
WHEN SAFETY LOOKS SUSPICIOUS: FALSE POSITIVES IN BRAZILIAN FINANCIAL FRAUD DETECTION¹

Gabriel Ferreira

Independent Researcher

<https://orcid.org/0000-0002-5368-0156>

With

Apart Research

Abstract

Every day, millions of Brazilians rely on Pix and CPF — and scammers know it, projecting the same urgent, government-coded language as routine financial notices. We built a contrastive benchmark of Brazilian-Portuguese messages where harmful and benign versions of the same scenario share vocabulary and structure but differ only in intent, then tested an LLM fraud analyst on 1,461 labeled messages and the full harmful and benign pair set.

We expected the model to miss disguised fraud. Instead, it sees fraud everywhere. It catches real scams reliably (95.9% recall) but clears legitimate Pix/CPF messages only 19.5% of the time — correctly telling the two apart in just 15.3% of pairs. Surprisingly, ordinary benign messages score worse than every actual manipulation tactic (53.2% vs. 94%+), suggesting the model is reacting to fraud-coded words, not reasoning about intent. A preliminary pilot, spanning flagship LLM models, shows the same pattern at every capability level.

For a country that runs on Pix and CPF, a filter this trigger-happy isn't just inaccurate — it risks alert fatigue and eroded trust in legitimate messages, a more immediate threat than the usual "model misses the scam" framing. We release the dataset, an evaluation harness, and a full per-category breakdown

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

Pix, Brazil's instant-payment rail, and the CPF (Cadastro de Pessoa Física, the national taxpayer ID) are the two pieces of financial infrastructure most Brazilians interact with daily. They are also the two most exploited surfaces for financial fraud in the country: fake suspended Pix or irregular CPF notices, fake tax-debt ("Dívida Ativa") collection messages, and fake bank or government support contacts all route through the same vocabulary that legitimate Pix, tax, and benefits communication uses. Fraud in this space does not look like a jailbreak prompt or an adversarial suffix — it looks like an ordinary message about an ordinary financial obligation, sent at a moment (tax season, Black Friday, year-end) when that obligation is genuinely top of mind [1].

This is the threat model we evaluate: "Can a model deployed as a fraud analyst distinguish harmful from benign messages when written in the same cultural register?" AI Safety or jailbreak benchmarks generally do not test this distinction. A benchmark that only tests obvious scams will systematically overstate how safe a fraud-detection system is for users in markets like Brazil's, where the most damaging scams look exactly like typical bank or government correspondence.

We built a benchmark to test this directly, and the result reframes the problem. We expected — following the common framing that AI safety evaluations miss fraud embedded in culturally normal workflows — that the failure mode would be the model missing harmful messages dressed up as routine Pix/CPF communication (false negatives). Instead we find the opposite: **the model is highly sensitive to anything that mentions Pix, CPF, "regularização," or government benefits, and over-flags the legitimate half of our dataset almost as often as it correctly catches the harmful half.** Our main contributions are:

1. A contrastive benchmark of Brazilian-Portuguese fraud scenarios with 914 message pairs (Pix scams, CPF/tax scams, fake bank/government support, WhatsApp account takeover, lottery and benefit scams, and others), where each pair shares a scenario and vocabulary but differs in a labeled *distinguishing feature* that separates harmful intent from a benign equivalent.

2. An end-to-end evaluation pipeline that combines benchmark construction, contrastive evaluation, LLM-based fraud detection, and performance analysis, yielding per- category, difficulty, and technique analysis alongside pair-level and harmful- and benign-side accuracy.

3. An empirical analysis and results showing tested models' failure on culturally embedded fraud is a false-positive problem (e.g., over-flagging benign Pix/CPF/government messages), not a false-negative blind spot, the opposite of our initial intuition. Harmful messages are caught at a 95.9% rate; genuinely benign ones are cleared at only 19.5%. A preliminary check across four additional judge models (see Sec. 4) suggests this is not specific to one model.

2. Related Work

Most public LLM safety and red-teaming benchmarks are built around adversarial prompting — jailbreaks [2], prompt injection [3], [4], and refusal-bypass attempts — where the model itself is the target of the attack. Our benchmark targets a different failure surface: the model acting as a *content classifier* over messages written by a third party, where the harmful and benign classes are deliberately constructed to share vocabulary, structure, and urgency cues. This is closer to fraud/spam classification literature than to jailbreak literature, but most fraud-classification benchmarks are built on English-language phishing/spam corpora [5], [6] (e.g., generic "your account is locked" templates) and do not test region-specific financial infrastructure such as Pix or CPF-based fraud, which dominates real-world scam traffic in Brazil according to industry security write-ups [1], [7]–[9]. After an initial search, we are not aware of a public benchmark that pairs harmful and benign Brazilian-Portuguese financial messages on the same scenario to isolate whether a model is reasoning about actual harmful intent versus reacting to topic/vocabulary. This work addresses this specific gap:

When would someone use this over a generic safety benchmark? When evaluating fraud-detection or content-moderation systems for deployment in Brazil's Pix/CPF financial context, where false positives on legitimate messages carry real trust and product costs.

What does this benchmark show that others would not? That high fraud-detection accuracy can come purely from topic overfitting rather than genuine intent discrimination—a blind spot present in any benchmark that doesn't pair harmful and benign messages in the same scenario.

3. Methods

We present a benchmark for financial fraud detection in Brazilian Portuguese, built through a four-stage and fully reproducible pipeline: benchmark construction, contrastive evaluation, LLM-based fraud assessment, and fine-grained performance analysis. The benchmark tests whether language models can distinguish fraudulent messages from legitimate ones that share similar linguistic and contextual features. Unlike traditional scam datasets, our contrastive approach measures reasoning about *intent* rather than reliance on surface-level cues — by pairing harmful and legitimate communications that share the same scenario, vocabulary, and style (see more details on Appendix A.3–A.6).

3.1 Models, Tools, and Experimental Design

Models. All benchmark generation, contrastive-pair construction, and primary evaluation experiments were performed using the [liquid/lfm-2-24b-a2b](#) model through OpenRouter. This model was selected because it offers a favorable capability-to-cost ratio, allowing large-scale generation and evaluation within hackathon resource constraints. Unless stated otherwise, all results use [liquid/lfm-2-24b-a2b](#). A small exploratory comparison was conducted using five models

spanning free-tier and flagship offerings to assess whether observed behaviors generalized across capability tiers.

Tools. The benchmark, evaluation harness, and analysis pipeline were implemented in Python using a unified OpenRouter client. Results were stored as append-only *JSONL* artifacts to support resumability, fault tolerance, and reproducibility. Analysis was performed using standard data-science tooling and custom evaluation scripts.

Dataset. The benchmark consists of synthetic Brazilian-Portuguese financial communications covering common fraud categories observed in Brazil, including Pix scams, phishing, government impersonation, fake benefits, investment fraud, extortion, and related social-engineering attacks. The final benchmark contains 1,470 generated messages, while the contrastive corpus contains 914 harmful–benign pairs.

3.2 Key Design Decisions

Our benchmark design was centered around the following decisions:

1. **Contrastive evaluation rather than standalone classification.** Instead of evaluating messages in isolation, harmful messages were paired with benign counterparts describing the same scenario. This allows measurement of whether a model reasons about intent or merely reacts to fraud-associated vocabulary.
2. **Brazil-specific fraud scenarios.** Categories were selected to reflect scams commonly encountered by Brazilian users, particularly those involving Pix transfers, CPF-related communications, government benefits, and financial institutions.
3. **Manipulation-technique annotations.** Harmful examples were explicitly tagged with techniques such as fear, urgency, authority, greed, trust exploitation, scarcity, social proof, and emotional manipulation. This enabled analysis beyond category-level accuracy.
4. **Difficulty stratification.** Messages were generated at easy, medium, and hard difficulty levels to evaluate robustness as scam signals become less explicit.
5. **Structured model outputs.** The fraud analyst was required to return a structured JSON schema including harmfulness, confidence, risk category, risk score, evidence, and explanation, enabling both quantitative and qualitative analysis.

4. Results

4.1 Overall Performance Analysis

Across the 1,461 parseable benchmark messages, the fraud analyst achieved an overall accuracy of 79.5% (1,162/1,461). A few responses that could not be parsed, were excluded from analysis.

While this aggregate score suggests reasonably strong performance, contrastive evaluation reveals a substantial asymmetry between harmful and benign messages (see Table 1).

Table 1 — Contrastive accuracy: harmful side vs. benign side vs. full pair

Metric	Correct/Total	Accuracy
Harmful side (genuinely harmful messages correctly flagged)	860/897	95.9%
Benign side (genuinely benign messages correctly cleared)	176/902	19.5%
Full pair (both sides of the same scenario correct)	140/914	15.3%

The model correctly identifies most harmful messages (95.9%) but correctly clears only 19.5% of benign messages. Only 15.3% of contrastive pairs are classified correctly on both sides.

4.2 Per-Manipulation-Technique Performance Analysis

Results are computed over the full benchmark dataset (see Table 2)

Table 2 — Accuracy by labeled manipulation technique (dataset-wide, not pairs-only)

Technique	Correct / Total	Accuracy
fear	148 / 148	100.0%
urgency	89 / 89	100.0%
emotional_manipulation	30 / 30	100.0%
scarcity	41 / 41	100.0%
authority	194 / 195	99.5%
greed	173 / 179	96.6%
trust_exploitation	168 / 178	94.4%
social_proof	23 / 45	51.1%
none (benign, no manipulation)	295 / 554	53.2%

*Accuracy exceeded 94% for all manipulation techniques except **social_proof** (51.1%). The weakest category was benign messages (**none**, 53.2%); almost flat performance across difficulty levels (1.8-point spread).*

4.3 Per-Category Performance Analysis

The lowest-performing categories were *lottery_scam* (59.5%), *whatsapp_cloning* (60.8%), and *pix_scam* (66.7%). The highest-performing categories were *romance_scam*, *investment_scam*, and *fake_support_team* (all with 96.0%).

No statistically meaningful conclusions could be drawn from our per-category analysis.

5. Interpretation

The results suggest that the dominant failure mode is not failure to detect fraudulent content but to distinguish fraudulent from legitimate messages that share the same vocabulary. Three observations support this interpretation:

1. Harmful-side accuracy is consistently high (95.9%), while benign-side accuracy is substantially lower (19.5%).
2. Benign messages without any manipulation technique achieve lower accuracy (53.2%) than nearly every manipulation category.
3. Performance remains relatively stable across difficulty levels, despite the expectation that genuinely subtle fraud should be harder to identify.

LLM models rely heavily on fraud-associated lexical and contextual cues (e.g., Pix, CPF, government communications) rather than reasoning about the specific intent of the message.

6. Limitations

To assess the asymmetry observed in the primary evaluation model, we conducted a small pilot using four additional models while holding a small sample of 15 pairs per model. All five models showed a consistent asymmetry (i.e., they exhibited substantially higher harmful-side accuracy than benign-side accuracy). Although the pilot is too small to establish statistical significance, all

tested models exhibited substantially higher harmful-side than benign-side accuracy, suggesting that the effect is not unique to a single model. As any exploratory study, our work presents many limitations:

1. Benchmark generation and primary evaluation were performed using the same model family, introducing the possibility of generation–evaluation coupling.
2. Multi-model comparison uses only 10–15 contrastive pairs per model and should be interpreted as exploratory rather than confirmatory.
3. Most distinguishing-feature categories contain only one or two examples, preventing reliable fine-grained analysis.
4. The same model family both generates and evaluates benchmark content, which potentially introduces generation–evaluation coupling.

7. Future Work

Future work should evaluate larger model sets, incorporate human-authored examples, and perform significance testing on full-scale multi-model evaluations. For example:

1. Extend the preliminary five-model pilot in 4 into a full sweep over the complete contrastive-pair set (and ideally over models not involved in generating the dataset) to confirm with statistical confidence whether the false-positive-dominant pattern generalizes across capability tiers;
2. Include real-world Brazilian scam/legitimate message pairs rather than relying solely on synthetic generation;
3. Extend the contrastive-pair methodology to other Global South financial-infrastructure contexts (e.g. M-Pesa in Kenya, UPI in India) to test whether topic-triggering false positives are a general pattern for LLM- based fraud classifiers trained on US/EU-centric infrastructure.

8. Conclusion

We built a contrastive benchmark of Brazilian-Portuguese financial fraud scenarios to test whether an LLM fraud analyst can distinguish genuinely harmful Pix/CPF/government-themed messages from benign ones that share the same vocabulary and structure.

We expected to find the model missing fraud disguised as routine financial communication. Instead, we find the model is good at catching real fraud (95.9% recall) but poor at clearing legitimate communication that merely resembles it (19.5% specificity), with a full pair-discrimination accuracy of only 15.3%. Benign content with no manipulation tactic at all scores among the worst of any technique bucket (53.2%, second only to social proof fraud at 51.1%), a strong evidence that the dominant failure mode here is over-triggering on surface vocabulary rather than missing disguised intent. A preliminary check across multiple judge models, from a small free-tier to

current flagship LLM models, shows the same qualitative asymmetry at every tier tested, suggesting this is not an idiosyncrasy of one model.

For deployment in markets where Pix- and CPF-linked communication is part of millions of people's daily financial life, this asymmetry is the safety-relevant finding worth further study.

9. References

- [1] ABES, "**Quase 7 milhões de Tentativas de Fraude Foram Registradas no 1º Semestre de 2025; Setor Bancário é Principal Alvo**," ABES, Oct. 8, 2025. [Online, Visited on 2026-06-21].
<https://abes.org.br/quase-7-milhoes-de-tentativas-de-fraude-foram-registradas-no-1o-semester-de-2025-setor-bancario-e-principal-alvo/>
- [2] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "**JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models**," in Proc. of NeurIPS, Datasets and Benchmarks Track, 2024.
- [3] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "**Formalizing and Benchmarking Prompt Injection Attacks and Defenses**," in Proc. of USENIX Security Symposium, 2024.
- [4] Q. Zhan, Z. Liang, Z. Ying, and D. Kang, "**InjecAgent: Benchmarking Indirect Prompt Injections in Tool-integrated Large Language Model Agents**," in arXiv preprint arXiv:2403.02691, 2024.
- [5] P. Mendes, E. Maia, and I. Praça, "**MeAJOR Corpus: A Multi-source Dataset for Phishing Email Detection**," arXiv preprint arXiv:2507.17978, 2025.
- [6] R. Toth, N. Gruschka, and T. Bisztray, "**The Phish, The Spam, and The Valid: Generating Feature-rich Emails for Benchmarking LLMs**," arXiv preprint arXiv:2511.21448, 2025.
- [7] TI Inside, "**Scams via Pix (Brazil's Instant Payment System) Are Projected to Reach 28 million Cases in Brazil by 2025, according to a report**," TI Inside Online, Nov. 21, 2025. [Online, Visited on 2026-06-21].
<https://tiinside.com.br/en/21/11/2025/golpes-via-pix-somam-28-milhoes-de-casos-no-brasil-em-2025-aponta-relatorio/>
- [8] ScamWatchHQ, "**Brazil scams 2026: The Pix Fraud Paradox**," 2026. [Online, Visited on 2026-06-21].
<https://scamwatchhq.com/brazil-scams-2026-pix-fraud-phantom-hand-pcc/>
- [9] QED Investors, "**The Frontlines of Fraud: How Brazil is Becoming a Global Testbed for Financial Crime Prevention**," 2025. [Online, Visited on 2026-06-21].
<https://www.qedinvestors.com/blog/the-frontlines-of-fraud-how-brazil-is-becoming-a-global-testbed-for-financial-crime-prevention>

Appendix A

A.1. Code, Data, and Reproducibility

The project is designed to be reproducible end-to-end. All benchmark generation, pair construction, evaluation, and analysis stages are implemented as independent, resumable commands. Intermediate artifacts are stored as JSONL files, allowing interrupted runs to continue without recomputing completed work. A researcher with (i) access to the source code hosted on GitHub, (ii) an OpenRouter API key, and (iii) an evaluation runbook, can regenerate the benchmark, reproduce all reported metrics, and evaluate alternative models using the same evaluation harness. The repository, hosted on Github, includes all generation scripts, evaluation prompts, analysis scripts, intermediate datasets, and final result artifacts.

A.2. LLM Usage Statement

Claude (Anthropic) was used to help explore the repository, compute and cross-check the quantitative results reported here directly from the project's generated JSONL/JSON output files, and draft sections of this report based on my outline story notes and the template rules. All numeric claims in this report were read directly from [docs/evaluation_report.json](#) (regenerated against the full 914-pair [docs/contrastive_pairs.jsonl](#)) and the underlying JSONL result files rather than generated or estimated by the LLM.

A.3. Benchmark Construction

The benchmark consists of synthetic Brazilian-Portuguese messages spanning a broad range of fraud and social-engineering scenarios, including Pix scams, phishing, government impersonation, fake benefit notifications, loan fraud, investment scams, job scams, extortion attempts, misinformation campaigns, and related categories. Each message is annotated with:

- a fraud category,
- a binary label (harmful or benign),
- a difficulty level (easy, medium, or hard), and
- a manipulation technique for harmful examples (e.g., fear, urgency, authority, greed, trust exploitation, scarcity, social proof, or emotional manipulation).

Benchmark generation yielded 1,470 labeled messages. Difficulty levels were designed to vary the explicitness of fraudulent

A.4. Contrastive Pair Construction

To isolate semantic intent from surface-level cues, we build a contrastive set: for each harmful message, a benign counterpart is generated preserving the same scenario, vocabulary, tone, and

structure, so the two differ mainly in intent rather than obvious fraud indicators. Each pair is annotated with a distinguishing feature — the minimal semantic difference between versions (e.g., an external verification link vs. an official banking channel, or a payment request vs. an explicit no-payment statement). This yielded 914 contrastive pairs across dozens of fraud scenarios.

A.5. LLM-Based Assessment Setup and Performance

Messages were evaluated by an LLM fraud analyst that returned structured classifications and risk assessments. After excluding malformed responses, evaluation covered 1,461 benchmark messages and 897 harmful / 902 benign contrastive examples. We report overall accuracy, category-, difficulty-, and technique-level breakdowns, and contrastive metrics including harmful-side, benign-side, pair-level, and distinguishing-feature accuracy.

A.5.1. Unsuccessful Approaches and Iterations

Our early experiments with standalone classification and unpaired benchmarks provided limited insight into model reasoning, motivating the contrastive-pair design. The direct generation occasionally produced safety refusals that were filtered from the dataset. A multi-model generation/evaluation setup was considered but deferred due to resource constraints.

A.6. Example benchmark messages

A.6.1 Harmful message example

*"URGENTE – Receita Federal: Prezado(a) [Nome], CPF *123.456.789-*. Foi identificada uma pendência fiscal em seu cadastro. Regularize imediatamente para evitar suspensão do CPF e bloqueio bancário. Acesse: receitafederal-consulta.org"*

Tax/CPF fraud, urgency + authority technique: impersonates Brazil's federal tax authority, uses the recipient's real name and partial CPF for credibility, threatens CPF suspension and bank blocking, and links to a non-`.gov.br` lookalike domain — a textbook combination of the fraud signals identified in our scenario design: urgency, authority impersonation, real personal data, and an unofficial link).

A.6.2. Benign (Harmful-equivalent) message example

*"Banco do Brasil: Olá, [Nome]. Confirmamos que sua chave Pix (CPF *123.456.789-*) está ativa e funcionando normalmente. Não é necessária nenhuma ação de sua parte. Em caso de dúvidas, acesse o aplicativo oficial do banco ou ligue para nossa central de atendimento."*

Same scenario family, contrastive design: same bank-notification register and partial-CPF reference as A.1, but confirms normal status, requests no action, contains no urgency or suspension threat, and points only to the official app/phone channel with no external link.

A.6.3. Fraud Analyst Output Schema

Each evaluation returns a JSON object of the form:

```
{
  "harmful": true,
  "confidence": 0.95,
  "risk_category": "phishing",
  "risk_score": 92,
  "manipulation_techniques": ["urgency", "authority"],
  "evidence": [...],
  "explanation": "..."
}
```

A.7. Data Dictionary

Each file in [docs/](#) is line-delimited JSON except [evaluation_report.json](#), which is a single nested JSON object. Field values are restricted to the fixed vocabularies below; any value outside these lists indicates a generation or parsing error.

1) [docs/benchmark_dataset.jsonl](#): one row per generated message (1,470 rows).

Field	Type	Meaning / allowed values
id	string (UUID)	Unique identifier for the message.
category	string	Scam/communication category. One of: pix_scam, phishing, romance_scam, investment_scam, impersonation, job_scam, lottery_scam, extortion, misinformation, whatsapp_cloning, fake_profile_urgency, fake_support_team, ...
label	string	Ground-truth class: harmful or benign.
difficulty	string	How disguised the message is: easy, medium, or hard.
seed	string	Short Portuguese description of the scenario used to prompt generation (e.g. "Notificação falsa de suspensão do Pix").
index	integer	Position of this message within its (category, label, difficulty) generation batch; not a global rank.
content	string	The Brazilian-Portuguese message text itself — this is what the fraud analyst evaluates.

manipulation_technique	string	Persuasion tactic used. One of fear, urgency, authority, greed, trust_exploitation, scarcity, social_proof, emotional_manipulation for harmful rows, or none for benign rows (benign messages contain no manipulation tactic by construction).
------------------------	--------	--

2) [docs/contrastive_pairs.jsonl](#): one row per harmful/benign pair (914 rows). [harmful](#) and [benign](#) are nested objects with the same shape as a dataset row minus [id](#)'s global uniqueness constraint.

Field	Type	Meaning / allowed values
category	string	Same vocabulary as Table A.7.1; shared by both sides of the pair.
scenario	string	Same as seed in the dataset file — the shared scenario description.
harmful	object	The harmful-side message (id, category, label="harmful", difficulty, technique, content). Note the key is "technique" here, not "manipulation_technique".
benign	object	The benign-side message (id, category, label="benign", difficulty, technique, content) — same scenario and vocabulary, intent removed.
distinguishing_feature	string	Free-text label naming the specific cue that separates the two sides (e.g. pix_verification_link); not drawn from a fixed enum, used only for qualitative grouping in Sec. 4.

3) [docs/benchmark_dataset_results.jsonl](#) and [docs/contrastive_results.jsonl](#): one row per analyst verdict (1,461 and 1,799 rows respectively, the latter covering both sides of every pair). Same schema in both files.

Field	Type	Meaning / allowed values
id	string (UUID)	Matches the id of the source message in the dataset/pairs file.
ground_truth	string	harmful or benign, copied from the source message's label.
raw_response	string	The fraud analyst's full raw model output (may include Markdown code fences) before JSON parsing.
parsed	object or null	The parsed verdict (null if raw_response could not be parsed as JSON — these are the "unparseable" entries referenced in Sec. 4). When present, it contains the fields below.
parsed.harmful	boolean	The analyst's verdict — true if it judged the message harmful. Compare against ground_truth to score correctness.
parsed.confidence	float (0–1)	Analyst's self-reported confidence in its verdict.

parsed.risk_category	string	Free-text risk label assigned by the analyst (e.g. phishing); not a fixed enum, model-generated.
parsed.risk_score	integer (0–100)	Analyst's numeric risk score.
parsed.manipulation_techniques	array of strings	Techniques the analyst believes are present; free-text, not constrained to the dataset's manipulation_technique vocabulary (the analyst can name techniques like credential_harvesting not in the generation taxonomy).
parsed.evidence	array of strings	Verbatim excerpts from the message the analyst cites as evidence.
parsed.explanation	string	Analyst's free-text rationale for the verdict.

4) [docs/evaluation_report.json](#): single object, the output of the analysis stage.

Field	Type	Meaning
overall.correct overall.total overall.accuracy overall.unparseable	int int float int	Dataset-wide accuracy (parsed.harmful == (ground_truth == "harmful")) and count of rows where parsed was null.
contrastive.pairs_evaluated contrastive.pairs_total	int int	How many of the 914 pairs had both sides parseable vs. the full pair count.
contrastive.pair_accuracy	float	Fractions of pairs where both the harmful and benign side were classified correctly .
contrastive.harmful_side contrastive.benign_side	object	Each contains correct, total, accuracy for that side only.
contrastive.by_distinguishing_feature	object	Keyed by distinguishing_feature string; each value gives per-feature pair accuracy (sparse — most keys have 1–2 pairs, see Limitations).
by_category	object	Keyed by category; each value gives correct/total/accuracy for that category, dataset-wide.
by_difficulty	object	Keyed by easy/medium/hard; same correct/total/accuracy shape.
by_manipulation_technique	object	Keyed by the technique vocabulary above (plus none for benign/no-technique rows); same correct/total/accuracy shape .

5) [docs/sweep/*.json](#): per-model outputs from the five-model pilot (Sec. 4, Table 3 source), one file per model, same [contrastive.*](#) shape as [evaluation_report.json](#) but computed over the 10–15 pairs sampled for that model rather than the full pair set.

Field	Type	Meaning
One file per model. Same contrastive.* shape as docs/evaluation_report.json (Table A.7.4), but computed over the 10–15 pairs sampled for that model rather than the full pair set.		