

Grading the Graders

We grade SI systems, and we grade the benchmarks those systems are measured against. This page is one audit, done in the open: a public benchmark graded by our method, and our own corpus graded by the same method, on the same page.

<div>Subject</div> <div>FinanceBench, the public benchmark published by Patronus AI</div>	<div>Sample</div> <div>30 items, seeded and stratified, graded against the full SEC filings</div>	<div>Method</div> <div>Two-pass blind audit with an ABSTAIN band and adjudication (house method v1.0)</div>
<div>Findings delivered</div> <div>October 16, 2026, privately to the publisher, ahead of publication</div>	<div>Published</div> <div>October 30, 2026</div>	<div>Status</div> <div>Final, subject to the amendments ledger in section 7</div>

EXECUTIVE SUMMARY

AITrailblazer audited a 30-item sample of FinanceBench, grading every gold answer against the full SEC filing the item cites. The roll-up: 27 CLEAN, 2 AMBIGUOUS, 1 WRONG_GOLD, so 90.0% of the sampled gold answers stand exactly as published. A second panel grades our own SEC filing corpus by the identical method, with our disagreements and relocks printed at the same size. This report is the first in the Benchmark Audit Series; its amendments ledger (section 7) is a standing part of the document.

PDF edition of Audit Report 2026-001 · Canonical version: aitrailblazer.com/grading-the-graders

Contents: What we do, and why this page exists · 1. The method · 2. The flaw taxonomy · 3. Panel one: a public benchmark sample · 4. Panel two: our own corpus · 5. What the two panels show together · 6. Close · 7. Amendments and corrections

AITrailblazer - October 30, 2026. Every figure in this paper carries its counts and its sample size.

What we do, and why this page exists

AITrailblazer is an independent evaluation practice. We grade SI systems that read financial documents, and we grade the benchmarks those systems are scored against. The work is a service: benchmark audits and eval construction, fixed fee, findings to the owner first.

Why grade a benchmark? Because buyers choose SI systems on benchmark scores, and a benchmark score is only as good as the benchmark's own answers. Those answers are almost never re-checked against the source documents. One wrong gold answer quietly misgrades every system ever measured against it, and nobody downstream can see it happen. So the audit starts one level down: grade the graders.

What follows is one such audit, published in full. Panel one grades a 30-item sample of FinanceBench, the public benchmark published by Patronus AI. Panel two grades our own SEC filing corpus by the identical method, with our own disagreements printed at the same size as theirs.

1. The method

Two grading passes, run separately, blind to each other, under the same written rubric, with an explicit ABSTAIN band: a grader who cannot settle an item from the evidence says so instead of forcing a label.

Splits and abstains go to a fresh adjudicator who sees the two classifications anonymized, never their origin.

Ground truth is a construction: every item is verified against the full source filing, exhibits included, never against an excerpt. A discrepancy that exists only in an excerpt is classified as an excerpt artifact and is never counted as a benchmark flaw.

Agreement is the quality metric: raw agreement and Cohen's kappa are reported together, with the class breakdown, because a bare kappa on a skewed sample misleads in both directions.

2. The flaw taxonomy

Each audited item receives one final class: CLEAN, WRONG_GOLD, AMBIGUOUS, EVIDENCE_MISMATCH, CONTAMINATED, STALE, or DIFFICULTY_MISLABELED. EXCERPT_ARTIFACT is reported separately and is never counted as a flaw.

Severity roll-up: percent clean, percent fixable (ambiguous, evidence mismatch, difficulty mislabeled), percent invalid (wrong gold, contaminated, stale).

3. Panel one: a public benchmark sample

The sample is 30 items drawn from the public set of FinanceBench (published by Patronus AI), stratified 10 / 10 / 10 across its three question types, drawn by a seeded, reproducible procedure. Rates describe this sample only; the gated set was not accessed and nothing here is extrapolated to it.

Table 1: FinanceBench sample, final classes (n = 30)

Final class	Count	Share
CLEAN	27	90.0%
AMBIGUOUS	2	6.7%
WRONG_GOLD	1	3.3%
All other classes	0	0.0%

Roll-up: 90.0% clean, 6.7% fixable, 3.3% invalid (n = 30).

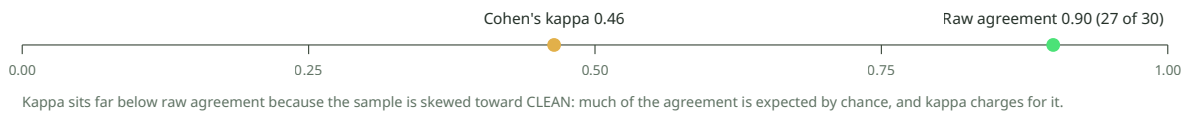
THE SAMPLE, ITEM BY ITEM



Each square is one audited item from the public FinanceBench set, graded against the full SEC filing. n = 30.

Grader agreement: the two passes agreed on 27 of 30 items (90.0%); Cohen's kappa is 0.4643. The kappa reads low against 90% raw agreement because the sample is heavily skewed toward CLEAN. The three disagreements were settled by the adjudicator as two AMBIGUOUS and one CLEAN. The single WRONG_GOLD item was called WRONG_GOLD by both passes independently; it was never a disagreement.

AGREEMENT, TWO WAYS OF COUNTING IT



4. Panel two: our own corpus, graded the same way

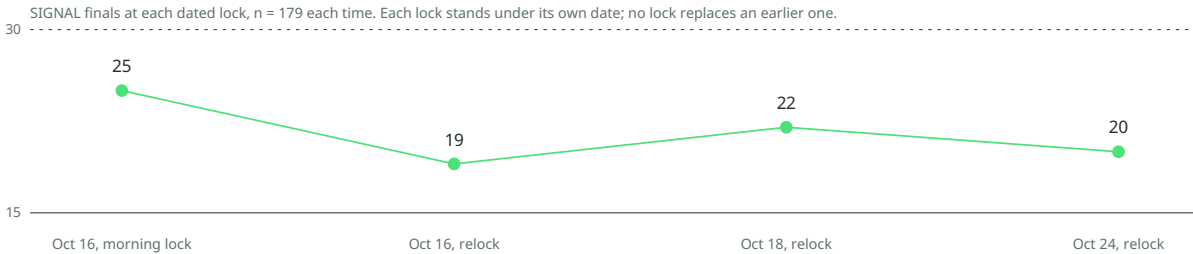
The same protocol runs on our own SEC filing corpus, the DeltaSignal Decider corpus: 179 filings, full filing packages, graded blind by two passes, splits adjudicated.

- Pipeline passes: agreement 164 of 179 (91.6%), kappa 0.6976.
- Morning lock, 2026-10-16: finals 25 SIGNAL / 152 NO SIGNAL / 2 ABSTAIN.
- Third independent grading, 2026-10-16: 12 SIGNAL / 160 NO SIGNAL / 7 ABSTAIN; agreement with the morning finals 160 of 179 (89.4%), kappa 0.5408.
- The 19 splits were adjudicated blind; the relock is recorded as a dated pair, never a replaced number: 25 SIGNAL at the morning lock, 19 SIGNAL at the relock.

Table 2: Decider corpus, the dated pair (n = 179)

Record	SIGNAL	NO SIGNAL	ABSTAIN
Morning lock, 2026-10-16	25	152	2
Relock, 2026-10-16 (after third grading + adjudication)	19	157	3

THE LOCKED RECORD, KEPT AS A SERIES



The corpus has been relocked twice since under the same protocol: 22 SIGNAL / 155 NO SIGNAL / 2 ABSTAIN on 2026-10-18, and 20 SIGNAL / 157 NO SIGNAL / 2 ABSTAIN on 2026-10-24. Each lock stands on record under its own date; no lock replaces an earlier one.

5. What the two panels show together

Blind two-pass grading with an abstain band finds real flaws at low rates in a mature public benchmark (3.3% invalid, n = 30) and produces live disagreement at the judgment edge of our own corpus (19 of 179 adjudicated). Both results are the method working, not failing.

Kappa without its breakdown is a quote waiting to be used against the auditor; every figure above carries its counts.

6. Close

AITrailblazer runs this audit as a service: <https://aitrailblazer.com/evaluation-services>

7.Amendments and corrections

This report is amended in the open. If anything in the method or the aggregates is wrong, write to constantine@aitrailblazer.com. A correction that holds up against the source documents is recorded here with its date, and the text it changes is amended with the change marked. Entries in this ledger are appended, never edited or removed.

Date	Entry
2026-10-30	Ledger opened. This report was published today. No corrections have been received.

AITrailblazer Audit Report 2026-001 · Version 1.1 · October 30, 2026 ·
The canonical version of this report is this page; the PDF edition is a dated snapshot of it.
