

TDC FLN 2026

Human-in-the-Loop is the Best Place for GenAI

Radamés Roriz

"Let's make one more GenAI brainstorming meeting"

MIT · STATE OF AI IN BUSINESS

95% of organizations are getting zero return

We've seen dozens of demos this year. Maybe **one or two** are genuinely useful. The rest are wrappers or science projects.

The few that work share one discipline, something I call **Engineering of Thinking**.

Source: [State of AI in Business 2025](#), MIT NANDA



The GenAI Divide
STATE OF AI IN
BUSINESS 2025

MIT NANDA

Aditya Challapally
Chris Pease
Ramesh Raskar
Pradyumna Chari
July 2025

WHAT I'LL COVER

In the next 30 minutes, I'll show you...

PART 1

Why They Fail

Expecting magic from the model

PART 2

Where It Fits

The human-in-the-loop

PART 3

How To Use It

Engineering of thinking

Spoiler: the model you pick matters far less than how you use it.

ABOUT ME

Radamés Roriz

Staff Software Engineer @ KnowBe4

Rails Contributor (v6 · v7 · v8)

16 Years Shipping Production Software

After years building with AI, I'm tired of brainstorm meetings about how to use it. Here's what actually works.

in [linkedin.com/in/radames-roriz](https://www.linkedin.com/in/radames-roriz)

roriz.dev



PART 1

Why They Fail

The world runs on deterministic code

WHY THEY FAIL

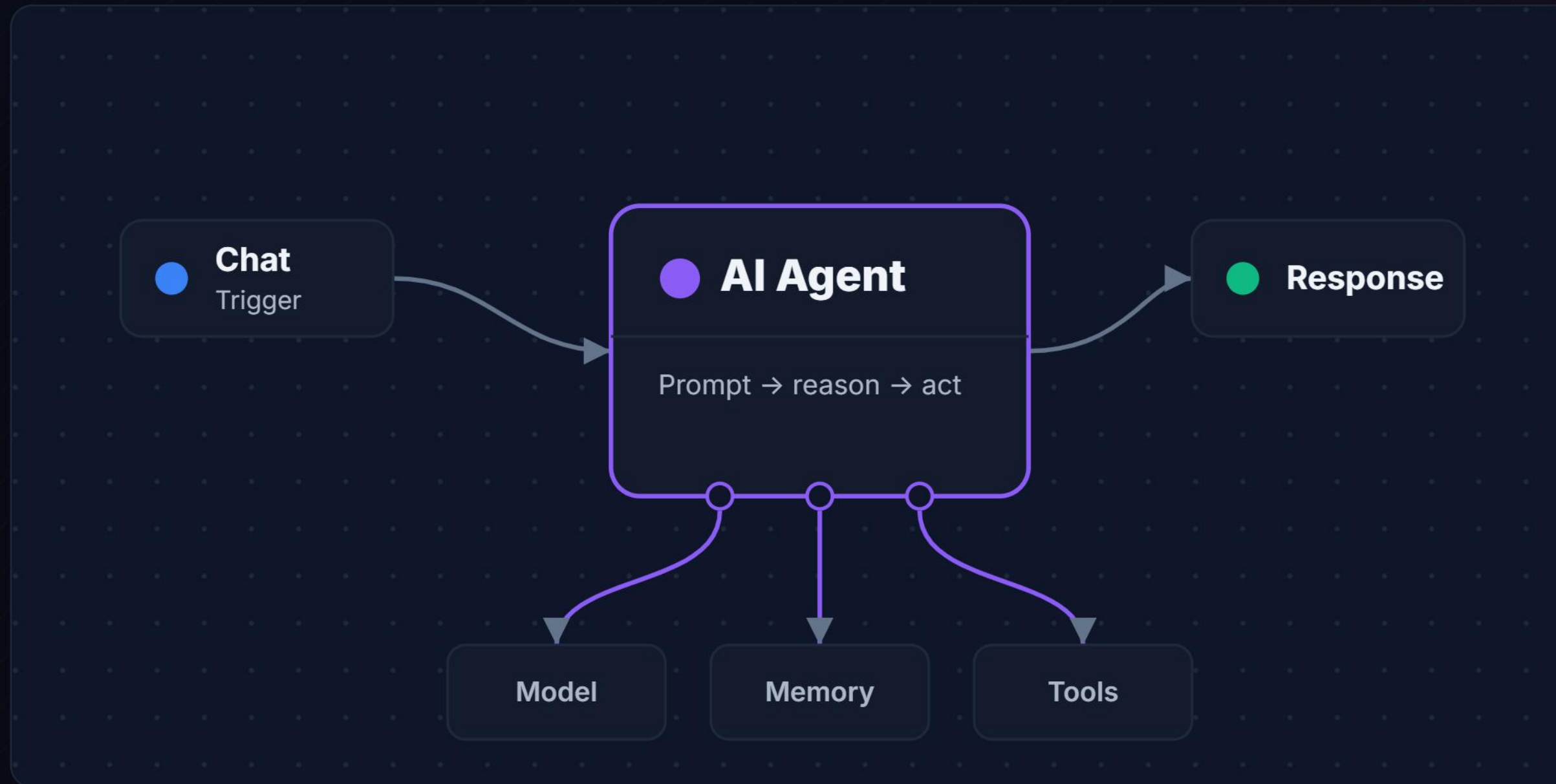
The World's Most Valuable Companies

End of 2021, before the GenAI bubble

#	COMPANY	WHAT THEY BUILD	MARKET CAP	ENGINEERING-FIRST?
1	Apple	Consumer devices	\$2,051B	✓ Engineering
2	Saudi Aramco	Oil & gas	\$1,920B	✗ Resources
3	Microsoft	Software & cloud	\$1,778B	✓ Engineering
4	Amazon	E-commerce & cloud	\$1,558B	✓ Engineering
5	Alphabet	Search & ads	\$1,393B	✓ Engineering

The most valuable companies on earth run on **deterministic engineering**, software and silicon that do exactly what they're told. Every time. At scale. Not one is a "GenAI company", and this isn't going anywhere.

The "Just Connect the Boxes" Illusion



It works for demos. It **breaks in production**. Every time.

WHY THEY FAIL

The Biggest Companies Already Know

META

Replacing **thousands** of human content moderators with AI

[Source: Financial Times](#)

AMAZON

AI + expert reviewers screen the marketplace, **15M** counterfeits seized in 2025

[Source: About Amazon](#)

ANTHROPIC

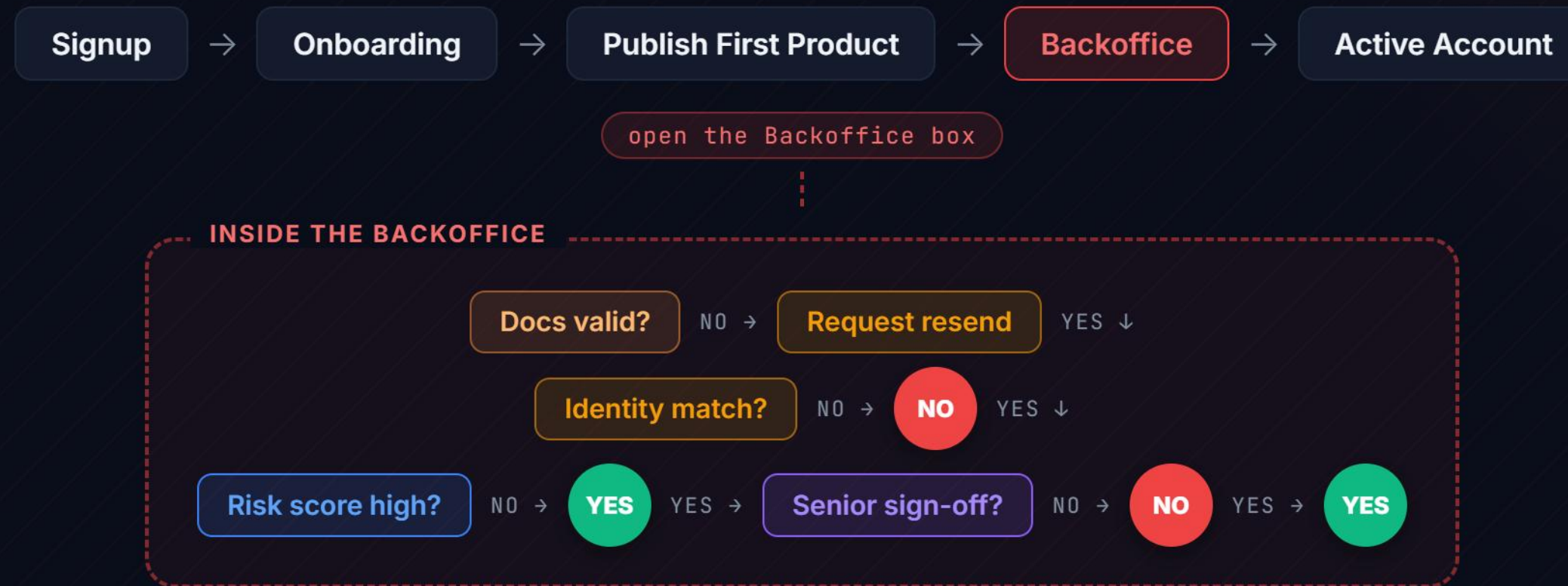
Claude Code grew from **autocomplete** to **full agents**, removing the developer's mechanical steps one at a time.

[Source: Anthropic](#)

These aren't side projects. These are core operations.

THE REAL QUESTION

Where Does AI Actually Belong?



That tangle inside one box is the **Human-in-the-Loop**.

The bottleneck isn't the software.
It's the **human box** inside the software.

PART 2

Where It Fits

Three reasons the human box exists, and no, it's not all agents.

THE FRAMEWORK

Three Reasons the Human is There

The test: **who signs?**

RESPONSIBILITY: AI WORKFLOW

The law requires it

Human signs. AI only suggests.

JUDGMENT: COPILOT

The spec is subjective

AI drafts. Human owns the result.

CHAOS: AGENT

The rules change daily

AI acts. Human sets policy & audits.

Most cases mix all three. Pick by the constraint that **dominates**.

RESPONSIBILITY

Online Prescription

1 · ID Verification → 2 · Assessment → 3 · Treatment Plan → 4 · Prescription Generation → 5 · Digital Signature → 6 · Transmission



AI **Critiques** the plan after it's written, **never blocks**. Flags a dosage too high, or a drug interaction.

A deterministic pipeline. The doctor writes the plan, signs, and is accountable, AI only critiques one step.

RESPONSIBILITY → AI ROLE

AI Workflow

Keep the human in charge, make them faster and sharper.

Right data, right time

Pull in the extra context, or strip the noise, so the reviewer sees only what matters.

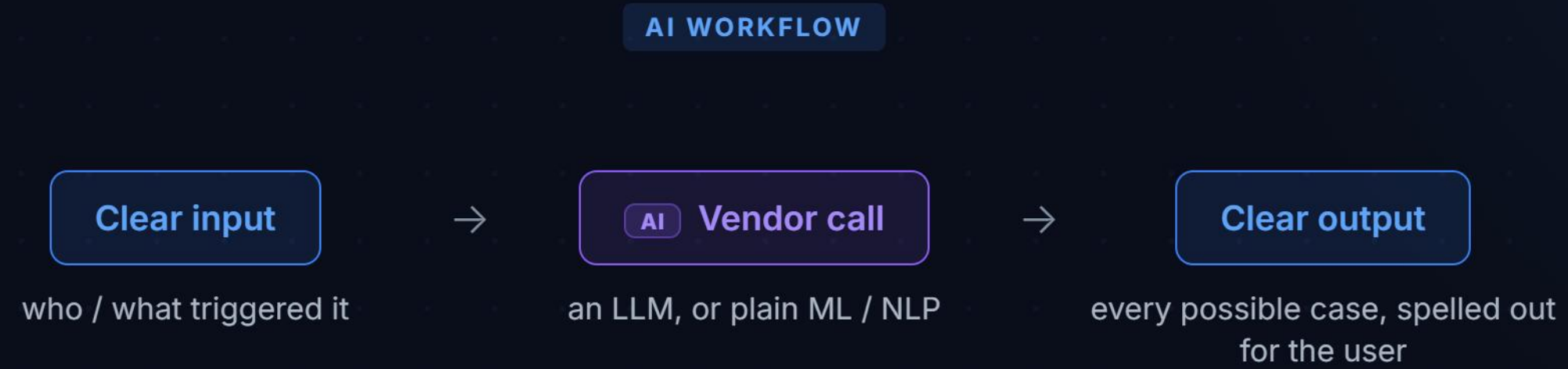
Surface anomalies

Flag what looks off, and why, before the human even starts looking.

Frictionless interface

Ask in plain language, skip the ten internal screens and legacy forms.

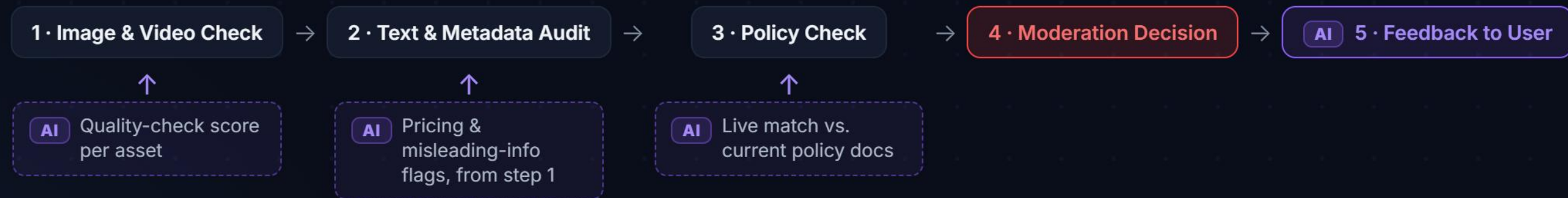
The human still decides, just with everything laid out in front of them.



Run it or skip it, **no consequences either way**. Worst case, the user just gets an error.

JUDGMENT

E-commerce Product Moderation



The human owns the decision (step 4). AI scores, flags and explains, everywhere but the call itself.

JUDGMENT

The Judgment Trap

Product: "Premium Leather iPhone Case Slim Fit"

MODERATOR A

✓ Approves

Clean title, good photo, correct category.

MODERATOR B

× Rejects

"The Apple logo is visible, that violates our brand-guideline policy."

Who's right? **Both.** The spec doesn't exist, it's all context. Monday-you disagrees with Friday-you.

JUDGMENT → AI ROLE

Copilot

15 → 2 minutes per review

Same reviewer, same judgment, **a fraction of the time.**



Think of it as an **extended AI workflow**. Instead of a single shot, it becomes a context that **accumulates over time** and gets more relevant with it.

CHAOS

Reading the ENEM Edital

EDITAL ENEM · MATRIZ DE REFERÊNCIA – MATEMÁTICA, COMPETÊNCIA 5

"Modelar e resolver problemas que envolvem variáveis socioeconômicas usando representações algébricas. H19 – Identificar a relação entre grandezas. H20 – Interpretar gráfico cartesiano que represente relações entre grandezas. H21 – Resolver situação-problema cuja modelagem envolva conhecimentos algébricos."

A FUTURE ENGINEER STUDIES

Functions (core)

Algebraic modeling

Graph interpretation

A FUTURE LAWYER STUDIES

Graph reading (also in Humanas)

Basic proportions

Skip deep algebra

Same content. **Different study plans.**

And the topics overlap: "reading a graph" lives in Math, Geography and Science at once.

CHAOS

The Chaos Trap

- ! You **can't enumerate every "important" topic**, it depends on the student's goal.
- ! The matriz is **abstract**. One "habilidade" hides a dozen concrete topics to study.
- ! Topics never split cleanly: one skill spans Math, Geography and Science at once.

Every student needs a **different study plan**. And the emphasis shifts every year.

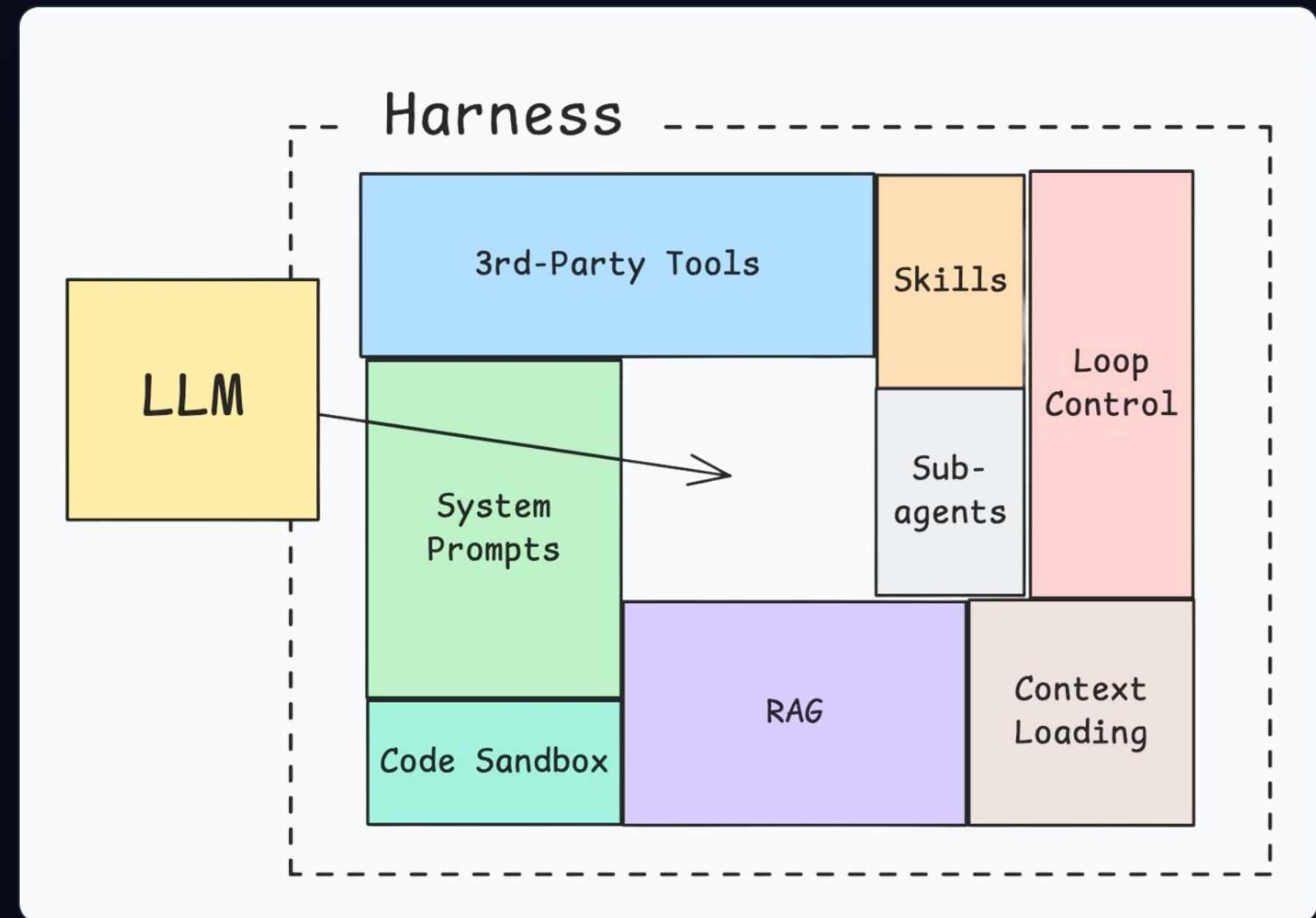
CHAOS → AI ROLE

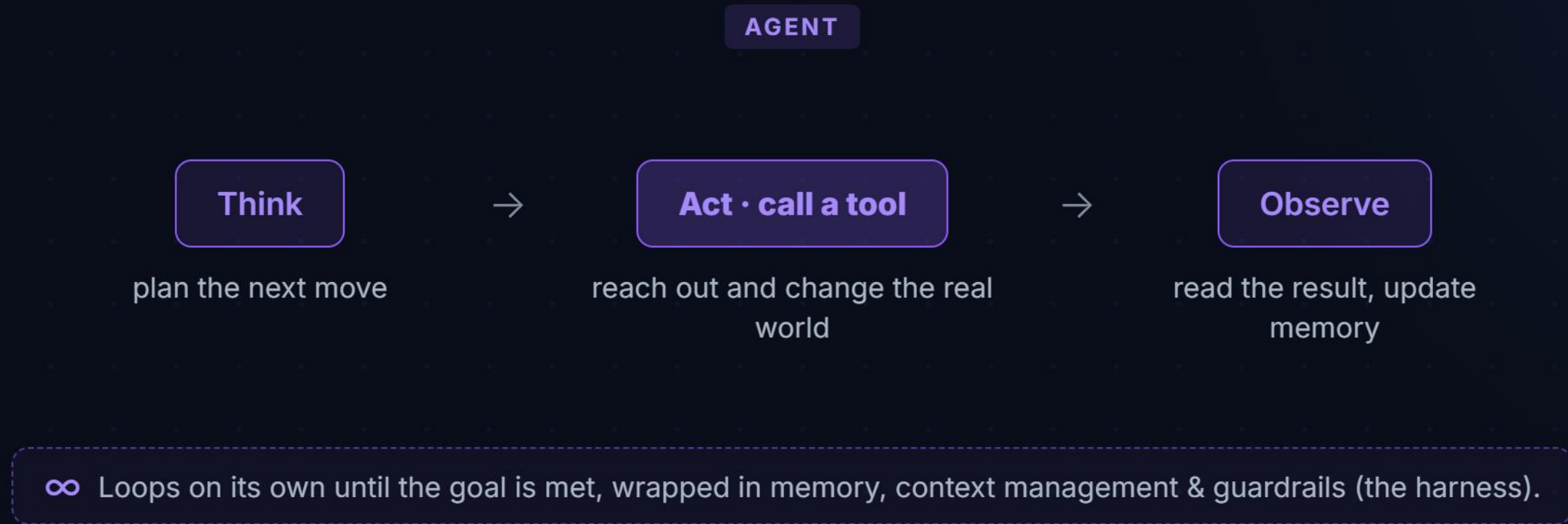
Agent

GenAI thrives in chaos.

Parse the matrix. Map abstract skills to concrete topics.

Re-plan for every student profile, and it never skims.

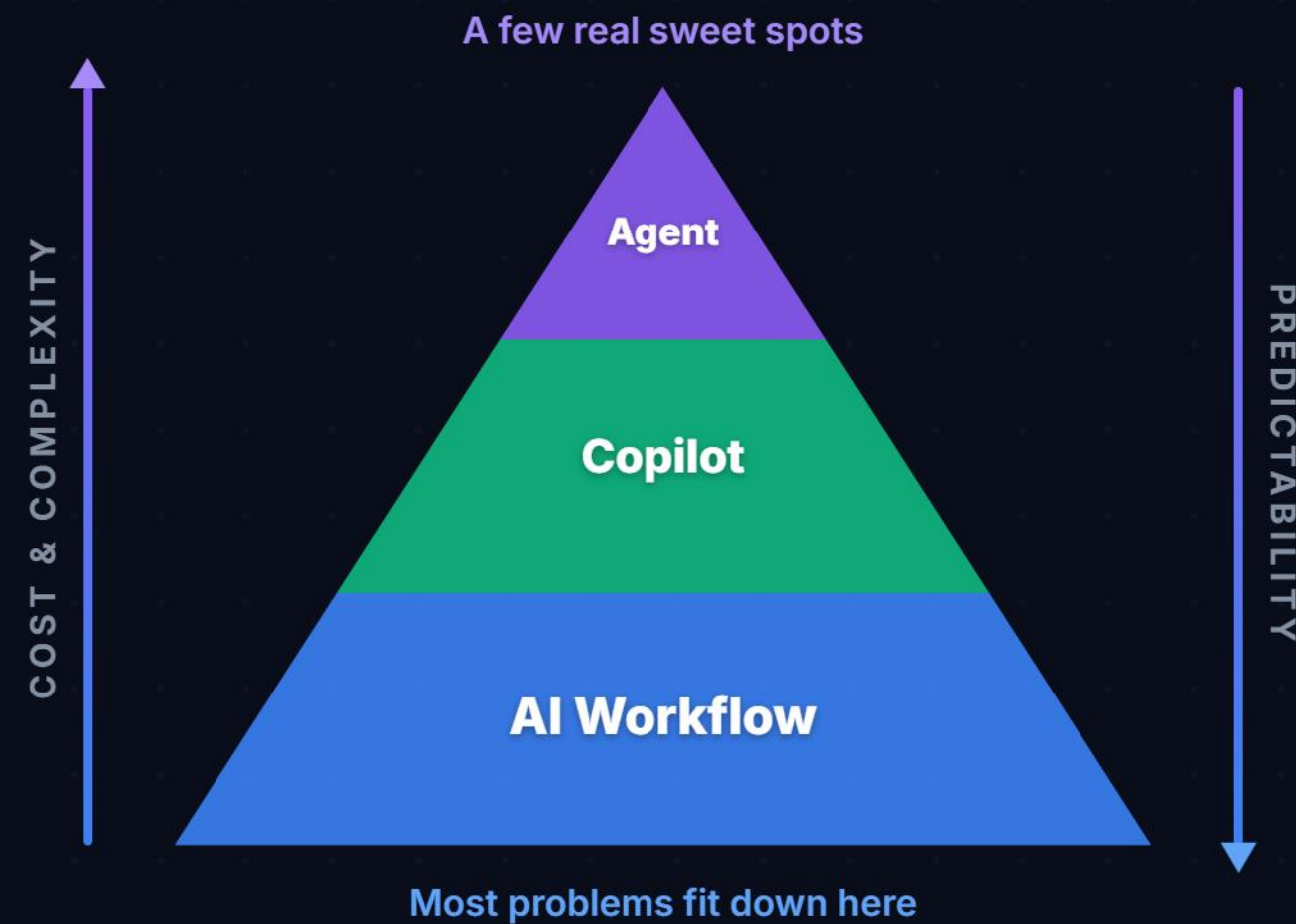




The **action is the whole definition.** If it can't call external tools, it isn't an agent.

THE TRAP

Could an **Agent** do a job an **AI Workflow** already handles? **Yes**. And that's the problem.



FINDING THE SWEET SPOT

The Product Team's Calls

1 · CHOOSE THE AUTONOMY TIER

Pick the simplest tech level on the pyramid. Favor Workflows and Copilots over Agents, so the system stays predictable and pays off.

2 · DEFINE SYSTEM BOUNDARIES

Draw the line between what runs automatically, where AI assists, and where a human has to make the call.

3 · BUILD TRUTH TABLES

Build ground-truth tables that check AI outputs against how your best operator actually decides.

Finding the sweet spot doesn't happen by accident. It's **core product work**.

PART 3 · HOW TO USE IT

It's just ask claude, right?

Engineering of thinking



THE REAL CULPRIT

The Operator's Secret

Remember the ENEM edital? It's not just "read the PDF"

- 1 Skims the matriz to see which areas carry the most weight
- 2 Figures out the student's goal and target course
- 3 Maps each abstract competência to concrete topics to study
- 4 Merges and splits overlapping skills by judgment, flags the ambiguous ones
- 5 Diffs against past exams to see which topics actually get tested

A **methodology that lives in the operator's head**, tacit knowledge never written down.

THE REAL CULPRIT

The Pitch Nobody Wrote Down

TWO REPS, SAME SCRIPT

Same training, same deck, same terms. But each one paces it their own way and skips whatever the room already gets. Both close.

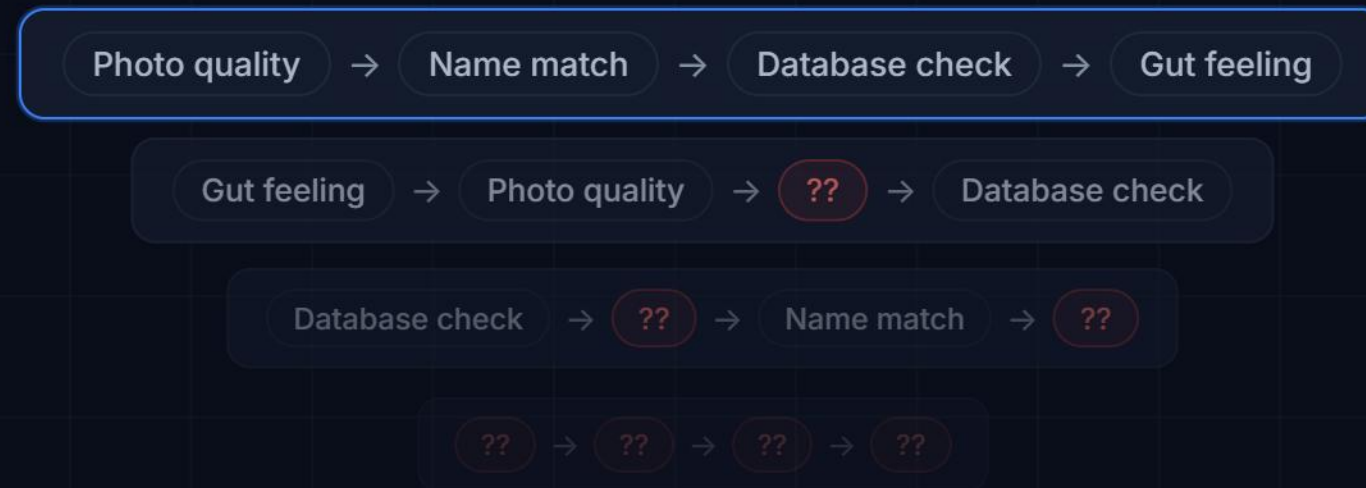
A SENIOR CLOSER

Barely opens the deck and runs a completely different play, reading the room as it goes. Also closes.

Yet each one stays consistent in their own style, pitch after pitch.

That personal *method* is the real gold: tacit, hard to write down. Structuring it is the job.

Give the AI the goal but not the method, and one task doesn't have **two** paths like our moderators. It has almost **infinite** ones.



✗ ∞ every run, a different path

**"LLMs don't improvise well.
They **hallucinate**."**

PART 3 · HOW TO USE IT

The Solution

Beat 2: engineering of thinking

THE SOLUTION

The Postmortem Analogy

Production crashes. You walk into the war room. "What happened?"

WITHOUT STRUCTURE

"It was the network."
"No, the database."
"Actually, the deploy."
"I think it was my fault..."

Random. Unreliable. Inconsistent.

THE POSTMORTEM RITUAL

- 1 Construct the timeline
- 2 Identify critical failures
- 3 Define action items

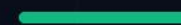
The output converges.

Two different engineers will produce slightly different timelines.

But the critical failures will be the same.

The action items will overlap.

The structure limits the paths available to "freak out" or go off on tangents.



This is exactly what we need to do with GenAI.

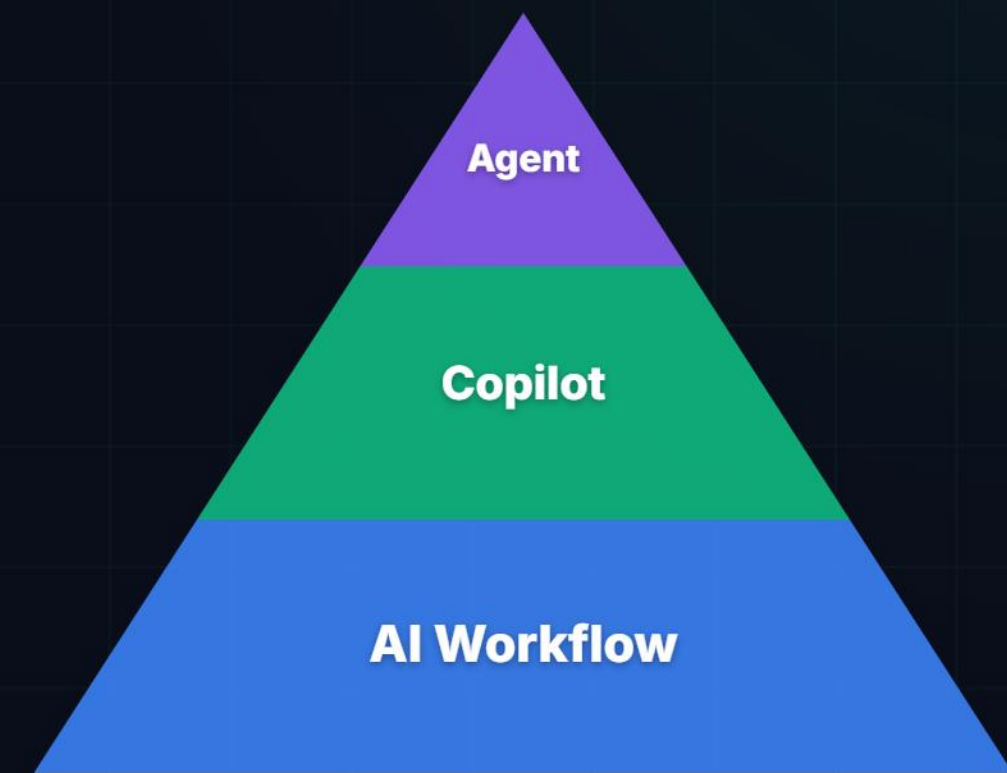
THE SOLUTION

Engineering of Thinking

Instead of asking the AI to "do the task,"
we ask it to "execute the process."

Take the invisible methodology from your operator's head
and turn it into a **rigid sequence of steps** the AI must follow.

It's not prompt engineering. It's **process engineering**.



THE PLAYBOOK

What the Product Team Actually Does

1

Learn the method

Get Claude's baseline plan

→

Shadow the operator & annotate the gaps

→ THE HIDDEN METHOD

2

Build the eval set

Decompose into verifiable steps

→

LLM-judge + human critique

→ VALIDATED DATASET

Verifiable = one step in, a checkable output out. Those pairs are your first gold dataset.

3

Improve forever

Loop to raise accuracy, ideally forever

→ RELIABLE PIPELINE

THE GOLD DATASET

What an Eval Set Looks Like

THE EVAL SET

INPUT	EXPECTED OUTPUT	JUDGE	PASS
H2O (graphs)	functions, data reading	LLM + human	✓
Competência 5	algebra, modeling	rubric	✓
Cross-subject skill	split by discipline	LLM	✗

87% pass · 200 cases · grows every week

THE TRUTH TABLE

Judge vs. human over 200 cases · human / judge

true / true
148
agree · pass

true / false
12
judge missed it

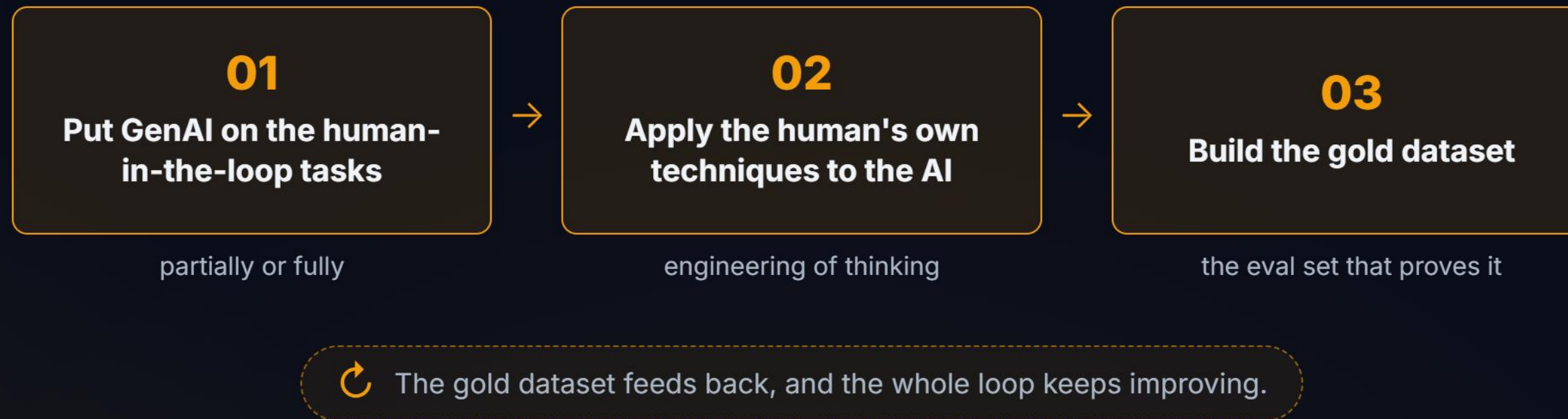
false / true
14
false pass

false / false
26
agree · fail

Diagonal = the judge agrees with the human. **87%**

THE RECAP

It Comes Down to This



That's it.

And notice: almost none of this is technical. It's a **business call**, not an engineering one.

THE VERDICT

Stop waiting for a better model.

Even if the next model is **twice as good**, it still won't know your operator's **process**.

The model was never the bottleneck.

Today's model is already enough. The only lever left is the **right place**, and the **right technique**.



ONE MORE THING

Liked this path? Dig deeper.

If this way of thinking clicked, take a look at **ActiveGenie**, my open-source library that ships engineering-of-thinking patterns out of the box.

Debate

Council

Ranking

...and more

github.com/Roriz/active_genielinkedin.com/in/radames-rorizroriz.dev

Thank you. I hope I can help you all.

