Lethal trifecta test (10 minutes)

Three questions to ask about every agent, and how to break one leg.

Developer **QA** **Product**

ASK THREE QUESTIONS

Leg	Question	Support examples
Private data	What sensitive data can this agent reach?	Customer records, order history, internal notes, credentials
Untrusted content	What text can an outsider put in front of it?	Tickets, forms, emails, chat messages, attachments, web pages
Outbound path	How can anything leave?	Replies, emails, API calls, notes, links or images that load from a URL

DECIDE

- All three present: **redesign** before launch
- Two present: add controls and test the third
- One present: normal controls

BREAK ONE LEG

- **Remove the outbound path** for this task, or limit replies to the verified requester with links and images stripped
- **Remove access to private data** for the step that reads untrusted text
- **Keep untrusted text away** from the tool-using agent: a reader agent summarizes it into a short structure
- **Require human approval** for any send or write

NOT ENOUGH ON ITS OWN

- A prompt that says to ignore hidden instructions
- A detection product that catches most attacks: a data leak needs only one that gets through
- Trusting that the model will notice

KEEP EVIDENCE

- Record the three answers in the design spec
- Log tool calls
- Add injection cases to the evaluation set and re-run them on every change