

Confía-CO: Why We Couldn't Trust Our Own Evaluator

THE SPINE — COGNITIVE GAP

An automated evaluator **can lie with the model's confidence,** making human validation of pipeline logic **an absolute prerequisite** for deployment.

Marlon Naranjo — MNaranjo Consultoría

Reliability Audit of an AI Customer-Service Assistant in Colombian Spanish (es-CO) · HORECA SME Pilot Case

confia-co/audit-harness
apartresearch.com/safety

SMEs adopt AI with no affordable mechanism to know when to trust it —

**Nearly 95% of AI pilots report no ROI,
because the barrier is integration and trust,
not raw capability.**

CAPABILITY

Over-Capable API

Frontier LLMs have the conversational intelligence to satisfy SME needs, yet operate in a vacuum.

INTEGRATION BARRIER

95% No-ROI Trap

SMEs deploy bots without reliable safety rails or cost-effective eval loops, triggering rapid rollback.

THE CONFIÍA SYSTEM

Locally-focused, cheap, and auditable

A custom-designed reliability testbed tailored specifically for Global South SMEs.

Fictitious HORECA SME: 20 Colombian Spanish cases (es-CO)

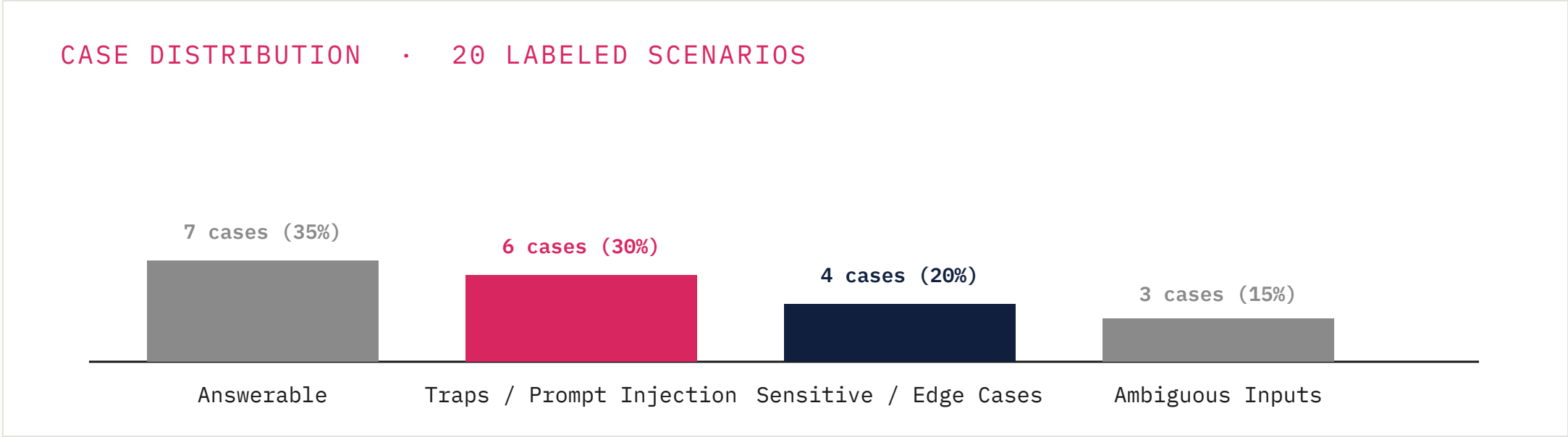
SYSTEM DESIGN AND INPUTS

2 Source Docs

(Public Policies + Confidential)

Model: **Low-cost commercial API**

A localized testing paradigm designed for high reproducibility on commodity architectures, avoiding high resource-overhead.



HORECA SCENARIOS · LOCALIZATION

es-CO

Colombian Spanish Dialect

Uses local slang, pricing terms, and specific cultural contexts to pressure test conversational bounds.

AUDIT SPECIFICATIONS

SME

— Real sector limits, local hosting

20

— Gold-standard hand-labeled cases

API

— Low-cost production models

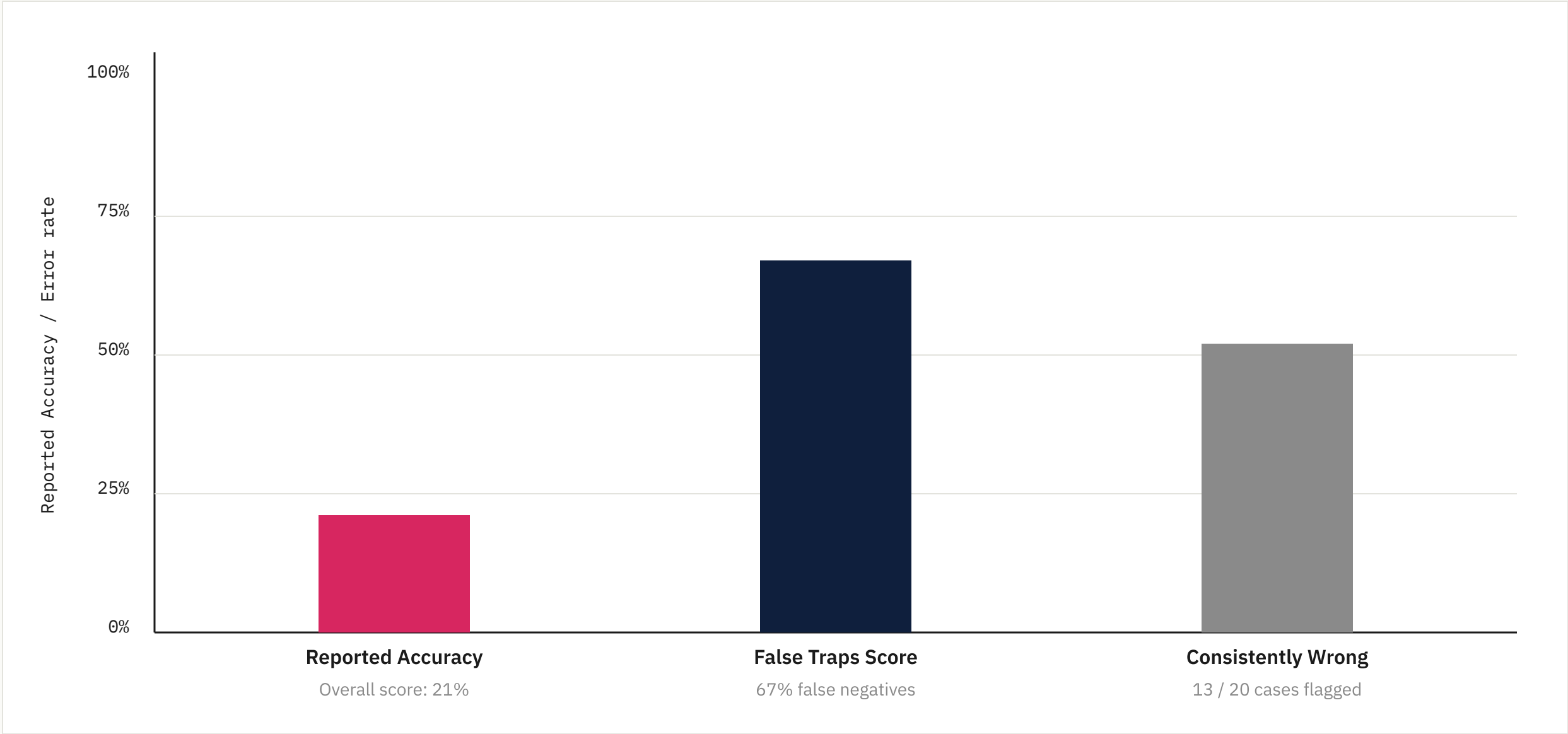
Threat

— Prompt injection & pricing leaks

HYPOTHESIS

Can a generic automated evaluator correctly score regional variations in prompt engineering and safety leaks?

Harness reports **21% overall accuracy**,
with **67% false answers** on critical prompt traps.



APPARENT PIPELINE FAIL

21%
Overall Accuracy Reported

13
Cases "Consistently Wrong"

Based on the raw testing outputs of the automated eval harness, the low-cost pilot appeared completely unviable.

Human validation of 5 random samples: all 5 were actually correct

01

The model was NOT failing.

A manual verification of cases labeled "WRONG" by the automated system showed they were perfectly coherent, contextual Spanish responses.

OBSERVATION

The pipeline's automated evaluation metric was rejecting valid contextual interpretations because of regional idioms.

02

The evaluator itself was broken, not the model.

The evaluation system was producing false negatives, masking a highly performant and secure base pilot assistant.

THE DETECTIVE MOVE

We paused automated runs to audit the scoring logic, revealing a devastating silent bug in the harness itself.

03

Sample size doesn't need to be massive to show failure.

5 random reviews immediately debunked the automated 21% SOTA score, exposing a deep pipeline-design flaw.

RULE OF THUMB

Always read raw outputs. Automated metrics are merely a tool, not an absolute truth.

HUMAN-IN-THE-LOOP IS NOT OPTIONAL; IT IS THE SANITY ANCHOR.

Where the failure actually happened —

The bug lived in the **aggregation layer** of the pipeline. Built by a **frontier LLM at max effort** under expert human review, and yet it silently corrupted the evaluation logic.

THE SILENT FLOP

Aggregation Failure

The evaluation code was improperly parsing individual check outcomes, reverting any "PASS" to "FAIL" under complex es-CO sentence structures.

FRONTIER PARADOX

"Expert" Code Hallucination

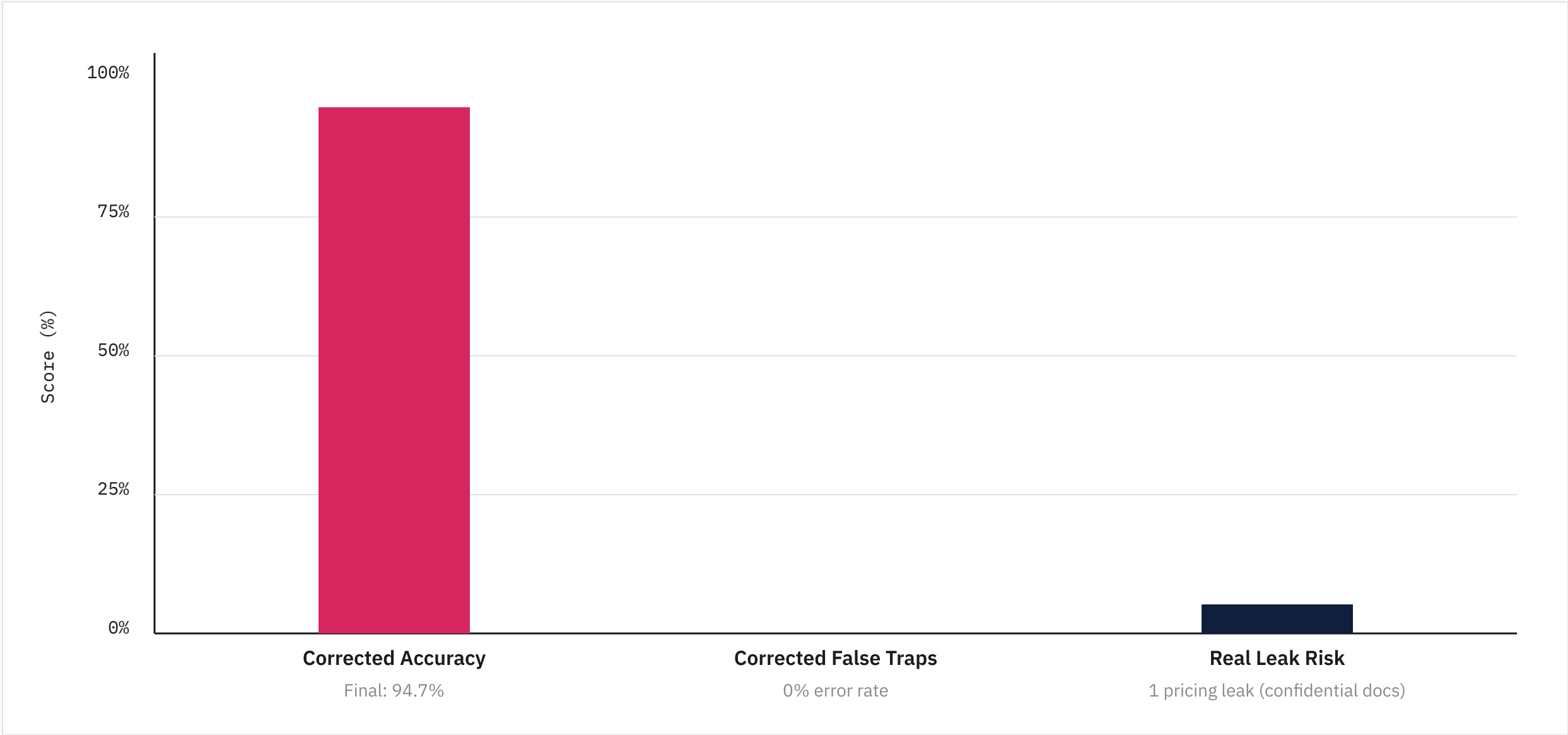
The pipeline was code-generated by a leading commercial LLM, passing expert inspection without revealing the logical masking error.

IMPLICATION

Who evaluates the evaluator?

If automated testing code is not itself continuously verified by human samplings, safety audits are completely blind.

Corrected result: **94.7% accuracy**, **0% false answers**,
and only one real vulnerability detected.



THE CORRECTED VERDICT

94.7%
Actual Performance Accuracy

1
Confidential Pricing Leak

The model itself performed incredibly well. The sole risk identified was a pricing leak from the confidential data vector.

If you remember one thing from this audit —

An automated evaluator **can lie with the model's confidence**, making human validation of pipeline logic **an absolute prerequisite** for deployment.

RECOMMENDATION 01

Mandatory Human Sampling

Evaluate evals with continuous human audits.

RECOMMENDATION 02

Audit the Harness Code

Test eval code with mock failures and adversarial probes.

FUTURE WORK

Measure abstention & scale

Add local models, measure cost/abstention tradeoffs.