AI Agent Evaluation

How you can create your own evaluation kit ?

Follow
Bhavishya Pandit

Repost

Introduction

AI agents promise transformative productivity but how do you know if they're actually working?

- 63% of AI agents fail to complete tasks optimally (Stanford, 2024)
- Dynamic workflows make traditional testing inadequate
- Poor tool selection or reasoning loops waste resources

Agents aren't just LLMs, they combine:

· Tool calls
· Multi-step reasoning
· Real-world actions

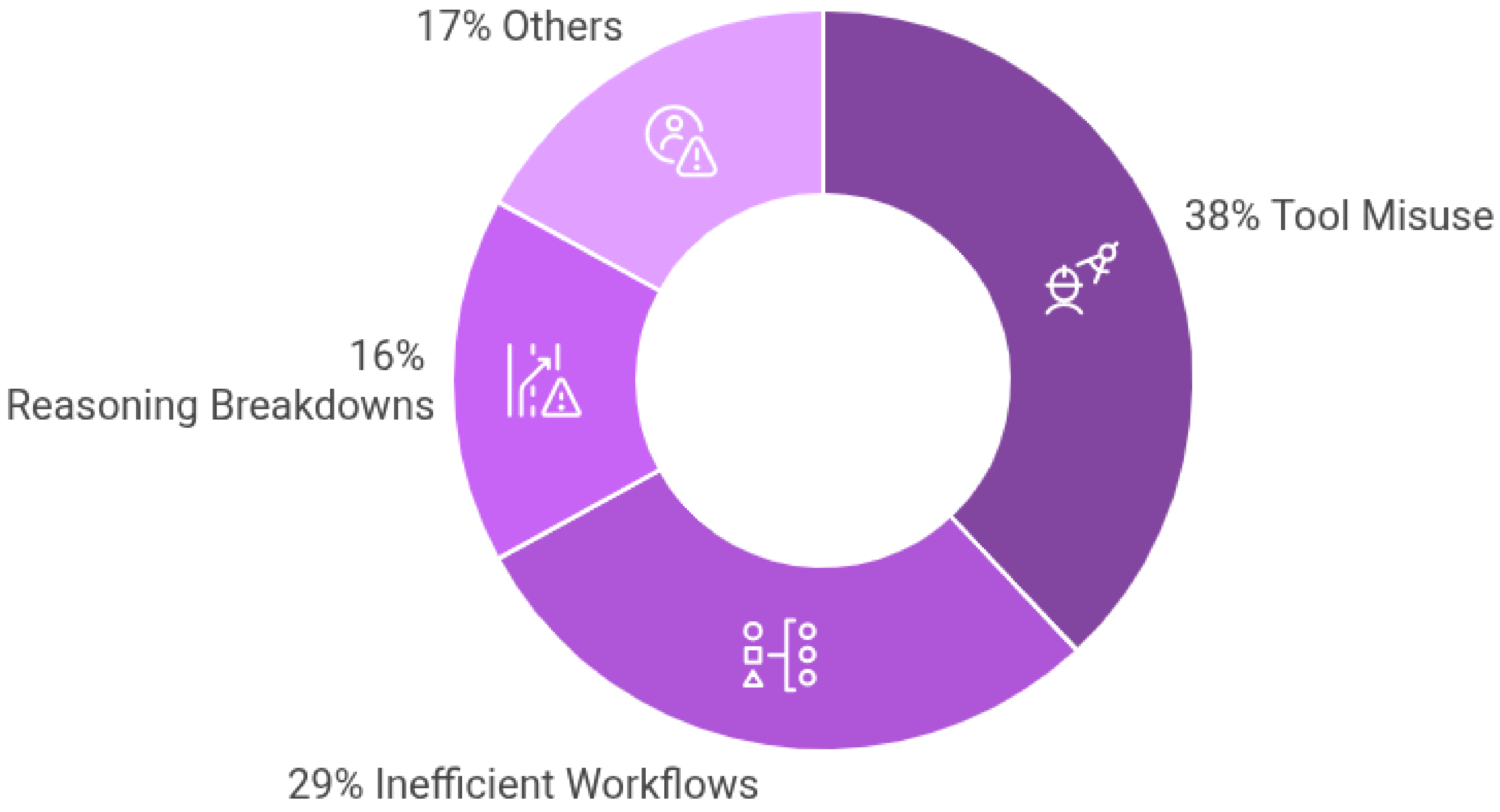
A LLM and an AI agent are completely different things:

Basic Chatbots	**AI Agents**
Single Responses	Multi-step workflows
No Tools	API/Tool integrations
Static	Dynamic decision-making

This makes AI agents evaluation complex. In this post we will learn about evaluating AI agents, the steps involved and everything in between.

Agent Failure Patterns

Before jumping into frameworks we need to understand common pain points orgs face with their agents. After evaluating 1200+ production agents, we see consistent pain points:



Each unresolved failure pattern costs enterprises:

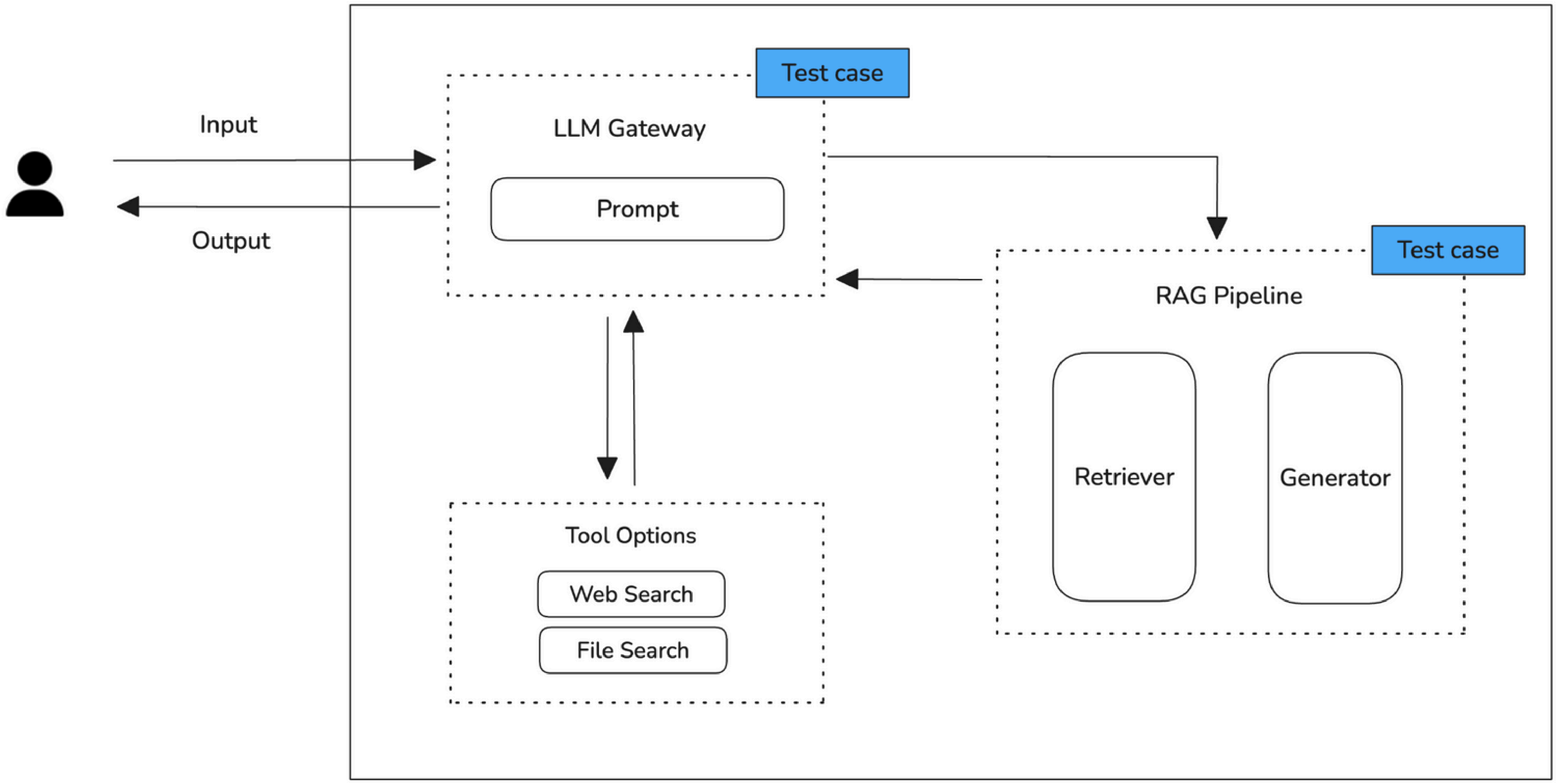
- $18-35 in wasted compute/resources per incident
- 22% higher developer maintenance hours
- 3.4x more user complaints

Considering the diversity of agents, it is not possible to implement one strategy for all. For Agent evaluation the process depends on the agent itself.

Component Level Evaluation

LLM agents are evaluated at two distinct levels:

- **End-to-end evaluation:** Treats the entire system as a black box, focusing on whether the overall task was completed successfully given a specific input.
- **Component-level evaluation:** Examines individual parts (like sub-agents, RAG pipelines, or API calls) to identify where failures or bottlenecks occur.

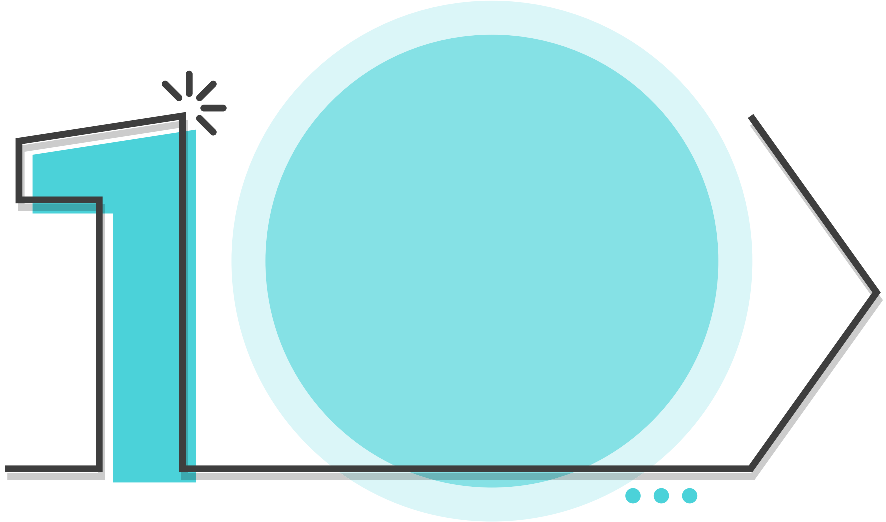


This layered approach helps diagnose both surface-level and deep-rooted issues in agent performance.

Lets talk about a standard 3-Level approach for agent evaluation.

3-Level Evaluation Framework

Evaluating AI agents requires inspecting what they deliver AND how they get there. A popular 3-level framework can help achieve that:

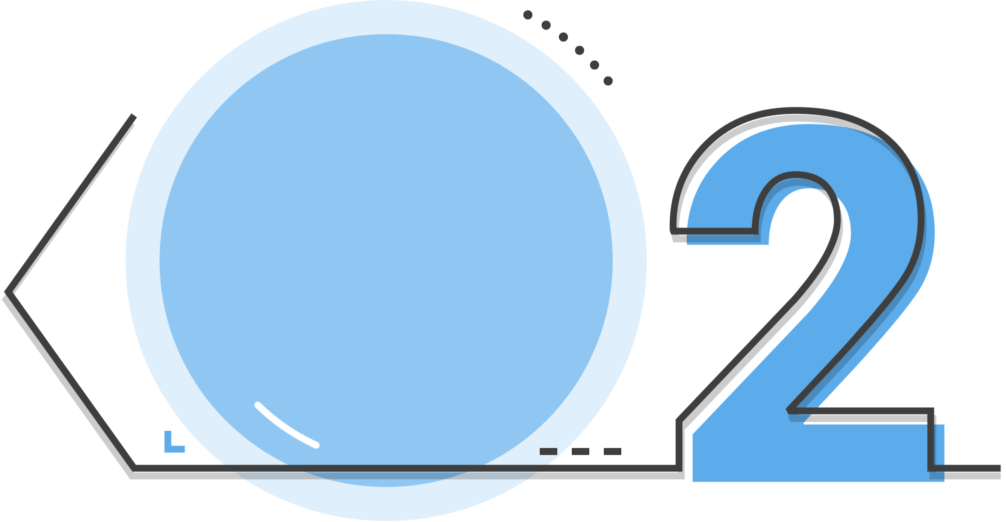


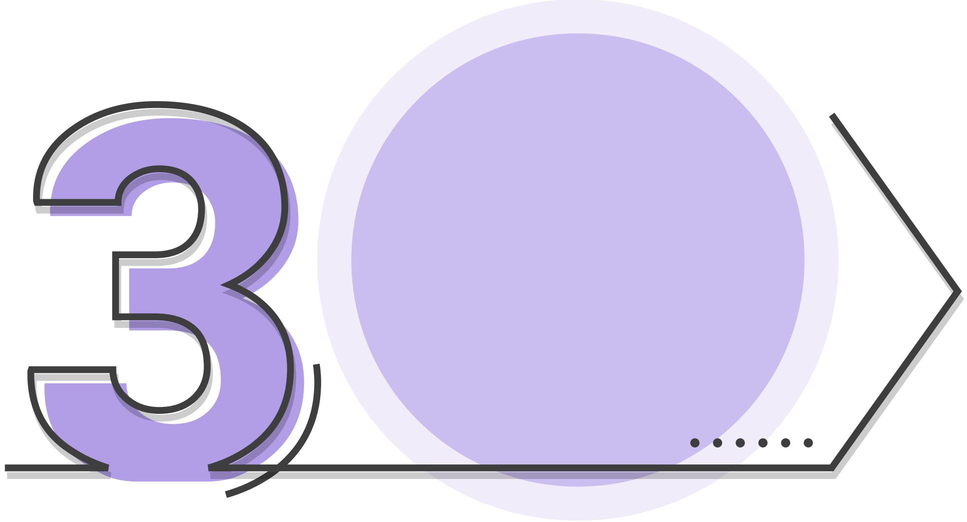
Output Accuracy

Tests if the final answer is correct

Step-by-Step Validation

1. **Verifies crtical decision points:**
2. **Intent classififcation accuracy**
3. **Tool selection logic**





Full Workflow Audit

Maps the agent's path against optimal benchmarks

This framework can be used on build evaluation strategies for specific AI agents. Lets learn how we can use this framework to build our own Agent evaluation kit

Build Your Own Evaluation Kit

Evaluating AI agents isn't one-size-fits-all. What works for a customer service bot might miss critical issues in a data analysis agent. Here's how to build evaluations that actually fit what you're building. Follow this 4 step process:

1. Agent’s Purpose

- Customer support? Focus on response accuracy
- Data analysis? Prioritize tool output validation
- Research tasks? Emphasize source quality checks

2. Metrics Selection

- Tool Correctness (for API-heavy agents)
- Task Completion (end-to-end validation)
- Reasoning Coherence (multi-step workflows)

3. Select tools

- DeepEval for pre-built metrics
- LangSmith for tracing workflows
- Golden datasets for baseline testing

4. Gradual Implementation

- Output validation
- Add step checks
- Full trajectory analysis

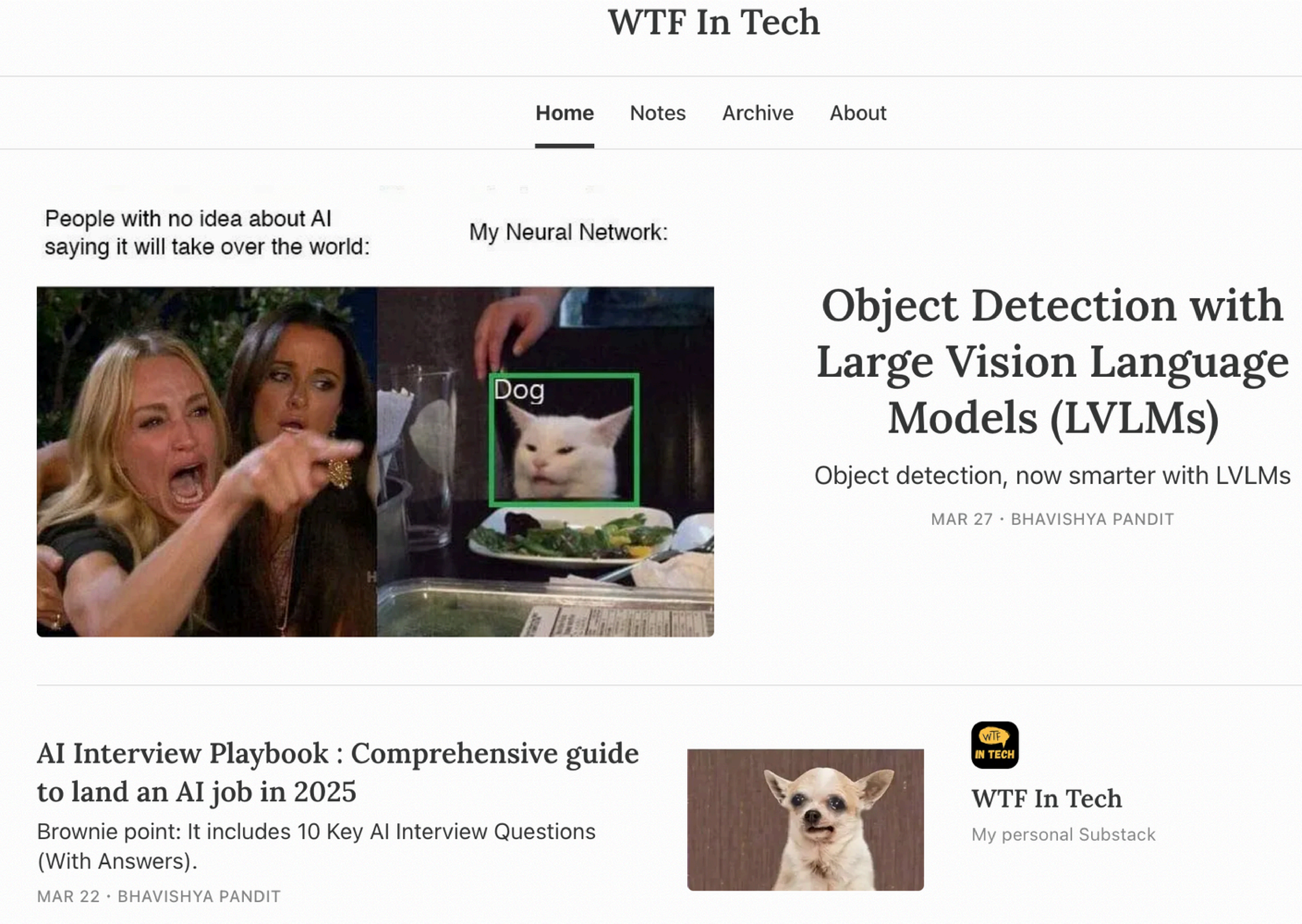
Evaluation process can be customized based on the agent, just build on the **3-level framework.** Hope you learned something new . ✌️

Stay Ahead with Our Tech Newsletter! 🚀

👉 Join 1.1k+ leaders and professionals to stay ahead in GenAI!
🔗 https://bhavishyapandit9.substack.com/

Join our newsletter for:

- Step-by-step guides to mastering complex topics
- Industry trends & innovations delivered straight to your inbox
- Actionable tips to enhance your skills and stay competitive
- Insights on cutting-edge AI & software development


Home
Notes
Archive
About
People with no idea about AI saying it will take over the world:
My Neural Network:
Dog

WTF In Tech

Object Detection with Large Vision Language Models (LVLMs)

Object detection, now smarter with LVLMs

MAR 27 · BHAVISHYA PANDIT

AI Interview Playbook : Comprehensive guide to land an AI job in 2025

Brownie point: It includes 10 Key AI Interview Questions (With Answers).

MAR 22 · BHAVISHYA PANDIT

WTF In Tech

My personal Substack

💡 Whether you're a developer, researcher, or tech enthusiast, this newsletter is your shortcut to staying informed and ahead of the curve.

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST