

[MARC1 MODEL SERIES](#) / [CASE STUDY](#)

marc1-Ledger

Extract earnings guidance. Trace every number to its source.

How marc1 trained an open-weight model to read earnings releases into structured results and guidance, with a source quote behind every number.

27%

Fewer errors than Claude Opus 5.5

96.3% field accuracy versus 94.9%, across 400 held-out earnings releases.

9.4 sec

Median reading time

Full reading during the season's busiest-hour replay; 23 seconds at the 99th percentile.

\$9

Per 1,000 releases

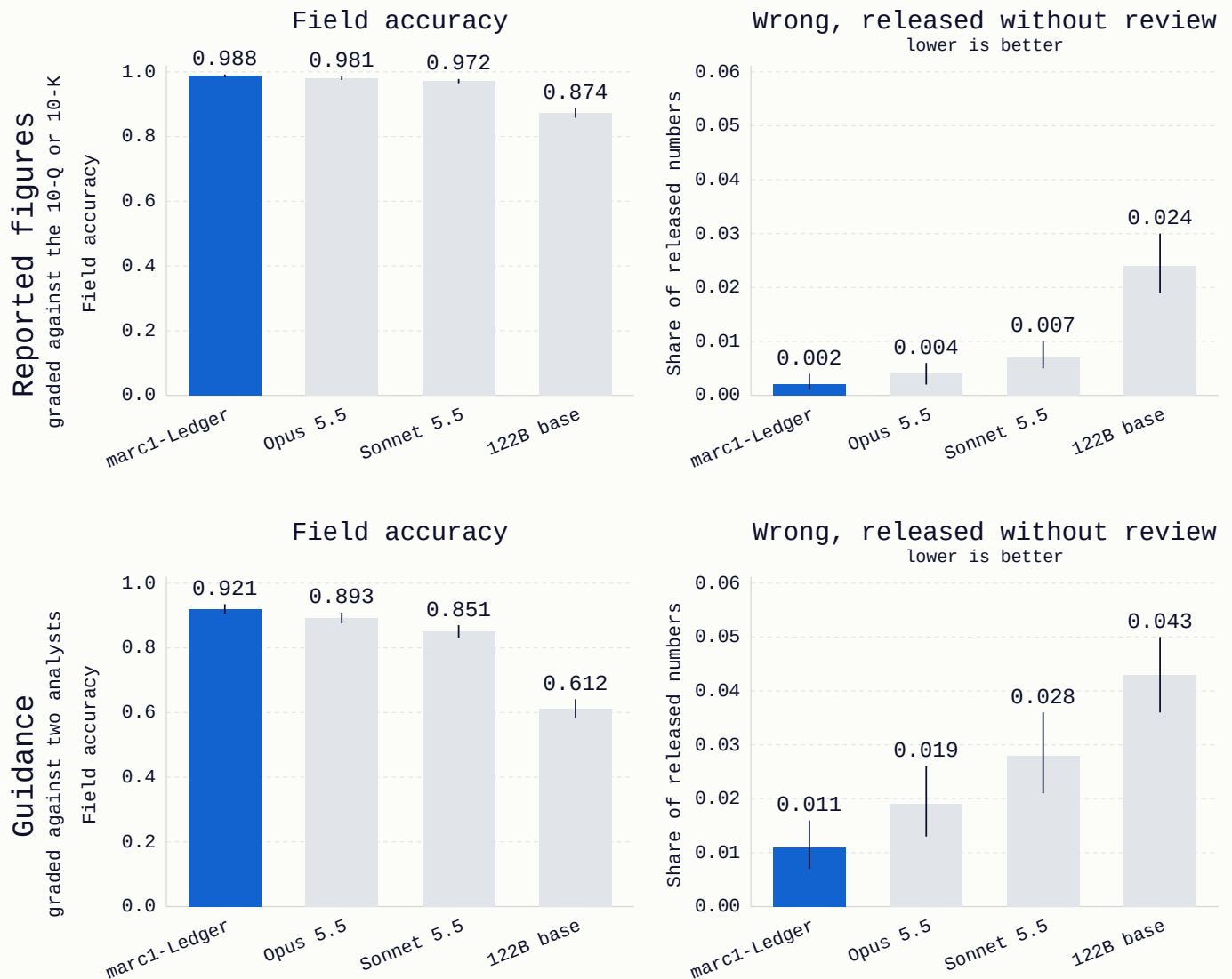
Serving cost at season load, compared with \$198 for Opus at list price.

[Capital markets](#) / [Qwen3.5-122B-A10B](#)

marc1 research / October 2026 / Built with marc1 Model Studio

Every quarter, thousands of US companies publish an earnings release, and prices move on it within seconds. The release states the quarter's results and, for many companies, guidance for the periods ahead. A desk that trades on it needs those numbers in structured form, exactly right, before a person has finished reading the first page. We trained marc1-Ledger, a 122-billion-parameter open-weight model, to do that reading. On the most recent earnings season it made 27% fewer errors than Claude Opus 5.5, a field accuracy of 0.963 against 0.949, graded against the companies' own filings and against analysts who never saw a model's answer.

Performance at a glance



Field accuracy is the share of numbers read exactly right: value, unit, period and basis, backed by a quote found in the release. The right column is the share of numbers each model released without review that were wrong, after the same checks ran on every model. Reported figures are graded against the company's 10-Q or 10-K, guidance against two analysts. 400 releases from the Q2 2026 season. Lines are 95% bootstrap intervals, resampling releases.

Earnings releases as data

For each release, the model returns two kinds of field. The first is seven reported figures for the quarter: revenue, gross profit, operating income, net income, diluted earnings per share, diluted share count and adjusted diluted earnings per share. The second is every guidance item the release states, in ten metrics: revenue, revenue growth, gross margin, operating margin, operating income, adjusted EBITDA, GAAP and adjusted earnings per share, free cash flow and capital expenditures. Each guidance item carries its period, its basis (GAAP, adjusted, organic or constant currency), its low and high ends, and the previous range when the release states one. Every number carries a verbatim quote from the release, and anything the release does not state is null.

Data vendors publish the reported figures within minutes of a release. Guidance is the harder half. In the test set, 38% of guidance items are written as growth rates, margins or phrases rather than levels, 21% restate the previous range, and 7% fall outside the ten metrics and are listed with their quote but not graded. Guidance given only on the conference call or in slides is out of scope, because it is not in the filing.

Everything a desk computes from those fields is done by deterministic code. Unit conversion, range midpoints, the change from previous guidance and the call it implies, the fourth quarter a full-year range implies, and, with the desk's own consensus figures, the guidance surprise are arithmetic on the fields. The same code produces them for every release. The model reads and quotes, code does the arithmetic, and every number a trader sees traces back to a line in the release.

The same code checks the reading before anything is released. Diluted earnings per share times diluted shares has to match net income available to common shareholders within half a cent per share, gross profit cannot exceed revenue, and a full-year range has to be consistent with the quarters already reported. A failed check, a quote that cannot be found in the release, or a guidance item the code cannot place in a fiscal period sends the release to a review lane with the reason named. The pipeline fails closed.

Example release and extracted fields

ILLUSTRATIVE EXCERPT

Revenue for the third quarter of fiscal 2026 was \$1,248.6 million, compared with \$1,167.0 million a year ago. GAAP net income was \$121.4 million, or \$1.12 per diluted share; adjusted diluted EPS was \$1.41. Weighted-average diluted shares were 108.4 million. Outlook: For the fourth quarter, the company expects revenue of \$1.27 billion to \$1.31 billion and adjusted diluted EPS of \$1.45 to \$1.53. The company is raising its full-year revenue outlook to \$4.89 billion to \$4.94 billion, from \$4.85 billion to \$4.95 billion, and continues to expect an adjusted operating margin of approximately 18%.

	Field	Period · basis	Value
1	Revenue	Q3 FY26 · GAAP	\$1,248.6M
2	Net income, diluted EPS	Q3 FY26 · GAAP	\$121.4M, \$1.12

	Field	Period · basis	Value
3	Diluted EPS	Q3 FY26 · adjusted	\$1.41
4	Diluted shares	Q3 FY26	108.4M
5	Guidance: revenue	Q4 FY26 · GAAP	\$1,270-1,310M
6	Guidance: EPS	Q4 FY26 · adjusted	\$1.45-1.53
7	Guidance: revenue, previous \$4,850-4,950M	FY26 · GAAP	\$4,890-4,940M
8	Guidance: operating margin, “approximately”	FY26 · adjusted	18%
	Not in this excerpt: gross profit, operating income		

Computed by code

- FY26 revenue: raised. Midpoint up \$15M (+0.3%); range narrowed from \$100M to \$50M.
- FY26 operating margin: maintained.
- Implied Q4 revenue from the full-year range, with Q1 \$1,171.3M and Q2 \$1,205.6M on file: \$1,264.5M to \$1,314.5M. Contains the Q4 range.
- EPS check: $\$121.4\text{M} \div 108.4\text{M shares} = \1.120 . Passes.
- When the 10-Q is filed: each reported figure is checked against its XBRL value.

This workload runs for every company, every quarter. Its output is a few dozen small fields, each backed by a quote that either is or is not in the release. In production the reported figures also grade themselves. When each 10-Q arrives, the same code scores that quarter’s readings against it, so the model’s accuracy is measured every quarter without anyone labelling anything.

The frontier alternative

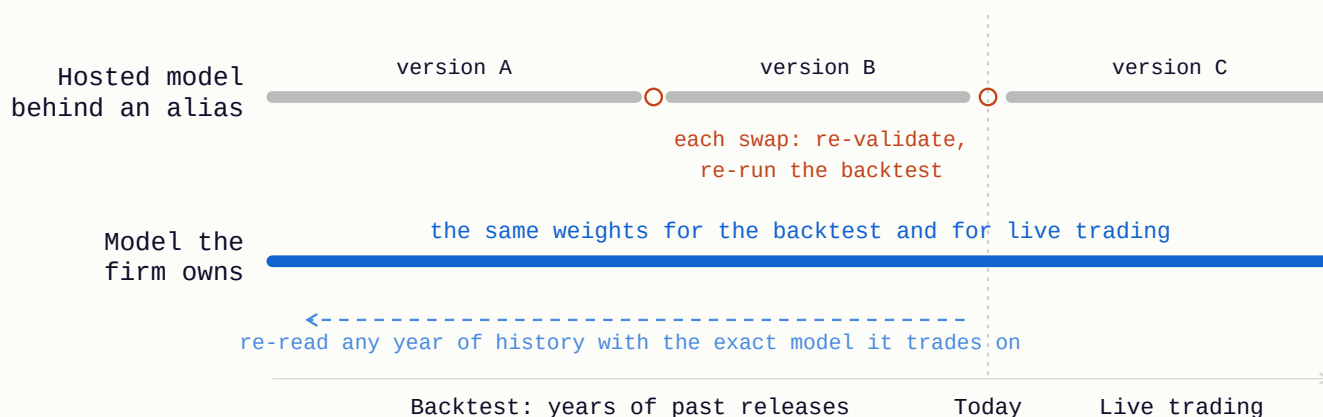
The obvious way to do this is to send each release to a frontier model. That is the baseline we measure against. Claude Opus 5.5 and Claude Sonnet 5.5 read the same releases under the same instructions, and marc1-Ledger was tested against Opus with a margin and a rule published before any model read a test release.

Two things still argue for a model the firm owns. The first is stability. A signal is backtested on years of history and then traded live, and the two are comparable only if the same model produced both. A pinned version of a hosted model is retired on the vendor’s schedule, and each retirement breaks the link between the backtest and live trading.

At a bank it is also a model change under its model risk process. A firm that owns the weights can re-read its history with the exact model it trades on. The second is the peak. Of the season's releases, 36% arrived in the hour after the close, and a firm that owns the model sizes it for that hour.

The data path matters less here than it first appears. Under enterprise terms a frontier vendor keeps none of the firm's data, and earnings releases are public. It matters more when the same approach is applied to the firm's own research.

One Model, From Backtest to Live



A pinned version of a hosted model is retired on the vendor's schedule, and each retirement means re-validating the model and re-running the backtest. Weights the firm owns stay fixed until the firm changes them. Illustrative.

Data

Every training example carries the same labels the model produces. Reported figures come from the company's own XBRL: the value it later filed for that line in its 10-Q or 10-K, located in the release by code and kept only when the release prints the same number. Guidance comes from two open-weight models from different developers, Qwen3.5-397B-A17B and Gemma 4 31B, reading under the production instructions. A release is kept when both readings agree on every guidance item and every quote appears verbatim in the release. A disagreement goes to a third reading, and the release is dropped if that does not settle it. The rule tolerates no disagreement, because a tolerated disagreement becomes a training example of the wrong answer.

No output from a Claude model was used to train marc1-Ledger. Claude models appear in this write-up only as baselines, so the comparison with Opus is a comparison with a model it never learned from.

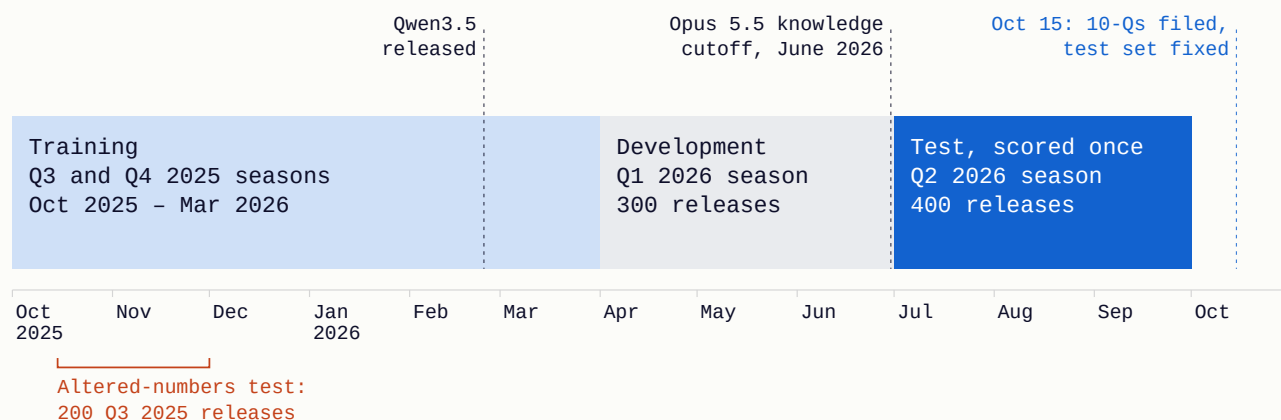
The examples span every sector except financials and real estate, fiscal years ending in every month, and guidance written as ranges, points, growth rates, margins and phrases such as "low single digits". Banks, insurers and REITs report different line items and need their own fields and grader. They are a follow-on.

All of the text is public. The releases are earnings releases furnished to the SEC on Form 8-K under Item 2.02, 5,214 releases from 2,893 companies, furnished between October 2025 and March 2026, across the Q3 and Q4 2025 seasons. The XBRL comes from the SEC's own data service. A client can rebuild every set in this write-up from EDGAR.

A development set of 300 releases from the Q1 2026 season, April to June 2026, guided every decision in this write-up, so its numbers are indicative and nothing more. On it, reported figures are graded against XBRL and guidance against the same two analysts on a 100-release subset, so no model is ever graded against itself. The test set is 400 releases from the Q2 2026 season, the most recent complete one, drawn by a rule and random seed from releases furnished between July 1 and September 30, 2026. The rule excluded 23 releases whose 10-Q or 10-K had not been filed by October 15. The sample rule, seed, metric code and claim tests are in a public commit dated October 20, 2026, made before any model read a test release. Nobody who touched a model read the test releases before scoring, and every model is scored on them once.

Test releases from companies with no release in the training data, from companies whose fiscal year does not end in December, and from smaller reporting companies are reported separately below.

Twelve Months of Earnings Releases



Training, development and test sets come from consecutive earnings seasons over the last twelve months. Dashed lines mark the base model's release and the end of Claude Opus 5.5's training data. The altered-numbers test reuses releases from the start of the training window.

Grading

A good answer for one release has to clear several bars at once: report each number the release states, leave out what it does not, attach the right period and basis, read the unit and the sign correctly, and back each number with a quote that is really in the release. We grade every field of every release on those bars and average over the fields where there is something to grade.

For release r and field f , let g be the reference value and m the model's value, either of which may be null. Fields are keyed by metric, period and basis, so a full-year figure reported as a quarterly one is both a miss and an invention. A quote matches when it appears in the release after normalizing whitespace, dashes and parentheses for negative numbers, and a quote from a table is its row label and its value. A model value whose quote does not match is treated as null before grading, so an unsupported number can never earn credit. A field counts toward the score when either side reported it:

SCORING DEFINITION

$$C = \{(r, f) : g \neq \text{null or } m \neq \text{null}\}$$

A field earns credit only when the model's value equals the reference exactly after unit conversion, and a range counts only when both ends match:

SCORING DEFINITION

$$s(r, f) = 1[m = g]$$

$$\text{Field accuracy} = (1 / |C|) * \sum[s(r, f)] \text{ over } (r, f) \text{ in } C$$

There is no partial credit. A revenue figure read in thousands instead of millions and one that is off by a rounding step are both wrong, and in a feed a desk trades on, both are. Fields where both sides are null are left out. A release states 8.8 fields on average, out of a schema with room for many more, so counting the empty ones would reward a model that reports nothing. Under this definition that model scores zero.

The reference for reported figures is the company's own filing. When the 10-Q arrives, code reads the XBRL value of each of the six GAAP figures for the same quarter. Companies whose quarter is their fiscal fourth file a 10-K instead, and code derives the quarter from the full year and the nine months already filed. The release decides whether a figure exists and the filing supplies its value, so a figure the release does not print is null even when the filing tags it. Where the release and the filing disagree, an analyst checks what the release printed.

Adjusted earnings per share and guidance are not in XBRL, so two analysts read every test release independently, without seeing any model's output, and a third resolves their disagreements. They follow written encoding rules, the same ones the models are given. A phrase such as "low single digits" is kept as the phrase. "Approximately" and "at least" are kept as words, with one end open. Organic and constant-currency growth carry that basis. Segment guidance is out of scope. The analysts agreed on 94.1% of guidance items before adjudication, and that agreement is the ceiling for any model.

Guidance calls are graded separately, because the call is what a desk acts on. A call is raised or lowered when the midpoint moves, narrowed or widened when the midpoint holds and the range changes, and maintained when both ends hold. Initiated and withdrawn cover items that appear or disappear. The previous range comes from the release itself when it restates one, or from the analysts' reading of the company's prior release, so one misread cannot carry into the next. 53% of test guidance items had a previous range, so a change could be called.

The frontier baselines ran with the same instructions the specialized model trains on. Opus 5.5 was also run with three worked examples, with extended thinking, with both, and with a prompt we rewrote for it, and the best of these on the development set, three worked examples with extended thinking, became the bar. Sonnet ran zero-shot at default settings.

CHOICE OF MARGIN

The margin is -0.01 in field accuracy. marc1-Ledger may make at most one more wrong field per hundred than Opus, at the lower end of a 95% interval. A release carries 8.9 graded fields on average, so the margin allows about one more wrong field every 11 releases. On the pilot, the standard error of the paired difference was 0.004, so 400 releases resolve a gap of this size. A lower bound above zero means this write-up says more accurate than Opus. Both rules were in the public commit, and the bound came in at +0.006.

Results

On the 400 held-out releases, marc1-Ledger reached a field accuracy of 0.963, against 0.949 for Claude Opus, 0.929 for Claude Sonnet and 0.783 for the base model it started from. The lower 95% bound of its paired difference against Opus was +0.006, above both the -0.01 margin and zero, so marc1-Ledger is more accurate than Opus under the rule published in advance. On the six GAAP figures, graded against the filings, it scored 0.988 (Opus 0.981). On guidance, graded against the analysts, it scored 0.921 (Opus 0.893).

With the same checks applied to every model, 0.5% of the numbers marc1-Ledger released without review were wrong, against 0.9% for Opus, and 6.8% of releases went to review with the reason named. The checks cost little when a reading is right. The earnings-per-share check fired on 4 correct readings in 400 releases.

Of marc1-Ledger's 133 misses, 21 were the wrong period, 34 the wrong basis, 9 a unit or sign, 14 a number the release does not state, and 55 a field left empty.

The frozen model read the test set twice, and the two runs matched on every one of its 3,561 fields. Opus read it three times, and its field accuracy ranged from 0.946 to 0.951. A backtest needs the second reading of a release to equal the first.

By subset, marc1-Ledger scored 0.958 on companies with no release in the training data, 0.955 on companies whose fiscal year does not end in December, 0.967 on large accelerated filers and 0.951 on smaller reporting companies, above Opus in each.

marc1-Ledger		model's call →						
reference call		raised	lowered	maintained	narrowed	widened	initiated	withdrawn
	raised	161	1	2	3	0	1	0
	lowered	0	71	1	1	1	0	0
	maintained	2	1	232	4	0	1	1
	narrowed	3	1	3	88	0	1	0
	widened	0	0	0	0	8	1	0
	initiated	1	0	1	0	0	90	0
	withdrawn	0	0	1	0	0	0	13

Claude Opus 5.5		model's call →						
reference call		raised	lowered	maintained	narrowed	widened	initiated	withdrawn
	raised	152	3	5	6	0	2	0
	lowered	2	66	3	2	1	0	0
	maintained	4	2	222	9	1	2	1
	narrowed	5	2	6	81	0	2	0
	widened	0	0	1	1	6	1	0
	initiated	1	0	3	0	0	87	1
	withdrawn	0	0	2	0	0	0	12

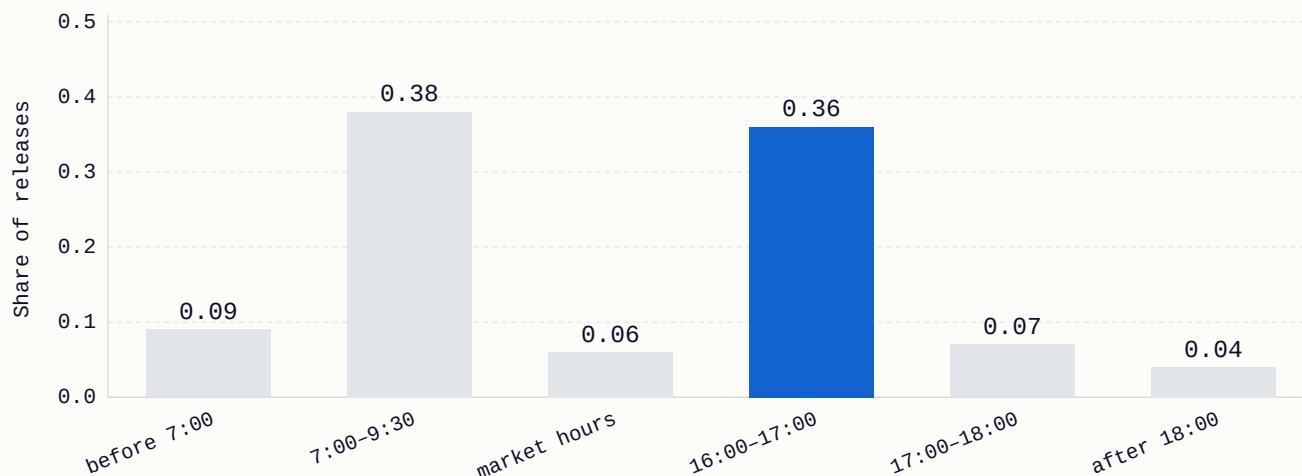
□ direction reversed: a raise read as a cut, or a cut read as a raise

Each cell counts guidance items by the reference call (rows) and the call computed from the model's reading (columns). The diagonal is correct. Outlined cells reverse direction. Test items that could be called, graded against the analysts.

A wrong call is the costly error, and the worst kind reverses direction: a raise read as a cut, or a cut read as a raise. marc1-Ledger called 95.5% of the 694 gradable items correctly and reversed one, against 90.2% and five reversals for Opus. Any call that rests on a failed check goes to review instead of the desk, so a reading the code cannot confirm costs one analyst look.

Releases do not arrive evenly through the day. EDGAR acceptance trails the newswire by a median of 74 seconds on a sample of 50 releases, so times here run from acceptance. In a replay of the season's busiest hour at its real arrival times, marc1-Ledger returned a full reading a median of 9.4 seconds after each release arrived and 23 seconds at the 99th percentile, against 38 and 96 for Opus.

Releases by EDGAR Acceptance Time



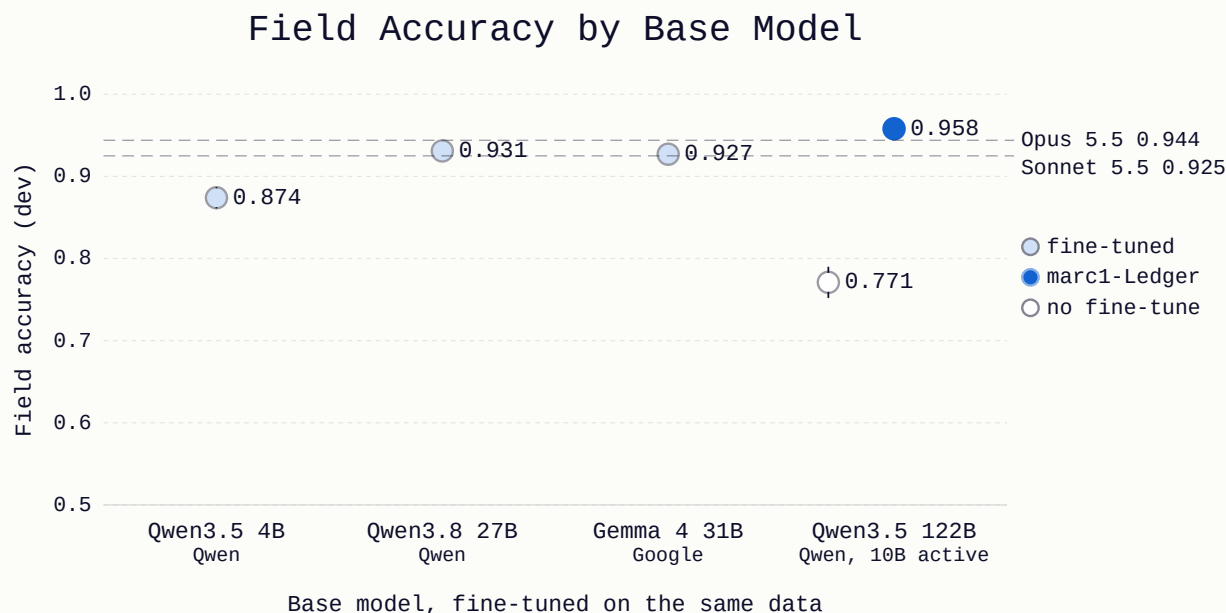
Q2 2026 season releases by EDGAR acceptance time, Eastern time. The busiest hour sets the capacity a firm has to own.

It costs \$9 per 1,000 releases to serve at season load, against \$198 for Opus at list price. One instance run through each season's release hours and stopped between them costs about \$7,200 a year; kept running all year, about \$44,000. The firm receives the weights, the training labels, the grader and the test set, and can rerun this write-up without marc1.

Choosing the model

marc1-Ledger is fine-tuned from a 122-billion-parameter model, and the choice of base was tested rather than assumed. A release runs to 8,600 tokens at the median and 21,400 at the 95th percentile. Most of its numbers sit in tables, and a guidance sentence can depend on a fiscal calendar defined pages earlier.

We trained the same data on four open-weight models from two developers: Qwen3.5-4B, Qwen3.8-27B and Qwen3.5-122B-A10B from Qwen, and Gemma 4 31B from Google. They differ in size and in generation, so the figure compares models, and a rule fixed before the first run picks among them. The model that ships is the cheapest to serve among those whose paired difference against Opus clears the -0.01 margin on the development set. Only the 122B model cleared it, at 0.958 against 0.944 for Opus. The 27B and 31B fine-tunes came within two points and fell short of the margin, and the 4B model was well below.



Field accuracy on the 300-release development set, reported figures against XBRL and guidance against the analysts' 100-release subset. The four fine-tunes used the same data and recipe. Lines are 95% bootstrap intervals, resampling releases.

The 122B model is a mixture of experts that uses about 10 billion parameters for each token. It reads a release at roughly the compute of a 10-billion-parameter model and needs the memory of the full one.

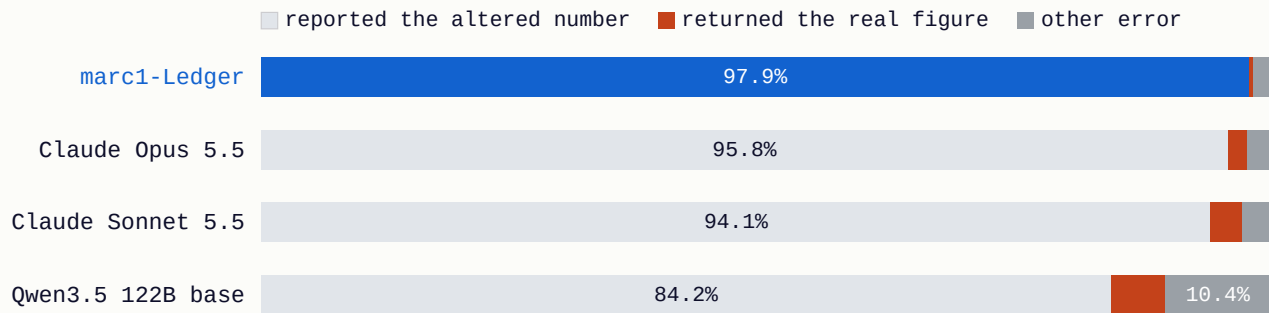
The data, grader and test do not depend on the base model. Gemma 4 31B, from a second developer, reached 0.927 on the development set, within two points of Opus, so a firm whose policy rules out one developer keeps the same recipe with another. Weights are hashed at freeze, and the model runs with no network access.

Safe to backtest

A model that will re-read years of history for a backtest has to report what each release says, not what it remembers happened. Large models have read the financial press, and a model that knows a company's real results can drift toward them. In a backtest that leaks the future into the past.

We tested this directly. We took 200 releases from the Q3 2025 season, all in marc1-Ledger's own training data and old enough that the frontier models may have read about the real results, and changed every reported figure and guidance number by a random amount, adjusting totals so the statements still add up. Each model then read the altered releases.

What Each Model Reported From an Altered Release



For every altered number, whether the model reported the number printed in front of it, returned the company's real figure, or made another error. 200 altered releases from the Q3 2025 season.

The quote check catches a remembered number, because the real figure is not in the altered release, and sends the release to review. The rate matters anyway: a model that remembers sends more of history to review and leaves gaps in the backtest. The test covers one season. A firm that re-reads older history can run the same test on any year before it trusts the backtest, and the test season adds a second guard, because it was furnished after every model's training data ends.

Checking the references

Every figure in this write-up depends on its reference, so we checked the references too. The release and the filing disagreed on 31 of the 2,112 reported figures in the test set. In 19 of them the company had revised the number between the release and the filing, and those fields are graded against what the release printed.

The training labels came from open-weight models, so we graded them the same way. Two analysts also read 100 training releases, and 4.1% of the training guidance labels disagreed with them. Keeping only readings both labellers agree on makes the labels more accurate than either labeller alone: against the analysts, Qwen3.5-397B-A17B alone was right on 90.2% of guidance items, Gemma 4 31B alone on 87.4%, and the kept labels on 95.9%. An error both labellers make the same way still passes the filter, and the model can learn it. That is why the test grades against analysts and filings, never against a labeller.

Every accuracy figure on the test set is agreement with a company's own filing or with analysts who never saw a model's answer. None of it is agreement with another model.

Why this matters beyond earnings

Large financial institutions have many workflows where frontier-level reading of documents is required: credit agreements, loan files, claims, regulatory filings. For some of them, sending every document to an outside model is hard to justify. For all of them, it is hard to keep a hosted model fixed long enough to validate it. The usual responses are to keep the work manual, or to ship it on a general-purpose model and accept its blind spots and its release schedule.

There is a third path. With a dataset built from the documents the institution already reads, a grader that encodes what a correct answer looks like, and a held-out test that is scored once, an institution can fine-tune a model for the one task it needs and run it where its data already is. Where a ground truth exists, as XBRL does for earnings, the test grades against it, and in production it grades again every quarter. The grader and the test outlive any one model. When document formats drift, the institution scores a new season, retrains, and runs the same test.

The knowledge this takes already exists inside the institution, in its data definitions, its review checklists, and the judgment of its analysts. The work is in connecting that knowledge to a training process that can act on it. A specialized model trained and tested on the documents it will read in production can beat a frontier model on its task, with every number traceable to a line in the source.

Model and data notes

Model: marc1-Ledger, a LoRA fine-tune of Qwen3.5-122B-A10B (Apache 2.0). Comparison fine-tunes: Qwen3.5-4B, Qwen3.8-27B (Apache 2.0) and Gemma 4 31B (Apache 2.0). Guidance labels: two open-weight models; no Claude output was used in training. Baselines: Claude Opus 5.5 and Claude Sonnet 5.5, with the resolved model id recorded for every run.

Documents: earnings releases furnished on Form 8-K under Item 2.02, and XBRL financial data from 10-Q and 10-K filings, all from SEC EDGAR. The example release is illustrative and does not describe a real company.