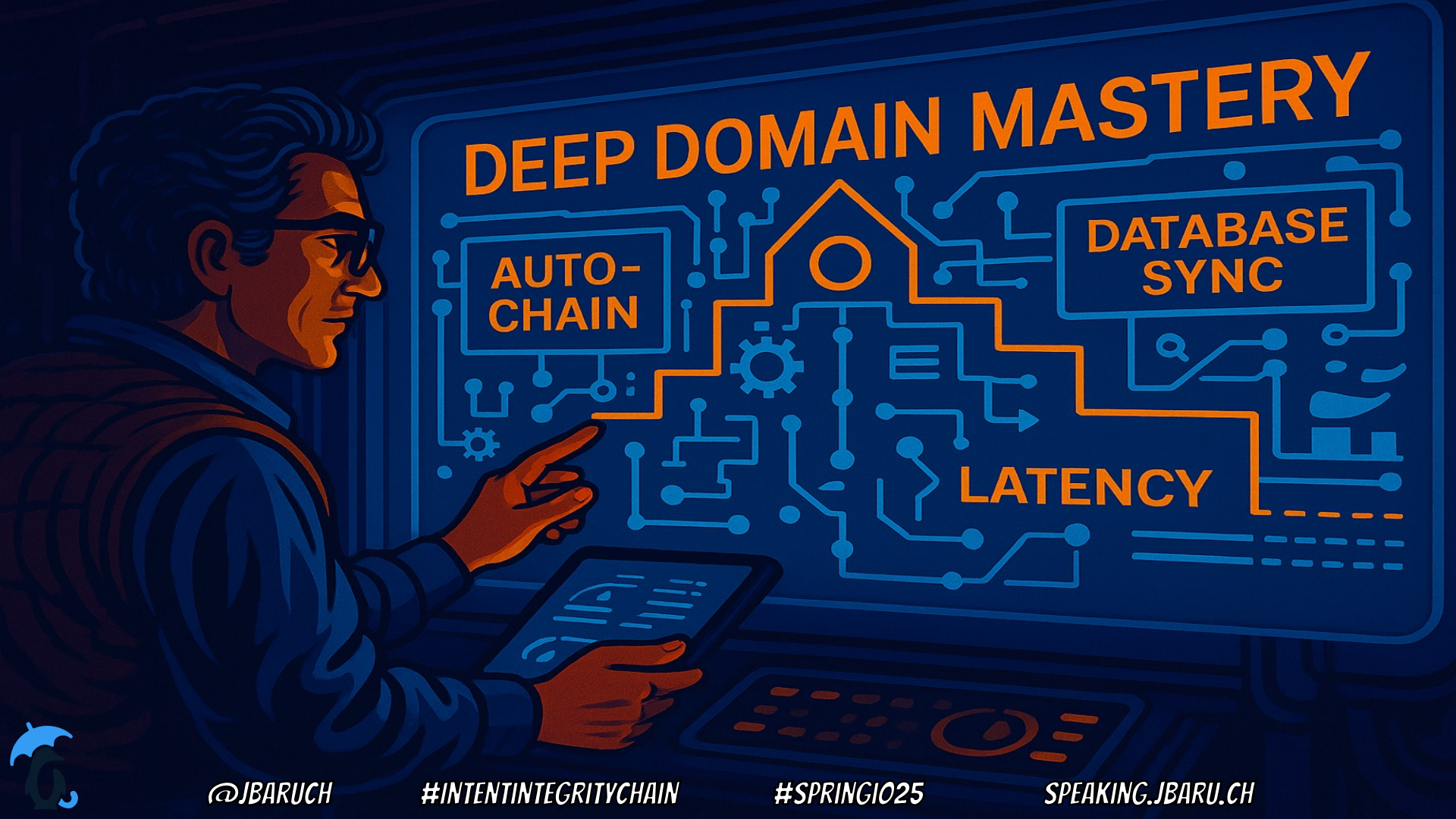
@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH





# DEEP DOMAIN MASTERY

AUTO-CHAIN

DATABASE SYNC

LATENCY



@JBARUCH

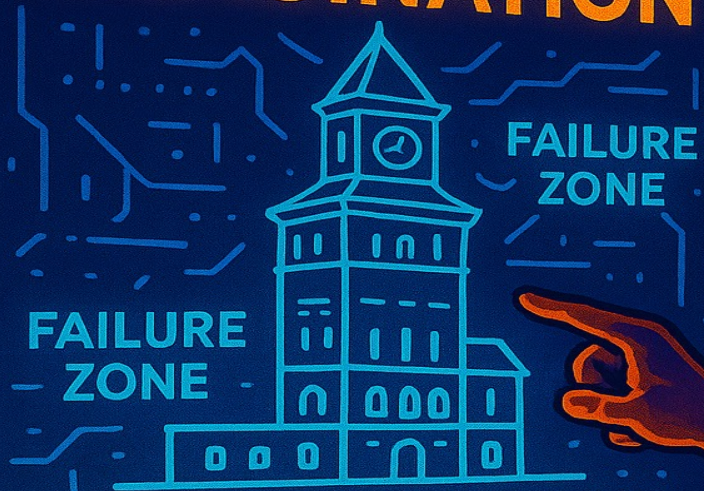
#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# STRUCTURED IMAGINATION



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# GUARDRAILS FOR SAFETY



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

# THE ART OF POSSIBLE



**DOMAIN  
MASTERY**

**STRUCTURED  
IMAGINATION**

**PURPOSEFUL  
GUARDRAILS**

@JBARUCH

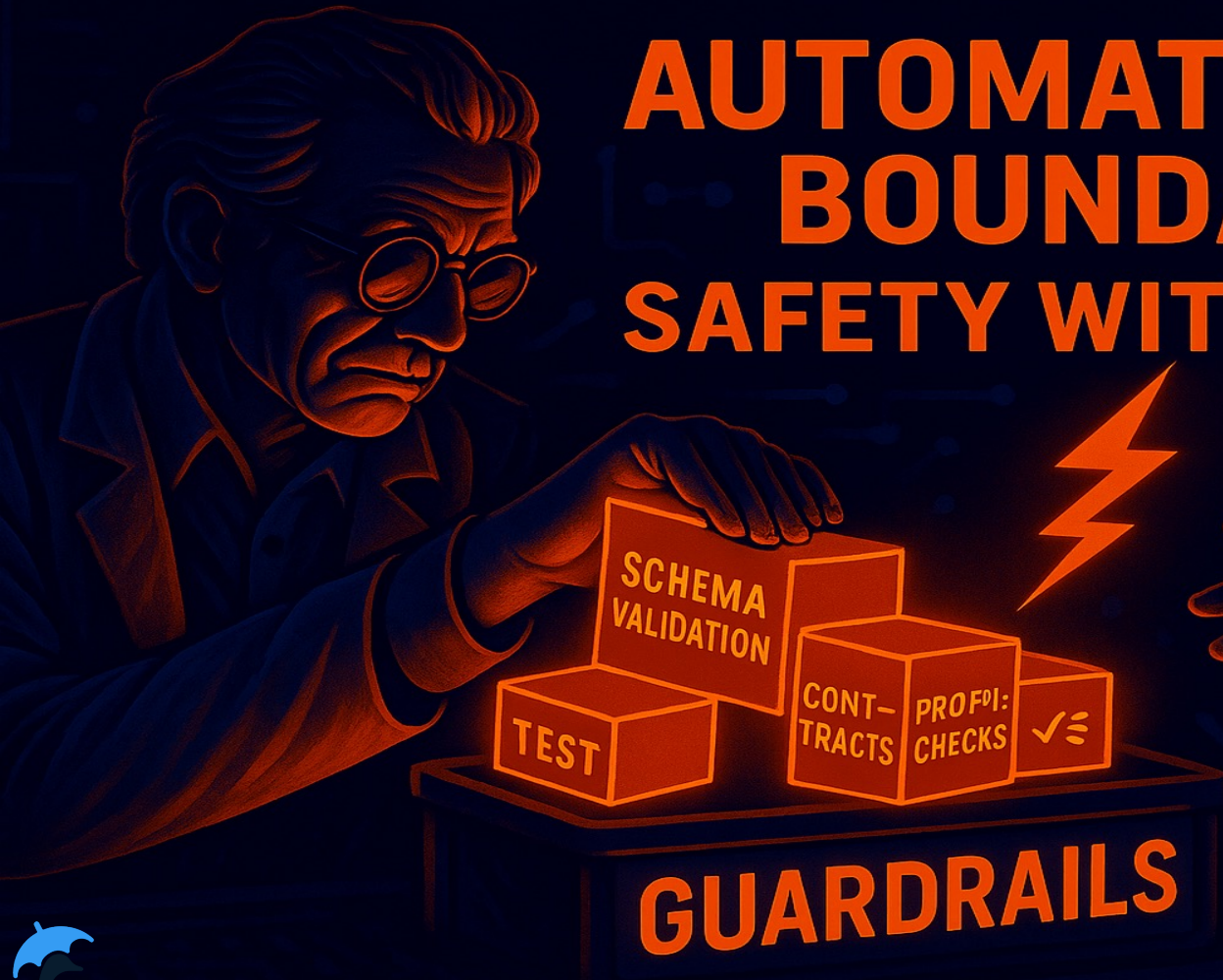
#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# AUTOMATION WITH BOUNDARIES. SAFETY WITH TEETH.



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

# LET'S EXPRESS INTENT IN TESTS...

TDD?

## EXECUTABLE INTENT

```
{  
  _____  
  _____  
  _____  
  _____  
}  
{  
  _____  
  _____  
}
```

✓  
**ALWAYS  
UP-TO-DATE**

✓  
**GENERATES  
CONSENSUS**

✓  
**EXECUTABLE**

✓  
✓  
✓  
✓  
**PARSED BY  
THE MACHINE**



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# DEVELOPERS ARE BIASED FOR ACTION



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

# TEST-FIRST. WITHOUT TRYING.



```
testProcess  
Orders() ..f  
assert  
Checks() ;
```

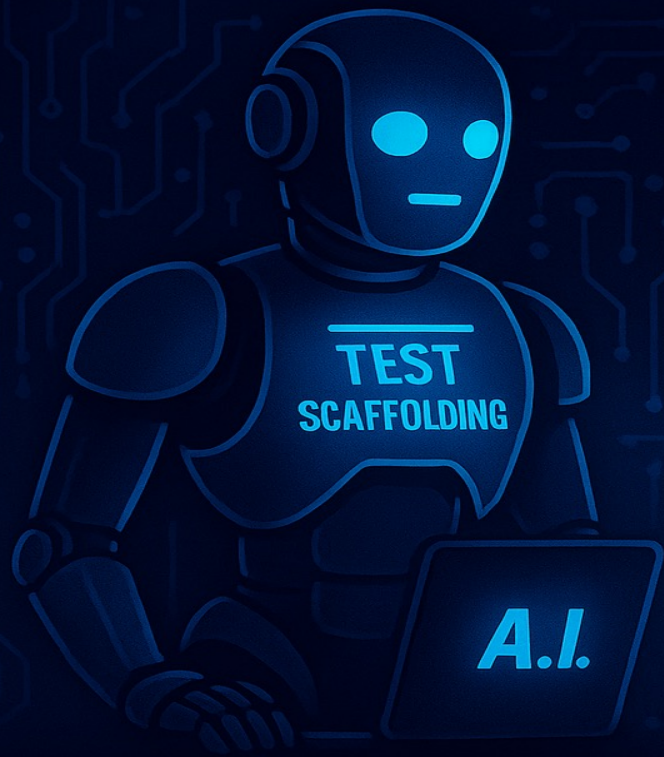
TestOrders.java

```
testProcess  
Orders();  
assertChecks(
```





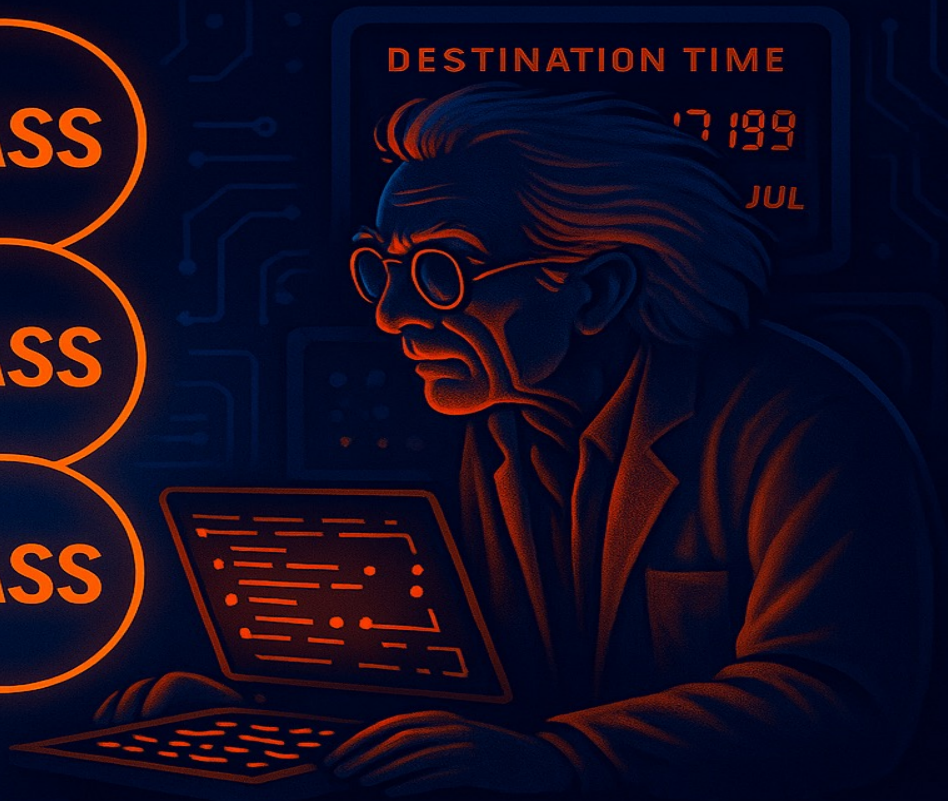
# DISCIPLINE BECOMES A SIDE EFFECT



PASS

PASS

PASS



@JBARUCH

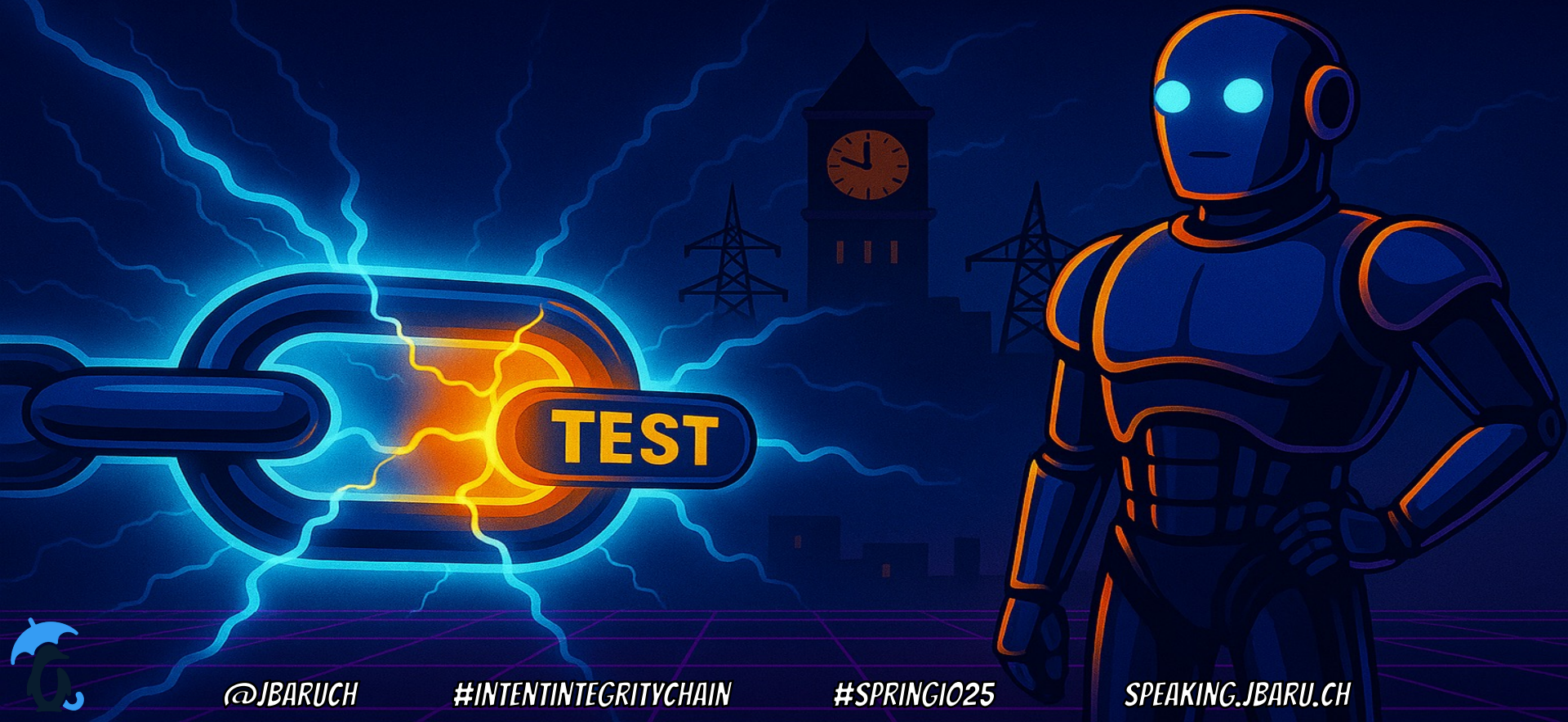
#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# TDD IS THE FIRST LINK



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# THEY DIDN'T REVIEW THE TESTS... BECAUSE THEY COULDN'T READ THEM



**PRDDUCT  
MANAGERS**



# DESCRIBE EXPECTED BEHAVIOR. IN STRUCTURED PLAIN ENGLISH



**GIVEN**  
the account is active

**WHEN**  
the user submits a transfer

**THEN**  
the balance is updated





# BDD REQUIRED DISCIPLINE AND SYNTAX

GIVEN user clicks on

Unrecognized token

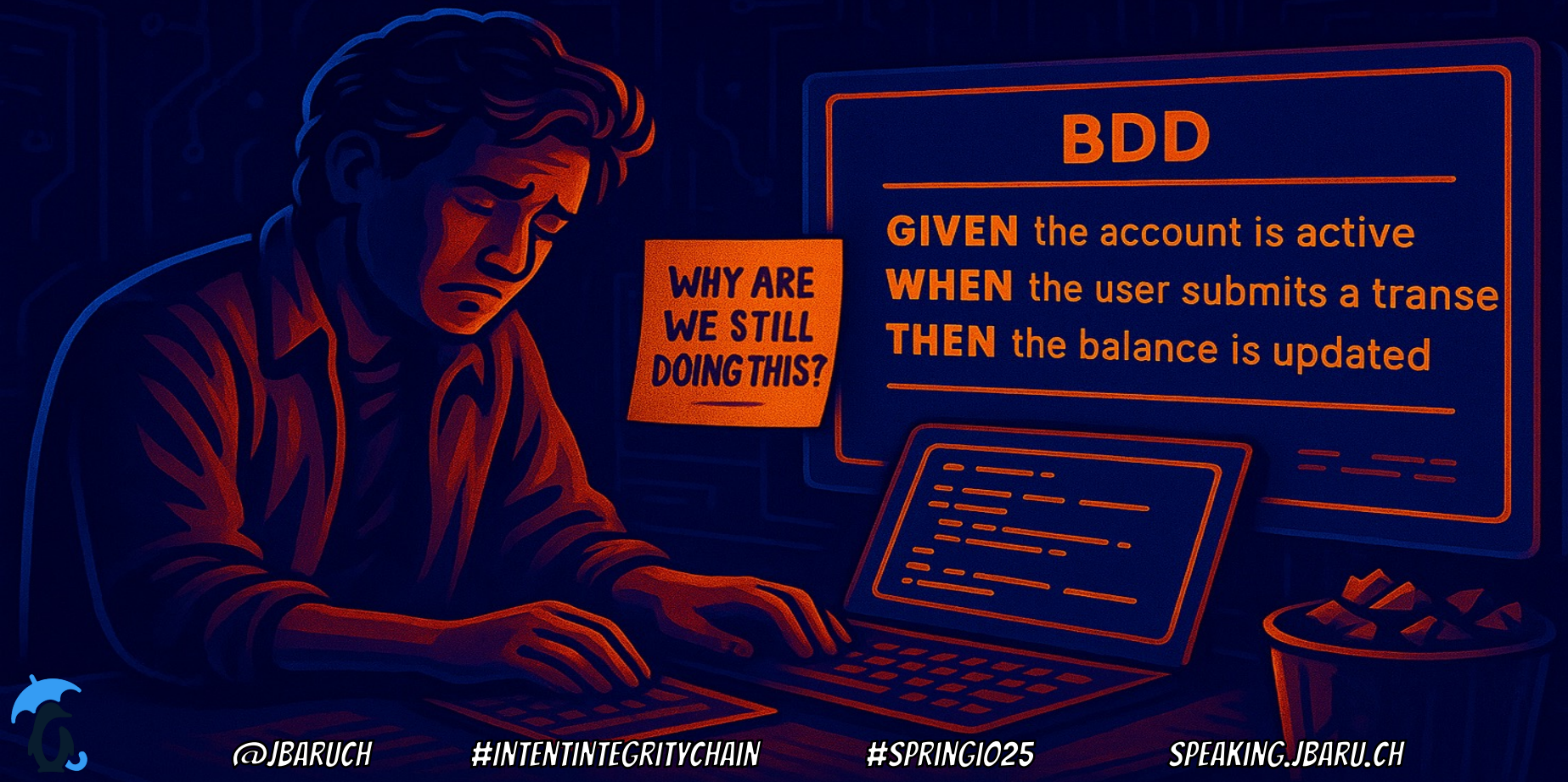
WHEN the (error) br...

Step mismatch

Missing THEN claus



# THE OVERHEAD OUTWEIGHED THE BENEFIT



## BDD

**GIVEN** the account is active  
**WHEN** the user submits a transe  
**THEN** the balance is updated

WHY ARE  
WE STILL  
DOING THIS?



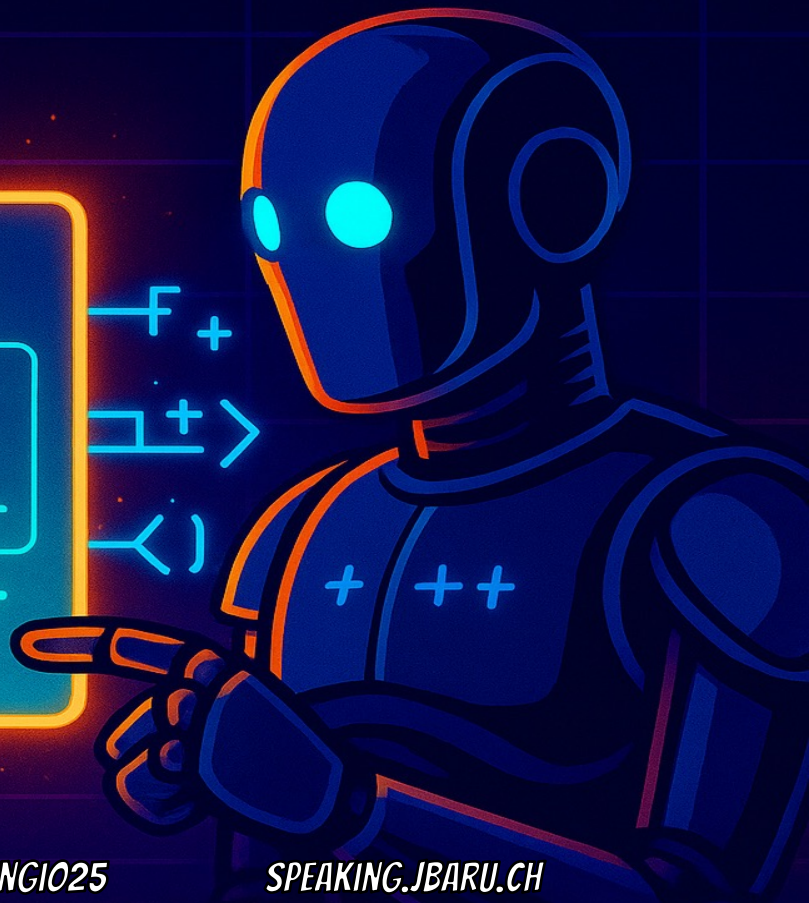
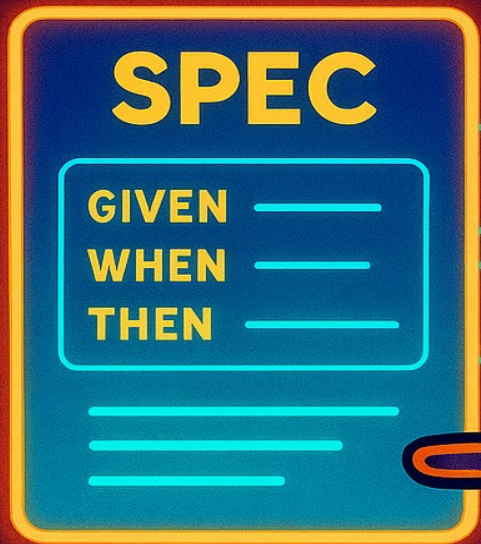
# WHO NEEDS GHERKIN WHEN YOU CAN VIBECODE?



Handle payments unless  
the card is expired...  
then retry mabe?



READABLE BY HUMANS. EXECUTABLE BY MACHINES.



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



# PROMPT



# SPEC

- GIVEN
- WHEN
- THEN



# INTENT INTEGRITY CHAIN





# ***INTENT INTEGRITY CHAIN***

- x HUMANS WRITE PROMPT***
- x MONKEYS GENERATE SPEC***
- x MACHINE CREATES TESTS***
- x MONKEYS WRITE CODE***

# NEVER TRUST A MONKEY



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH





**ARTIFACT**

**CREATED**

**CAN WE  
TRUST IT?**

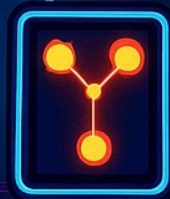
**PROMPT**



**SPEC**



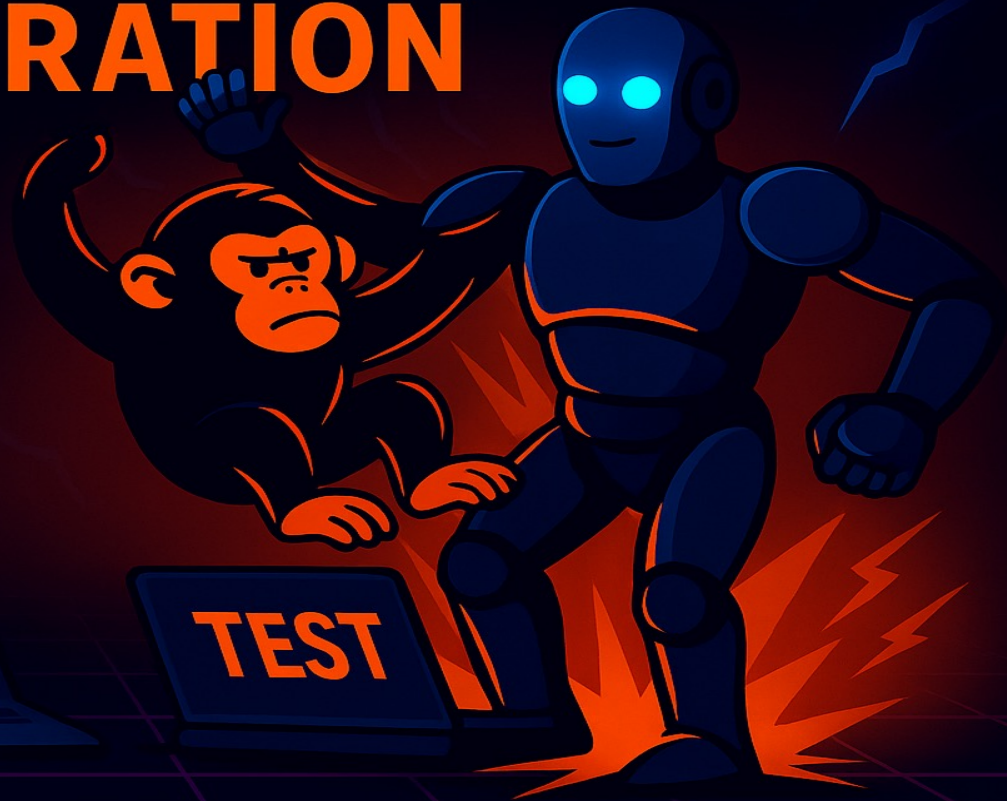
**TEST**



# DETERMINISTIC TEST GENERATION

PROMPT

SPEC



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



**ARTIFACT**

**CREATED**

**CAN WE  
TRUST IT?**

**PROMPT**



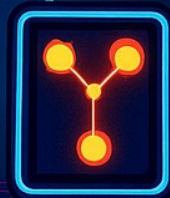
**SPEC**



**TEST**



**CODE**



# TESTS ARE LOCKED



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



**ARTIFACT**

**CREATED**

**VALIDATED**

**PROMPT**



**SPEC**



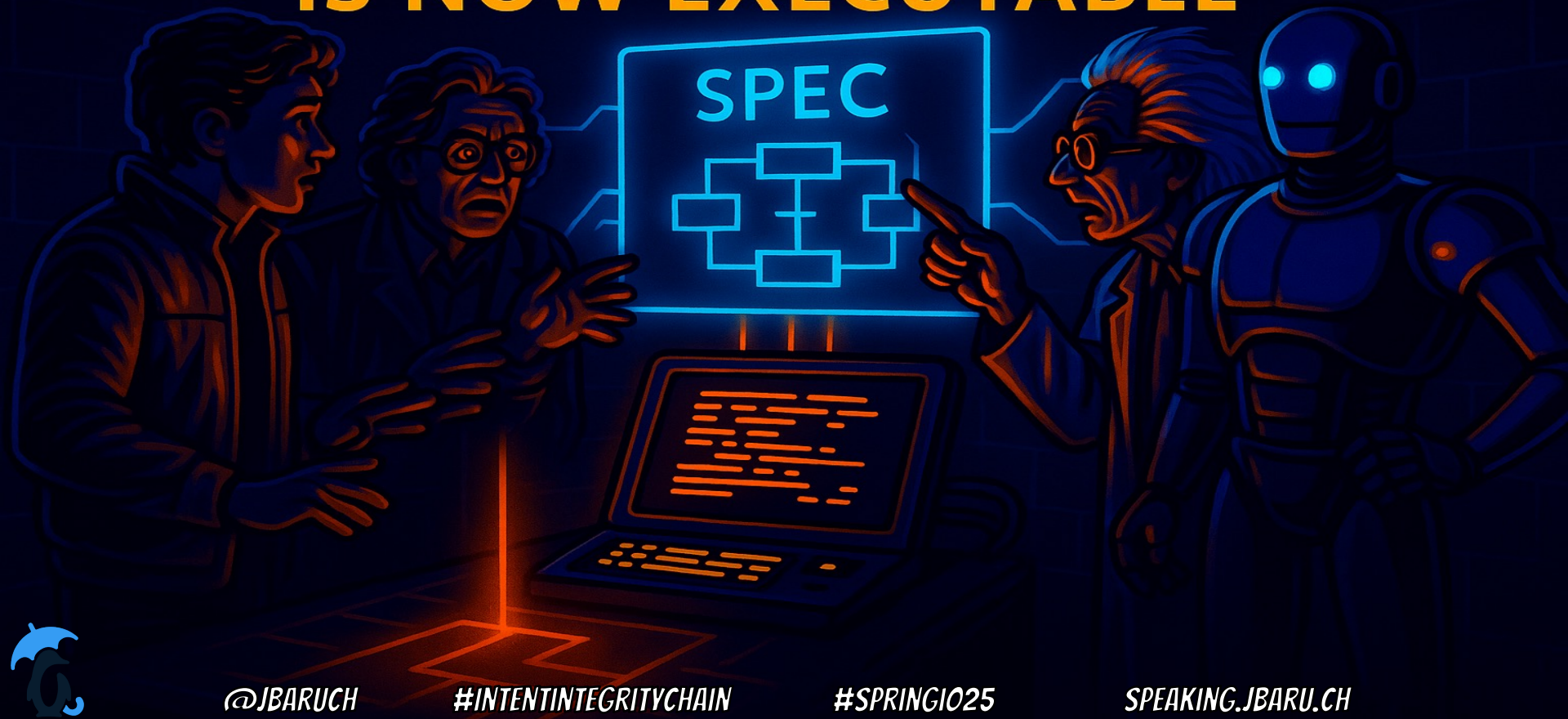
**TEST**



**CODE**



# SHARED UNDERSTANDING IS NOW EXECUTABLE



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH





# ONLY WHAT'S IN SPEC GETS TO TESTS



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH

# ONLY WHAT PASSES THE TESTS STAYS IN CODE



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH





AppleIntelLIJ IDEAFileEditViewNavigateCodeRefactorBuildRunTools

ic-demo

main

Project

ic-demo

.gradle

.idea

.junie

guidelines.md

prompt.md

gradle

External Libraries

Scratches and Consoles

Search Everywhere Double

Go to File

Recent Files

Navigation Bar

Drop files here to open them

Current File

Junie

Type your task here, press Enter to send prompt

Code

Ask

Brave Mode

Examples

Generate project-specific guidelines

Create a list of tasks to improve the project

Analyse requirements and create an improve...

Implement the improvements according to the t...

ic-demo

.junie

guidelines.md

Windsurf Pro (Pre-Release)

**ARTIFACT**

**CREATED**

**VALIDATED**

**PROMPT**



**SPEC**



**TEST**



**CODE**







**Andrej Karpathy** ✓  
@karpathy



The hottest new programming language is English

9:14 PM · Jan 24, 2023 · **7.3M** Views



1.1K



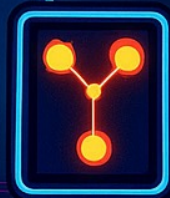
6.9K



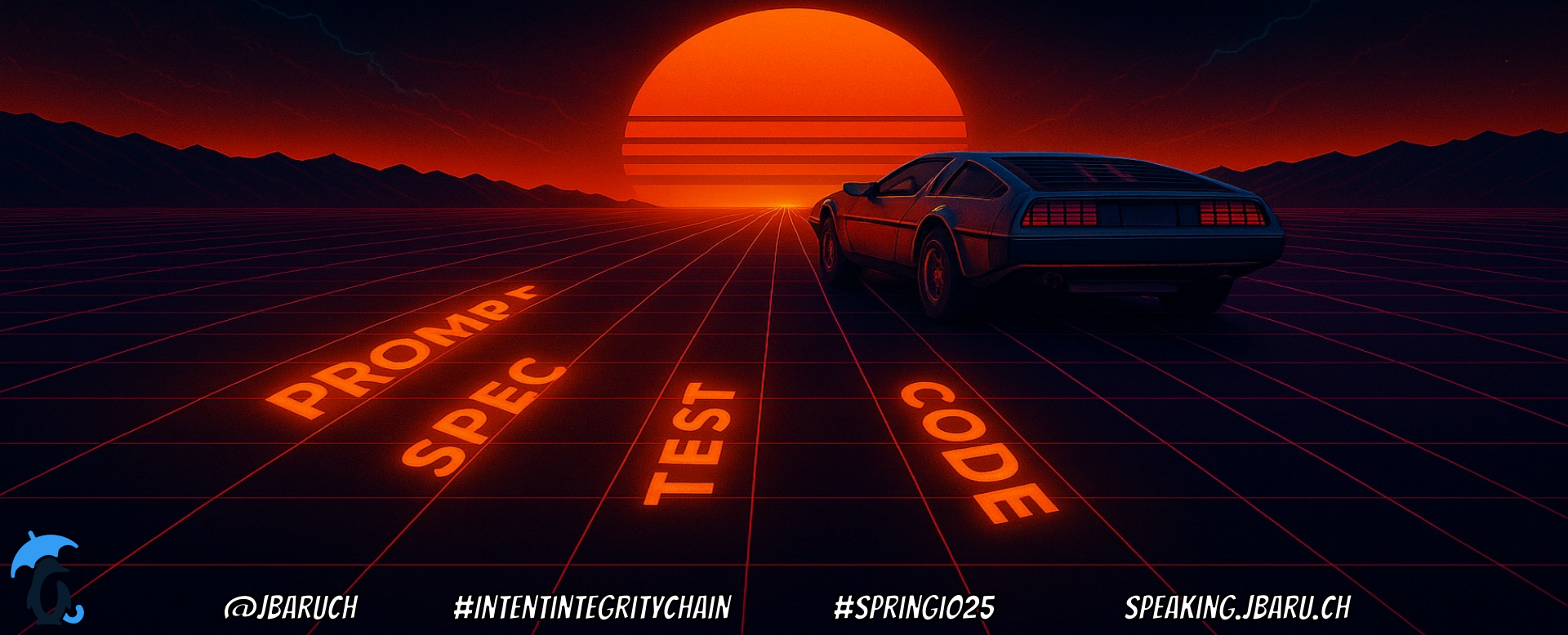
44K



4.2K



# IT'S NOT ABOUT PROMPTING. IT'S ABOUT ALIGNMENT.



@JBARUCH

#INTENTINTEGRITYCHAIN

#SPRINGIO25

SPEAKING.JBARU.CH



- x EVERY ABSTRACTION LEAP  
STARTED WITH PANIC AND ENDED  
WITH PROGRESS***
- x MASTERY MATTERS MORE THAN  
EVER***
- x SHARED UNDERSTANDING IS NOW  
EXECUTABLE***





# THANK YOU!

- x @JBARUCH
- x #SPRING1025
- x INTENTINTEGRITYCHAIN
- x SPEAKING.JBARU.CH

