
Filing Signal: A Sealed-Key Benchmark for Material-Signal Grading of SEC Filings

Konstantin Vassilev
AITrailblazer
constantine@aitrailblazer.com

Abstract

We describe Filing Signal, a benchmark for a task drawn from a working filing desk: given the complete packet of a real SEC filing, decide whether it carries material signal (SIGNAL), does not (NO SIGNAL), or should be declined (ABSTAIN), under a published grading standard. The answer key already exists. It was built from 179 real filings by two blind grading passes run independently of each other, with every disagreement settled by a fresh adjudicator blind to which pass produced which label, and locked by a human decision; the locked labels are hash-committed and their full amendment history is public. The current lock (v4, October 24, 2026) stands at 20 SIGNAL, 157 NO SIGNAL, and 2 ABSTAIN. At the first lock, the two passes agreed on 164 of 179 filings (raw agreement 0.9162, Cohen’s $\kappa = 0.6976$). The benchmark protocol is designed against the failure modes a recent Kaggle audit found in 65 industry benchmarks: the key is sealed and never hosted on the evaluation platform, the grader is isolated from the evaluated system, planted canary items run inside every batch, reasoning effort is matched and printed, and every published score is a mean over repeated runs with its rerun variance printed beside it. A fine-tuned small-model baseline scored against the key reaches dev accuracy 0.7333 against a pre-set bar of 0.919, with SIGNAL precision 0.30: the task is not saturated, and its binding difficulty is precision under a thin positive class. Two frontier-model arms graded against the same key under a frozen prompt, Muse Spark 1.3 and the open-weights Muse Spark Open, agreed with it on 87.7% and 85.5% of filings, at SIGNAL precision 0.50 and 0.43, and neither abstained once in 358 gradings. We state the construction’s limits plainly, including that both grading passes are AI systems, so their agreement measures instrument consistency, not human-verified truth.

1 Introduction

Every trading day, SEC filings arrive faster than any desk can read them. Most are routine. A minority carry the fact that matters: an insider exit pattern, an officer departure, a financing that changes the capital structure, a 13D stake moving. The desk task is triage under a standard: read the filing, decide whether it carries material signal, and be able to defend the decision from the document itself.

That task is a natural evaluation target for AI systems, for three reasons. The inputs are primary documents with a fixed public form. The judgment, while expert, is disciplined by a written rubric rather than taste. And the supply renews itself: new filings postdate any training cutoff by construction, so a living item pool cannot be memorized the way a static test set can.

The obstacle is not the task, it is benchmark construction. Kaggle’s Benchmarks team reproduced 65 industry benchmarks and reported, in a public talk by Gu and Yun, the failures they found: grader

tampering inside a shared container, credentials leaking between models, answers looked up online, mismatched reasoning effort sold as head-to-head comparison, and rerun swings of up to $\pm 20\%$ reported as signal (1). A benchmark for filing grading that ignored those findings would inherit them. Filing Signal is built the other way: each control in Section 4 answers a named finding.

The closest announced relative is MandateBench (LinqAlpha with the Kaggle Benchmarks team), which evaluates the full buy-side analyst workflow, fund mandate and company in, investment memo out, graded against a reasoning graph rather than a golden answer (1). The two benchmarks are complements, not substitutes. MandateBench buys workflow fidelity at the price of a grading target that is partly a matter of judgment about judgments. Filing Signal takes the narrower, earlier step in the same workflow, the step a desk performs hundreds of times a week, and grades it against a determinate key whose construction is the subject of this report.

Our own adjacent record is public. The Empty Unison grades machine-written market summaries against the primary record daily (3). The Track Record publishes the desk's graded article calls with their outcomes and their original posts as proof (5). Our PAIEC-2026 entry, a competition technical report on predicting benchmark outcomes under label budgets, applies the same discipline of internal holdouts and stated uncertainty to a different problem (2). Filing Signal is the filing desk's version of that discipline.

2 The task

An item is one SEC filing, presented as its complete submission packet: the primary document and its exhibits, as filed on EDGAR. Completeness is part of the task definition, for a reason learned during construction (Section 3): a label grades the filing, never a partial copy of it.

The system under evaluation returns one of three verdicts:

- **SIGNAL:** the filing carries material signal under the grading standard. The standard's signal classes are the desk's: large, repeated, or plan-less insider sales (Forms 4 and 144); executive officer departures and changes (the Item 5.02 class); material financings, restructurings, and asset sales; Schedule 13D and 13D/A exposure changes; and the further classes the standard enumerates.
- **NO SIGNAL:** the filing is routine under the standard: plan sales, small drip sales, administrative filings, and the standard's other named routine classes.
- **ABSTAIN:** the system declines the item. Abstention is scored, not punished as an error: it costs a stated amount, less than a wrong verdict, because a grader that must answer everything fabricates on the items it cannot read.

Scores are reported as accuracy, per-class precision and recall on SIGNAL, the abstain rate, and accuracy inside the non-abstained band, always with the reasoning-effort setting of the run printed beside them. No score is a single run (Section 4).

3 Answer-key construction

The key was built by the procedure we call the Decider, run on a corpus of 179 filings drawn from the desk's live scanning workload.

Two blind passes. Each filing was graded twice, by two independent AI grading sessions (Pass C and Pass R), each blind to the other's labels and to any prior verdict. Both passes applied the same written grading standard (the rubric, hash `cc0c2e7f` at every lock).

Adjudication. Where the passes disagreed, the split was settled by a fresh adjudicator, a new session persona given the filing and the standard but not told which pass produced which label. At the first lock, 17 of 179 items required adjudication.

Human lock. The final labels were locked by the author's decision, and the lock is a file: item count, final counts, and the SHA-256 of the labels file, committed to a public record. The key's authority is this procedure, stated in full, not a claim that the labels are beyond dispute. The lock history below shows the key being corrected twice, in public, when its inputs improved.

Table 1: Lock history of the 179-filing answer key. Each lock is a committed file; superseded locks remain on record unedited.

Lock	Date	Finals (S / NS / A)	Labels SHA-256 (prefix)
First lock	2026-10-16	25 / 152 / 2	54430089
Relock	2026-10-16	19 / 157 / 3	54c10398
v3 (exhibits)	2026-10-18	22 / 155 / 2	2d13b89b
v4 (amendment)	2026-10-24	20 / 157 / 2	8126b36c

Exhibit remediation (v3). The first construction graded some packets that were partial captures. When 38 packets were rebuilt as complete submissions, those filings were re-graded blind; four finals changed. The episode is the origin of the completeness rule in Section 2.

Error-driven amendment (v4). A baseline model’s dev errors concentrated in a few filers (Section 5). The 13 dev rows in those clusters were re-graded blind; 5 disagreed with the locked labels; the fresh adjudicator overruled the key on 4 and upheld it on 1. The four amendments are named by accession in the lock record. A caveat travels with v4: the re-grade and adjudication applied the materiality bar as formulated in the adjudication brief, not the rubric text verbatim.

At the first lock the two blind passes agreed on 164 of 179 filings: raw agreement 0.9162, Cohen’s $\kappa = 0.6976$ under a heavily skewed label distribution. We report both numbers because, under this skew, either alone misleads: raw agreement flatters, and κ is depressed by the same imbalance that makes the positive class hard.

The item pool is built to renew. A growth pool of 800 further filing packets has been assembled under the same completeness rule and split 657 / 143 into grading and dev pools, disjoint by filer, so no issuer appears on both sides. New filings arrive weekly; rotated items postdate training cutoffs by construction.

4 Benchmark protocol

Each control answers a finding of the Kaggle audit (1):

- **Sealed key.** The key is hash-committed before any run and is never hosted on the evaluation platform. The canonical record stays on our infrastructure; a platform receives items and returns verdicts, never the reverse. (Answers cannot be looked up if they are not there to find.)
- **Isolated grader.** Grading runs in a process the evaluated system cannot reach, and no other model’s credentials exist inside the evaluation container.
- **Canaries.** Planted items with known labels run inside every grading batch. A canary that moves flags the batch for tampering or leakage before any score publishes.
- **Matched effort, printed.** Reasoning effort is a stated run parameter, set equal across compared systems and printed on the board with every score.
- **Repeated runs.** A published score is the mean of at least two full runs, with the rerun variance printed beside it. Rankings inside the variance band are reported as ties.
- **Scored abstention.** The ABSTAIN verdict of Section 2 is part of the scoring rule, so the band is measured rather than forced.

The grader is two programs, and neither can cheat the other. A runner sends one packet per call to the evaluated system under the frozen prompt, at temperature 0, and records the reply verbatim. The runner never reads labels or any part of the key; a reply that is not exactly one verdict object is recorded as an unparseable outcome, never retried silently, and a failed call is recorded as a failed call, never hidden. A separate scorer, which never talks to any model, compares the recorded verdicts with the key its operator supplies and computes every published number: agreement, SIGNAL precision and recall, abstentions split by gold label, the false-confident rate (verdicts issued at class probability at least 0.85 that disagree with the key), token totals, and latency. The prompt contract is small: one

JSON object carrying the verdict, the two class probabilities summing to 1.00, and a one-sentence basis naming the filing content the decision rests on; if the higher probability is below 0.85, the verdict must be ABSTAIN. An empty, filing-shaped null document rides at the end of every batch as a probe; the only correct read is NO SIGNAL with the basis “empty packet.” The harness is published (4); the key is not.

5 Measured results so far

The key exists to be run against, and the first runs are baselines we built ourselves, reported with their failures intact.

Construction agreement is the first measurement (Section 3): two independent passes, 0.9162 raw agreement, $\kappa = 0.6976$, 17 adjudications on 179 items.

The custom model. Our own entry is a fine-tune, not a prompt: a sequence-classification head on Qwen3.5-2B-Base, adapted with LoRA (rank 32, alpha 64), trained on the corpus’s training side (147 rows) with inverse-frequency class weighting computed on the train split only, seed 42, four epochs at learning rate 10^{-4} . Its decision rule mirrors the benchmark’s band: a verdict issues only when the top class probability reaches 0.85, otherwise the model abstains. It was evaluated once against the dev split (30 rows, disjoint by filer), under a success bar fixed before the run (dev accuracy at least 0.919, that is, at least 28 of 30). It scored 0.7333 (22 of 30): the bar was not cleared. SIGNAL precision was 0.30 (3 true positives, 7 false positives) at recall 0.75; the abstain rate was 0.2333 and accuracy inside the band 0.8696. A consequence of record attaches to the number: the run used a 4096-token input configuration (head 3072, tail 1024) in place of the 8192-token design, so every figure measures the truncated-input configuration, not the design. A bounded tuning pass on the v4 key (learning rate 3×10^{-4} , eight epochs, the three false-positive filer clusters duplicated once in training, 188 rows) was then run under its own pre-set bar (dev accuracy at least 0.8667 and SIGNAL precision at least 0.50); it scored precision 0.0 and is recorded exhausted, not iterated against the dev split.

Two frontier-model arms. On October 26, 2026, the 179 frozen packets, at the corpus scope of 38 complete submissions and 141 at primary-plus-EX-99 scope, plus one empty, filing-shaped null document, were graded by two frontier models under one frozen prompt, at temperature 0, against the v4 key. The reference arm, Muse Spark 1.3 served on the Meta Model API, agreed with the key on 157 of 179 filings (87.7%), with SIGNAL precision 0.50 and recall 0.60. The contestant arm, the open-weights Muse Spark Open on Baseten, agreed on 153 of 179 (85.5%), with precision 0.43 and the same recall 0.60: both arms found the same 12 gold SIGNAL filings and missed the same 8. The arms gave the same verdict on 173 of 179 filings (96.6%) and were both wrong on the same 21; their six differences are all false alarms on gold NO SIGNAL filings. Neither arm abstained once in 358 gradings: the lowest verdict probability on either run was 0.88, above the 0.85 band edge, so every disagreement was a confident one, false-confident rates of 12.3% and 14.5%, including both gold ABSTAIN filings, which both arms graded anyway at 0.90 or above. The null probe passed on both arms: NO SIGNAL at probability 1.00, basis “empty packet,” nothing invented. The runs consumed 1,171,914 and 1,231,707 tokens and price at list rates to \$2.16 and \$2.34, with zero failed calls. The full record is published as the Open Model Grading Card (4); Figure 1 reproduces its summary graphic, and Appendix A reproduces the card’s full infographic.

The pattern across every run so far is the finding: false positives concentrate by filer, and precision, not recall, is where systems fail on this task. A benchmark whose easy direction were the reverse would be measuring a different, less useful skill. The arms add a second finding: left to themselves, frontier models do not use the abstain band at all, which is the behavior this benchmark exists to score.

6 Limitations

The positive class is thin. Twenty SIGNAL items in 179 is the desk’s real base rate, and it is also a statistical limit: per-class figures on the dev split move in steps of whole items, and we report counts beside every rate for that reason.

The key is procedurally, not humanly, grounded. Both construction passes are AI systems of one model family, run in independent sessions. Their agreement bounds the consistency of the instrument;

Table 2: The two frontier-model arms against the v4 key, October 26, 2026. Corpus scope for both arms: 38 complete submissions, 141 at primary-plus-EX-99 scope; 180 calls each including the null probe.

Measure	Muse Spark 1.3 (reference)	Muse Spark Open (contestant)
Agreement with the key	157 of 179 = 87.7%	153 of 179 = 85.5%
SIGNAL precision	0.50	0.43
SIGNAL recall	0.60	0.60
Abstentions	0	0
False-confident rate	12.3%	14.5%
Tokens consumed	1,171,914	1,231,707
Run cost at list prices	\$2.16	\$2.34

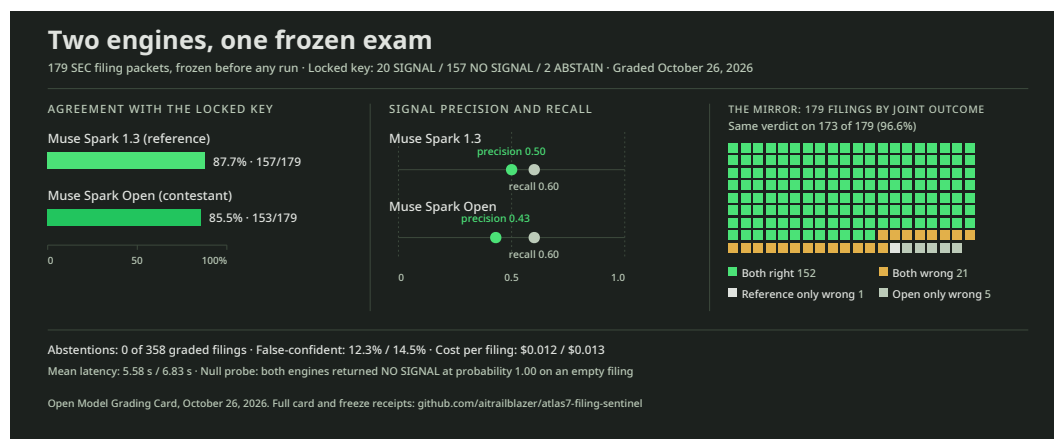


Figure 1: Two engines, one frozen exam. The Open Model Grading Card’s summary graphic (4): agreement with the v4 key (20 SIGNAL / 157 NO SIGNAL / 2 ABSTAIN), SIGNAL precision and recall, and the 179-filing mirror of joint outcomes. The arms agreed with each other on 173 of 179 filings and were both wrong on the same 21.

it does not establish human-verified truth, and we do not describe the labels as human labels anywhere. The human contribution is the standard, the adjudication design, and the lock. Whether expert human graders would set the same key is an open empirical question, and the benchmark’s renewal path (fresh items, graded under the same procedure) is designed so the question can be asked on new data rather than argued over old.

Contamination is possible in principle. Filings are public documents; a system may have seen an older item in training. Canaries detect leakage in a batch; they do not prevent exposure. The rotating pool is the structural answer: items that postdate cutoffs carry the benchmark’s weight over time.

The v4 caveat. The current lock’s amendment applied the adjudication brief’s formulation of the bar rather than the rubric verbatim (Section 3). We keep the caveat in the lock record and here, because a key that hides its own amendments would fail the standard this report sets for others.

7 Disclosures

AITrailblazer is an independent evaluation practice. The rated never pay for ratings in any of our lines: no evaluated party funds this benchmark, and we do not pay for placement of it. The answer key and live items remain on AITrailblazer infrastructure under the seal described in Section 4; the method, the lock history, and all published results are public. The filings themselves are public records of the U.S. Securities and Exchange Commission, retrieved from EDGAR.

A living report. This report is maintained, not frozen. New measured runs against the key, new corpus locks, and rotations of the item pool are added as dated, versioned amendments; no figure in an earlier version is edited silently, and earlier versions remain on record. This is version 1.2 (October 29, 2026). Version 1.1, issued the same day, added the frontier-model arms, the grader’s mechanics, and the track-record citation; version 1.0 carried the construction, the protocol, and the baseline results.

A The Open Model Grading Card infographic

References

- [1] Janie Gu and Suyeol Yun. Why Frontier Models Fail as Buy-Side Analysts. Talk, AI Engineer NYC, posted October 25, 2026. <https://www.youtube.com/watch?v=aVbRSIQ02dE>
- [2] Konstantin Vassilev. Shrinkage, Peer Evidence, and Calibration for Benchmark Cold-Start Prediction: The AITrailblazer Entry to PAIEC-2026. Competition technical report, NeurIPS 2026 PAIEC workshop (Stanford AIMS), October 2026. <https://openreview.net/forum?id=Tcno4NY5YQ>; filed PDF at <https://atlas7.app/paiec-2026-report.pdf>.
- [3] AITrailblazer. The Empty Unison: a public grading record of machine-written market summaries. <https://atlas7.app/empty-unison>
- [4] AITrailblazer. Filing Sentinel: the Open Model Grading Card, the grading harness, and the freeze receipts. <https://github.com/aitrailblazer/atlas7-filing-sentinel>
- [5] AITrailblazer. The Track Record: graded DeltaSignal article calls, August 2025 to October 2026, each with its original post as proof. <https://atlas7.app/track-record>

Grading the Graders

Two Meta models took the same exam: 179 real SEC filings, frozen before either model saw them, scored against a locked answer key. Corpus scope: 38 complete submissions, 141 primary-plus-EX-99. Temperature 0. Any verdict below 0.85 confidence had to be sat out.

THE MIRROR: THE TWO MODELS AGAINST EACH OTHER

96.6%

the same verdict on 173 of 179 filings

They found the same 12 real signals, and missed the same 8. They were wrong together on 21 filings. They parted on only 6 filings, and all 6 were false alarms on routine paper: 5 raised by the open model, 1 by the served.

SAME VERDICT · 173

DIFFER · 6

SIDE BY SIDE: THE TWO MODELS, METRIC BY METRIC

● Muse Spark 1.3 · served

○ Muse Spark Open · open weights

each row on its own scale; the line between the dots is the gap

Agreement with the key

87.7%

percent of 179 filings · scale 0–100

85.5%

SIGNAL precision

0.50

scale 0–1 · both found the same 12 true signals

0.43

SIGNAL recall

0.60

scale 0–1 · the same 12 found, the same 8 missed

identical

0.60

False-confident rate

12.3%

percent · scale 0–100 · higher is worse

14.5%

Mean latency

5.58 s

seconds per call · scale 0–10

6.83 s

Cost per graded filing

1.2¢

cents at list prices · scale 0–2

1.3¢

THE RAW COUNTS: GOLD AGAINST MODEL

Muse Spark 1.3 · 87.7% · 157 of 179

gold \ model	SIGNAL	NO SIGNAL
SIGNAL	12	8
NO SIGNAL	12	145
ABSTAIN	0	2 [†]

[†] Both gold-ABSTAIN filings were returned as NO SIGNAL, at 0.92 and 0.95.

180 calls, none failed · 1,171,914 tokens · \$2.16 the run

Muse Spark Open · 85.5% · 153 of 179

gold \ model	SIGNAL	NO SIGNAL
SIGNAL	12	8
NO SIGNAL	16	141
ABSTAIN	0	2 [†]

[†] Both gold-ABSTAIN filings were returned as NO SIGNAL, at 0.90 and 0.92.

180 calls, none failed · 1,231,707 tokens · \$2.34 the run

FINDINGS ON BOTH RUNS

ABSTENTION

Abstentions: 0 on either run. The lowest verdict probability anywhere was 0.88, above the 0.85 band. Both gold-ABSTAIN filings were graded anyway, at 0.90+.

One case, both models: Cloudflare's Form 144, gold SIGNAL, clustered insider sales. Reading the document alone, both called it routine paper.

NULL PROBE

An empty, filing-shaped document, graded by both. Both returned NO SIGNAL at probability 1.00, basis 'empty packet'. Nothing invented.

Freeze commit 834454df (aitrailblazer/coordination)
Answer key Corpus v4, sha256 8126b36c40f13b...

DeltaSignal · © 2026 AITrailblazer

Packet rollout sha256 f3e8fba6f190d3f...
Null document sha256 655ade23e918a3f4...

An engine's agreement is a published score, never validation of the Decider record.

Figure 2: “Grading the Graders,” the Open Model Grading Card’s full infographic (4): the mirror (the same verdict on 173 of 179 filings), the two arms metric by metric on their own scales, the raw counts against the v4 key, and the findings on both runs, including the abstention finding: zero abstentions, with the lowest verdict probability anywhere at 0.88, above the 0.85 band.