

How to Test an AI Agent Before You Trust It With Real Work

A simple guide for Singapore businesses, written in plain English. No jargon required.

Five stops on this checklist

01 Two testing stages

Before launch, and every week after

02 Five things every test must check

Not one overall "good or bad" score

03 Building a fair test set

Easy cases, hard cases, attack cases

04 The "AI judge"

How a second AI scores the first one

05 Singapore rules to test against

IMDA, MAS, and PDPA

Test it like a new hire

1

Before it starts work. Run it against a stack of practice cases with known right answers, checked by a real expert.

2

While it's working. Pull a random sample of real conversations every week. Real customers find edge cases no test writer thought of.

5

things every agent test must check

Score each one separately. A slow answer and a leaked customer record are not the same kind of mistake.

01 Does it finish the job?

At least 95 out of 100, in the fewest steps possible

02 Does it use its tools correctly?

Right format every time, wasted calls under 2%

03 Does it stick to the facts?

Matches its source at least 92 out of 100

04 Does it stay safe?

Zero serious safety mistakes allowed. Not one.

05 Does it know when to ask for help?

Catches at least 99 out of 100 cases that need a human

Three kinds of cases

60%

Everyday cases. Real requests, checked and approved by someone who knows the job.

25%

Tricky cases. Missing info, expired passwords, slow systems, confusing requests.

15%

Attack cases. People trying to trick the agent into breaking its own rules.

A second AI grades the first

The judge reads the full conversation, writes out its reasoning step by step, then gives a score from 1 to 5. A score of 4 or 5 is a pass.

Its "creativity" is turned all the way down, so it gives the same score to the same conversation, every single time it is asked.

Run the tests every time you change:

- 01** The instructions given to the agent
- 02** The AI model itself, or its version
- 03** The tools or systems it can call
- 04** The documents it searches for facts

Don't build this from scratch

[IMDA · AI Verify](#)[MAS · FEAT Principles](#)[PDPA](#)

Singapore already has real standards to test against. Build your checklist around them, not from a generic template.

Test it like you mean it.

Before an agent touches money, health records, or personal data, it should already have passed every check on this list.



Terence Kok · AI & Innovation Strategist

terencekok.com

[SAVE](#) · [COMMENT](#) · [FOLLOW](#)