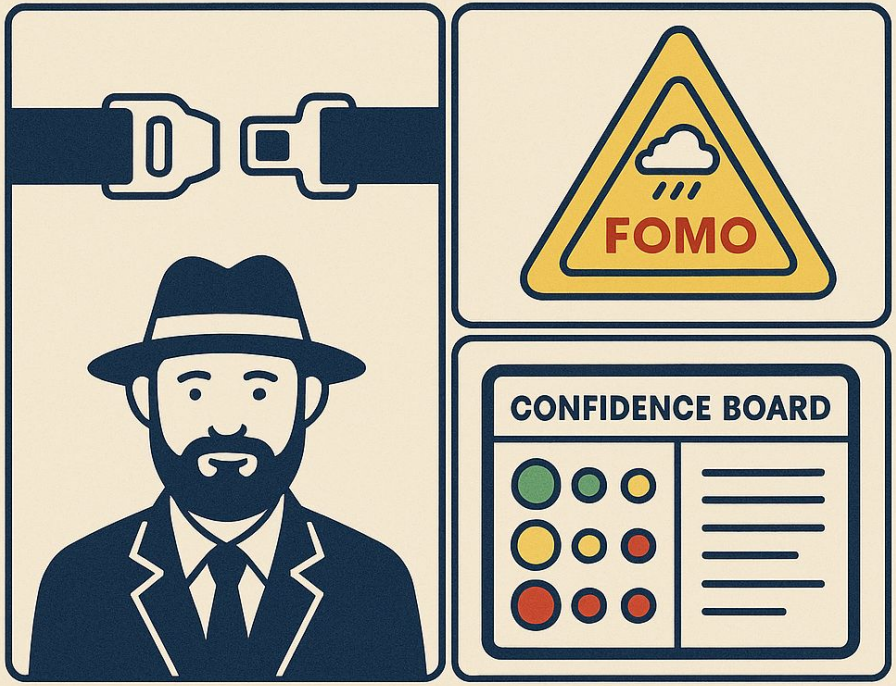
Martin Hynie
LaunchPT.com

PLEASE FASTEN YOUR GUARDRAILS



CONSEQUENCES > VIBES

Confidence isn't a feeling;
it's evidence we can show."

FOMO GRAVITY

HYPE picks directions; evidence sets speed.

Build lenses… Not laws or limits.
Use them to notice “more”.

ACT I: WHY THINGS FAIL NOW

INTERACTION

DO NOT INCORRECTLY INTERACT

INFRASTRUCTURE

BROKEN SYSTEMS MAY OCCUR

INSTITUTION

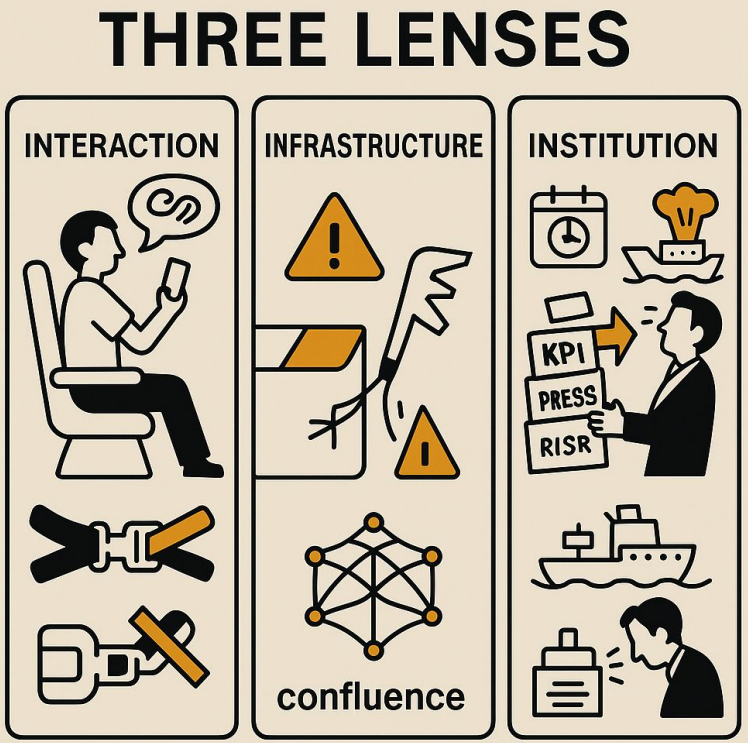
AVOID OVERWHELMING PRESSURE

Interaction: prompt/context drift, retrieval variance, misalignment.

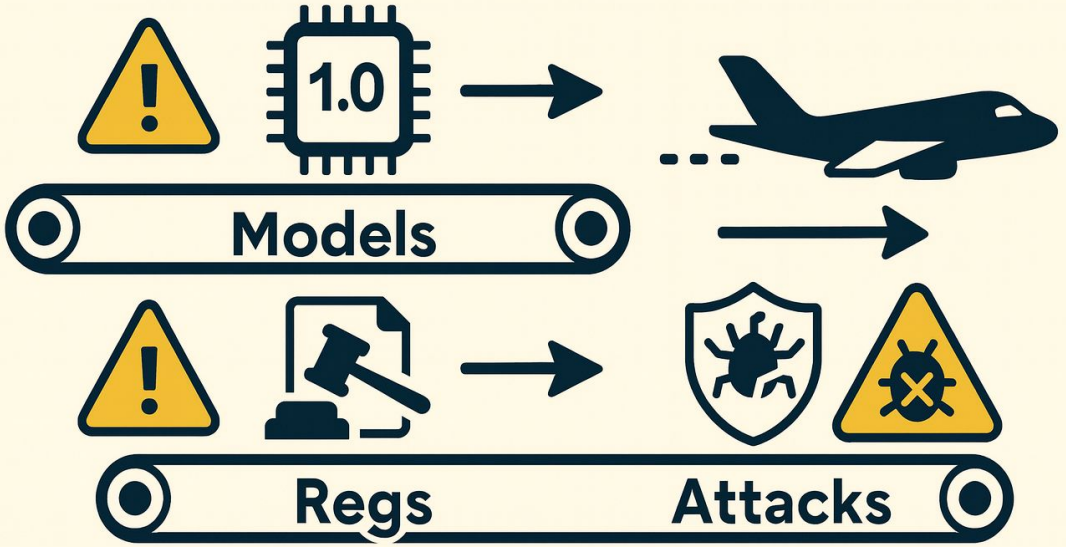
Infrastructure: fuzzy data lineage, stale evals (constant risk), degraded guardrails, missing traces.

Institution: incentives, process debt, systemic shipping pressure, fomo

Complexity 101: outcomes emerge from interactions, not single rca

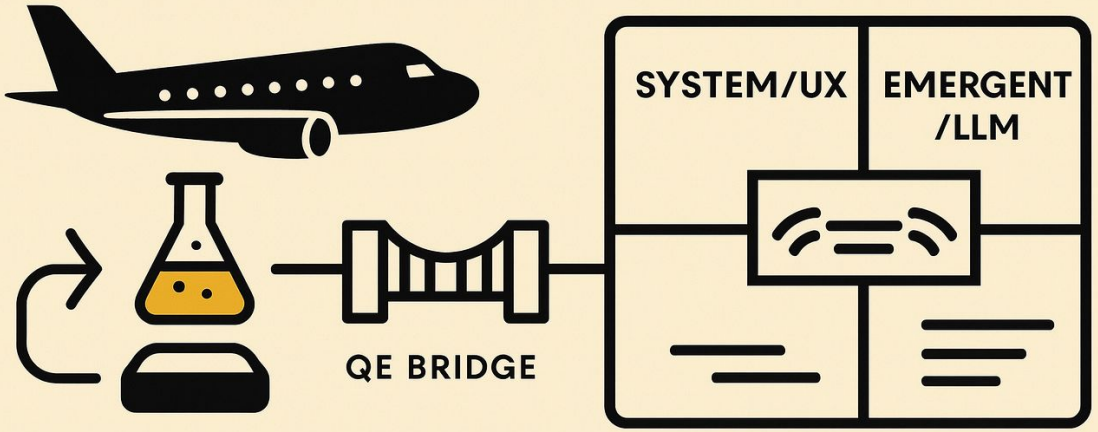


NON-STATIONARY WORLD

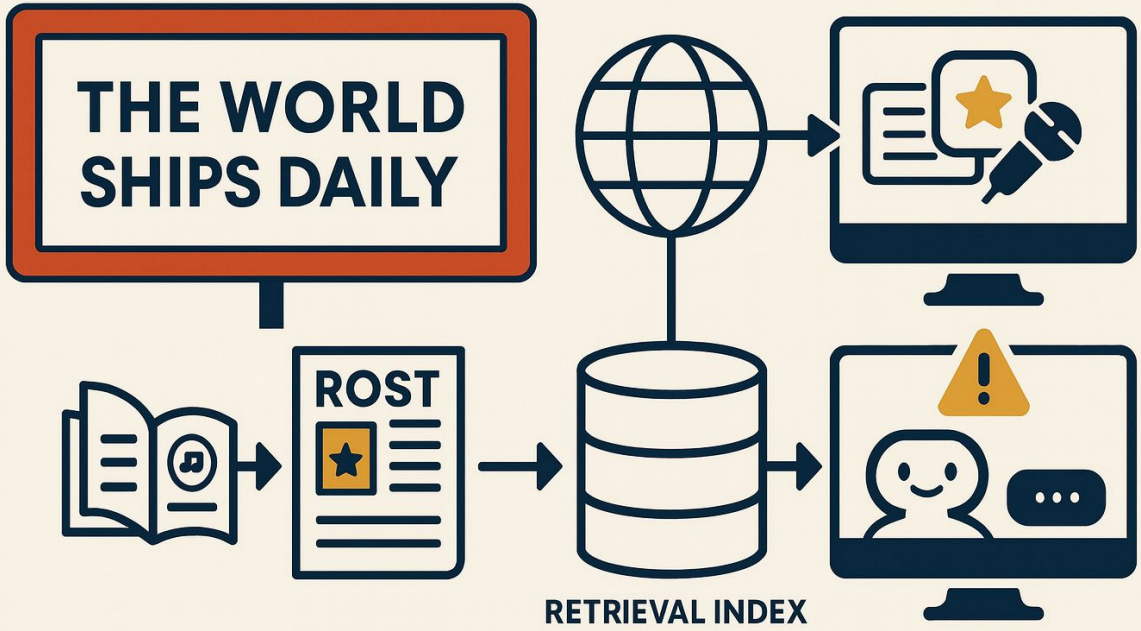


Fixing yesterday's cause ≠ shrinking tomorrow's risk surface.

ACT II—BUILD TO EXPLORE
(AND TEST ON TWO PLANES)



REALITY DRIFT

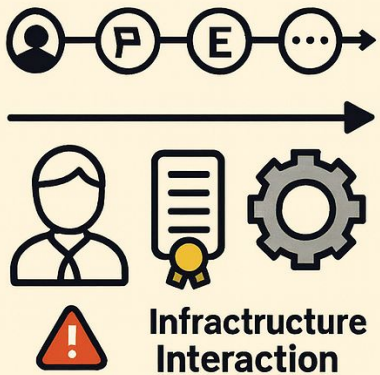


"YOUR MODEL DIDN'T CHANGE. THE WORLD DID."

LEARNING REVIEWS > BLAME

"If an incident doesn't create new prcbes, we paid tuition and learned nothing."

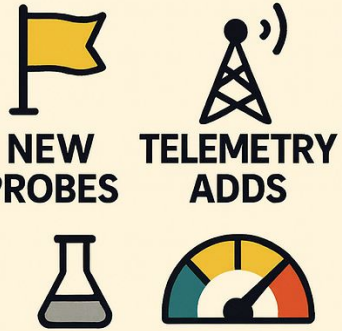
CONTRIBUTING-FACTOR TIMELINE



CONTROLS CATALOG

- PRESENT
- ABSENT
- DEGRADED

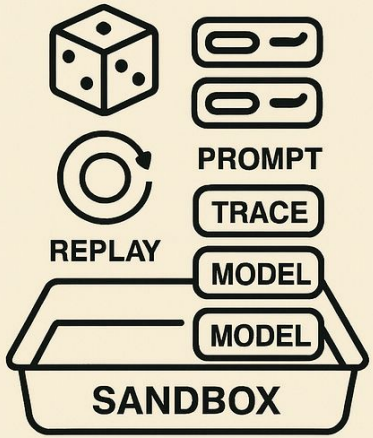
OUTPUTS



BUILD TO EXPLORE

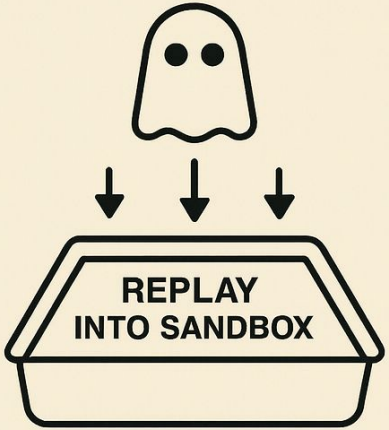
Don’t just test the system–buiid little systems to explore it.

SANDBOX LAB

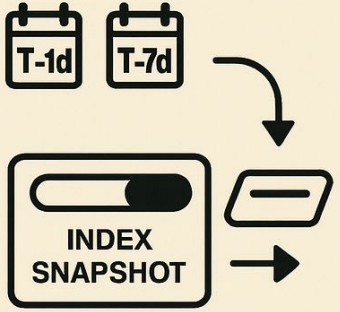


SANDBOX LAB

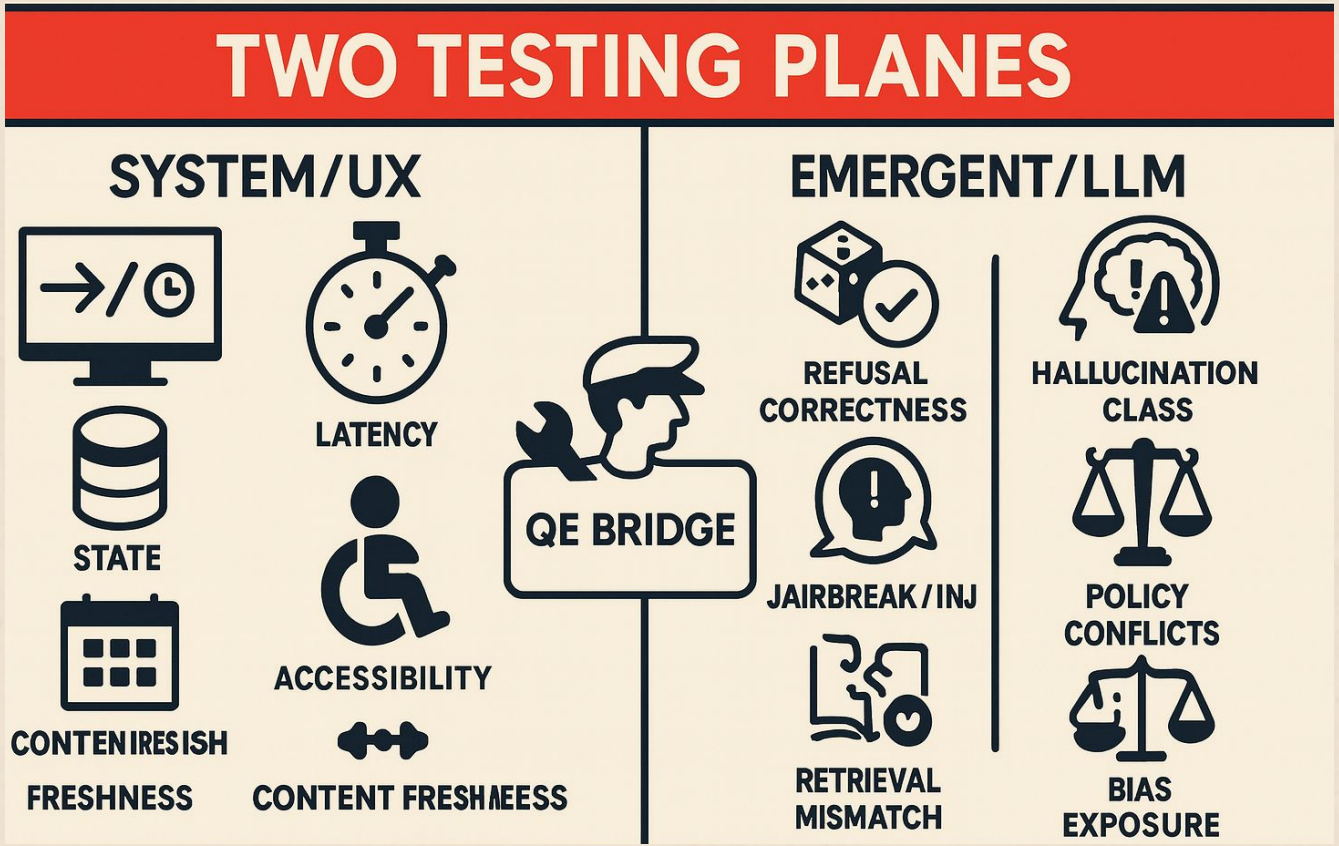
SHADOW TRAFFIC

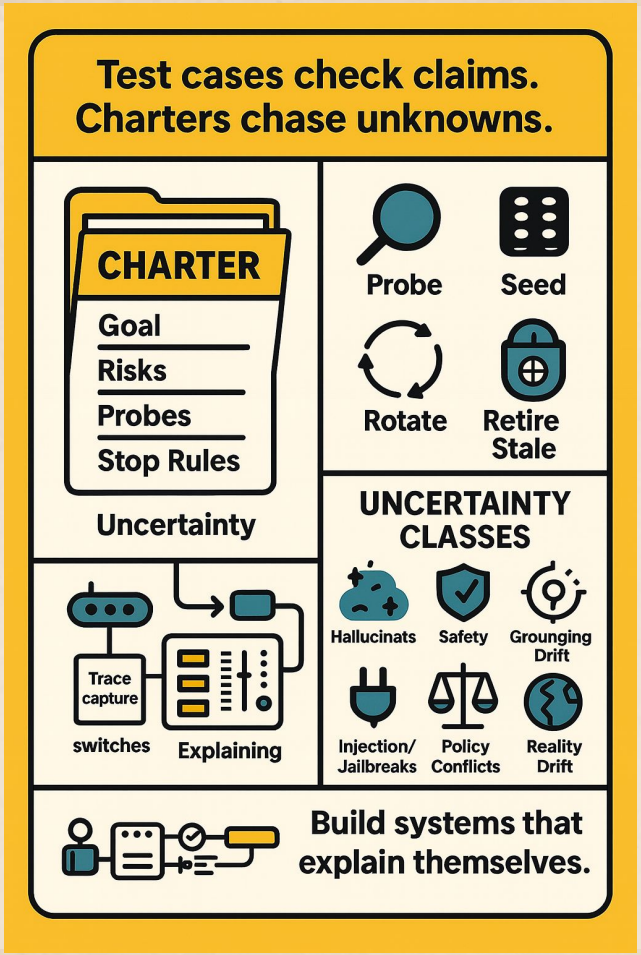


SCENARIO SWITCHES



SCENARIO
SWITCHES


TWO TESTING PLANES
SYSTEM/UX
LATENCY
STATE
ACCESSIBILITY
CONTENIRESISH
FRESHNESS
CONTENT FRESHMEESS
QE BRIDGE
EMERGENT/LLM
REFUSAL
CORRECTNESS
HALLUCINATION
CLASS
JAIRBREAK/INJ
POLICY
CONFLICTS
RETRIEVAL
MISMATCH
BIAS
EXPOSURE



As a tester, you have to advocate for building the system in such a way that these charters are explorable by design.

YOU...

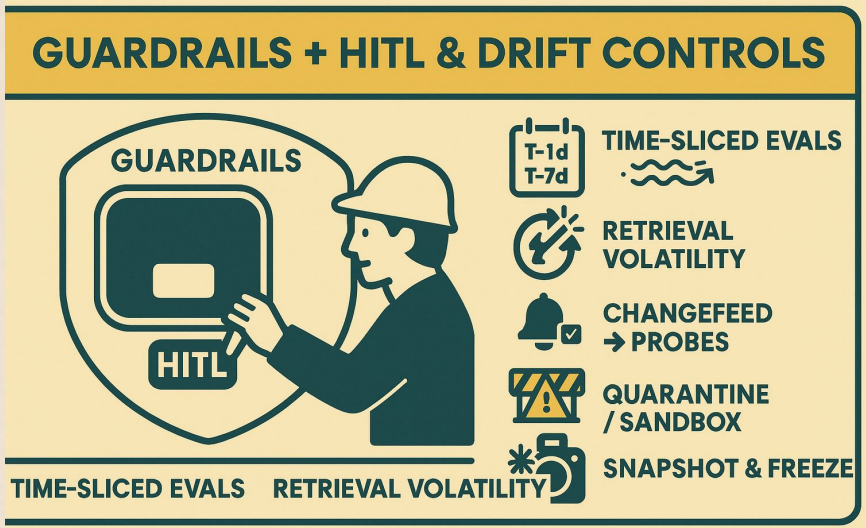
have to be part of the design process.

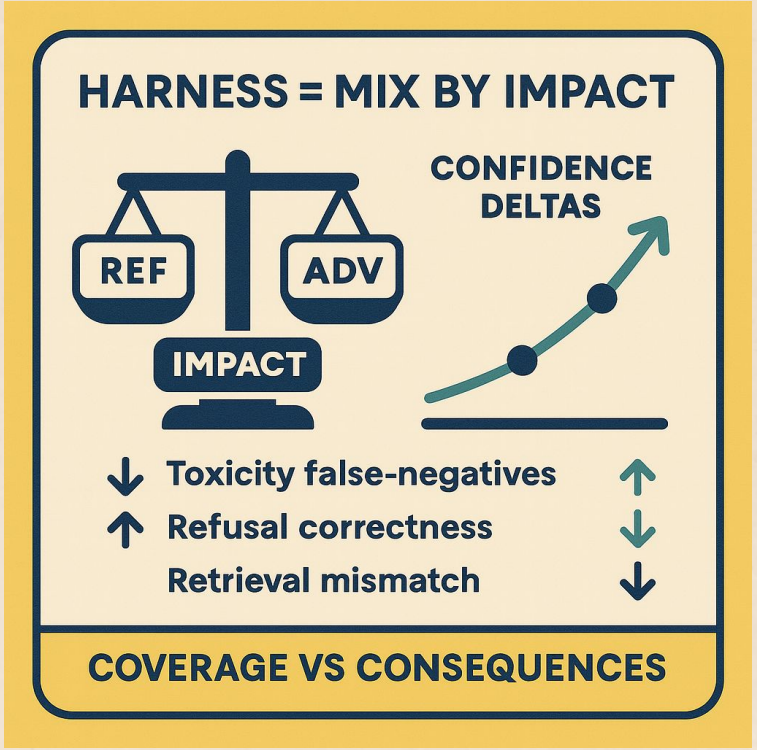
No downstream tool will save you if you did not build a system expected to explain itself.

Reference + adversarial sets weighted by impact.

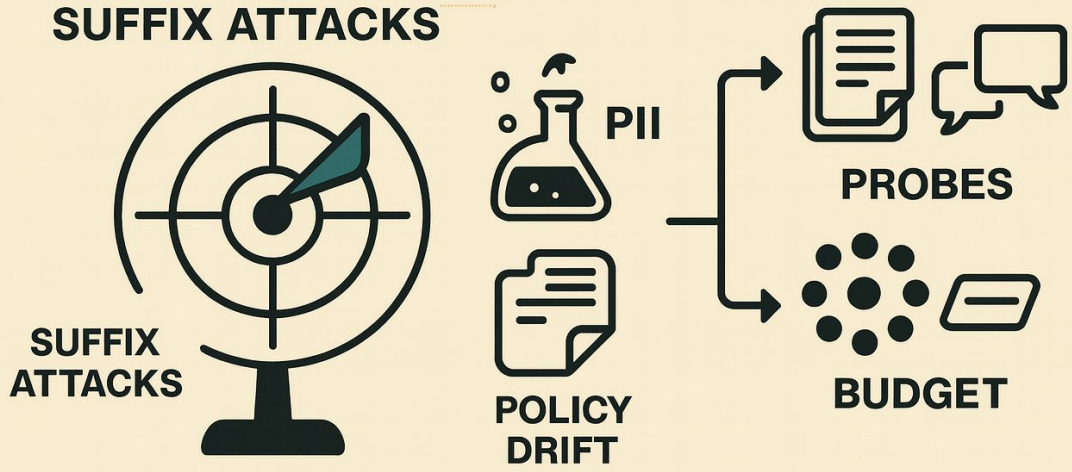
Track confidence deltas over time

Use data blast-radius constraints… so one trending source can't flip everything





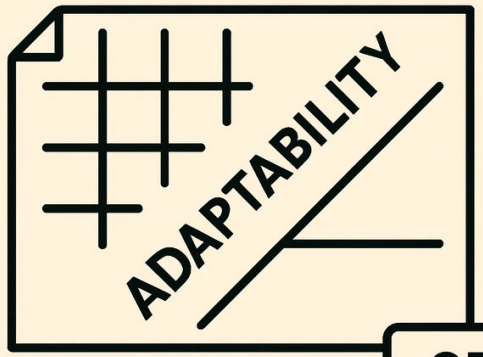
ADJACENT FIELDS VIGILANCE



You're not security, but you are early warning.

"You don't have to know everything—just what changes your r risk today."

ACT III –
QE = ARCHITECTURE & EXPLAINABILITY
+ GENEVAL LEADS DESIGN



QE BRIDGE

QE = ARCHITECTURE & EXPLAINABILITY

shape the system

Routing layers & kill switches

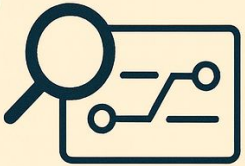
ADAPTABILITY

Poiicy-as-code

Prompt/trace capture

Uncertainbility hooks

Explainability hooks



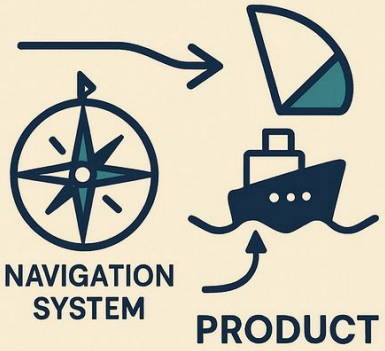
Explainabiliity hooks

We can't shrink the unknowns—but we can shape the system that meets them.

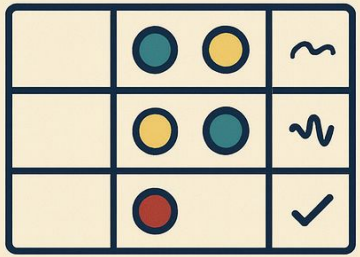
GenEval (By Design)

“Evaluation leads design.

NAVIGATION SYSTEM

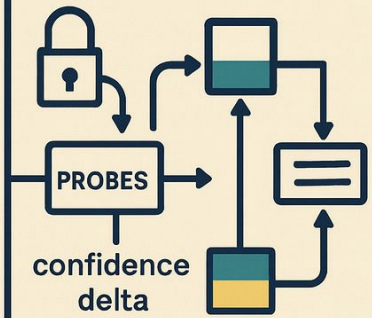


CONFIDENCE BOARD

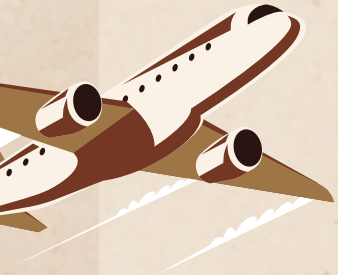


evolves even
when product doesn’t

BY-DESIGN LOOP



FELEASE
GATE



CASE VIGNETTE:

INCIDENT → PROBES → GUARDRAILS → BUDGETS → TEEMETRY

INCIDENT

UNACCEPTABLE OUTPUT

PROBES

PROBE CLASS

REF + ADV PARAPHRASES

GUARDRAILS & ROUTING

YOUTH/ DISTRESS

NEAR-ZERO

BUDGETS

TRACE

POLICY APPLIED

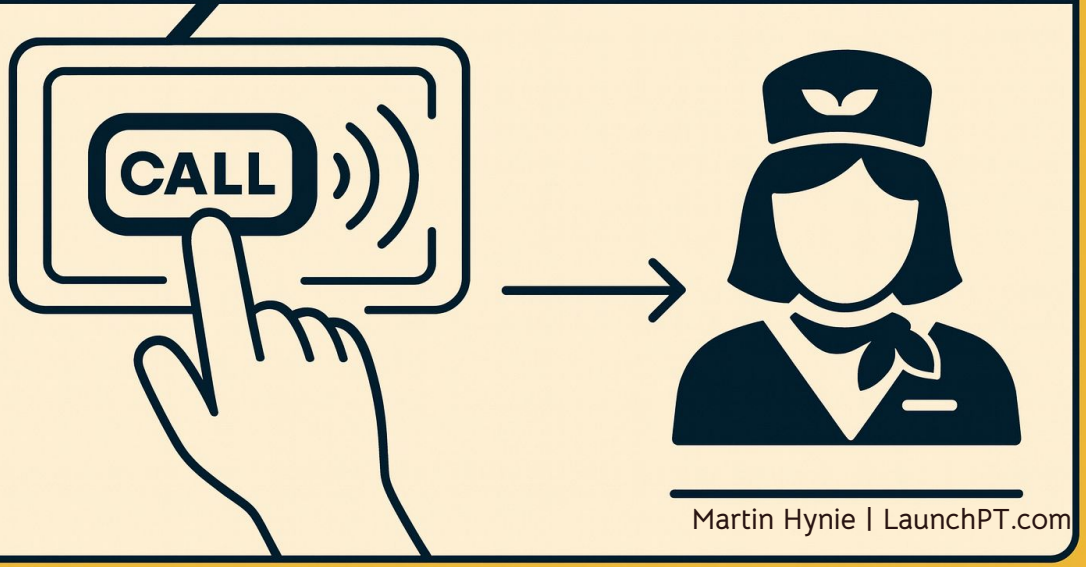
CONFIDENCE

TELEMETRY

MODEL HARDENS? TIGHTEN PROBES... WORLD SHIFTS? TIME-SLICE EVALS, VOLATILITY SLO, QUARANTINE, SNAPSHOT/FREEZE.


MAKE CONFIDENCE VISIBLE
MAKE CONFIDENCE VISIBLE
ARTIC OCEAN
PACIFIC OCEAN
OCEANIA
AUTRALIA

Q & A



Martin Hynie | LaunchPT.com