SAFECHAT

BUILDING AI-POWERED SAFETY SYSTEMS AT SCALE IN A REAL-TIME MARKETPLACE

Bruna Pereira • Software Engineer

About Me

- **Based in São Paulo 🇧🇷**
- **Lead Trust & Safety engineering team at DoorDash**
- **Background in Fintech**

DoorDash

Marketplace, quick interactions

Consumers

Dashers

Merchants

Two metrics matter equally:

PEOPLE ARE SAFE

PEOPLE FEEL SAFE

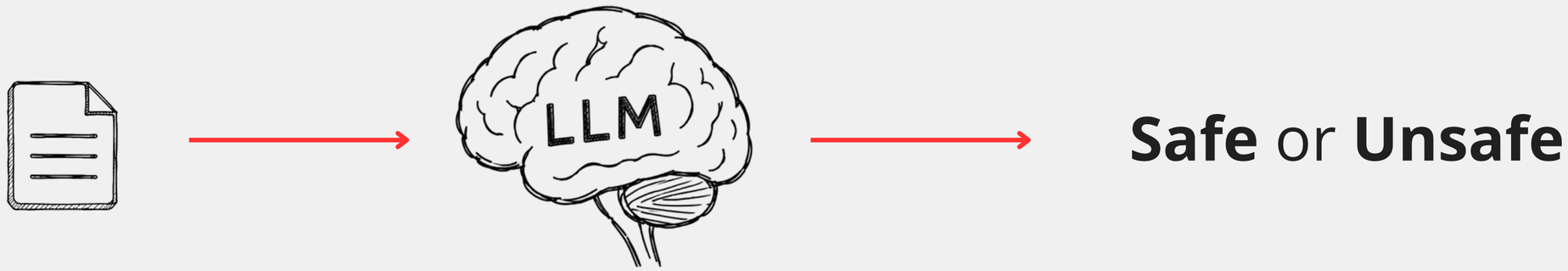
The Scale

4M+

400K+

200K+

The naive answer



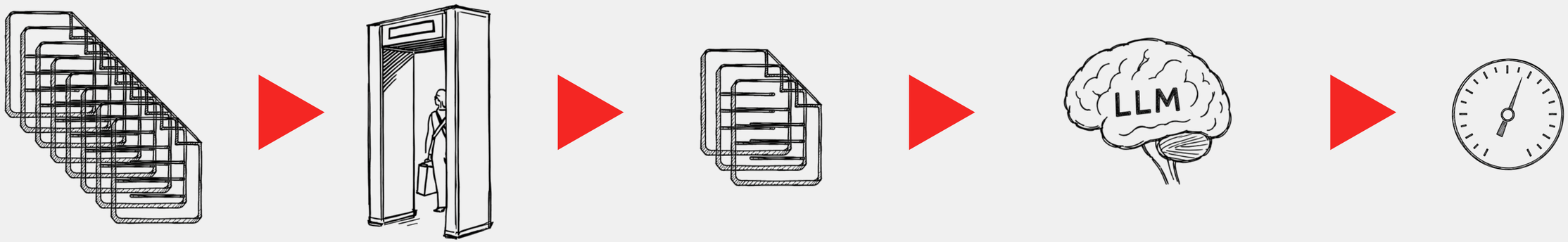
Listen First

Understand your data.

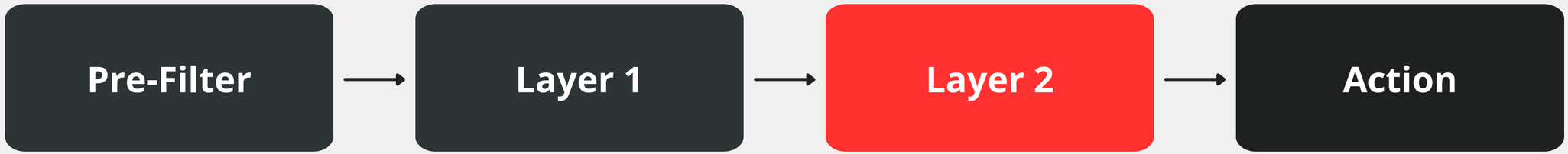
A small filter

Fast. Cheap. One job: clear the obvious safe.

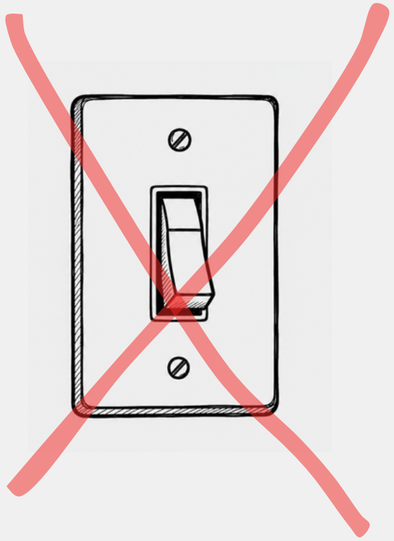
LLM, only when you have to



The Two-Layer Pipeline



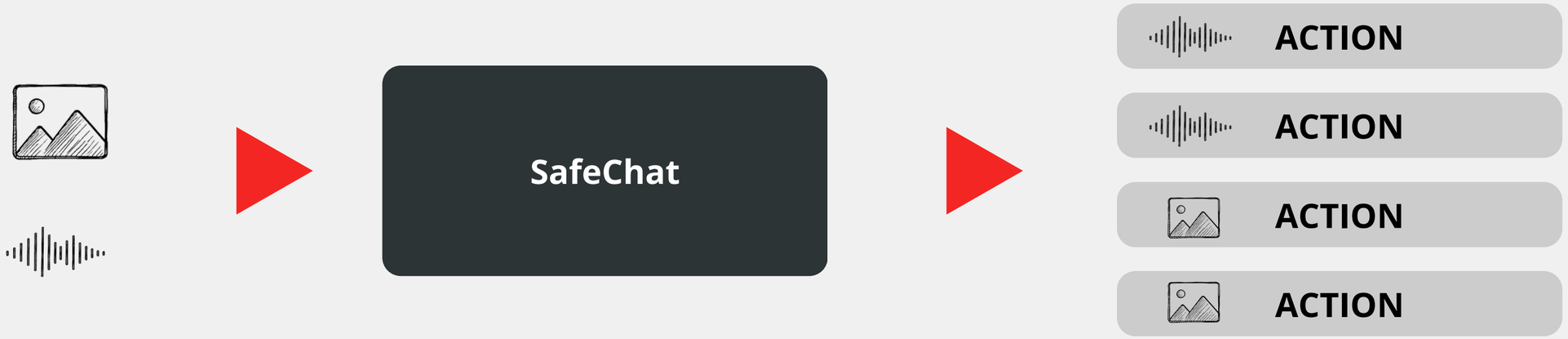
Severity beats binary



Action is proportional

Severity		Action
low	▮ ▯ ▯ ▯	censor
mid	▮ ▮ ▯ ▯	block
high	▮ ▮ ▮ ▯	offer cancel
very high	▮ ▮ ▮ ▮	auto cancel

Voice & Image

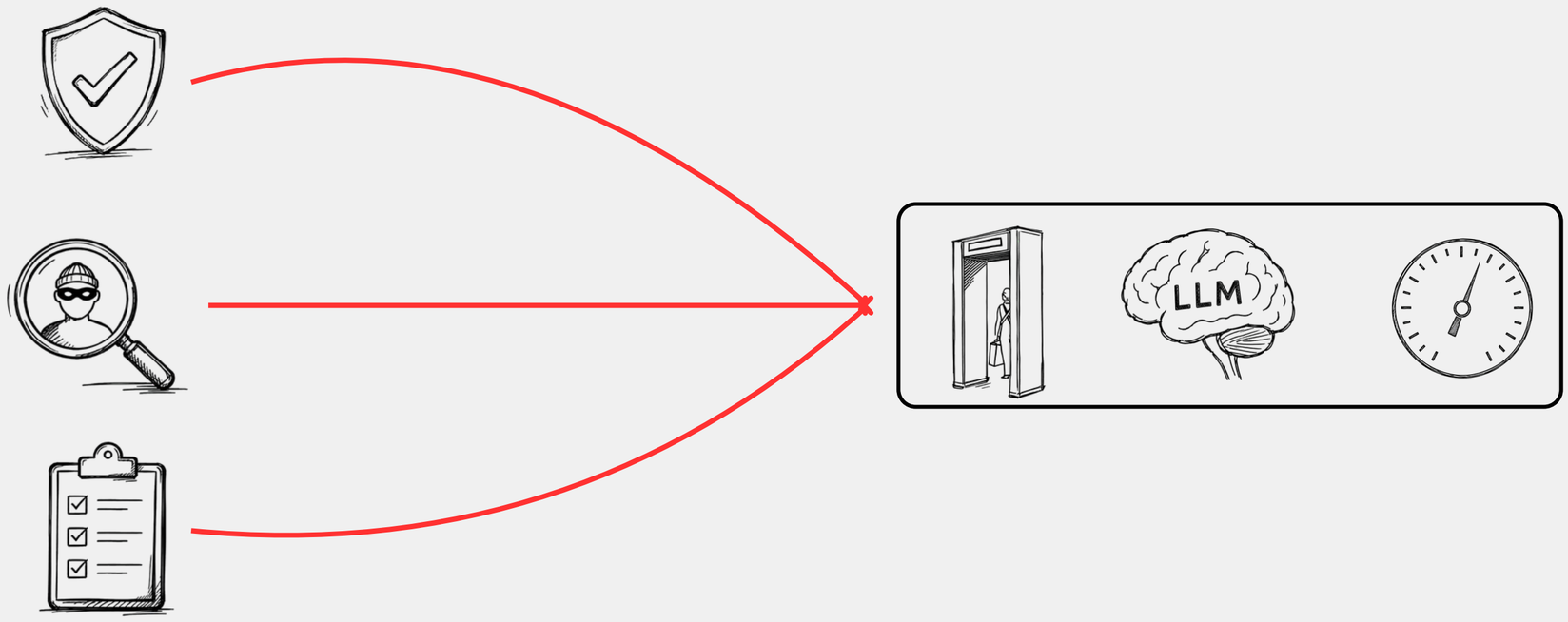


The Result

-50% **incidents**
due to verbal abuse

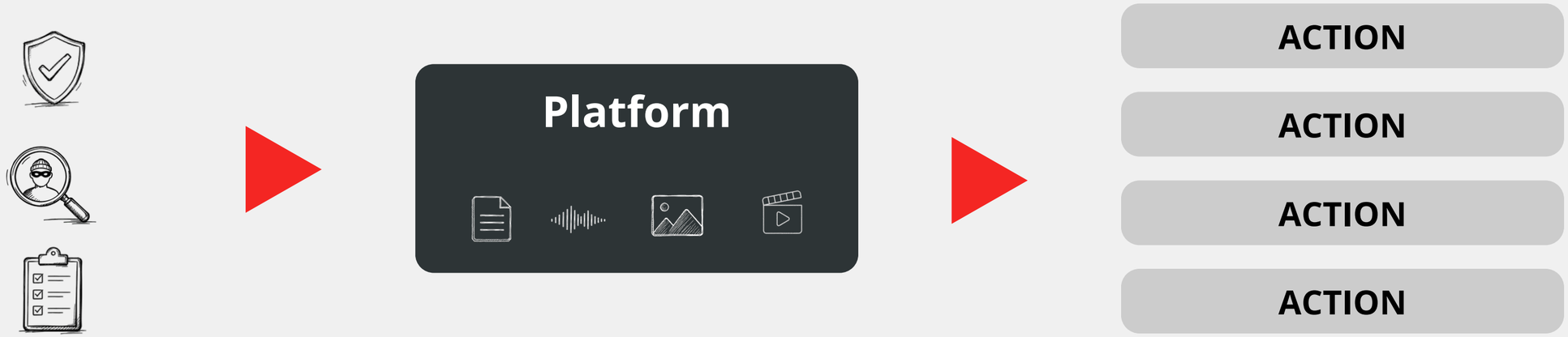

so we threw
it away

The pattern is the asset

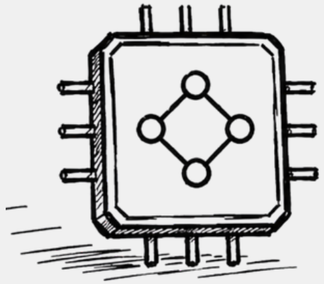


Content-agnostic moderation platform

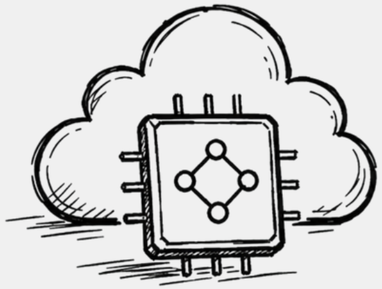
Teams bring the meaning. Platform brings orchestration.



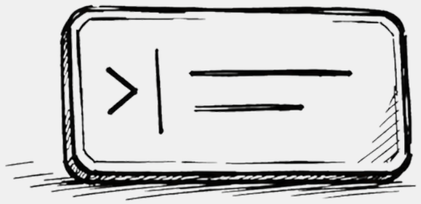
Three kinds of model



Internal model

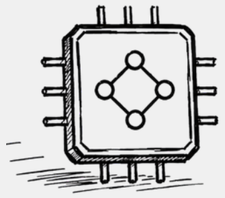


External model



External prompt

Internal models



1.

Trained and hosted internally.

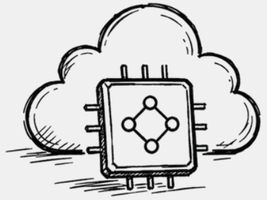
2.

Cheap & fast

3.

Pre-defined input and output schema

External models



1.
Commodity tasks

2.
Served through vendor APIs

3.
Available for any agent to use

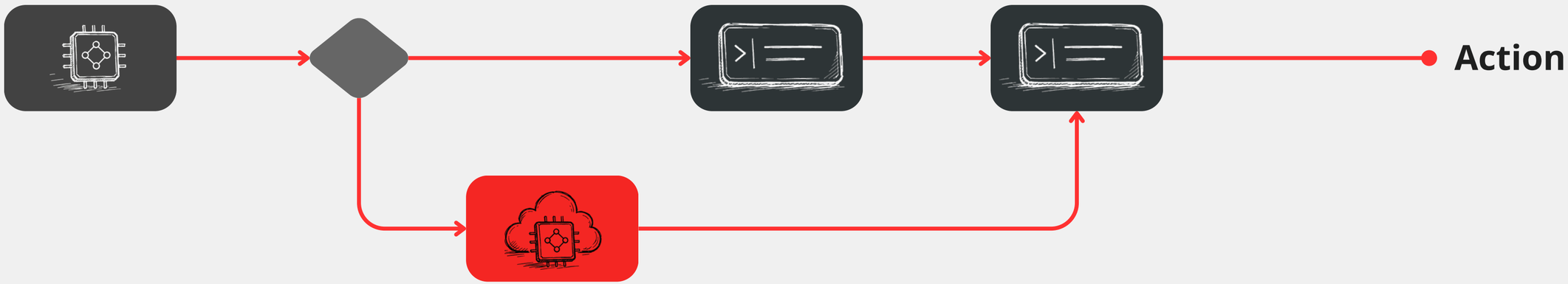
External Prompts

1.
Uses the LLM gateway

2.
Dynamic input and output schemas

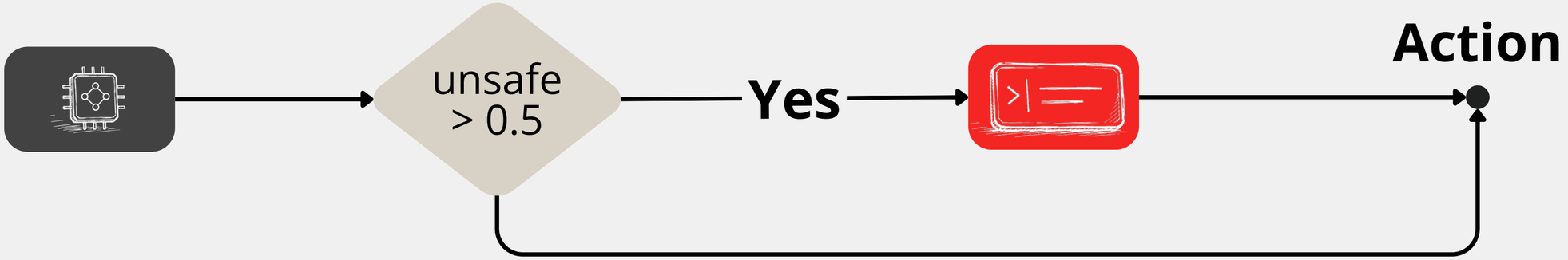
3.
Declarative fallback and retry

Composing into agents



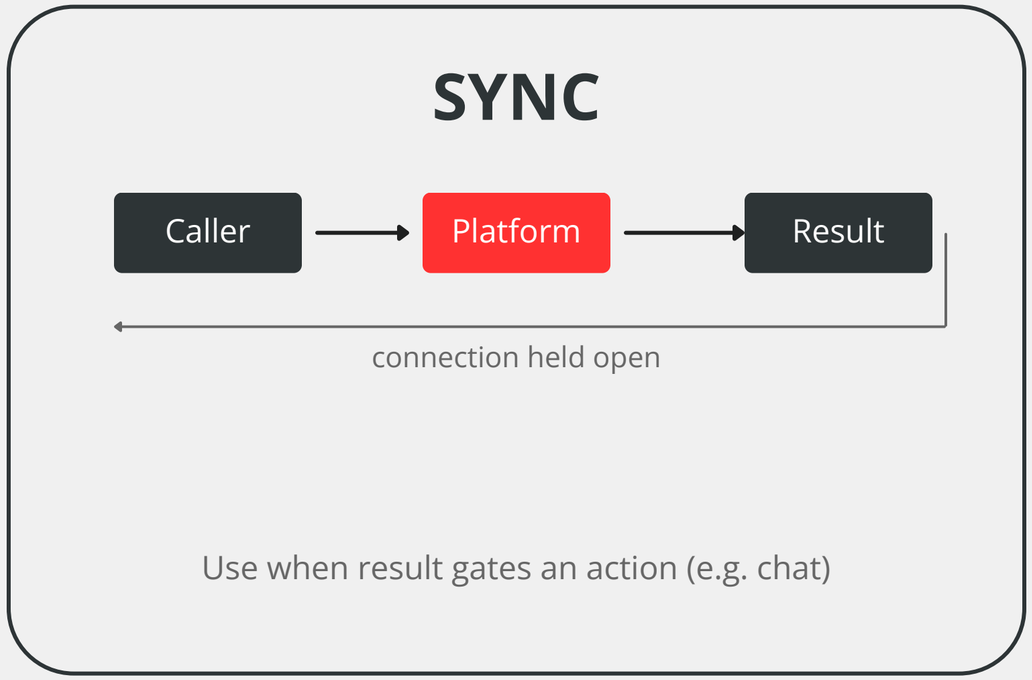
Conditional flows

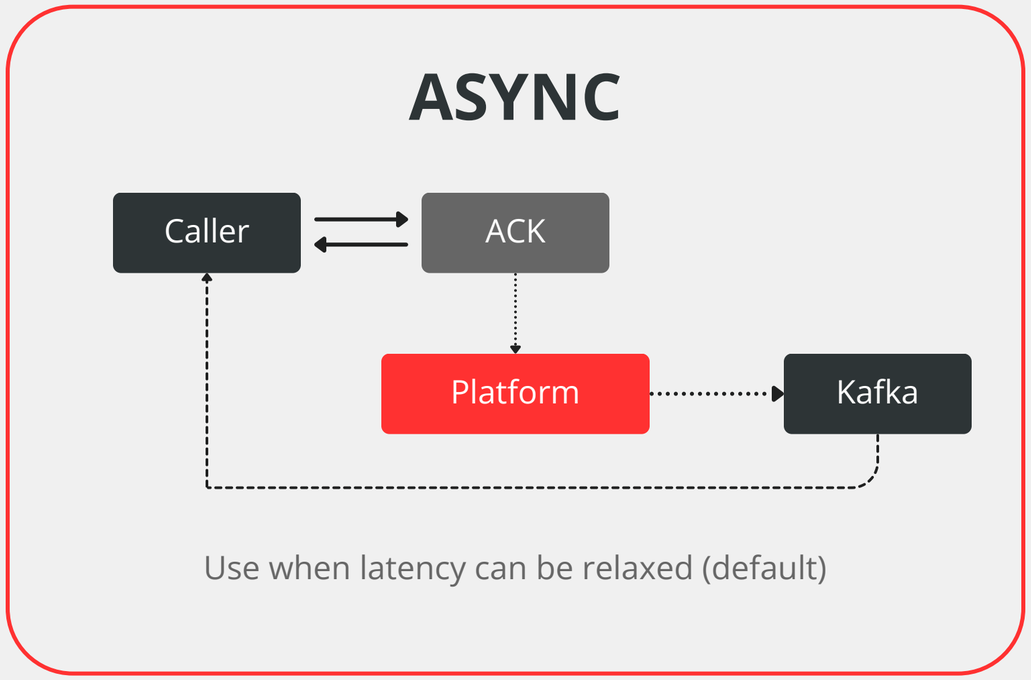
Express conditions between steps



Wait, or fire and forget

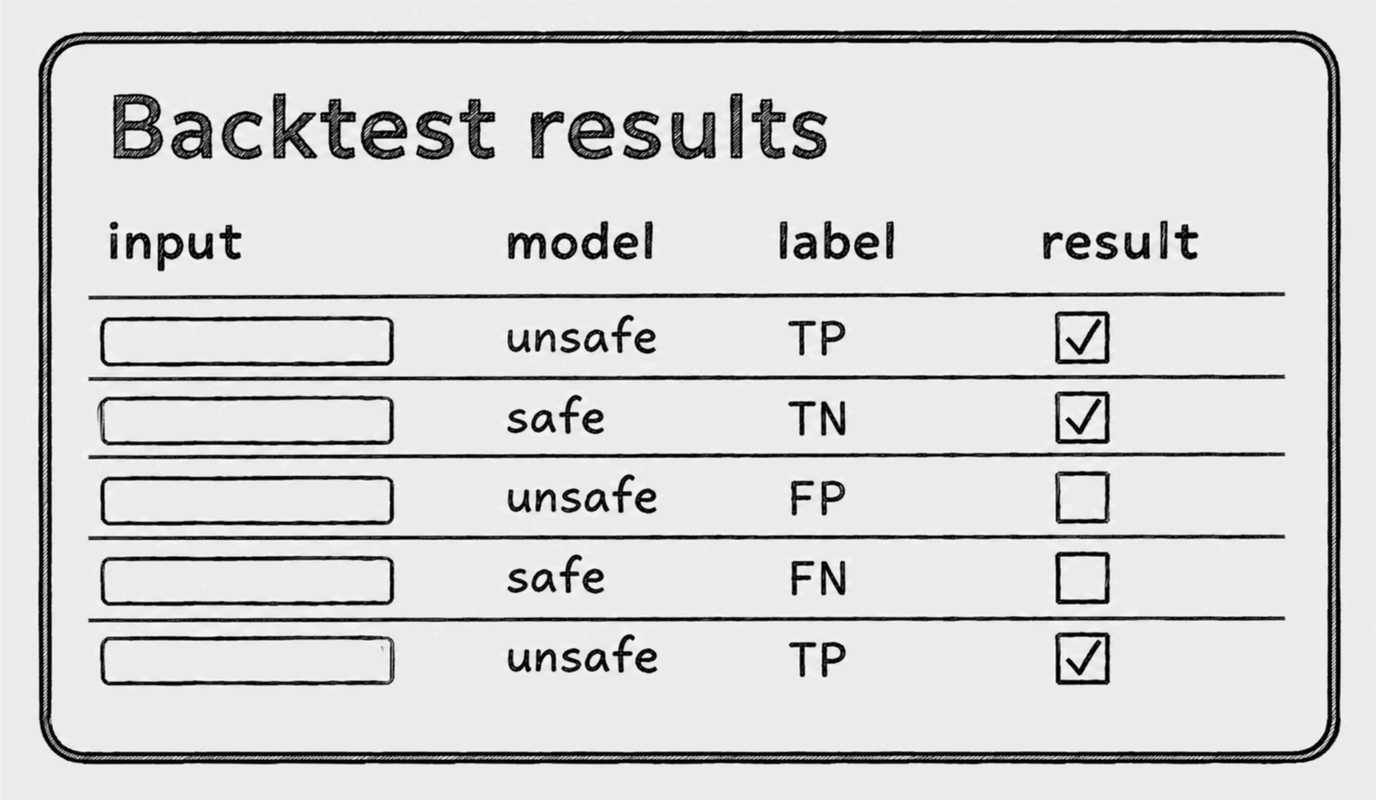
Two modes for every agent: hold the line, or let it go.





Test Before Trust

Backtest prompts and models on historical data before launch.



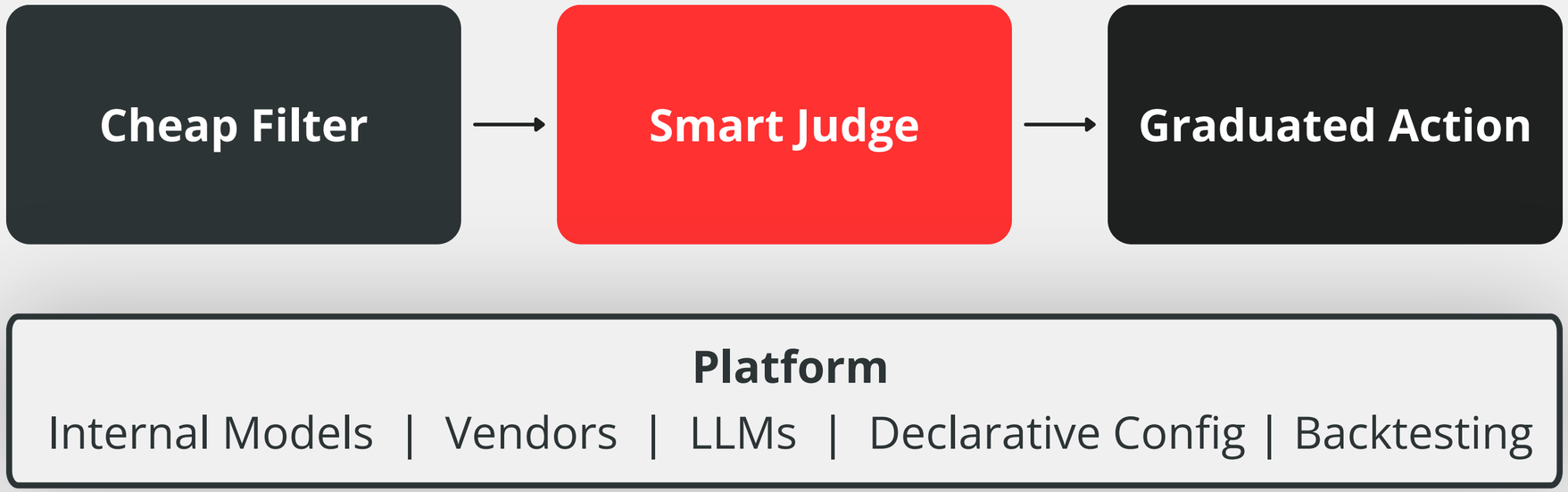
Three Lessons

1.
Put a cheap model in front of the LLM

2.
Ask for severity, not labels

3.
Know when to throw a system away

The Whole Talk in One Shape



Thank You!

Questions?

Bruna Pereira

Software Engineer, Trust & Safety Lead

DoorDash, São Paulo

bruna.pereira@doordash.com