



The Evolution of GenAI Evaluation: Implementing and Challenges of LLM-as-a-Judge

Stanley Chou | Director of Security Engineer | 2025.04



GenAI Trend

Shifting from
cloud, edge to
anywhere

Perception,
reasoning to
adaptive



GenAI is
becoming the
modern IT
backbone

What if GenAI is insecure, unsafe, inaccurate, unfair, or unethical?

Discriminate
without regret

Abuse critical
systems

Jeopardize
individual rights
and freedom

Exceed human
controls



Rigorous evaluation is essential to prevent harmful outcomes and build trustworthiness

Problems



Costly

Irreproducible

Inconsistent

Adversarial
Attack

Solutions



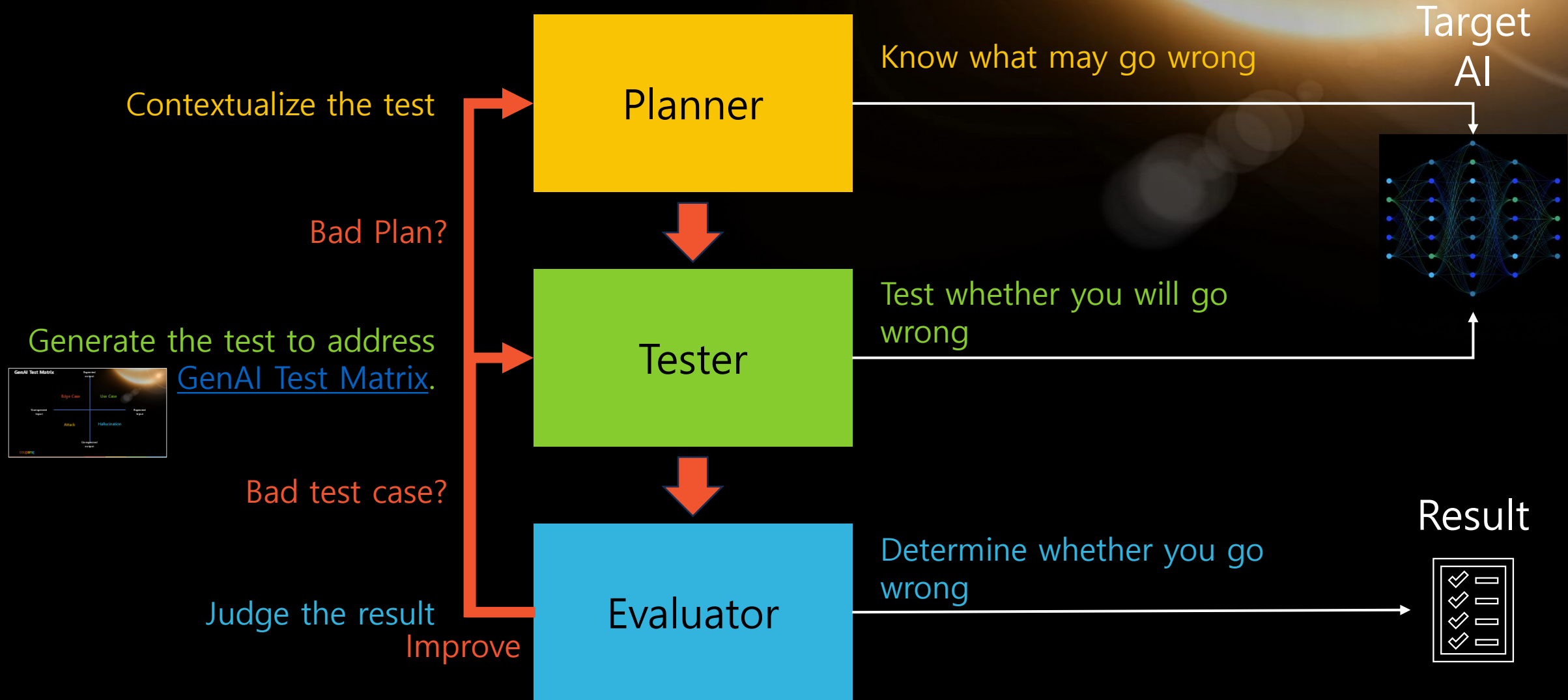
Automation

Monte Carlo
method

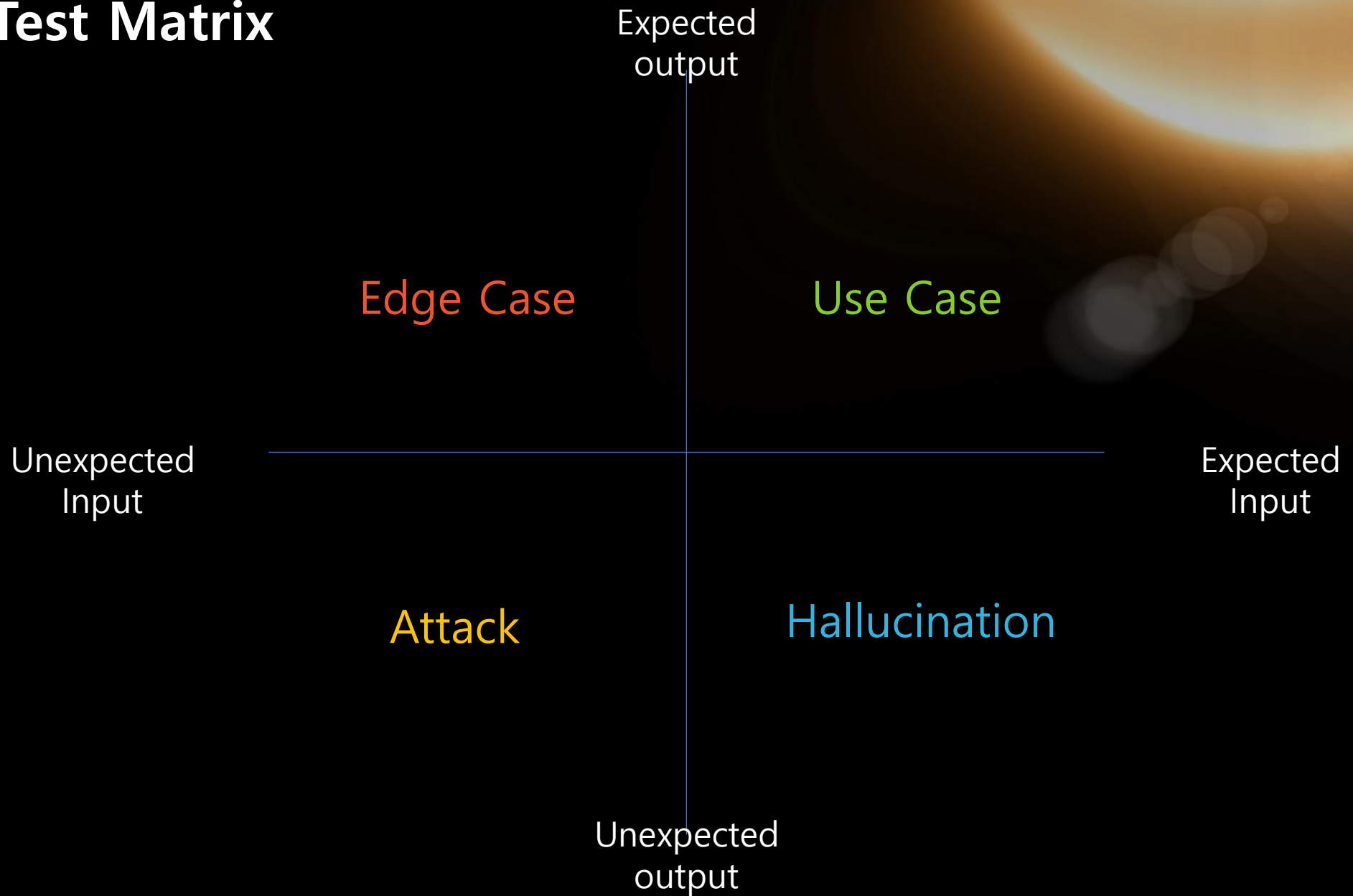
Single source of
baseline with
majority vote

Include Attack
Tactics,
Techniques, and
Procedures in
the test

Conceptual System



GenAI Test Matrix



KPIs

Planner

Did you plan the right test?

F1 score

→ Threats - Use Case Alignment

Tester

Was the test representative and enough?

Attack Successful Rate

→ Coverage
→ Depth

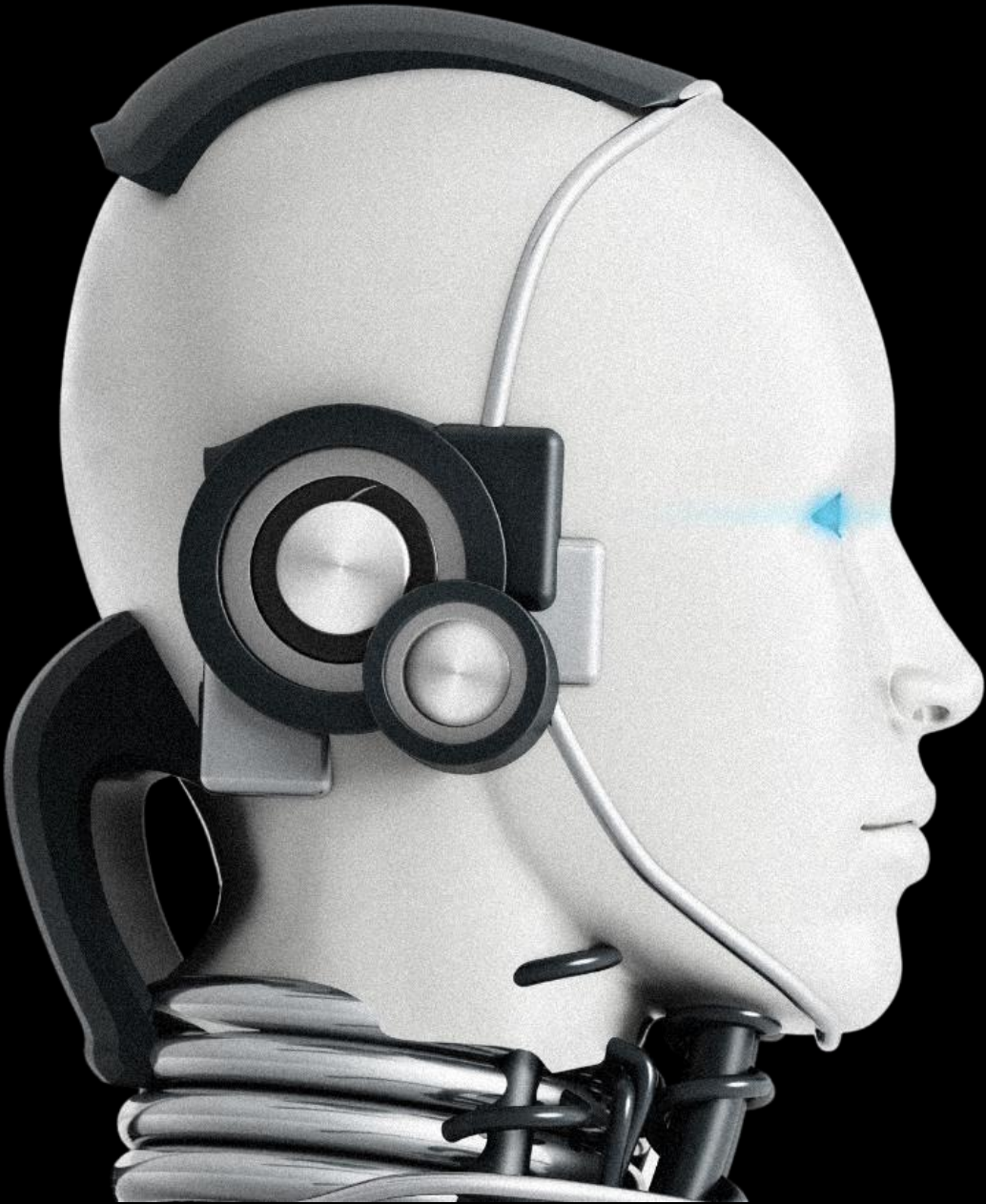
Evaluator

Did we evaluate the response correctly?

F1 score

→ Consistency
→ Perplexity
→ Toxicity
→ Truthfulness
→ etc.

Future Works



Adaptive baseline for various threats and use cases.

Extend the LM evaluation to multi-model.

Federated Learning to Improve the system across distributed environment.



Copyright © 2025 Coupang, Inc. All rights reserved. All Coupang trademarks, Coupang logos and service marks displayed herein are property of Coupang, Inc. and/or its affiliates (collectively, “Coupang”), registered in the U.S. and other countries. Any other company mentioned herein is merely for identification purposes. Coupang acknowledges that the company name may be a registered trademark of the company and recognizes that any such trademark is owned solely and exclusively by such company.

The information contained herein are based on the author’s own individual experience as an employee and are not representative of any views or opinions of Coupang. Coupang has not verified, and it makes no representation as to, the adequacy, fairness, accuracy, or completeness of any information contained herein.