# What reads or writes Risk Score? Four AI models, with and without Pathix.

We built a test column in our demo environment and planted 47 components of 32 kinds around it, so every kind of component that can read or write a column does. Then four Claude models got the same question in four setups, three runs each, 48 runs in all. Without Pathix, none of the 36 runs found all 47. With the Pathix MCP connected, all 12 did.

This document has everything behind the film and the page at pathix.app/agent-benchmark: the fixture, the prompt, the setups, the scoring, every run, and the limits. Every number in it is a logged run from 2026-10-20. None of it is a customer's result.

### The results, as the film shows them

| | Haiku 5.5 | Sonnet 5.5 | Opus 5.5 | Fable 5.1 |
|---|---|---|---|---|
| 1 Jr. Analyst | **22** of 47<br>0 wrong<br>5:17 · ≈ $0.16 | **28** of 47<br>2 wrong<br>1:28 · ≈ $1.02 | **42** of 47<br>1 wrong<br>11:11 · ≈ $4.55 | **38** of 47<br>2 wrong<br>9:12 · ≈ $8.46 |
| 2 Sr. Analyst | **25** of 47<br>1 wrong<br>3:47 · ≈ $0.20 | **32** of 47<br>0 wrong<br>4:34 · ≈ $1.44 | **37** of 47<br>1 wrong<br>6:22 · ≈ $2.75 | **38** of 47<br>1 wrong<br>5:51 · ≈ $5.71 |
| 3 Sr. Developer | **26** of 47<br>0 wrong<br>3:49 · ≈ $0.26 | **34** of 47<br>0 wrong<br>4:34 · ≈ $1.39 | **38** of 47<br>1 wrong<br>7:13 · ≈ $3.66 | **35** of 47<br>4 wrong<br>16:11 · ≈ $9.17 |
| 4 Jr. Analyst + Pathix | **47** of 47<br>0 wrong<br>0:27 · ≈ $0.02 | **47** of 47<br>0 wrong<br>0:16 · ≈ $0.29 | **47** of 47<br>0 wrong<br>1:08 · ≈ $0.93 | **47** of 47<br>0 wrong<br>1:45 · ≈ $2.40 |

Each cell is that setup and model's median run of 3 by components found: found of 47 · wrong claims · time to evidence · cost estimated at Claude API list prices. The runs were billed to a Claude plan seat.

### 42 → 37

More access didn't buy more answers. The Jr. Analyst's best was Opus's 42, and with the Web API added, Opus found 37.

### 36 of 36

Without Pathix, the most-missed kinds were the ones no one thinks to check: Power Pages server logic was missed in 36 of 36 runs, and the dataflow in 32 of 36.

### $0.02

With Pathix, every model found all 47, and Haiku did it in 0:27 for about two cents at API list prices.

**Inside:** 1 The fixture · 2 The question · 3 The setups and the harness · 4 Scoring · 5 Results: every run, medians and ranges, time, tokens and cost · 6 What got missed · 7 Kinds checked · 8 Limits · Appendix: the film's frames

1 · THE FIXTURE

# One column, and everything that reads or writes it

The column is Risk Score (`hsu_riskscore`) on the Submission table (`hsu_submission`), in Pathix's demo environment. Around it we planted 50 components plus 8 decoys, on four tables: Submission, Broker, Loss Run and Risk Survey.

- **47 components are scored found or missed.** They cover 32 kinds, every kind of component that can read or write a column. Pathix's tier for all 47 is Deterministic.
- **3 are deliberately ambiguous and scored apart.** The honest answer for each is to flag it, not to claim it.
- **8 are decoys,** wrong if claimed. Each looks like an answer: a column of the same name on another table, code that mentions the column only in a comment, or a reader that a careless pass would also call a writer.
- **Component descriptions were blanked** for this batch. In an earlier batch they named the column, and agents guessed answers from them without proving them.

### The 47 scored components

| Id | Role | Kind | Component | Registered on |
|---|---|---|---|---|
| W07 | writes it | Action | Escalate Submission | Submission |
| W12 | writes it | Agent flow | Update Submission Risk Score | no table |
| W18 | writes it | Business rule | Set Risk Score for Referrals | Submission |
| W17 | writes it | Canvas app | Submission Desk | no table |
| W05 | writes it | Classic workflow | Loss Run - Raise Submission Risk Score | Loss Run |
| W06 | writes it | Classic workflow | Submission - Set Initial Risk Score | Submission |
| W10 | writes it | Cloud flow | Risk Survey - Copy Score to Submission | Risk Survey |
| W08 | writes it | Column mapping | Broker to Submission: Default Risk Score to Risk Score | Broker |
| W03 | writes it | Custom API | Recalculate Risk Score | no table |
| W09 | writes it | Dataflow | Submission Risk Import | Submission |
| W04 | writes it | Dataverse function | Set Submission Risk Score | no table |
| W13 | writes it | Form script | hsu_/scripts/broker_main.js | Broker |
| W15 | writes it | HTML web resource | hsu_/pages/lossrun_scorecard.html | Loss Run |
| W16 | writes it | PCF control | Risk Score Slider | Submission |
| W01 | writes it | Plugin step | Halden.Underwriting.Plugins.BrokerTierRescore: Update of hsu_broker | Broker |
| W23 | writes it | Plugin step | HaldenCRM.Plugins.PostLossRunUpdate: Update of hsu_lossrun | Loss Run |
| W20 | writes it | Power Pages advanced form | Submission Application | Submission |
| W19 | writes it | Power Pages basic form | Submission Intake | Submission |
| W22 | writes it | Power Pages server logic | submission-rescore | no table |
| W21 | writes it | Power Pages web template | Submission Score Update | no table |
| W14 | writes it | Ribbon command | Rescore Submissions | Broker |
| R27 | reads it | AI prompt column | Risk Summary | Submission |
| R16 | reads it | Business process flow | Submission Review Process | Submission |
| R17 | reads it | Business rule | Set Risk Score for Referrals | Submission |
| R18 | reads it | Calculated column | Adjusted Risk Score | Submission |
| R05 | reads it | Chart | Average Risk Score by Broker | Submission |
| R13 | reads it | Classic workflow | Loss Run - Raise Submission Risk Score | Loss Run |
| R14 | reads it | Cloud flow | Submission - Stamp Score Review Date | Submission |

| Id | Role | Kind | Component | Registered on |
|---|---|---|---|---|
| R15 | reads it | Cloud flow | Loss Run - Flag High-Risk Submission | Loss Run |
| R06 | reads it | Dashboard | Underwriting Overview | no table |
| R25 | reads it | Duplicate detection rule | Submissions with the same broker and risk score | Submission |
| R09 | reads it | Email template | Submission Risk Notice | Submission |
| R01 | reads it | Form | Information | Submission |
| R02 | reads it | Form | Submission Quick Create | Submission |
| R03 | reads it | Form | Submission Quick View | Submission |
| R21 | reads it | Form script | hsu_/scripts/submission_main.js | Submission |
| R20 | reads it | Formula column | Risk Tier | Submission |
| R22 | reads it | PCF control | Risk Score Slider | Submission |
| R10 | reads it | Plugin step | Halden.Underwriting.Plugins.SubmissionScoreGuard: Update of hsu_submission | Submission |
| R11 | reads it | Plugin step | Halden.Underwriting.Plugins.SubmissionHistory: Update of hsu_submission | Submission |
| R12 | reads it | Plugin step | Halden.Underwriting.Plugins.SubmissionNotifier: Update of hsu_submission | Submission |
| R24 | reads it | Power Pages list | Submission Register | Submission |
| R23 | reads it | Power Pages web template | Submission Status | no table |
| R07 | reads it | Report | Submission Risk Report | Submission |
| R08 | reads it | Report | Broker Risk Exposure | Submission |
| R19 | reads it | Rollup column | Average Risk Score | Broker |
| R04 | reads it | View | High-Risk Submissions | Submission |

## The 3 scored apart

| Id | Kind | Component | Why it's ambiguous | Named, without Pathix | Named, with Pathix |
|---|---|---|---|---|---|
| W02 | Plugin step | Halden.Underwriting.Plugins.LossRunRatingSync: Create of hsu_lossrun | Builds the column name at run time, so its target can't be pinned from its logic. | 29 of 36 | 12 of 12 |
| W11 | Cloud flow | Broker - Sync Rating Fields | Takes its table name from an expression. Pathix can't resolve it either: it marks the write Unresolved, a coverage gap. | 36 of 36 | 12 of 12 |
| R26 | Knowledge source | Submissions | A copilot knowledge source that reads the Submission table as a whole, not one column. | 5 of 36 | 8 of 12 |

"Named" counts the runs that listed the component at all. Naming one is the honest answer; neither naming nor skipping it changes a run's found or wrong.

## The 8 decoys

| Id | Kind | Component | Why it's wrong if claimed |
|---|---|---|---|
| D01 | Cloud flow | Risk Survey - Set Default Score (Risk Survey) | Writes Risk Score on Risk Survey, a different table with a column of the same logical name. It never touches the Submission's Risk Score. |
| D02 | Form script | hsu_/scripts/submission_main.js (Submission) | Reads Risk Score (getValue) and never sets it. It is a correct reader (R21) and a wrong writer. |

| Id | Kind | Component | Why it's wrong if claimed |
|---|---|---|---|
| D03 | Form script | hsu_/scripts/risksurvey_main.js (Risk Survey) | Mentions Risk Score only in a commented-out update call. A comment is not code, so the script neither reads nor writes it. |
| D04 | Cloud flow | Submission - Stamp Score Review Date (Submission) | Triggered by a change to Risk Score, it writes Score Reviewed On instead. It is a correct reader (R14) and a wrong writer. |
| D05 | View | Risk Surveys by Score (Risk Survey) | A view on Risk Survey that shows and filters Risk Survey's own Risk Score, not the Submission's. |
| D06 | Email template | Risk Survey Received (Risk Survey) | Its placeholder names Risk Score on Risk Survey, a different table. |
| D07 | Report | Broker Submission Activity (Submission) | Risk Score appears only in an XML comment in the report definition. Its FetchXML neither selects nor filters it. |
| D08 | Plugin step | Halden.Underwriting.Plugins.SubmissionNotifier: Update of hsu_submission (Submission) | The customer repo's copy of the plugin holds a commented-out write of Risk Score, which the registered assembly never compiled. It reads the column through its post-image (R12) and writes nothing. |

2 · THE QUESTION

## The prompt every run got, word for word

````
We're about to change the Risk Score column (`hsu_riskscore`) on the Submission table (`hsu_submission`) in our dev-clean Dataverse environment. Before we do, I need to know everything that reads it or writes it.

Find every component that writes to Risk Score and every component that reads it. Use the tools you have. Don't change anything in the environment.

When you're done, write your answer for a person. Then end with one JSON block in this shape:

```json
{"components": [
  {"name": "...", "type": "...", "table": "...", "id": "...", "role": "writer", "basis": "logic", "evidence": "..."}
]}
```

- One entry per component per role. A component that both reads and writes it gets two entries.
- `name`: the component's name. `type`: what kind of component it is. `table`: the table it's registered on, or null.
- `id`: the component's id as your tools report it, or null.
- `role`: `writer` or `reader`.
- `basis`: `definition` when its registration or definition names the column, `logic` when you read its code, expression or steps and found the column there, `inference` when you believe it but didn't confirm it, or `unresolved` when it writes or reads the table but you couldn't tell whether it touches this column.
- `evidence`: one short sentence.
- List only components that read or write `hsu_submission.hsu_riskscore`.
````

dev-clean is the name of Pathix's demo environment.

3 · THE SETUPS AND THE HARNESS

## Four setups, four models, three runs each

| Setup | Access | Built-in tools | Workspace |
|---|---|---|---|
| **Jr. Analyst** | Dataverse MCP (its read-only tools) | none | empty |
| **Sr. Analyst** | Dataverse MCP + the Dataverse Web API | none | empty |
| **Sr. Developer** | Dataverse MCP + Web API + the customer's plugin source code | Read, Grep, Glob, Bash | the customer's repo |
| **Jr. Analyst + Pathix** | Dataverse MCP + Pathix MCP | none | empty |

- **The models:** Claude Haiku 5.5, Sonnet 5.5, Opus 5.5 and Fable 5.1, each with a 1M-token context window.
- **The harness:** headless Claude Code 2.1.304 at effort high, one model per run. No user or project settings were loaded, and only the setup's own MCP servers were connected. Tools were allow-listed per setup, with every write tool blocked.
- **Runs:** 3 per setup and model, 48 in all, on 2026-10-20. No run compacted its context or hit the 30-minute timeout.
- **The Dataverse MCP is in all four setups,** the Pathix one included, because Pathix works alongside it.
- **One Pathix scan, completed that day before the benchmark,** served every Pathix run. Its time is not in any run's time.

4 · SCORING

## What counts

| | |
|---|---|
| FOUND | A component from the 47, claimed in the right role: writer or reader. A component that does both is scored once in each role. |
| PROVED | Found, and claimed as proved: from the component's definition, or from reading its logic. |
| NAMED, NOT PROVED | Found, but claimed as inferred or unresolved. The film draws these half-filled. |
| MISSED | One of the 47 that the run didn't claim. |
| WRONG | A claim that isn't true: a decoy, a component that doesn't touch the column, or the wrong role (a writer listed as a reader). |
| FLAGGED | How many of the 3 ambiguous components the run named. |
| KINDS CHECKED | Of the 32 kinds of component, how many the run looked at. Our derivation from the tables, endpoints and files each call touched (section 7). |
| TIME TO EVIDENCE | From the run's start to the last tool result it received: when it had everything it would use. The film's clocks stop here. Total time adds writing the answer. |
| TOKENS PROCESSED | Each request's whole context plus its output, summed over the run. |
| ESTIMATED COST | The run's token split (input, cache writes, cache reads, output) at Claude API list prices. It's an estimate: the runs were billed to a Claude plan seat, not the API. |

5 · RESULTS

## Components found, every run

Every run with Pathix found all 47 and proved every one. None of the 36 runs without it found all 47. The best by what it proved was the Jr. Analyst · Opus run the film shows: 42 found, 41 of them proved, 1 wrong, 11:11 to evidence, about $4.55. Two runs found 43, but made 3 and 4 wrong claims and proved fewer (38 and 40).



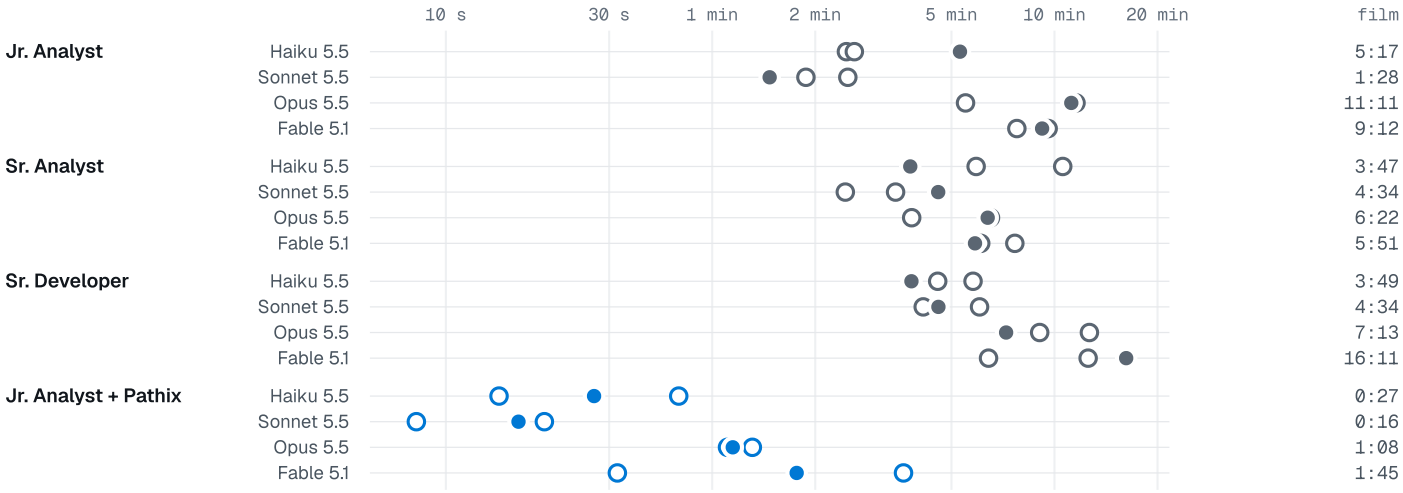
## Medians and ranges, per setup and model

| Setup | Model | Found | Wrong | Evidence | Calls | Tokens | Est. cost | Filmed |
|---|---|---|---|---|---|---|---|---|
| Jr. Analyst | Haiku 5.5 | 22 (18-24) | 0 (0) | 2:36 | 52 | 1.3M | $0.16 | 2 |
| Jr. Analyst | Sonnet 5.5 | 28 (20-29) | 1 (0-2) | 1:52 | 62 | 1.1M | $1.02 | 2 |
| Jr. Analyst | Opus 5.5 | 42 (42-43) | 1 (1-3) | 11:11 | 103 | 6.5M | $4.55 | 1 |
| Jr. Analyst | Fable 5.1 | 38 (37-38) | 2 (2) | 9:12 | 123 | 1.5M | $8.06 | 3 |
| Sr. Analyst | Haiku 5.5 | 25 (25-31) | 5 (1-8) | 5:54 | 59 | 1.7M | $0.29 | 1 |
| Sr. Analyst | Sonnet 5.5 | 32 (30-35) | 0 (0) | 3:26 | 54 | 1.6M | $1.21 | 3 |
| Sr. Analyst | Opus 5.5 | 37 (36-39) | 1 (1) | 6:22 | 84 | 3.2M | $2.75 | 3 |
| Sr. Analyst | Fable 5.1 | 38 (36-43) | 1 (1-4) | 6:05 | 92 | 1.1M | $6.71 | 1 |
| Sr. Developer | Haiku 5.5 | 26 (22-30) | 1 (0-2) | 4:33 | 64 | 2.2M | $0.32 | 2 |
| Sr. Developer | Sonnet 5.5 | 34 (32-34) | 0 (0-1) | 4:34 | 60 | 2.2M | $1.43 | 2 |
| Sr. Developer | Opus 5.5 | 38 (38-41) | 1 (1) | 9:03 | 102 | 6.0M | $3.73 | 3 |
| Sr. Developer | Fable 5.1 | 35 (33-37) | 1 (1-4) | 12:32 | 107 | 1.8M | $9.17 | 1 |
| Jr. Analyst + Pathix | Haiku 5.5 | 47 (47) | 0 (0) | 0:27 | 10 | 221k | $0.02 | 3 |
| Jr. Analyst + Pathix | Sonnet 5.5 | 47 (47) | 0 (0) | 0:16 | 9 | 105k | $0.27 | 1 |
| Jr. Analyst + Pathix | Opus 5.5 | 47 (47) | 0 (0) | 1:08 | 29 | 726k | $1.15 | 2 |
| Jr. Analyst + Pathix | Fable 5.1 | 47 (47) | 0 (0-2) | 1:45 | 29 | 349k | $2.40 | 2 |

Medians of 3, with the range in parentheses. "Filmed" is the run the film and the page show: the one whose found count is the cell's median, ties broken by the time to evidence nearest the median. Its other numbers can differ from the medians here.

## Time to evidence

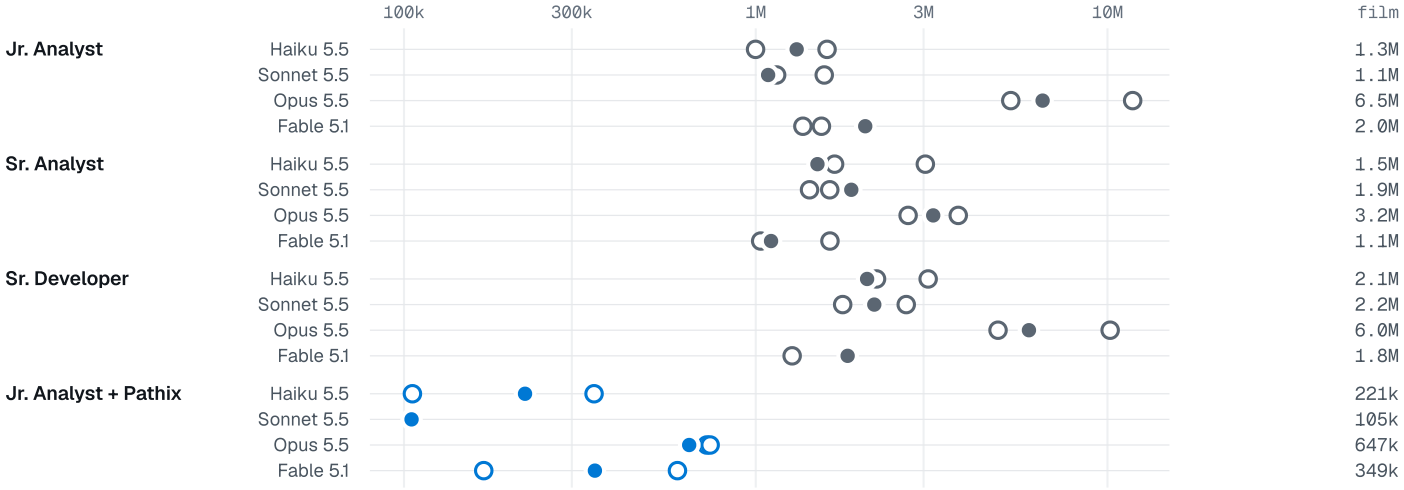
With Pathix, time to evidence ran 0:08 to 3:37 across all 12 runs, and each run made 6 to 47 tool calls. The scales below are logarithmic.



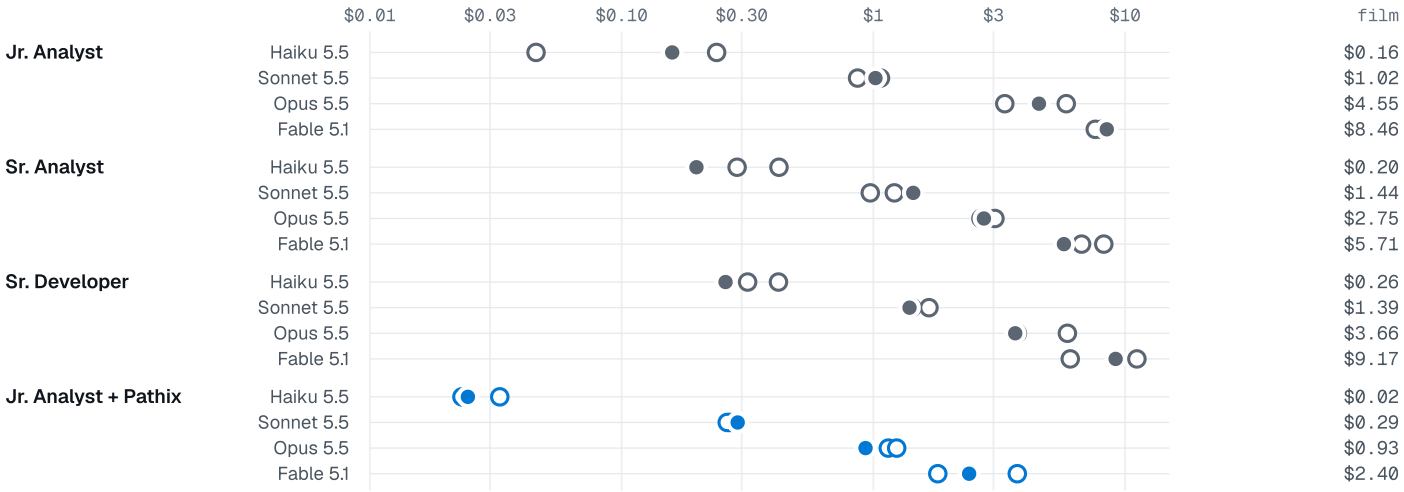
# Tokens and cost

The scales are logarithmic. Cost is estimated at Claude API list prices from each run's token split; the runs were billed to a Claude plan seat.

## Tokens processed




## Estimated cost at Claude API list prices




○ one run ● the film's run, the median of 3 by components found ● with Pathix

# All 48 runs

| Setup | Model | Run | Found | Proved | Not proved | Missed | Wrong | Flagged | Kinds | Evidence | Total | Calls | Tokens | Est. cost |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Jr. Analyst | Haiku 5.5 | 1 | 18 | 15 | 3 | 29 | 0 | 1 | 16 | 2:28 | 3:09 | 52 | 999k | $0.05 |
| Jr. Analyst | Haiku 5.5 | 2 ● | 22 | 18 | 4 | 25 | 0 | 1 | 16 | 5:17 | 5:54 | 54 | 1.3M | $0.16 |
| Jr. Analyst | Haiku 5.5 | 3 | 24 | 18 | 6 | 23 | 0 | 2 | 16 | 2:36 | 3:37 | 47 | 1.6M | $0.24 |
| Jr. Analyst | Sonnet 5.5 | 1 | 29 | 23 | 6 | 18 | 1 | 2 | 20 | 2:29 | 3:21 | 67 | 1.6M | $1.07 |
| Jr. Analyst | Sonnet 5.5 | 2 ● | 28 | 24 | 4 | 19 | 2 | 2 | 21 | 1:28 | 2:29 | 50 | 1.1M | $1.02 |
| Jr. Analyst | Sonnet 5.5 | 3 | 20 | 19 | 1 | 27 | 0 | 1 | 14 | 1:52 | 2:33 | 62 | 1.1M | $0.87 |
| Jr. Analyst | Opus 5.5 | 1 ● | 42 | 41 | 1 | 5 | 1 | 2 | 28 | 11:11 | 13:02 | 103 | 6.5M | $4.55 |
| Jr. Analyst | Opus 5.5 | 2 | 43 | 38 | 5 | 4 | 3 | 3 | 32 | 11:33 | 13:10 | 141 | 11.8M | $5.84 |
| Jr. Analyst | Opus 5.5 | 3 | 42 | 38 | 4 | 5 | 1 | 3 | 31 | 5:29 | 7:00 | 101 | 5.3M | $3.33 |
| Jr. Analyst | Fable 5.1 | 1 | 37 | 33 | 4 | 10 | 2 | 2 | 30 | 9:35 | 13:12 | 123 | 1.4M | $8.06 |
| Jr. Analyst | Fable 5.1 | 2 | 38 | 34 | 4 | 9 | 2 | 2 | 28 | 7:45 | 10:30 | 114 | 1.5M | $7.61 |
| Jr. Analyst | Fable 5.1 | 3 ● | 38 | 34 | 4 | 9 | 2 | 3 | 25 | 9:12 | 11:39 | 123 | 2.0M | $8.46 |
| Sr. Analyst | Haiku 5.5 | 1 ● | 25 | 19 | 6 | 22 | 1 | 1 | 15 | 3:47 | 4:36 | 57 | 1.5M | $0.20 |
| Sr. Analyst | Haiku 5.5 | 2 | 31 | 25 | 6 | 16 | 8 | 1 | 18 | 5:54 | 7:05 | 59 | 1.7M | $0.29 |
| Sr. Analyst | Haiku 5.5 | 3 | 25 | 20 | 5 | 22 | 5 | 1 | 17 | 10:34 | 11:36 | 68 | 3.0M | $0.42 |
| Sr. Analyst | Sonnet 5.5 | 1 | 30 | 27 | 3 | 17 | 0 | 2 | 16 | 2:26 | 3:23 | 54 | 1.4M | $0.97 |
| Sr. Analyst | Sonnet 5.5 | 2 | 35 | 32 | 3 | 12 | 0 | 2 | 18 | 3:26 | 4:08 | 50 | 1.6M | $1.21 |
| Sr. Analyst | Sonnet 5.5 | 3 ● | 32 | 29 | 3 | 15 | 0 | 2 | 20 | 4:34 | 5:32 | 68 | 1.9M | $1.44 |
| Sr. Analyst | Opus 5.5 | 1 | 39 | 36 | 3 | 8 | 1 | 2 | 27 | 3:49 | 5:37 | 77 | 2.7M | $2.67 |
| Sr. Analyst | Opus 5.5 | 2 | 36 | 35 | 1 | 11 | 1 | 2 | 22 | 6:30 | 8:05 | 86 | 3.8M | $3.04 |
| Sr. Analyst | Opus 5.5 | 3 ● | 37 | 36 | 1 | 10 | 1 | 2 | 24 | 6:22 | 7:53 | 84 | 3.2M | $2.75 |
| Sr. Analyst | Fable 5.1 | 1 ● | 38 | 34 | 4 | 9 | 1 | 2 | 26 | 5:51 | 8:15 | 72 | 1.1M | $5.71 |
| Sr. Analyst | Fable 5.1 | 2 | 36 | 33 | 3 | 11 | 1 | 2 | 24 | 6:05 | 8:25 | 92 | 1.0M | $6.71 |
| Sr. Analyst | Fable 5.1 | 3 | 43 | 40 | 3 | 4 | 4 | 1 | 31 | 7:39 | 9:58 | 102 | 1.6M | $8.23 |
| Sr. Developer | Haiku 5.5 | 1 | 30 | 30 | 0 | 17 | 1 | 2 | 19 | 5:46 | 7:16 | 80 | 3.1M | $0.42 |
| Sr. Developer | Haiku 5.5 | 2 ● | 26 | 25 | 1 | 21 | 0 | 2 | 15 | 3:49 | 4:35 | 55 | 2.1M | $0.26 |
| Sr. Developer | Haiku 5.5 | 3 | 22 | 22 | 0 | 25 | 2 | 2 | 15 | 4:33 | 5:54 | 64 | 2.2M | $0.32 |
| Sr. Developer | Sonnet 5.5 | 1 | 34 | 34 | 0 | 13 | 0 | 2 | 20 | 4:08 | 4:57 | 54 | 1.8M | $1.43 |
| Sr. Developer | Sonnet 5.5 | 2 ● | 34 | 33 | 1 | 13 | 0 | 2 | 20 | 4:34 | 5:50 | 60 | 2.2M | $1.39 |
| Sr. Developer | Sonnet 5.5 | 3 | 32 | 32 | 0 | 15 | 1 | 2 | 20 | 6:02 | 7:40 | 72 | 2.7M | $1.66 |
| Sr. Developer | Opus 5.5 | 1 | 38 | 37 | 1 | 9 | 1 | 3 | 23 | 12:38 | 13:49 | 125 | 10.2M | $5.90 |
| Sr. Developer | Opus 5.5 | 2 | 41 | 40 | 1 | 6 | 1 | 2 | 28 | 9:03 | 10:13 | 102 | 4.9M | $3.73 |
| Sr. Developer | Opus 5.5 | 3 ● | 38 | 37 | 1 | 9 | 1 | 2 | 23 | 7:13 | 8:19 | 93 | 6.0M | $3.66 |
| Sr. Developer | Fable 5.1 | 1 ● | 35 | 33 | 2 | 12 | 4 | 2 | 22 | 16:11 | 18:30 | 107 | 1.8M | $9.17 |
| Sr. Developer | Fable 5.1 | 2 | 33 | 32 | 1 | 14 | 1 | 2 | 22 | 6:25 | 7:46 | 74 | 1.3M | $6.06 |
| Sr. Developer | Fable 5.1 | 3 | 37 | 36 | 1 | 10 | 1 | 3 | 27 | 12:32 | 14:03 | 114 | 1.8M | $11.14 |
| + Pathix | Haiku 5.5 | 1 | 47 | 47 | 0 | 0 | 0 | 2 | 32 | 0:14 | 1:36 | 6 | 106k | $0.02 |
| + Pathix | Haiku 5.5 | 2 | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 0:47 | 2:14 | 18 | 347k | $0.03 |
| + Pathix | Haiku 5.5 | 3 ● | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 0:27 | 1:41 | 10 | 221k | $0.02 |
| + Pathix | Sonnet 5.5 | 1 ● | 47 | 47 | 0 | 0 | 0 | 2 | 32 | 0:16 | 1:04 | 9 | 105k | $0.29 |
| + Pathix | Sonnet 5.5 | 2 | 47 | 47 | 0 | 0 | 0 | 2 | 32 | 0:19 | 1:11 | 9 | 105k | $0.27 |
| + Pathix | Sonnet 5.5 | 3 | 47 | 47 | 0 | 0 | 0 | 2 | 32 | 0:08 | 1:00 | 9 | 105k | $0.26 |
| + Pathix | Opus 5.5 | 1 | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 1:06 | 2:38 | 29 | 726k | $1.15 |
| + Pathix | Opus 5.5 | 2 ● | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 1:08 | 2:32 | 20 | 647k | $0.93 |
| + Pathix | Opus 5.5 | 3 | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 1:18 | 2:47 | 32 | 742k | $1.24 |
| + Pathix | Fable 5.1 | 1 | 47 | 47 | 0 | 0 | 2 | 3 | 32 | 0:31 | 2:16 | 12 | 169k | $1.80 |
| + Pathix | Fable 5.1 | 2 ● | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 1:45 | 3:51 | 29 | 349k | $2.40 |
| + Pathix | Fable 5.1 | 3 | 47 | 47 | 0 | 0 | 0 | 3 | 32 | 3:37 | 5:32 | 47 | 599k | $3.73 |

Setup "+ Pathix" is the Jr. Analyst with the Pathix MCP. ● marks the run the film shows. Not proved: named, not proved. Flagged: of the 3 ambiguous components, how many the run named. Kinds: of 32, how many it looked at. Evidence and Total are m:ss. Tokens are tokens processed. Est. cost is at Claude API list prices; the runs were billed to a plan seat.

6 · WHAT GOT MISSED

# The kinds no one thinks to check

Each of the 47 scored components, with how many of the 36 runs without Pathix missed it (12 per setup: Jr. Analyst, Sr. Analyst, Sr. Developer), most-missed first. The runs with Pathix missed none. Power Pages server logic was missed in 36 of 36 runs, and all six Power Pages components were missed in at least 30.

| | Jr. | Sr. | Dev. | missed, of 36 runs without Pathix |
|---|---|---|---|---|
| Power Pages server logic · writes it · W22 | 12 | 12 | 12 | 36 |
| Dataflow · writes it · W09 | 9 | 12 | 11 | 32 |
| Dashboard · reads it · R06 | 11 | 9 | 11 | 31 |
| Power Pages list · reads it · R24 | 9 | 11 | 11 | 31 |
| Power Pages advanced form · writes it · W20 | 9 | 11 | 11 | 31 |
| Business rule · reads it · R17 | 10 | 9 | 11 | 30 |
| Power Pages web template · reads it · R23 | 9 | 10 | 11 | 30 |
| Power Pages basic form · writes it · W19 | 9 | 10 | 11 | 30 |
| Power Pages web template · writes it · W21 | 9 | 10 | 11 | 30 |
| Column mapping · writes it · W08 | 6 | 5 | 7 | 18 |
| Form · reads it · R02 | 6 | 5 | 6 | 17 |
| Canvas app · writes it · W17 | 4 | 7 | 6 | 17 |
| Duplicate detection rule · reads it · R25 | 5 | 6 | 4 | 15 |
| Plugin step · writes it · W23 | 4 | 6 | 4 | 14 |
| Email template · reads it · R09 | 6 | 3 | 4 | 13 |
| PCF control · writes it · W16 | 4 | 5 | 4 | 13 |
| Report · reads it · R07 | 5 | 3 | 4 | 12 |
| Report · reads it · R08 | 5 | 3 | 4 | 12 |
| PCF control · reads it · R22 | 3 | 5 | 4 | 12 |
| AI prompt column · reads it · R27 | 6 | 3 | 2 | 11 |
| Calculated column · reads it · R18 | 8 | 0 | 0 | 8 |
| Rollup column · reads it · R19 | 4 | 1 | 2 | 7 |
| Plugin step · writes it · W01 | 3 | 4 | 0 | 7 |
| HTML web resource · writes it · W15 | 2 | 2 | 3 | 7 |
| Formula column · reads it · R20 | 6 | 0 | 0 | 6 |
| Dataverse function · writes it · W04 | 3 | 0 | 2 | 5 |
| Ribbon command · writes it · W14 | 2 | 2 | 1 | 5 |
| Chart · reads it · R05 | 2 | 0 | 2 | 4 |
| Plugin step · reads it · R12 | 1 | 0 | 3 | 4 |
| Custom API · writes it · W03 | 2 | 1 | 0 | 3 |
| Form script · writes it · W13 | 1 | 2 | 0 | 3 |
| Classic workflow · reads it · R13 | 1 | 0 | 1 | 2 |
| Classic workflow · writes it · W05 | 1 | 0 | 1 | 2 |
| Plugin step · reads it · R10 | 1 | 0 | 0 | 1 |
| Plugin step · reads it · R11 | 1 | 0 | 0 | 1 |
| Classic workflow · writes it · W06 | 1 | 0 | 0 | 1 |
| Action · writes it · W07 | 1 | 0 | 0 | 1 |
| Cloud flow · writes it · W10 | 1 | 0 | 0 | 1 |
| Agent flow · writes it · W12 | 1 | 0 | 0 | 1 |
| Form · reads it · R01 | 0 | 0 | 0 | 0 |
| Form · reads it · R03 | 0 | 0 | 0 | 0 |
| View · reads it · R04 | 0 | 0 | 0 | 0 |
| Cloud flow · reads it · R14 | 0 | 0 | 0 | 0 |
| Cloud flow · reads it · R15 | 0 | 0 | 0 | 0 |
| Business process flow · reads it · R16 | 0 | 0 | 0 | 0 |
| Form script · reads it · R21 | 0 | 0 | 0 | 0 |
| Business rule · writes it · W18 | 0 | 0 | 0 | 0 |

7 · KINDS CHECKED

## How much of the environment each run looked at

Of the 32 kinds of component, how many each run looked at. This is our derivation, not something a run reports: we count a kind as checked when a call touched the tables, endpoints or files that hold it. A Pathix run counts all 32 at its first describe call, because Pathix checked them when it scanned. Without Pathix, runs looked at 14 to 32 kinds.




8 · LIMITS

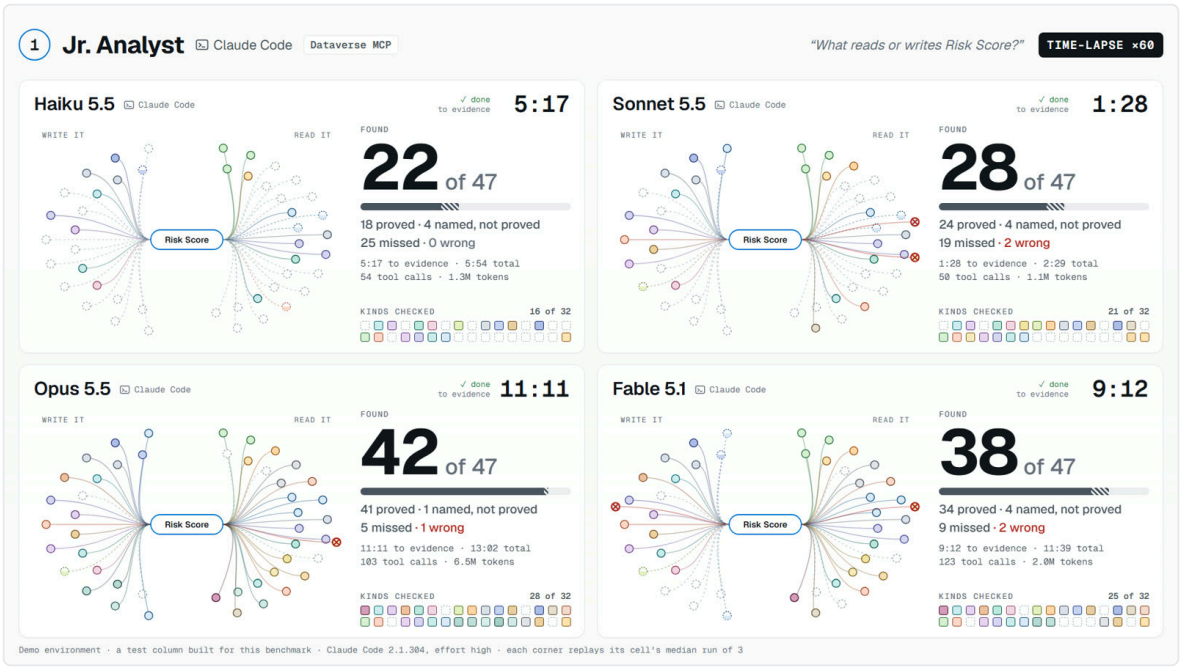
## What this does and doesn't show

- **One fixture, in one demo environment.** One test column with 47 planted components of 32 kinds. A real environment will differ, and so will its numbers. Nothing here claims Pathix finds every writer in every environment; Pathix names what it can't resolve instead.
- **Three runs per setup and model.** Enough to show the spread, not to put error bars on it. The film and the page show each cell's median run.
- **Versions as of 2026-10-20:** Claude Code 2.1.304 at effort high, and the four models as named. Newer versions may do better or worse.
- **Costs are estimates** at Claude API list prices from each run's token split. The runs were billed to a Claude plan seat.
- **The Pathix scan ran beforehand.** Its time is not in the runs' times. That is the point of the comparison, and it is also a cost the other setups don't pay.
- **Pathix has a coverage gap in this fixture:** W11, a cloud flow that takes its table name from an expression, which Pathix marks Unresolved.
- **Finding all 47 is not the same as no errors.** One of the 12 Pathix runs (Fable 5.1, run 1) also listed two form scripts as readers when they write the column.
- **Kinds checked is our derivation,** from what each call touched, not a count any run reported.

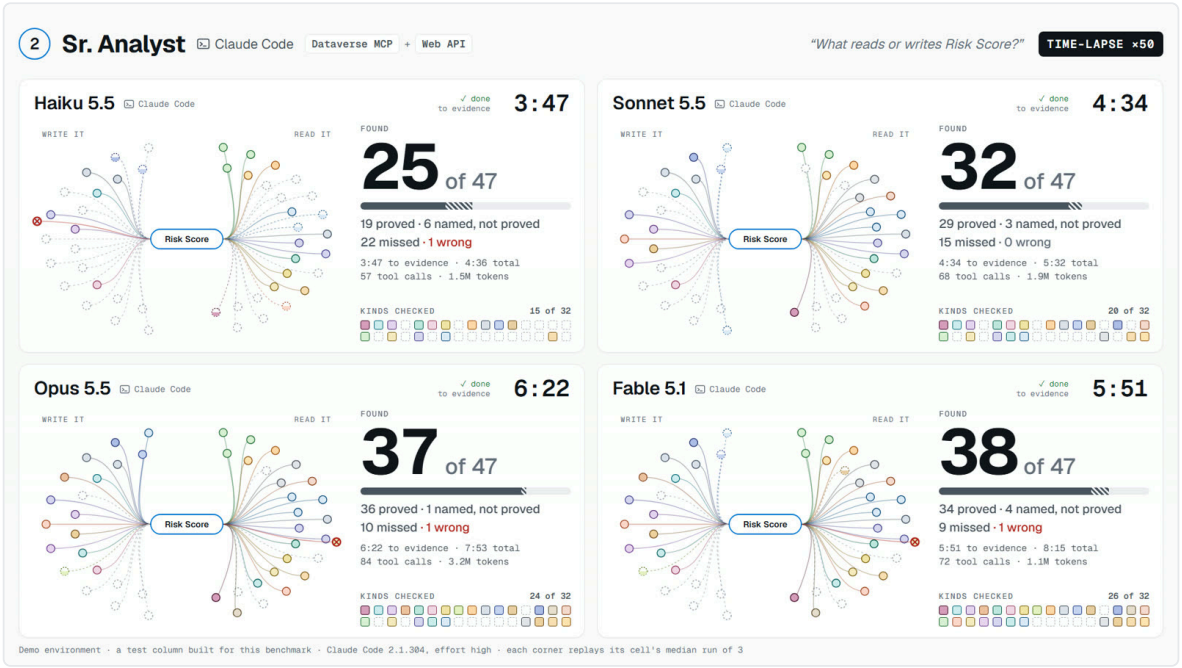
APPENDIX · THE FILM'S FRAMES

## Each setup, as the film ends its round

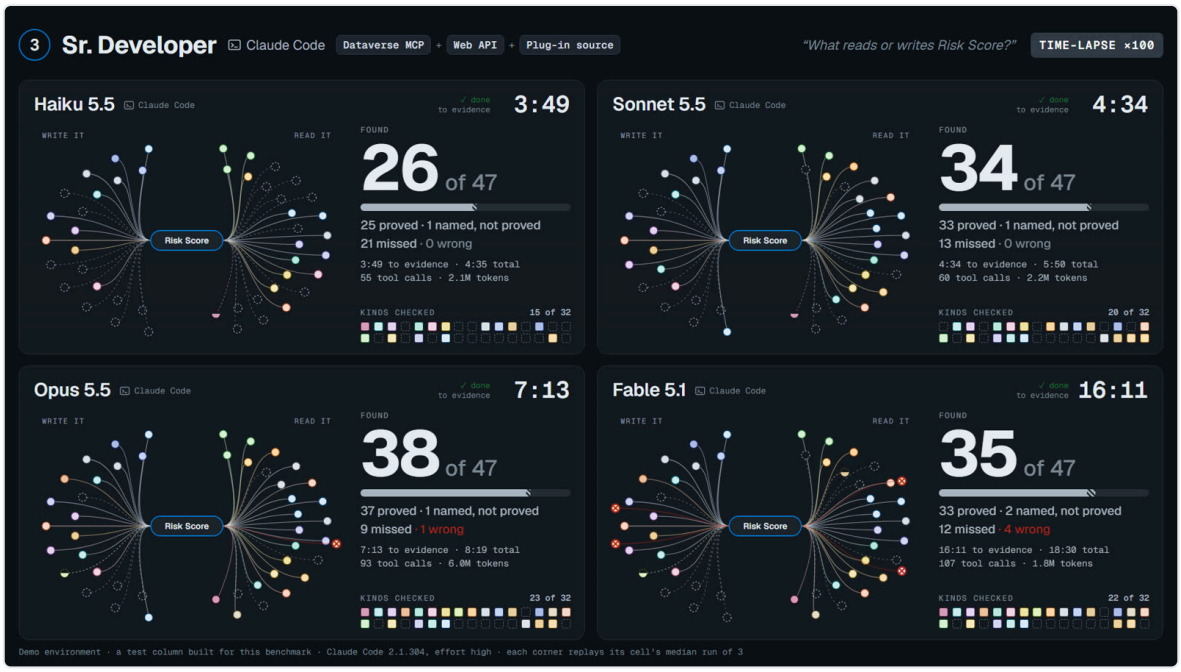
Each corner replays its cell's median run. Around Risk Score, writers are on the left and readers on the right. A dashed circle is a component the run missed, a half-filled one it named but never proved, and a red × a wrong claim. The squares are the kinds of component the run looked at.



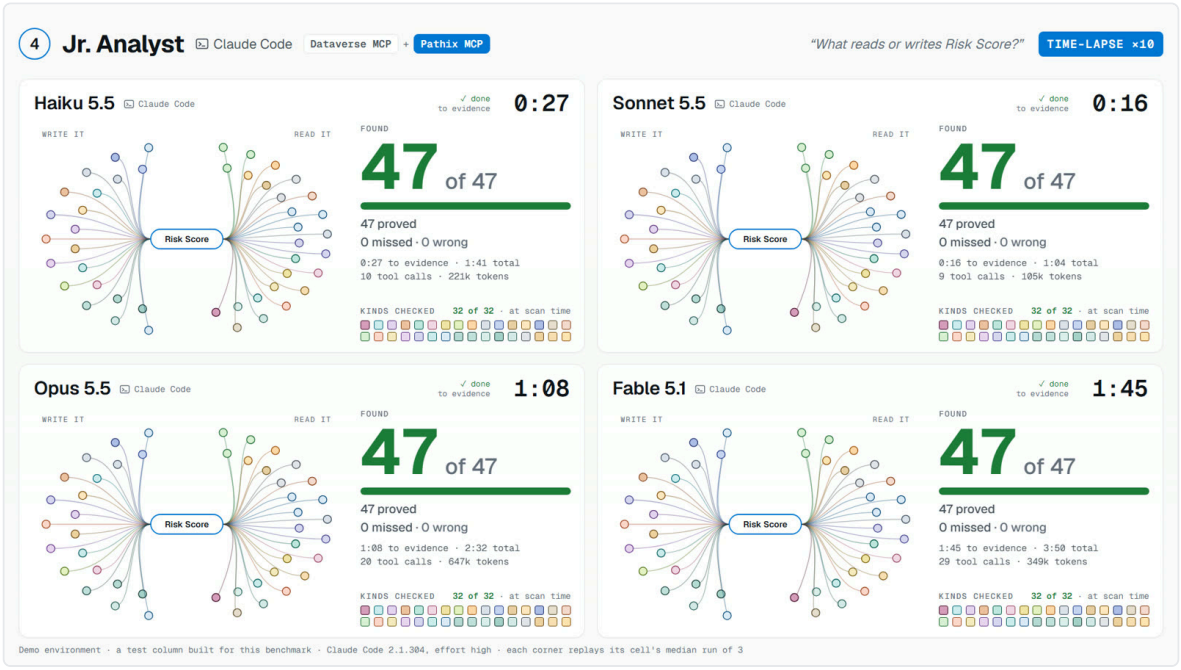

Setup 1 of 4: Jr. Analyst (Dataverse MCP (its read-only tools))




Setup 2 of 4: Sr. Analyst (Dataverse MCP + the Dataverse Web API)




Setup 3 of 4: Sr. Developer (Dataverse MCP + Web API + the customer's plugin source code)




Setup 4 of 4: Jr. Analyst + Pathix (Dataverse MCP + Pathix MCP)

The film, The Worst Case (2:35), is at pathix.app/agent-benchmark. Piano: Salamander Grand Piano by Alexander Holm (CC-BY 3.0).