BrowserStack

STACKCONNECT : JAKARTA EDITION

TESTING LLM BASED PRODUCT AND ITS OBSERVABILITY

PREPARED BY ALIFIA

ATC
Alif Test Consulting

Alifia

- Software Testing Consultant based in Yogyakarta.
  Active in ISQA Yogyakarta Chapter as volunteer
- Alif Test Consulting provide testing as service : R&D, Consulting, Outsourcing and Training.
- Disclaimer : I’m not an AI expert

More and more applications will include AI components

Why should we care to test AI part of the application?

1. Assess more risks coverage
2. Increase our value

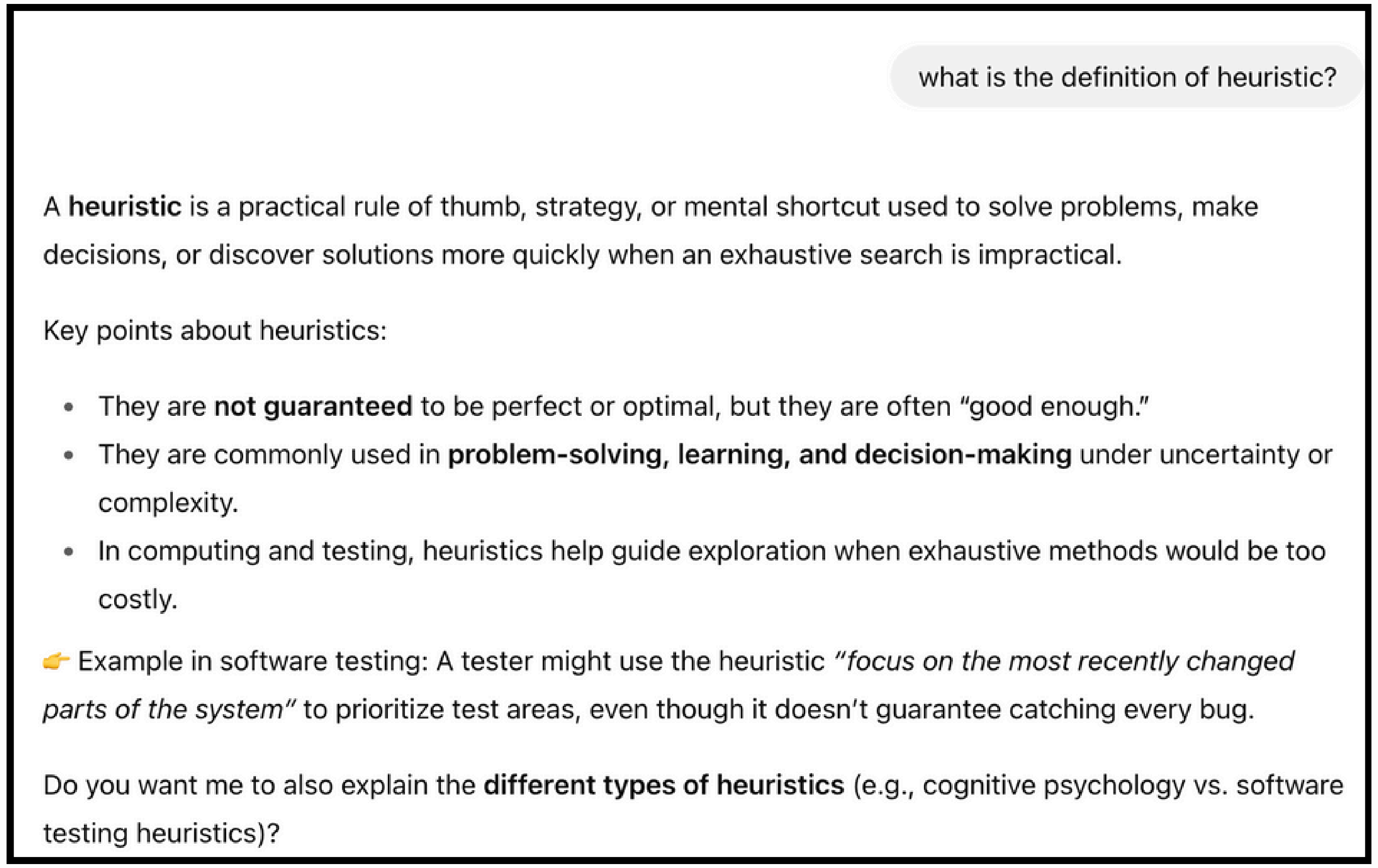
AI
is here to stay

LLM CHALLENGES

Non-Deterministic
No same output for the same input

Hallucination
make up fact/deviate from truth

Non Deterministic



what is the definition of heuristic?

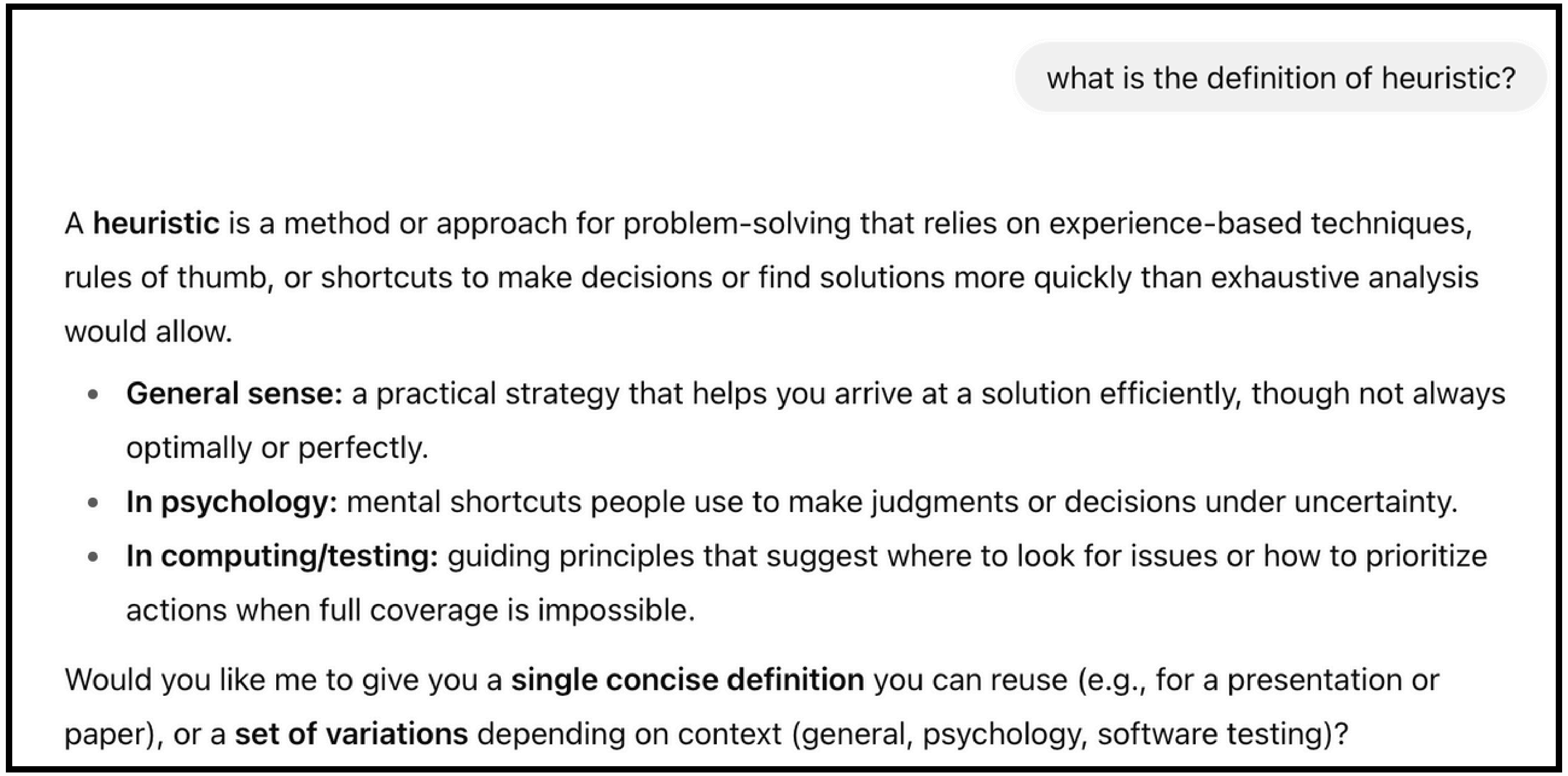
A **heuristic** is a practical rule of thumb, strategy, or mental shortcut used to solve problems, make decisions, or discover solutions more quickly when an exhaustive search is impractical.

Key points about heuristics:

- They are **not guaranteed** to be perfect or optimal, but they are often "good enough."
- They are commonly used in **problem-solving, learning, and decision-making** under uncertainty or complexity.
- In computing and testing, heuristics help guide exploration when exhaustive methods would be too costly.

👉 Example in software testing: A tester might use the heuristic *"focus on the most recently changed parts of the system"* to prioritize test areas, even though it doesn't guarantee catching every bug.

Do you want me to also explain the **different types of heuristics** (e.g., cognitive psychology vs. software testing heuristics)?



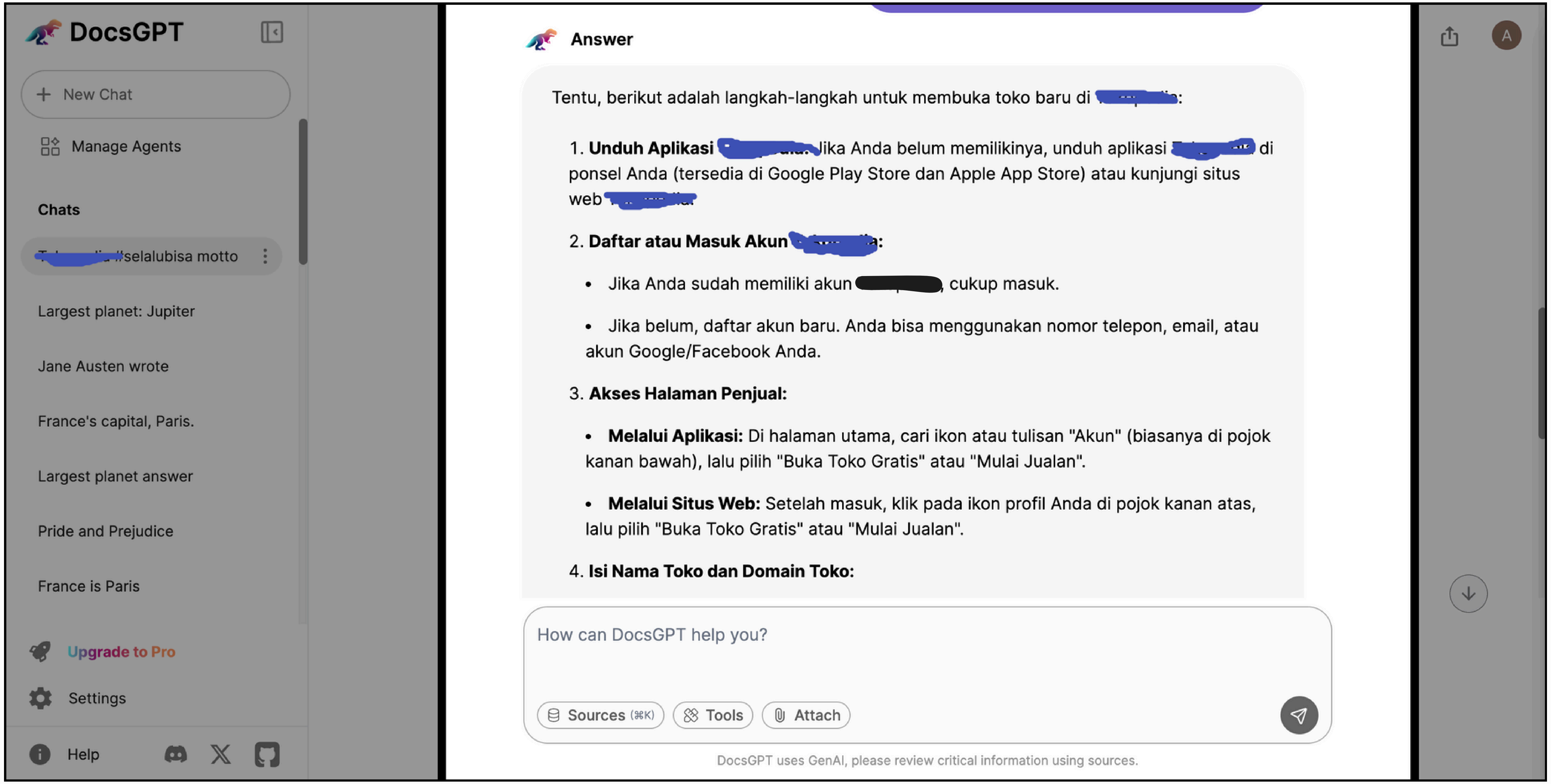
what is the definition of heuristic?

A **heuristic** is a method or approach for problem-solving that relies on experience-based techniques, rules of thumb, or shortcuts to make decisions or find solutions more quickly than exhaustive analysis would allow.

- **General sense:** a practical strategy that helps you arrive at a solution efficiently, though not always optimally or perfectly.
- **In psychology:** mental shortcuts people use to make judgments or decisions under uncertainty.
- **In computing/testing:** guiding principles that suggest where to look for issues or how to prioritize actions when full coverage is impossible.

Would you like me to give you a **single concise definition** you can reuse (e.g., for a presentation or paper), or a **set of variations** depending on context (general, psychology, software testing)?

Hallucination



Tentu, berikut adalah langkah-langkah untuk membuka toko baru di [redacted]:

1. **Unduh Aplikasi [redacted]** Jika Anda belum memilikinya, unduh aplikasi [redacted] di ponsel Anda (tersedia di Google Play Store dan Apple App Store) atau kunjungi situs web [redacted]
2. **Daftar atau Masuk Akun [redacted]:**
   - Jika Anda sudah memiliki akun [redacted], cukup masuk.
   - Jika belum, daftar akun baru. Anda bisa menggunakan nomor telepon, email, atau akun Google/Facebook Anda.
3. **Akses Halaman Penjual:**
   - **Melalui Aplikasi:** Di halaman utama, cari ikon atau tulisan "Akun" (biasanya di pojok kanan bawah), lalu pilih "Buka Toko Gratis" atau "Mulai Jualan".
   - **Melalui Situs Web:** Setelah masuk, klik pada ikon profil Anda di pojok kanan atas, lalu pilih "Buka Toko Gratis" atau "Mulai Jualan".
4. **Isi Nama Toko dan Domain Toko:**

LLM-AS-A-JUDGE

Evaluation method to assess the quality of text generated by LLM based app

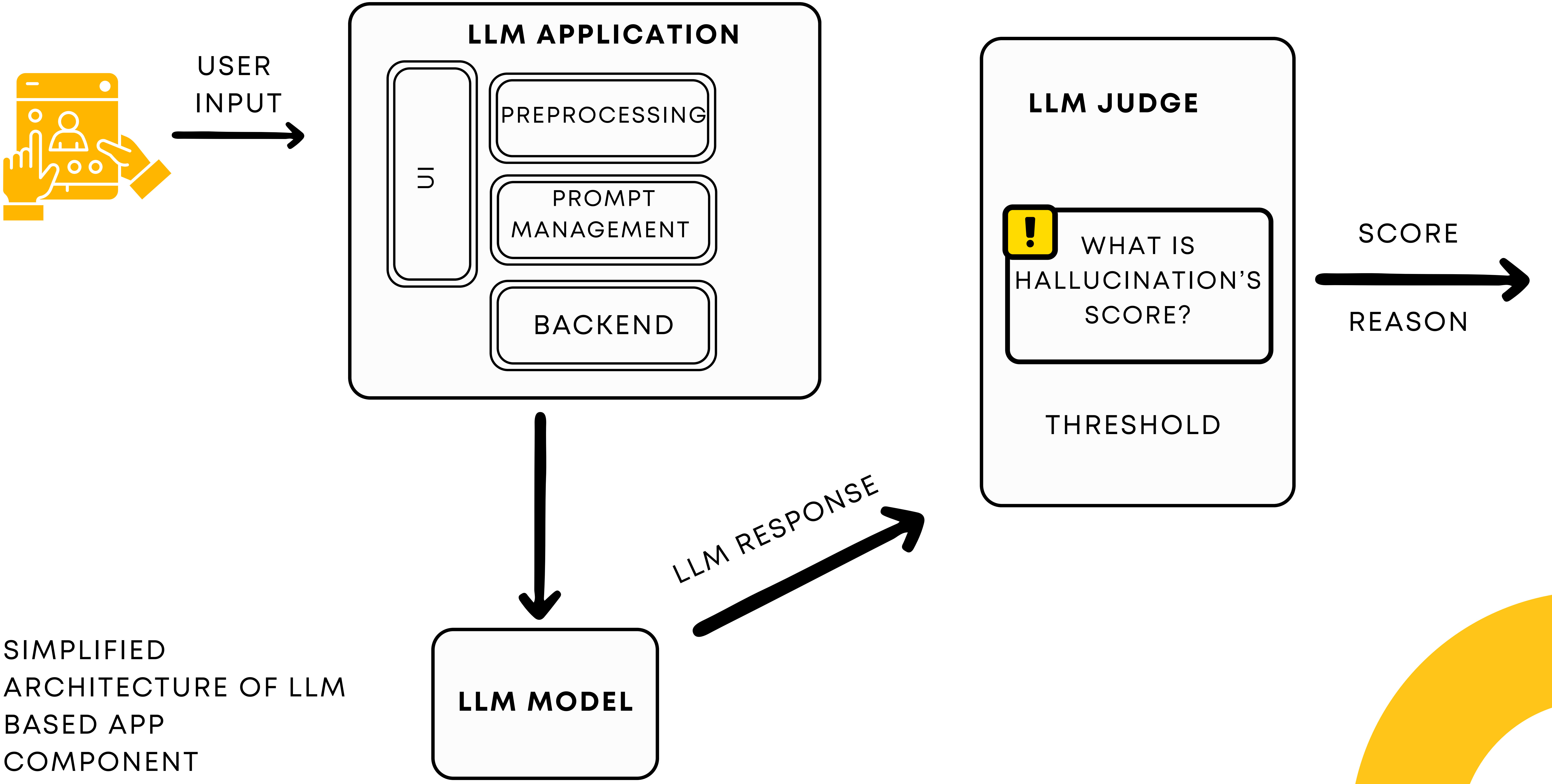
It’s a diverse technique that can be used as human judgement proxy with various different criteria.

Why this technique works?

“It’s easier to critique than to create“

Source : evidentlyai.com

Notes : this is just one of many testing technique that can be implemented



USER INPUT

LLM APPLICATION

UI

PREPROCESSING

PROMPT MANAGEMENT

BACKEND

LLM MODEL

LLM RESPONSE

LLM JUDGE

WHAT IS HALLUCINATION'S SCORE?

THRESHOLD

SCORE

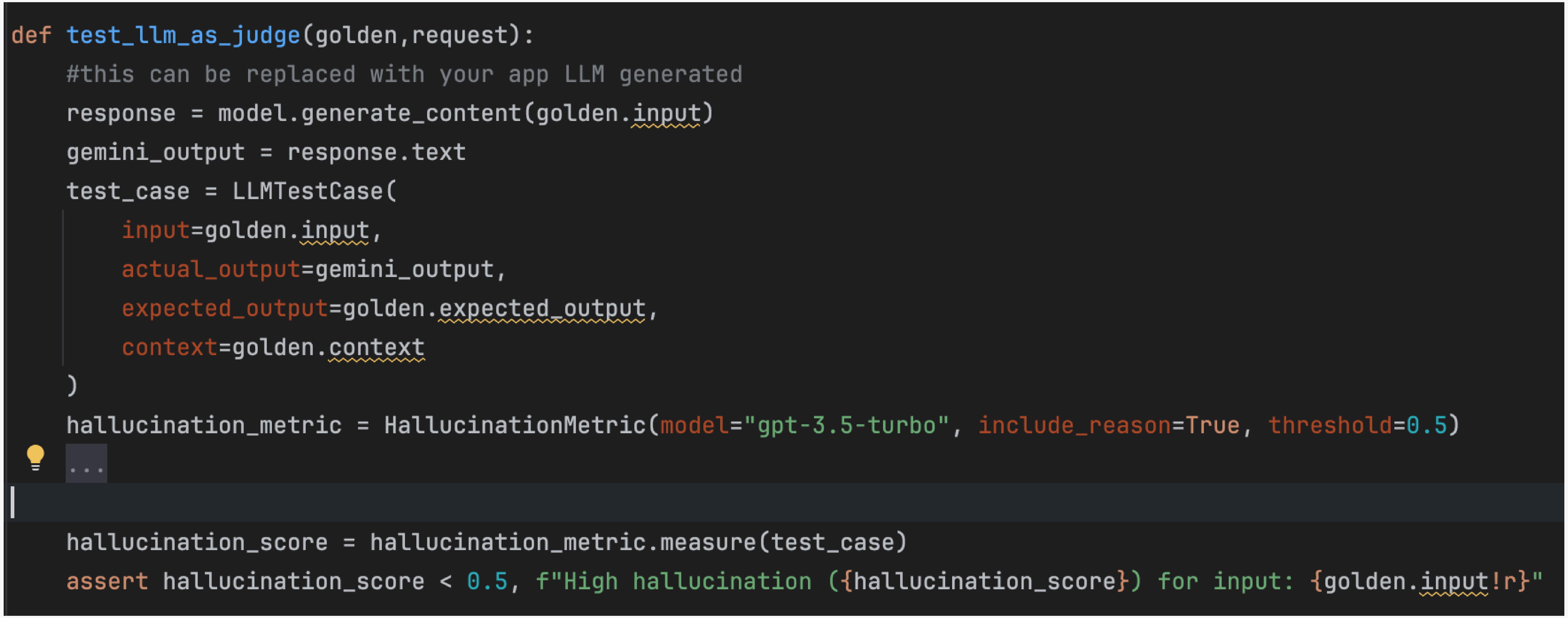
REASON

SIMPLIFIED ARCHITECTURE OF LLM BASED APP COMPONENT

INTEGRATE WITH BROWSERSTACK TEST REPORTING & ANALYTICS

KEY STRATEGIES FOR TESTING LLM BASED APPLICATION

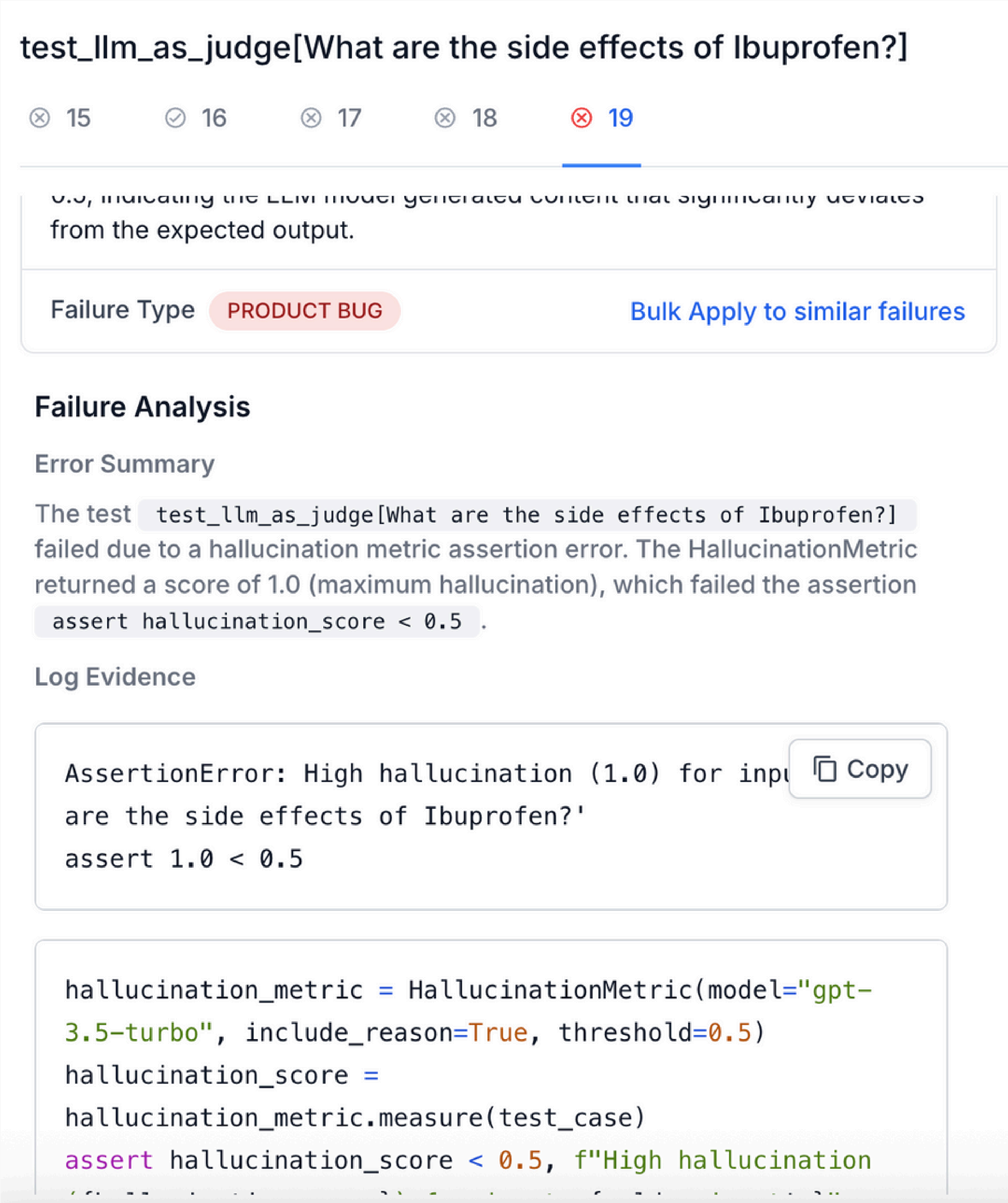
1. Balance combination between automated evaluation, observability, and human feedback
2. Budget conscious
3. Align business use case with testing strategy : No one-size-fit-all testing technique



```python
def test_llm_as_judge(golden,request):
    #this can be replaced with your app LLM generated
    response = model.generate_content(golden.input)
    gemini_output = response.text
    test_case = LLMTestCase(
        input=golden.input,
        actual_output=gemini_output,
        expected_output=golden.expected_output,
        context=golden.context
    )
    hallucination_metric = HallucinationMetric(model="gpt-3.5-turbo", include_reason=True, threshold=0.5)
    ...

    hallucination_score = hallucination_metric.measure(test_case)
    assert hallucination_score < 0.5, f"High hallucination ({hallucination_score}) for input: {golden.input!r}"
```

SIMPLE CODE TO TEST USING DEEPEVAL



test_llm_as_judge[What are the side effects of Ibuprofen?]

0.5, indicating the LLM model generated content that significantly deviates from the expected output.

Failure Type PRODUCT BUG

Failure Analysis

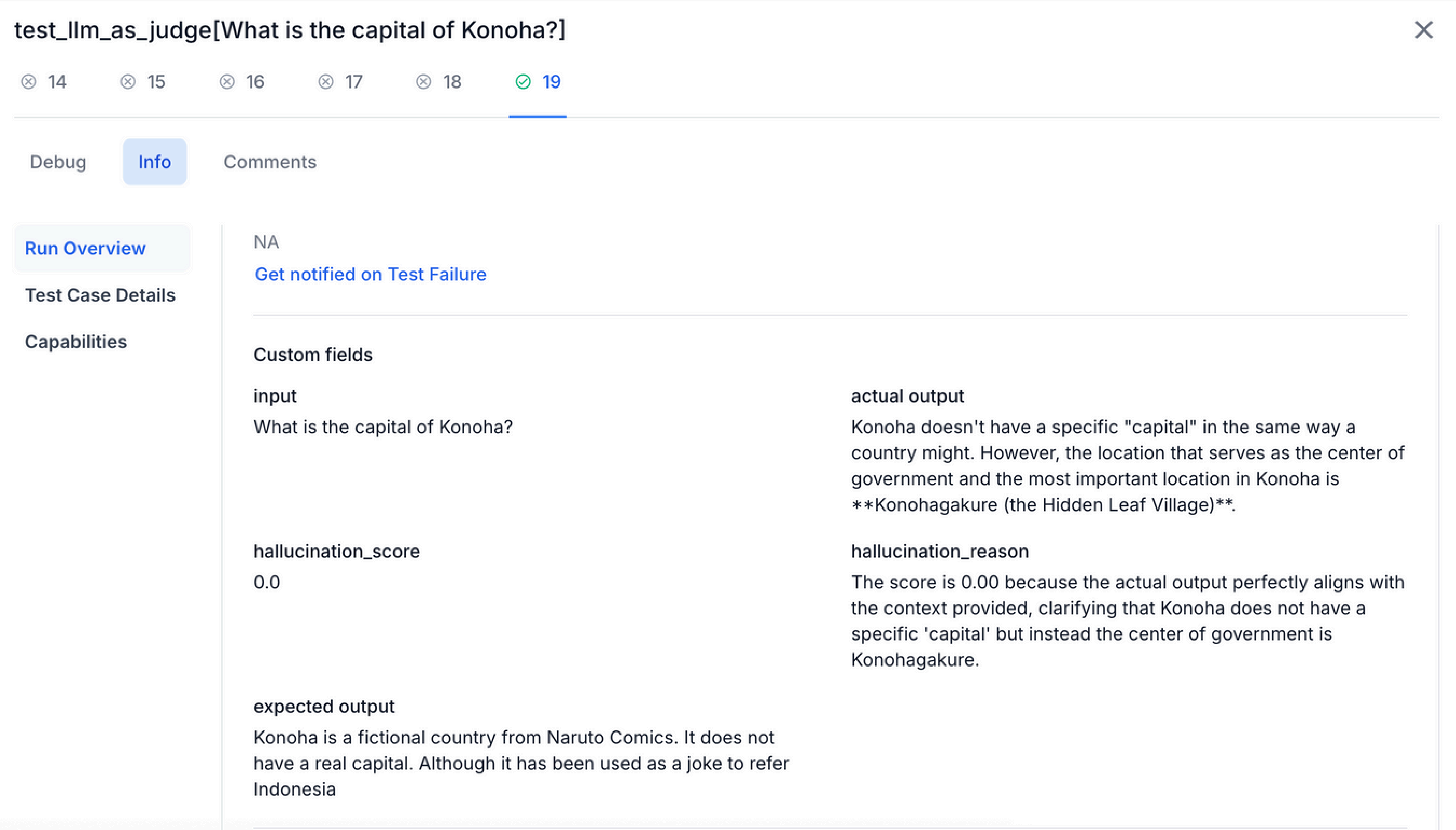
Error Summary

The test `test_llm_as_judge[What are the side effects of Ibuprofen?]` failed due to a hallucination metric assertion error. The HallucinationMetric returned a score of 1.0 (maximum hallucination), which failed the assertion `assert hallucination_score < 0.5`.

Log Evidence

```
AssertionError: High hallucination (1.0) for inpu
are the side effects of Ibuprofen?'
assert 1.0 < 0.5
```

```
hallucination_metric = HallucinationMetric(model="gpt-
3.5-turbo", include_reason=True, threshold=0.5)
hallucination_score =
hallucination_metric.measure(test_case)
assert hallucination_score < 0.5, f"High hallucination
```



test_llm_as_judge[What is the capital of Konoha?]

NA

Get notified on Test Failure

Custom fields

input
What is the capital of Konoha?

actual output
Konoha doesn't have a specific "capital" in the same way a country might. However, the location that serves as the center of government and the most important location in Konoha is **Konohagakure (the Hidden Leaf Village)**.

hallucination_score
0.0

hallucination_reason
The score is 0.00 because the actual output perfectly aligns with the context provided, clarifying that Konoha does not have a specific 'capital' but instead the center of government is Konohagakure.

expected output
Konoha is a fictional country from Naruto Comics. It does not have a real capital. Although it has been used as a joke to refer Indonesia

ADD OBSERVABILITY COMPONENT USING TEST ANALYTICS BY BROWSERSTACK

Thank you!

✉ **alifia@aliftestconsulting.com**

🌐 **aliftestconsulting.com**

in https://www.linkedin.com/in/alifia-/

ATC
Alif Test Consulting