March 2026

Martin Seeler
Sr Staff AI Engineer

# AI Evals Done Right

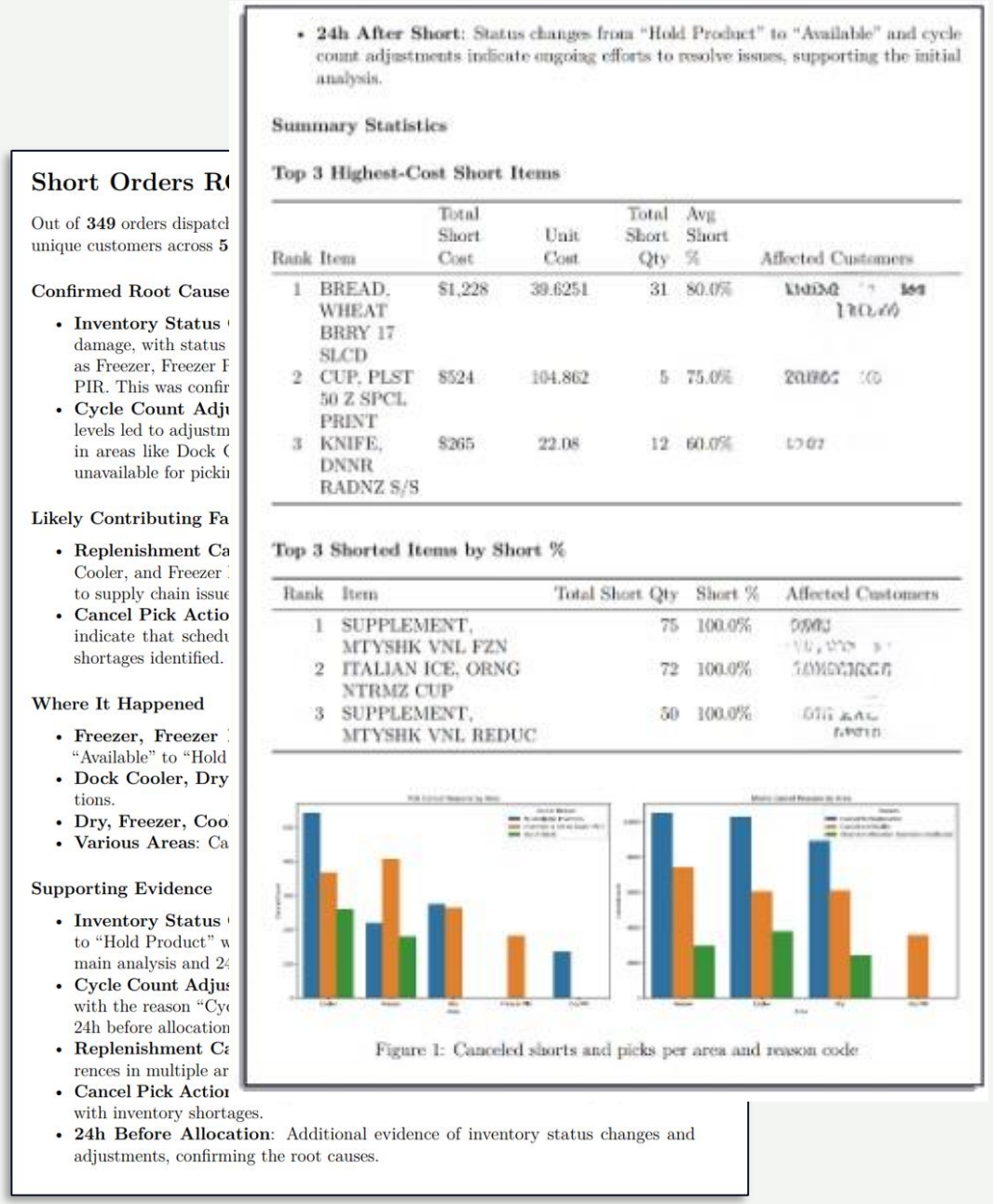
From Vibes to Confident Decisions

AI Studio

# Root Cause Analysis

- Quantitative Process Mining
- Qualitative LLM Interpretation

- Analyzes
  - Late Shipments
  - Order Lead Times
  - Labor Productivity
  - Short Orders
  - ...



Short Orders R

Out of **349** orders dispatcl
unique customers across **5**

**Confirmed Root Cause**

- **Inventory Status** damage, with status as Freezer, Freezer F PIR. This was confir
- **Cycle Count Adju** levels led to adjustm in areas like Dock C unavailable for picki

**Likely Contributing Fa**

- **Replenishment Ca** Cooler, and Freezer to supply chain issue
- **Cancel Pick Actio** indicate that schedu shortages identified.

**Where It Happened**

- **Freezer, Freezer** "Available" to "Hold
- **Dock Cooler, Dry** tions.
- **Dry, Freezer, Coo**
- **Various Areas**: Ca

**Supporting Evidence**

- **Inventory Status** to "Hold Product" w main analysis and 24
- **Cycle Count Adjus** with the reason "Cyc 24h before allocation
- **Replenishment Ca** rences in multiple ar
- **Cancel Pick Actio** with inventory shortages.
- **24h Before Allocation**: Additional evidence of inventory status changes and adjustments, confirming the root causes.

- **24h After Short**: Status changes from "Hold Product" to "Available" and cycle count adjustments indicate ongoing efforts to resolve issues, supporting the initial analysis.

**Summary Statistics**

**Top 3 Highest-Cost Short Items**

| Rank | Item | Total Short Cost | Unit Cost | Total Short Qty | Avg Short % | Affected Customers |
|---|---|---|---|---|---|---|
| 1 | BREAD, WHEAT BRRY 17 SLCD | $1,228 | 39.6251 | 31 | 80.0% | [illegible] |
| 2 | CUP, PLST 50 Z SPCL PRINT | $524 | 104.862 | 5 | 75.0% | [illegible] |
| 3 | KNIFE, DNNR RADNZ S/S | $265 | 22.08 | 12 | 60.0% | [illegible] |

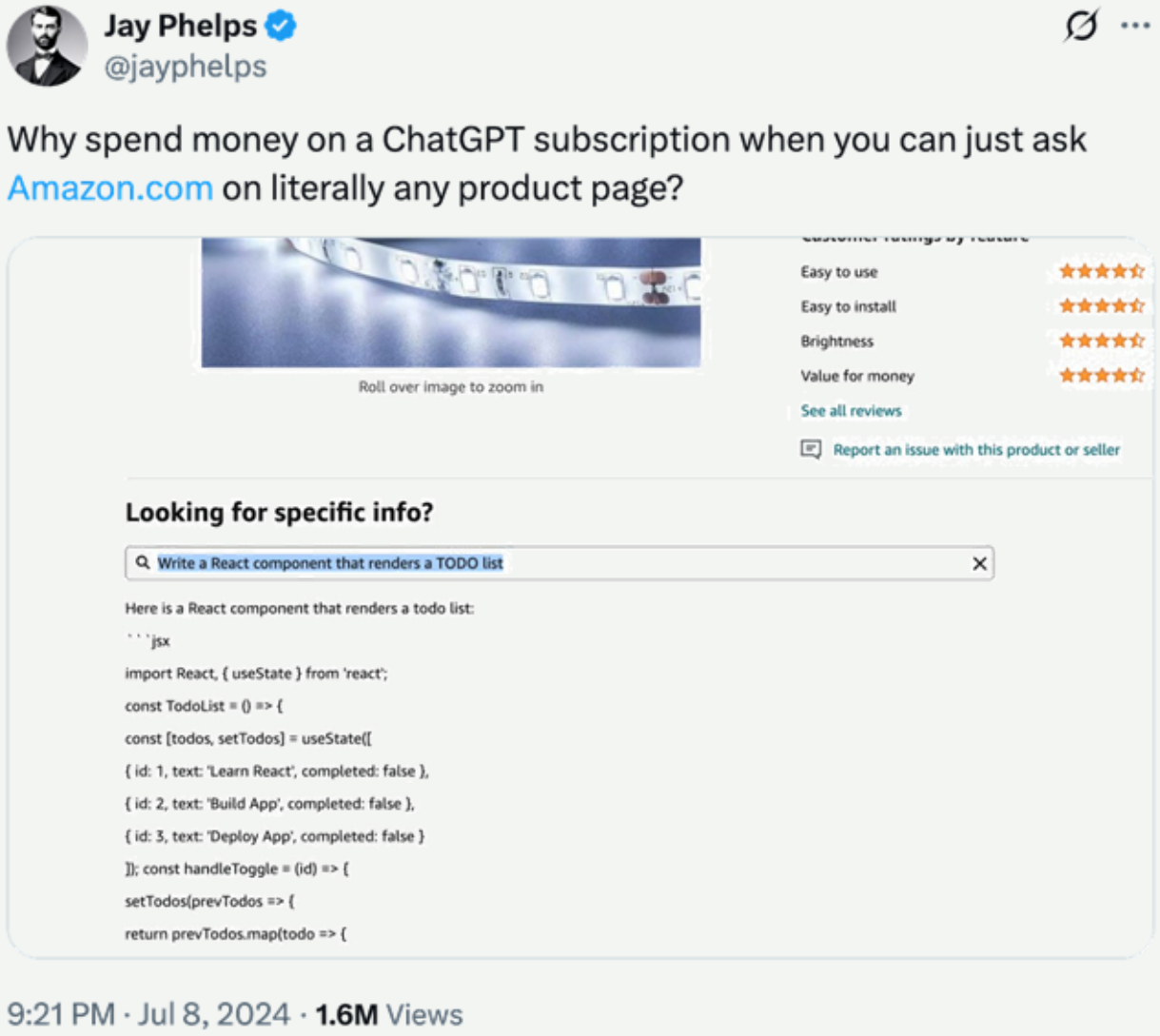
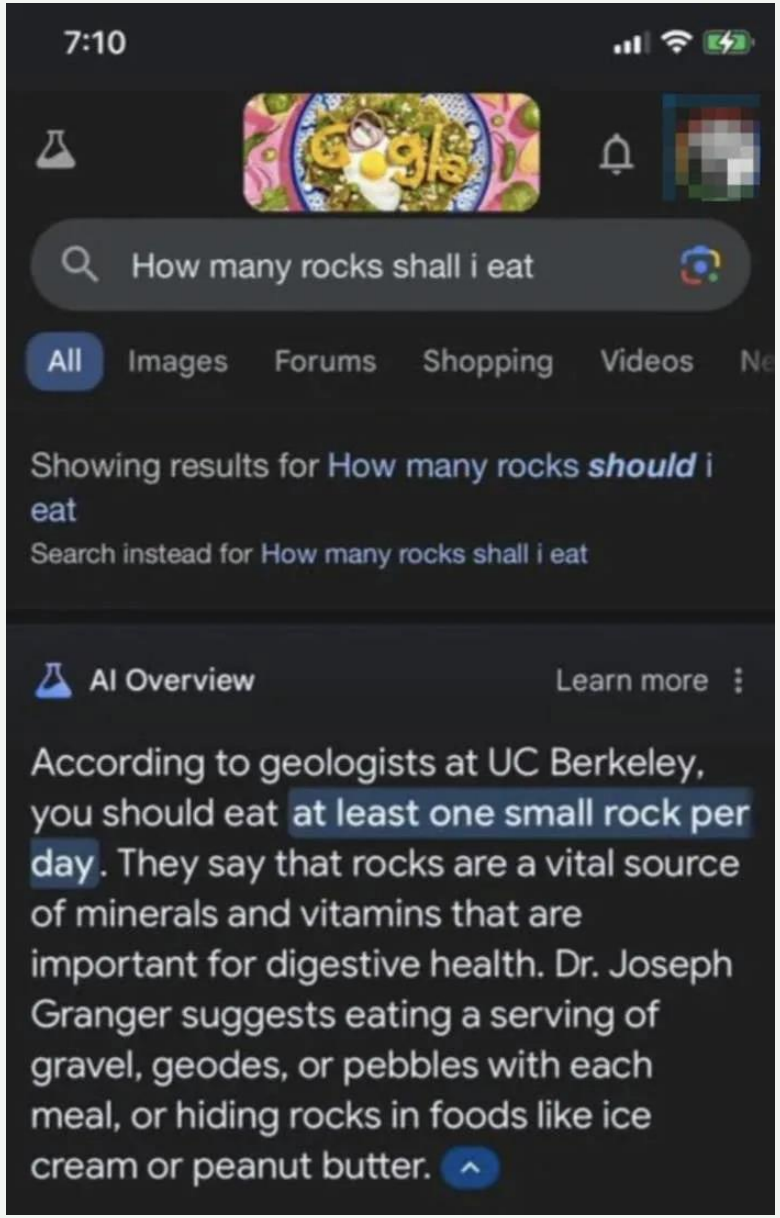
**Top 3 Shorted Items by Short %**

| Rank | Item | Total Short Qty | Short % | Affected Customers |
|---|---|---|---|---|
| 1 | SUPPLEMENT, MTYSHK VNL FZN | 75 | 100.0% | [illegible] |
| 2 | ITALIAN ICE, ORNG NTRMZ CUP | 72 | 100.0% | [illegible] |
| 3 | SUPPLEMENT, MTYSHK VNL REDUC | 50 | 100.0% | [illegible] |



Figure 1: Canceled shorts and picks per area and reason code

AI Studio


Jay Phelps
@jayphelps
Why spend money on a ChatGPT subscription when you can just ask Amazon.com on literally any product page?
Roll over image to zoom in
Easy to use
Easy to install
Brightness
Value for money
See all reviews
Report an issue with this product or seller
Looking for specific info?
Write a React component that renders a TODO list
Here is a React component that renders a todo list:
```jsx
import React, { useState } from 'react';
const TodoList = () => {
const [todos, setTodos] = useState([
{ id: 1, text: 'Learn React', completed: false },
{ id: 2, text: 'Build App', completed: false },
{ id: 3, text: 'Deploy App', completed: false }
]); const handleToggle = (id) => {
setTodos(prevTodos => {
return prevTodos.map(todo => {
9:21 PM · Jul 8, 2024 · 1.6M Views


7:10
How many rocks shall i eat
All
Images
Forums
Shopping
Videos
Showing results for How many rocks should i eat
Search instead for How many rocks shall i eat
AI Overview
Learn more
According to geologists at UC Berkeley, you should eat at least one small rock per day. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter.

AI Studio
```

# The Problem with Generic Evals

- False sense of measurement and progress
- Good starting point, but don’t correlate with real failures
- Miss domain-specific problems


AI Studio

# Questions

01 How can we find good evals?

AI Studio

# Questions

01 How can we find good evals?

02 How can we quantify the product quality and track improvements over time?

AI Studio

# Questions

01 How can we find good evals?

02 How can we quantify the product quality and track improvements over time?

03 What is the most critical task to tackle next?

AI Studio

# Questions


01 How can we find good evals?

02 How can we quantify the product quality and track improvements over time?

03 What is the most critical task to tackle next?

04 How can we be sure to release a new model?

AI Studio

# Error Analysis

- Identifies failure modes unique to your app

- Process
  1. Creating a Dataset (extract examples)
  2. Open Coding (write comments)
  3. Axial Coding (group same errors)
  4. Iterative Refinement (measure and repeat)


AI Studio

# 0. Collect Traces

- A trace is
  - one AI request, captured from start to finish
  - a user experiencing your product

- Start collecting traces early
- Track costs and durations

- Use tools like Langfuse, Braintrust, Arize Phoenix, LangSmith


AI Studio

# 1. Create Dataset

- How many?
  - Start with 20 traces
  - Stop at “theoretical saturation”, no new issues

- Which ones?
  - Random Sampling
  - Generic Evals (e.g., ”helpfulness”)
  - Outlier detection
  - Metric based (e.g., long outputs first)


AI Studio

# 2. Open Coding - Annotate your Traces

**100% of shipments are late**, yet the number of late shipments is **0**. This paradox highlights severe inefficiencies in warehouse operations, impacting labor productivity and customer satisfaction. The root cause analysis reveals critical areas for improvement ranked by their impact on labor productivity:

### Loading Delays (Most Significant)

- **Shipments are held up due to insufficient loading resources** or poor scheduling, causing significant delays in the loading process.

### Picking Inefficiencies (Moderate Significance)

- **Picking takes longer for late shipments**, often requiring re-sorting due to errors, which slows down the process.

### Waiting for Packing Space (Moderate Significance)

- **Orders are delayed waiting for packing space**, especially between packing and loading, due to poor coordination.

### Inventory Issues (Least Significant)

> - **Disorganized inventory or missing items lead to longer search times** during picking and packing.

AI Studio

# 2. Open Coding - Annotate your Traces

Our warehouse's **average worker productivity** stands at **14.52**, with a notable disparity between high and low performers. **High performers achieve a productivity of 18.76**, while **low performers manage only 2.12**. Alarmingly, **25.49% of our workforce are low performers**. The root cause analysis reveals critical operational bottlenecks affecting productivity levels:

### Picking to Shipping Delays (Most Significant)

- **Low performers take 70% longer** to move orders from picking to shipping, likely due to **packing congestion** or waiting for equipment like carts or packing stations.

### Inventory Update Delays (Moderate Significance)

- After shipping, **low performers take 16 minutes to update inventory**, compared to no delay for high performers. This is possibly due to **manual processes** or waiting for system updates.

### Receiving Inefficiencies (Least Significant)

- **Low performers repeat receiving operations 39 times**, while high performers do it less than once. This indicates a need for **repeated checks or corrections**, slowing down the process.

AI Studio

# 2. Axial Coding – Group Comments

- Group similar comments together by using
  - Embeddings
  - LLMs (ask ChatGPT)
  - Keywords
  - Manual
  - ...
- We reduced 168 comments to 19 failure modes
- Each failure mode is a perfect eval for your app

Avoid judgmental language like "low performers" or "lazy workers" that could be perceived as blaming individuals

"**Low performers**... I think HR would kill us with this wording!" (id: ecd25693)

"Alarmingly, **25.49% of our workforce are low performers**" - this is judgmental and not backed up by actual impact (id: 2fc636f7)

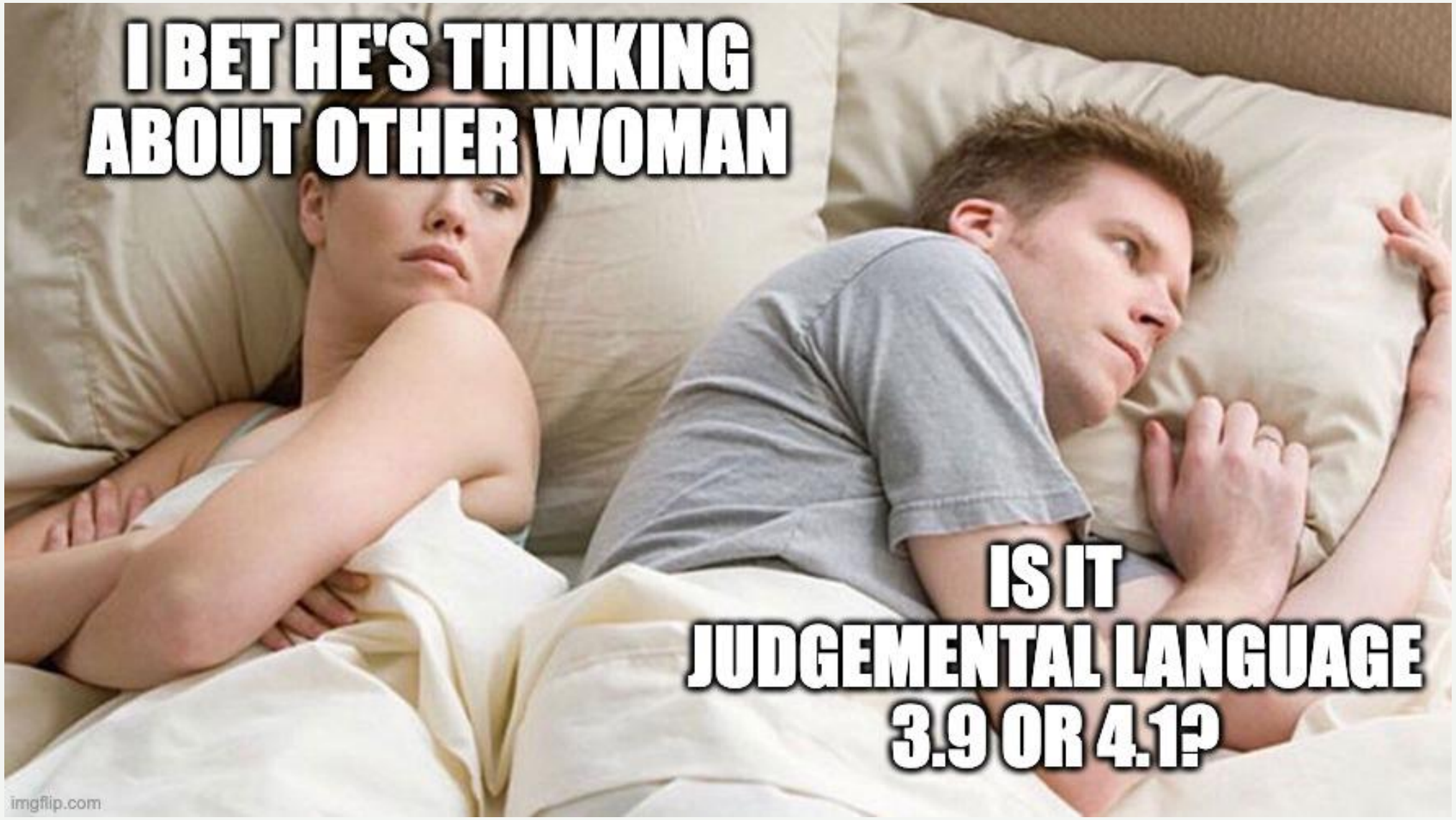
# Writing Evals

- Verifiables
  - Math
  - Code
  - Length / Regex checks
- Non Verifiables
  - Creative Writing
  - Tone of Voice
  - Subjective
- Use positive and negative examples from your comments

**Evaluation Prompt**

```
"You are evaluating an RCA (Root Cause Analysis) report for a

Check for HR-sensitive or unprofessional labels that blame ir
- Judgmental worker classifications (e.g., "low performers",
- Blame-oriented language (e.g., "workers are causing delays'
- Alarmist or emotionally charged terms (e.g., "alarmingly",
- Language that could be perceived as targeting or criticizir

Do NOT flag as errors:
- Objective, neutral descriptions (e.g., "workers with produc
- Factual process descriptions without blame (e.g., "delays c

The report FAILS (score 1) if it contains ANY judgmental, HR-
The report PASSES (score 0) if ALL language is objective, neu

Here is the report:
{{input}}"
```

AI Studio


I BET HE'S THINKING
ABOUT OTHER WOMAN
IS IT
JUDGEMENTAL LANGUAGE
3.9 OR 4.1?
imgflip.com

# Use Binary Scores

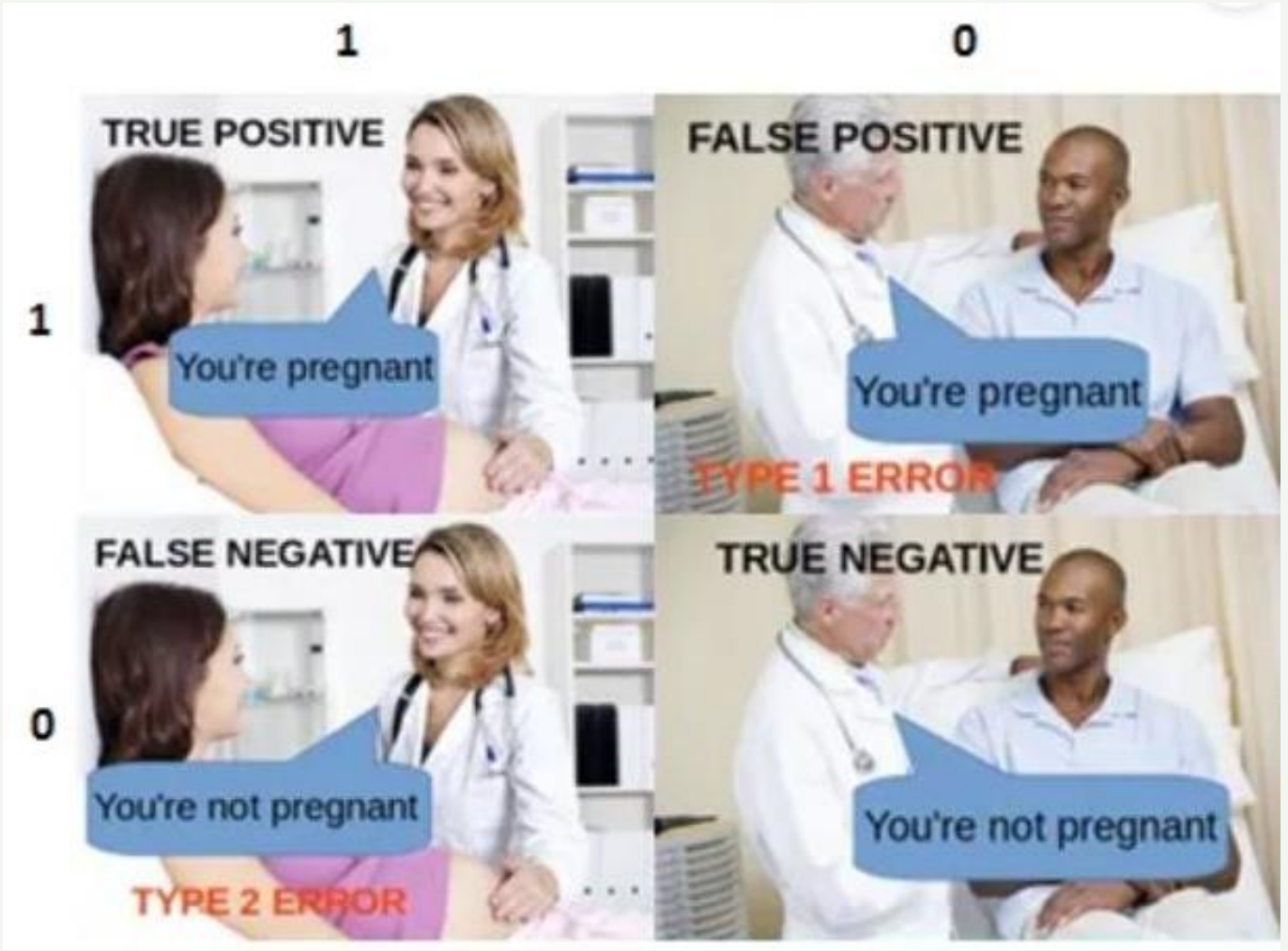
- *Likert Scale:* 1 to 5 👎
  - People default to safe middle values like 3 to avoid making decisions
  - For an LLM (as a judge), it's equally hard to distinguish
- *Binary*: Pass or Fail 👍
  - Forces clear decisions
  - 1 = Failure exists (fail)
  - 0 = No Failure (pass)

- You make binary decisions anyway!
  - what's the cutoff rate for "good enough"?


AI Studio

# Measure Judge Quality

- Use confusion matrix – your labels vs judge labels
- Tells you, how well your judge aligns with human judgments
- Focus on
  - True Positive Rate (TPR)
  - True Negative Rate (TNR)
- Errors hide in the long tail




AI Studio

# Same Model for App and Eval?

- Yes - Judge has a different task than your main LLM
- Models *can* exhibit bias when evaluating their own outputs

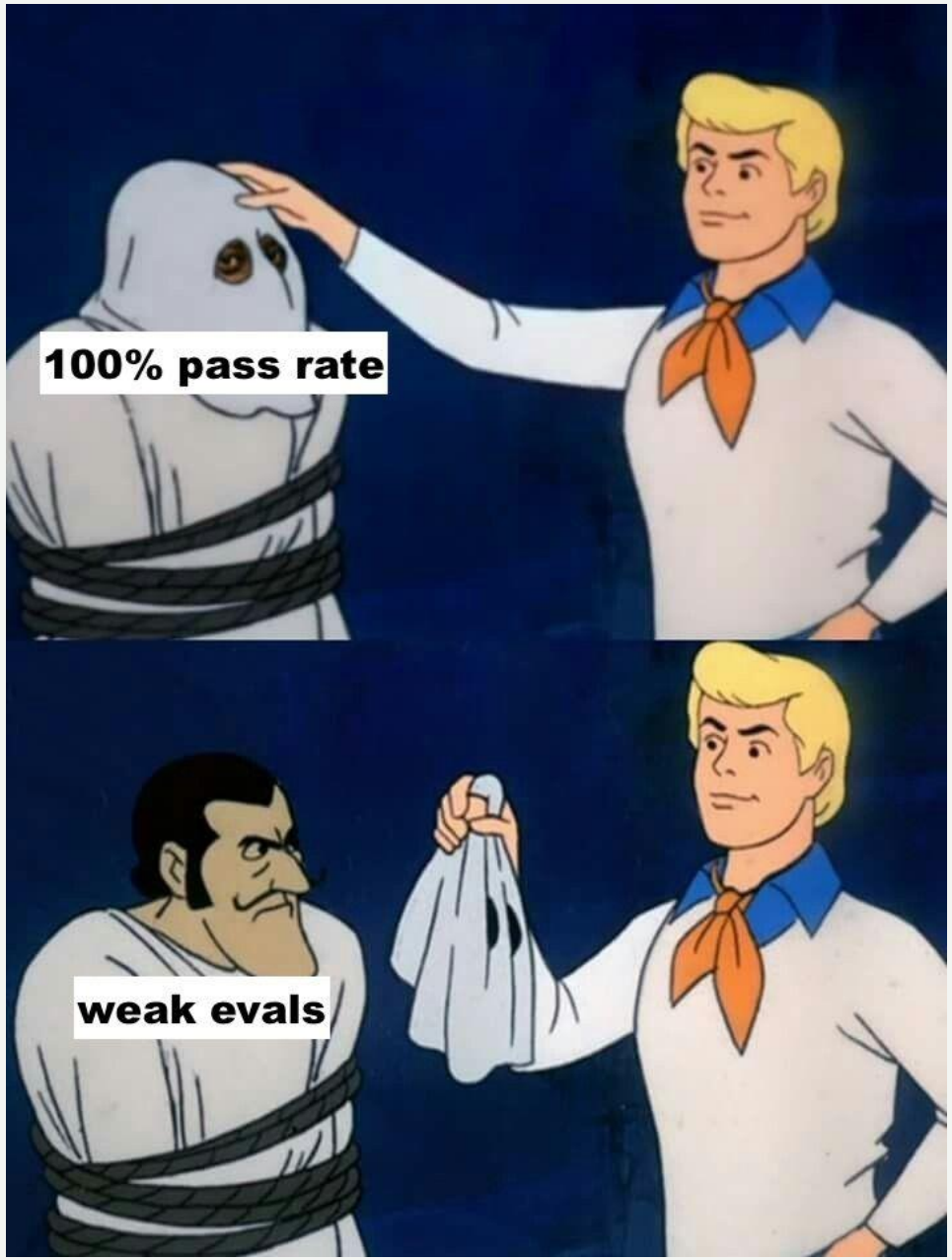
- Start with most capable models first to establish strong alignment with human judgement
- optimize for cost later


AI Studio

# 100% Pass Rate = 🚩

- Evals != Tests
- Your evals aren't challenging enough

- Teams that hit 100% often gamed their metrics rather than fixing problems
- if you're at 100%, what do you work on next?




AI Studio

# Taxonomy

## 01 Context Clarity

CC-01: Missing Units and Metrics
CC-02: Missing Scope and Period
CC-03: Missing Sample Size
CC-04: Missing Comparative Context
CC-05: Missing KPI Definition

## 02 Data Validity

DV-01: Unusual Patterns
DV-02: Contradictory Data
DV-03: Empty Dataset Reporting
DV-04: Inadequate Data Slicing

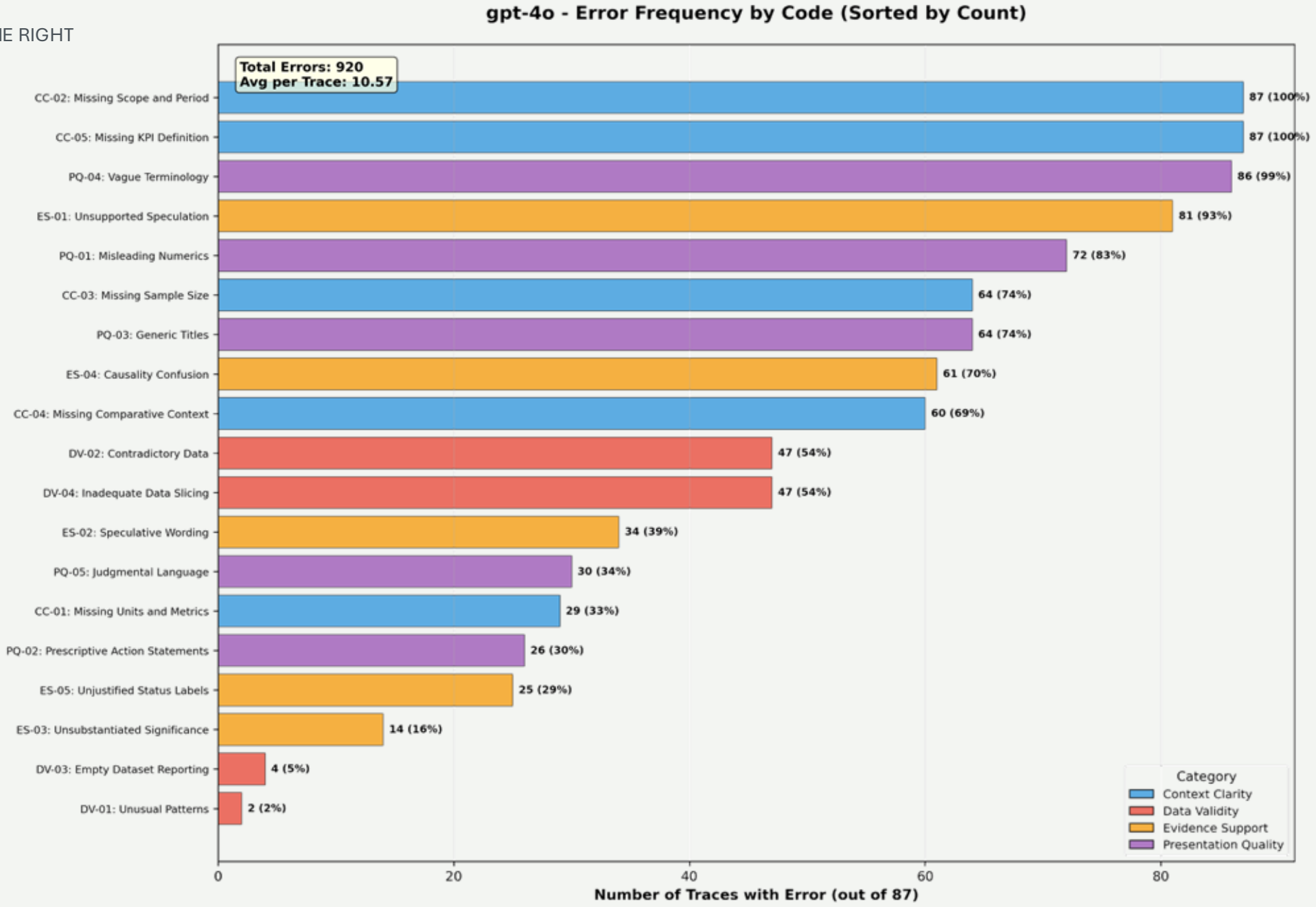
## 03 Evidence Support

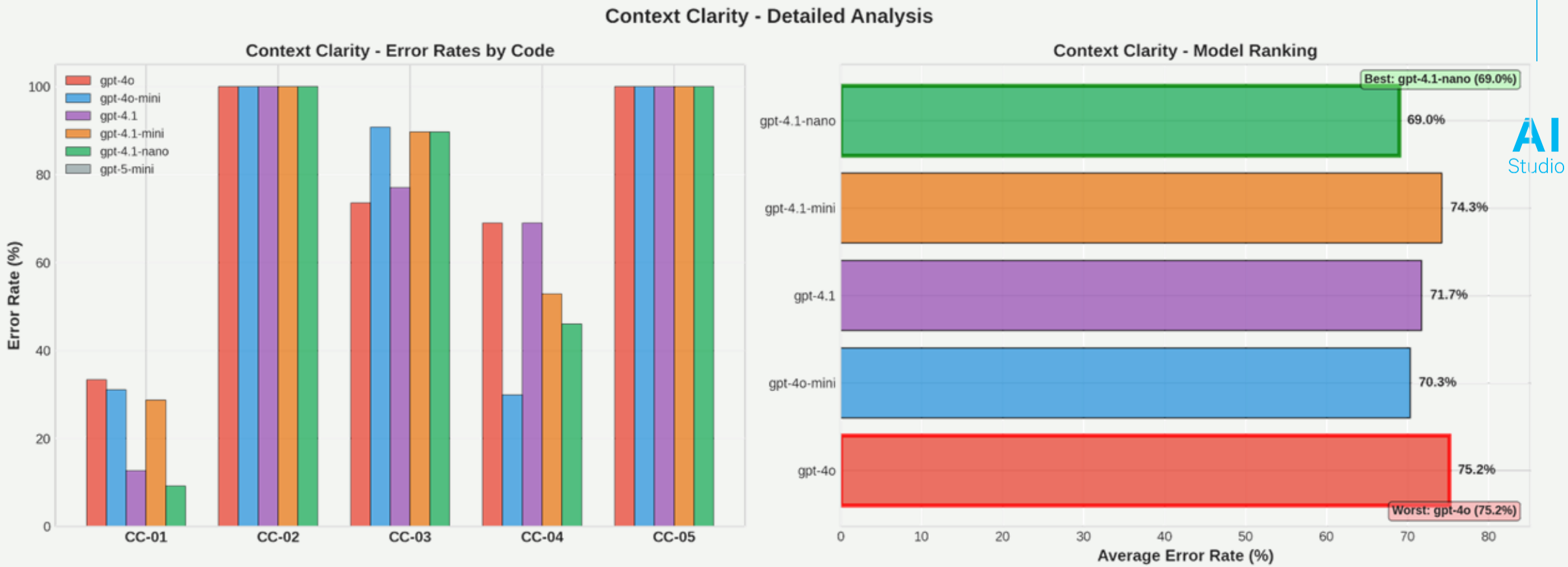
ES-01: Unsupported Speculation
ES-02: Speculative Wording
ES-03: Unsubstantiated Significance
ES-04: Causality Confusion
ES-05: Unjustified Status Labels

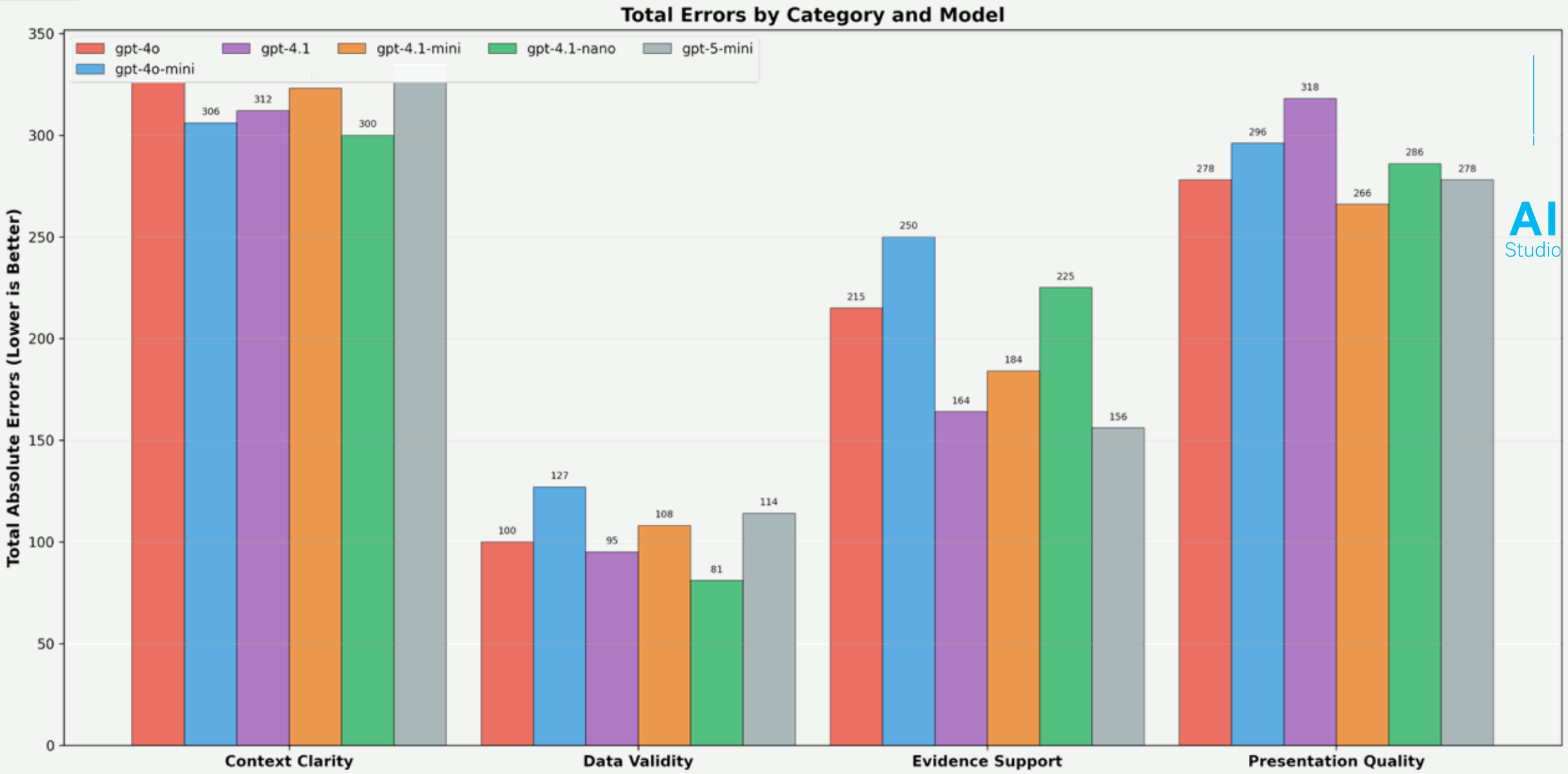
## 04 Presentation Quality

PQ-01: Misleading Numerics
PQ-02: Prescriptive Action Statements
PQ-03: Generic Titles
PQ-04: Vague Terminology
PQ-05: Judgmental Language

AI Studio


gpt-4o - Error Frequency by Code (Sorted by Count)
Total Errors: 920
Avg per Trace: 10.57
CC-02: Missing Scope and Period
CC-05: Missing KPI Definition
PQ-04: Vague Terminology
ES-01: Unsupported Speculation
PQ-01: Misleading Numerics
CC-03: Missing Sample Size
PQ-03: Generic Titles
ES-04: Causality Confusion
CC-04: Missing Comparative Context
DV-02: Contradictory Data
DV-04: Inadequate Data Slicing
ES-02: Speculative Wording
PQ-05: Judgmental Language
CC-01: Missing Units and Metrics
PQ-02: Prescriptive Action Statements
ES-05: Unjustified Status Labels
ES-03: Unsubstantiated Significance
DV-03: Empty Dataset Reporting
DV-01: Unusual Patterns
87 (100%)
87 (100%)
86 (99%)
81 (93%)
72 (83%)
64 (74%)
64 (74%)
61 (70%)
60 (69%)
47 (54%)
47 (54%)
34 (39%)
30 (34%)
29 (33%)
26 (30%)
25 (29%)
14 (16%)
4 (5%)
2 (2%)
Category
Context Clarity
Data Validity
Evidence Support
Presentation Quality
0
20
40
60
80
Number of Traces with Error (out of 87)


Context Clarity - Detailed Analysis
Context Clarity - Error Rates by Code
gpt-4o
gpt-4o-mini
gpt-4.1
gpt-4.1-mini
gpt-4.1-nano
gpt-5-mini
Error Rate (%)
0
20
40
60
80
100
CC-01
CC-02
CC-03
CC-04
CC-05
Context Clarity - Model Ranking
Best: gpt-4.1-nano (69.0%)
gpt-4.1-nano
69.0%
gpt-4.1-mini
74.3%
gpt-4.1
71.7%
gpt-4o-mini
70.3%
gpt-4o
75.2%
Worst: gpt-4o (75.2%)
0
10
20
30
40
50
60
70
80
Average Error Rate (%)


Total Errors by Category and Model
gpt-4o
gpt-4o-mini
gpt-4.1
gpt-4.1-mini
gpt-4.1-nano
gpt-5-mini
Total Absolute Errors (Lower is Better)
0
50
100
150
200
250
300
350
306
312
300
100
127
95
108
81
114
215
250
164
184
225
156
278
296
318
266
286
278
Context Clarity
Data Validity
Evidence Support
Presentation Quality
AI
Studio

# Result GPT-4.1-nano

GPT-4.1-nano **reduced Context Clarity errors by 6.2%** compared to our current model, while running **37% faster** and costing **96% less.**

6.2%

Better Context Clarity

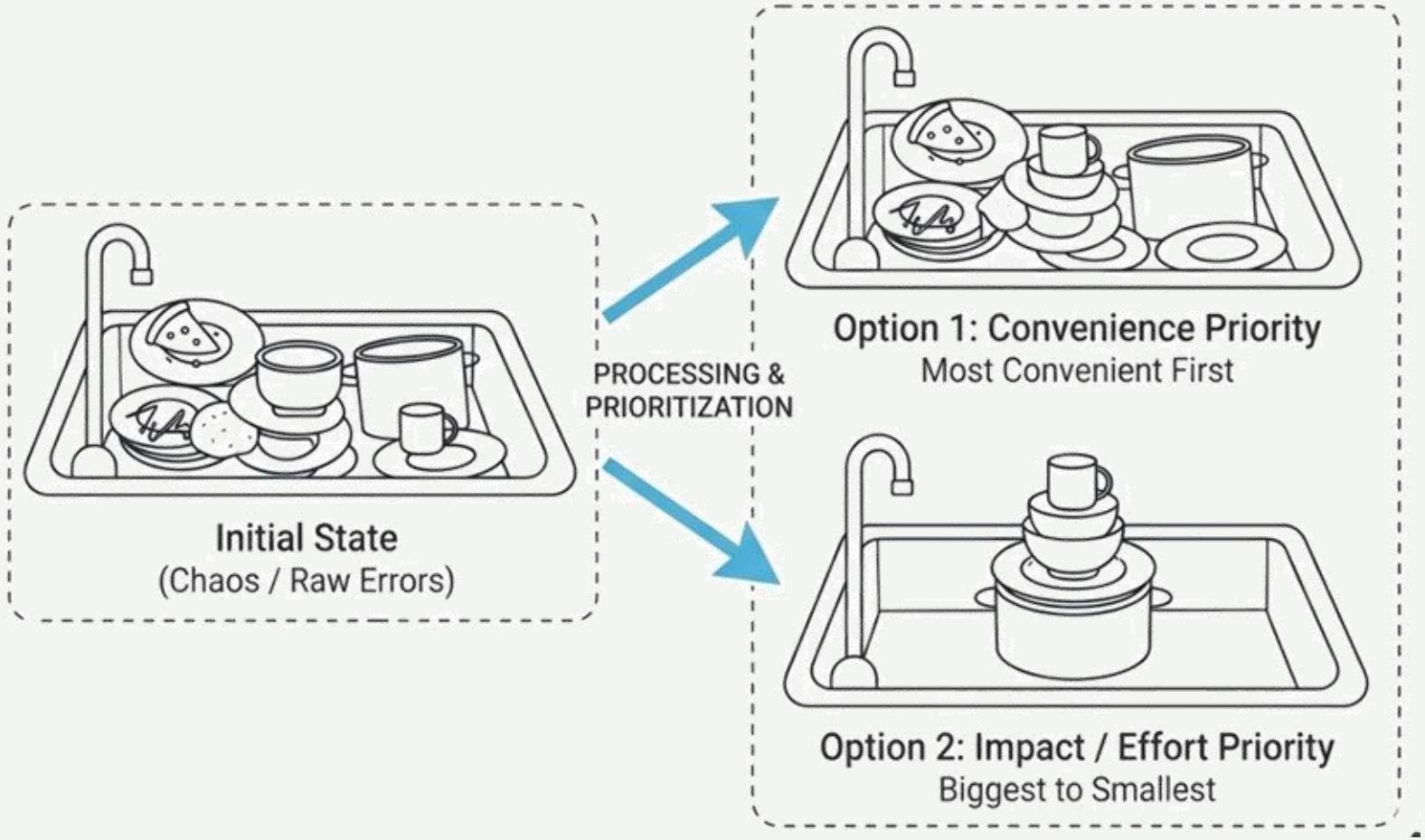
37%

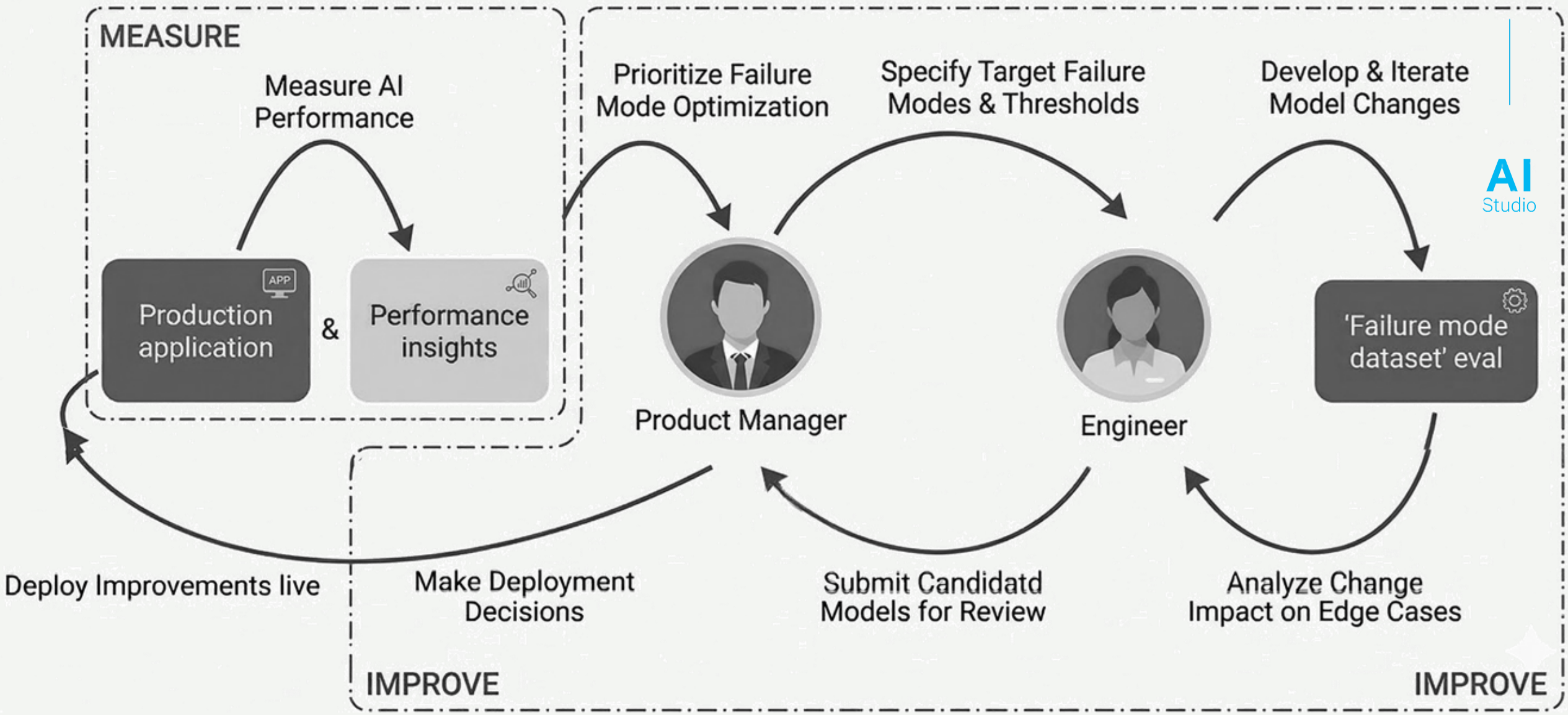
Faster

96%

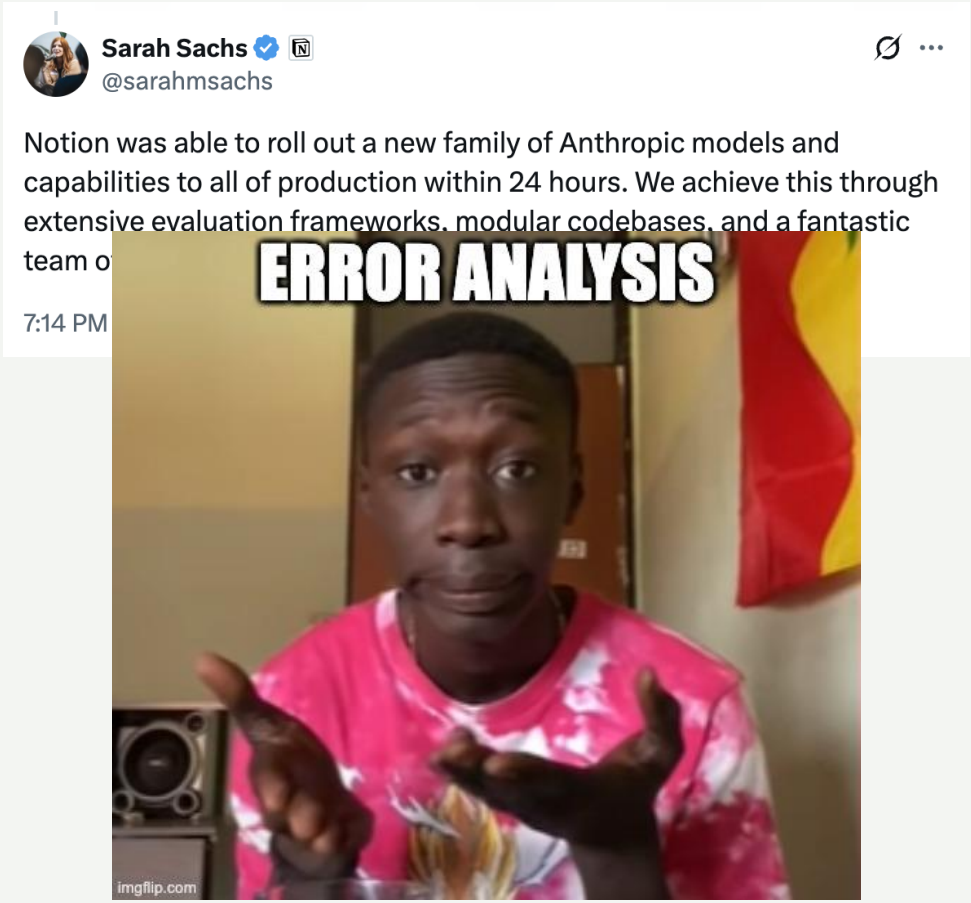
Cheaper

# Priority = Frequency × Impact × Effort




MEASURE
Measure AI Performance
Production application
&
Performance insights
Prioritize Failure Mode Optimization
Specify Target Failure Modes & Thresholds
Develop & Iterate Model Changes
AI Studio
Product Manager
Engineer
'Failure mode dataset' eval
Deploy Improvements live
Make Deployment Decisions
Submit Candidatd Models for Review
Analyze Change Impact on Edge Cases
IMPROVE
IMPROVE

# ~~Questions~~ Answers




- ✅ How can we find good evals?
  - With error analysis
- ✅ How can we quantify the product quality and track improvements over time?
  - Count failure modes / categories
- ✅ What is the most critical task to tackle next?
  - Prioritize failure modes for impact + cost + urgendc
- ✅ How can we be sure to release a new model?
  - Run evals, count failure modes, compare failures + duration

AI Studio

# In Depth Training

AI Studio

FEATURED IN Lenny's List

## AI Evals For Engineers & PMs

4.7 (765)

Hamel Husain ML Engineer with 20 years of experience

Shreya Shankar ML Systems & Applied AI Evals Researcher

View Syllabus

This course is popular.
25 people enrolled last week.

# Thank You!

AI Studio

## Let’s connect!

- github.com/MartinSeeler
- linkedin.com/in/martinseeler
- Martin.Seeler@blueyonder.com


BlueYonder
Move beyond boundaries™