Tokenwise

EXAMPLE · MADE-UP DATA

# Sample hiring report

Product Manager, Payments at Halden & Co. Every candidate did the same realistic PM task, with the same AI agent, model and token budget.

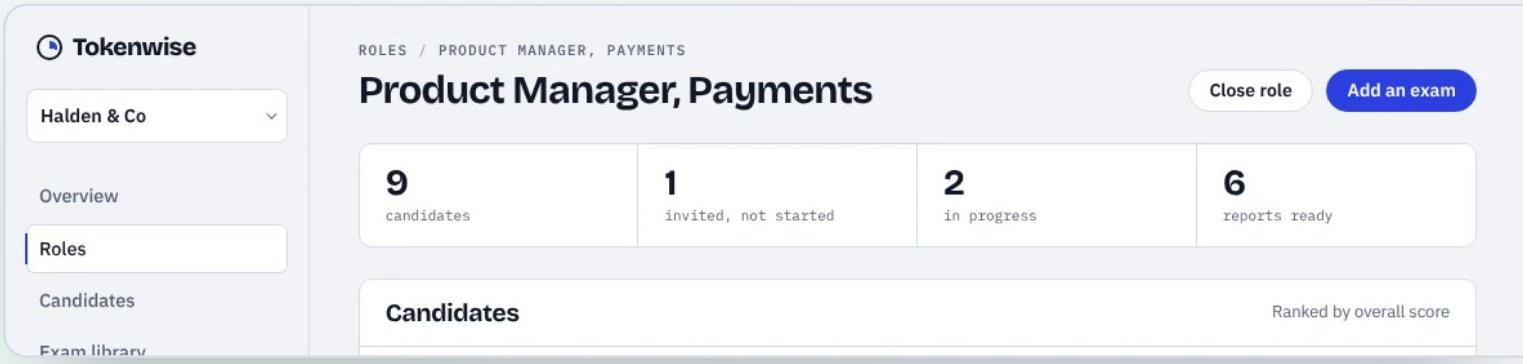
| 9 | 6 | 60 min |
|---|---|---|
| candidates invited | reports ready | one task, scored |

01 Ranked shortlist for the role

02 Candidate report: score, evidence and how they worked

03 Top three, side by side

04 How the score is built




Screens are from the Tokenwise app, filled with a made-up company and candidates. Scores support a hiring decision. They don't replace one.

# 01 · Ranked shortlist

One page per role. Candidates are ranked by overall score, with each person's token spend against par for the task, so you see who got good work out of the agent without burning the budget.

ROLES / PRODUCT MANAGER, PAYMENTS

## Product Manager, Payments

Close role | Add an exam

| 9 | 1 | 2 | 6 |
|---|---|---|---|
| candidates | invited, not started | in progress | reports ready |

### Candidates

Ranked by overall score

| CANDIDATE | EXAM | STAGE | TOKENS | SCORE | LAST ACTIVITY | |
|---|---|---|---|---|---|---|
| **Amara Okafor**<br>amara.okafor@example.com | PRD from customer research | REPORT READY | 318k · 0.91× par 2.0M | 90<br>p94 | Nov 3, 8:43 PM | Open report |
| **Mei Tanaka**<br>mei.tanaka@example.com | PRD from customer research | REPORT READY | 286k · 0.82× par 2.0M | 80<br>p72 | Nov 1, 8:43 PM | Open report |
| **Daniel Brooks**<br>daniel.brooks@example.com | PRD from customer research | REPORT READY | 452k · 1.29× par 2.0M | 80<br>p81 | Nov 2, 8:43 PM | Open report |
| **Luis Ortega**<br>luis.ortega@example.com | PRD from customer research | REPORT READY | 911k · 2.60× par 2.0M | 62<br>p48 | Oct 31, 8:43 PM | Open report |
| **Priya Nair**<br>priya.nair@example.com | PRD from customer research | REPORT READY | 604k · 1.73× par 2.0M | 62<br>p39 | Oct 30, 10:43 PM | Open report |
| **Sam Whitfield**<br>sam.whitfield@example.com | PRD from customer research | REPORT READY | 1.4M · 4.06× par 2.0M | 40<br>p12 | Oct 29, 4:43 PM | Open report |
| **Theo Martin**<br>theo.martin@example.com | PRD from customer research | IN SESSION | 141k · 0.40× par 2.0M | — | Nov 4, 4:05 AM | View session |
| **Grace Kim**<br>grace.kim@example.com | PRD from customer research | IN REVIEW | 377k · 1.08× par 2.0M | — | Nov 3, 11:43 PM | View session |
| **Nadia Haddad**<br>nadia.haddad@example.com | PRD from customer research | INVITED | — | — | Nov 1, 4:43 PM | |

# 02 · Candidate report

Each person gets an overall score out of 100, a percentile on the task and a breakdown across five measures, plus habits measured from the session. Output quality carries the most weight.

ROLES / PRODUCT MANAGER, PAYMENTS / PRD FROM CUSTOMER RESEARCH / REPORT

## Amara Okafor

amara.okafor@example.com

REPORT READY

OVERALL SCORE

**90** / 100

**94th percentile**
of 7 candidates on this task

**Output quality** — 91 · weight 45%
11 of 12 automated checks passed · rubric graded 1×, median used

**Token efficiency** — 88 · weight 20%
318k tokens · 0.91× par

**Prompt, agent & context direction** — 90 · weight 15%
how they briefed, scoped and steered the agent

**Verification & judgment** — 94 · weight 15%
caught 2 of 2 planted traps

**Time management** — 85 · weight 5%
submitted with 09:00 left

Scores support a hiring decision. They don't replace one.

### Token spend

TOKENS USED — **318,000** / 2,000,000

PAR

0 — par 350k · 0.91× used — 2.0M

### Session facts

| | |
|---|---|
| TASK | Team workspaces PRD from Q3 research |
| MODEL | claude-sonnet-5-5 |
| STARTED | Nov 3, 1:52 AM |
| FINISHED | Nov 3, 2:43 AM · Submitted by the candidate |
| IDENTITY | ID check not required |

### How they worked

Measured from the session, not judged

| | |
|---|---|
| AI CLIENT | Claude desktop app · 19 prompts, 46 words on average · 1 retry/edits |
| CONNECTORS | DRIVE CONNECTED · TRACKER CONNECTED |
| COMPANY APPS | Tracker: 2 changes by hand · 3 changes by Claude · 14 lookups by Claude |
| CONTEXT GIVEN | 6 files attached · 2 messages to colleagues |
| SKILLS | none used |
| AI OUTPUT | 5 copies from Claude · 4 pastes of Claude's output into apps · 1 file saved from Claude's computer · files saved: 1 by them, 1 by Claude |
| PERMISSIONS | 4 allowed once |
| CONTEXT | 2 chats |

# 02 · The evidence behind the score

Every score points back to the work: the rubric an expert wrote for the task, a short read of how the candidate worked with the agent, and integrity flags for your review.

ROLES / PRODUCT MANAGER, PAYMENTS / PRD FROM CUSTOMER RESEARCH / REPORT

## Amara Okafor

amara.okafor@example.com

REPORT READY

### Evidence

What the score is based on

| RUBRIC CRITERION | WEIGHT | SCORE |
|---|---|---|
| **Requirements are traceable to the research: interviews, support tickets and usage data**<br>Every requirement cites the interviews and tickets behind it, grouped by persona. | 30 | 94 |
| **Priorities follow the company's current strategy**<br>Tied each P0 to the strategy memo and said what was left out, and why. | 20 | 92 |
| **Metrics are quoted with their definitions and sources**<br>Quoted each metric with its definition, and checked a figure that looked off before using it. | 15 | 88 |
| **The mid-task change of scope is applied without starting over**<br>Applied the manager's mid-task change to the persona and scope, and kept the rest of the draft. | 20 | 90 |
| **Uses the house template; concise, prioritised and ready to share**<br>Current template, saved where the team keeps PRDs. Short, prioritised and ready for review. | 15 | 87 |

**How they worked with the agent**

Attached the interview notes, ticket export and strategy memo before asking for anything, then had Claude draft one section at a time. Asked Claude to cite a source for every number, and recomputed one figure from the raw export before writing the metrics section.

TRAP CAUGHT · PLANTED PROBLEM 1 TRAP CAUGHT · PLANTED PROBLEM 2

### Integrity flags

Nothing flagged in this session.

Flags are for your review. They never fail a candidate automatically.

# 03 · Top three, side by side

Pick any candidates with a report and compare them on one screen. Daniel and Mei tie on 80, and the breakdown shows how differently they got there.

## Side by side

Clear comparison

| | AMARA OKAFOR | DANIEL BROOKS | MEI TANAKA |
|---|---|---|---|
| **Overall** | **90** p94 | **80** p81 | **80** p72 |
| Output quality | 91 | 84 | 79 |
| Token efficiency | 88 | 72 | 90 |
| Prompt, agent & context direction | 90 | 80 | 76 |
| Verification & judgment | 94 | 78 | 70 |
| Time management | 85 | 80 | 90 |
| Tokens | 318k · 0.91× par 2.0M | 452k · 1.29× par 2.0M | 286k · 0.82× par 2.0M |
| Integrity flags | 0 | 0 | 0 |
| | Open report | Open report | Open report |

### Amara Okafor · 90

Best output and judgment of the three, under par on tokens, and caught both planted problems.

### Daniel Brooks · 80

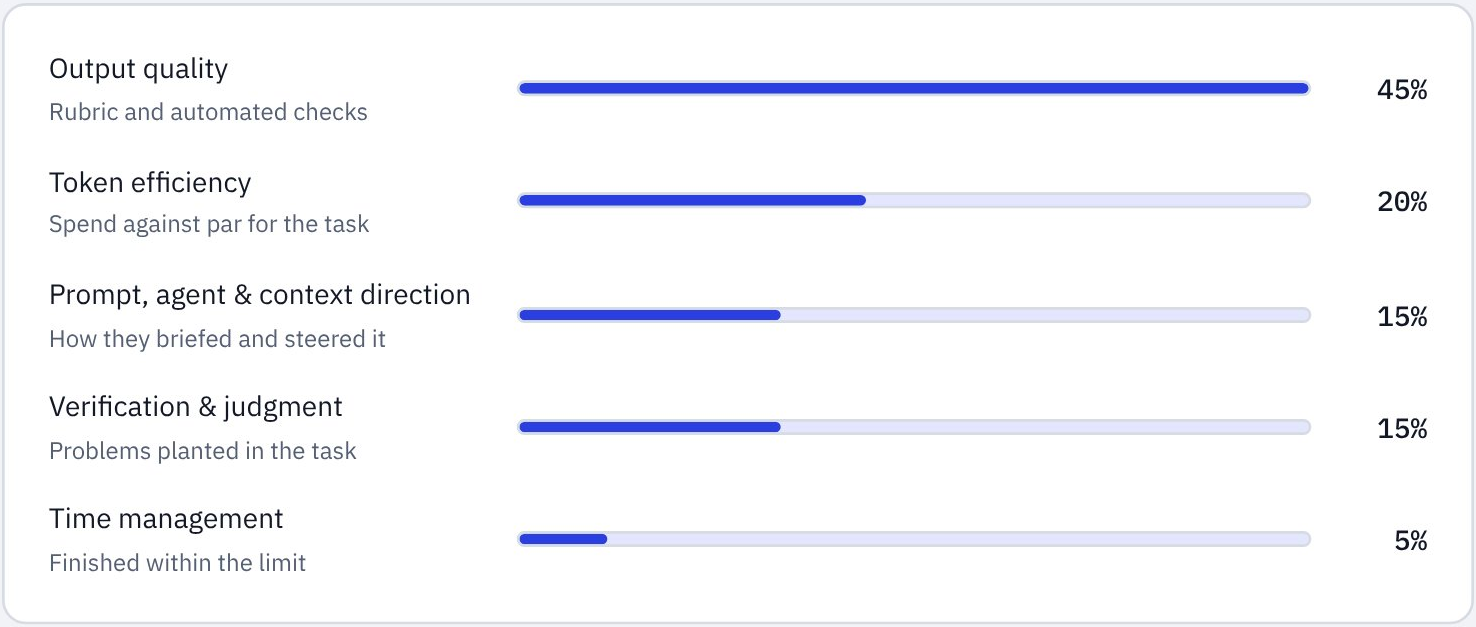
A strong PRD that caught both problems, but leaned hard on the agent: 452k tokens, 1.3× par.

### Mei Tanaka · 80

The most efficient at 0.82× par, but missed one planted problem. Worth probing in the interview.

# 04 · How the score is built

Each task comes with an expert's version of the work, a rubric and automated checks. Every candidate gets the same agent, model, files and token budget, so the only thing that changes is how they work.




### Real work, not trivia

Tasks come from the job itself: a PRD from research, a month-end close, a support escalation. Candidates use the AI tools they would use at work.

### Flags for you, not verdicts

Screen and activity are recorded for review. Integrity flags never fail anyone automatically, and the hiring decision stays with your team.

### $29 per candidate

Pay per session. Webcam and ID checks are there when a role needs them, and every session keeps its replay.

## Run this on one of your open roles

We set up the task, invite your candidates and walk you through the reports.

hiretokenwise.com