

AIOS INSTITUTE

AIOS Yield Model

How to use it. A ten minute tutorial.

Human, AI, or both? Work out, per unit of work, whether a person, AI, or a person using AI should do it, and be willing to accept keep it human.

aios.institute/resources/yield

1. What it answers, and when to use it

The Yield Model costs one unit of work three ways: a person alone, a person using AI, and AI alone. It returns one of four verdicts: Automate it, Augment the person, Keep it human, Not yet. All four are real answers. The page is built to say keep it human when that is true, and to show what that refusal costs.

Use it before a business case, not after. It takes a role, two numbers you know, and a few honest judgement calls. It needs no login and stores nothing: everything you enter lives in the link.

RESOURCE

Human, AI, or both?

Pick the role closest to the work and the answer appears at once: Automate it, Augment the person, Keep it human, or Not yet. Then give the two numbers that decide your case.

Support Agent

per resolved tier 1 ticket

Sales Development Rep

per researched lead

Accounts Payable Clerk

per processed supplier invoice

Recruiter

per screened CV

Issue Triage Engineer

per triaged bug report

Code Reviewer

per reviewed pull request

CI Failure Diagnostician

per diagnosed failed pipeline run

Security Analyst (vulnerability triage)

per triaged scanner finding

Bars: what a person, a person using AI, and AI alone cost per unit at our starting numbers.
Cheapest in gold. The verdict is not always the cheapest: this model will say keep it human, and show what that refusal costs.

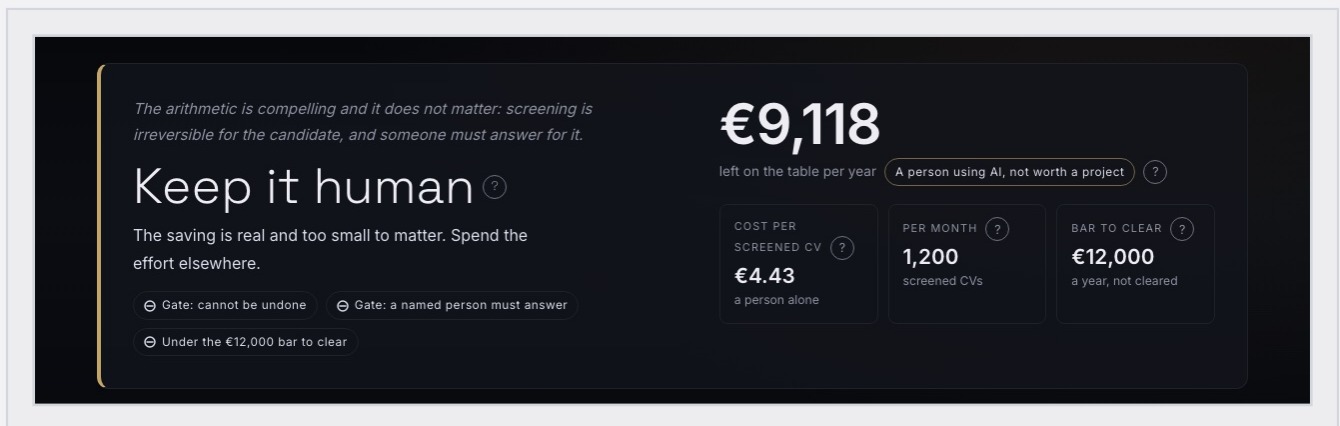
Try it: open aios.institute/resources/yield and pick the role closest to the work you have in mind.

2. The three minute walk

Pick a role: eight starting points, four business and four technical. The answer appears at once from our starting numbers. Read the verdict and the headline figure. Give the two numbers the page asks for; they are the two that decide this role. Type yours, or press Use this value if ours is close.

Check the five judgement calls. Unlock the Yield Card and share the link.

Everything else is optional: every number is editable, the model you use can be picked from a verified price list, and you can bring your own data.

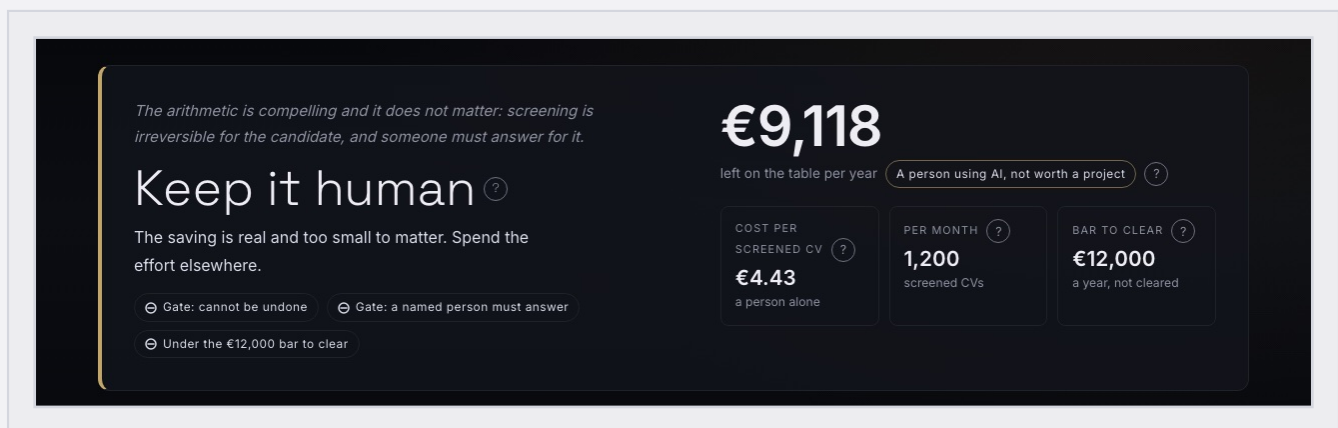


Try it: pick Recruiter, then Issue Triage Engineer. Two different verdicts from the same arithmetic.

3. Reading the verdict and the headline figure

The verdict word is the answer. The line under it says why in one sentence, and the chips name any judgement call that ruled a lane out. The headline figure is the money at stake per year: the saving of the chosen lane, or, for Keep it human, what the refused lane would have saved. Left on the table is not a recommendation. It shows what the judgement calls cost, on purpose.

Three tiles follow: cost per unit in the chosen lane, monthly volume, and the bar to clear, the yearly saving a project must exceed to be worth running.

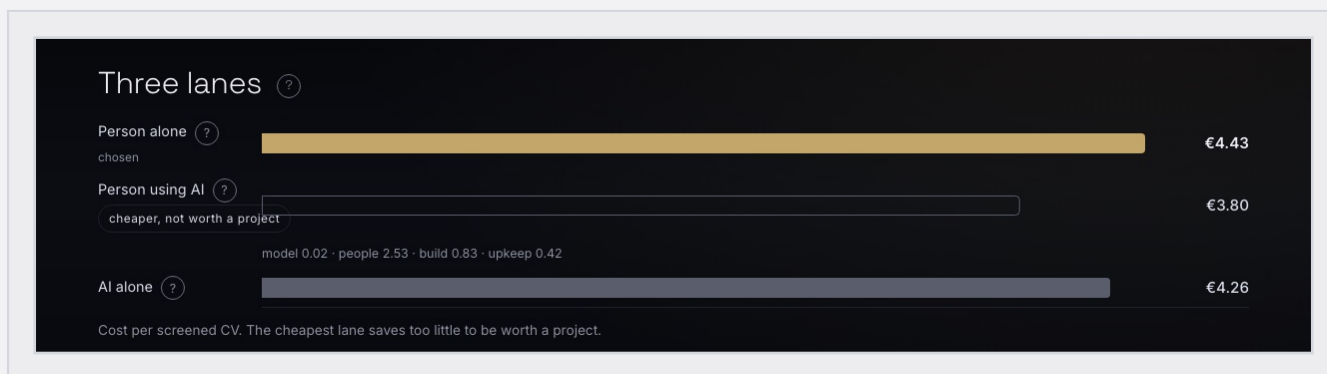


Try it: hover any ? on the page for what a figure is, what moves it and how to read it.

4. The three lanes

Each bar is the cost of one unit in that lane: people time, model runs, build and upkeep spread over the volume, and the cost of fixing bad outputs. The cheapest bar is gold. When a judgement call rules the cheapest lane out, the bar stays on the page with a chip saying so, and the verdict picks the cheapest lane that is permitted.

Under the lane the page argues about, the breakdown shows which term dominates. It is usually people time, rarely the model cost.

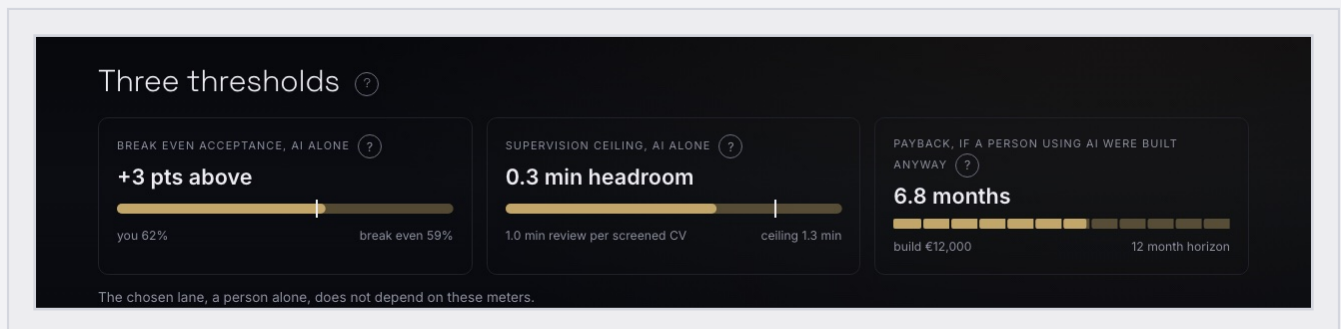


Try it: raise Review minutes in All numbers and watch AI alone grow while Person alone stays put.

5. The three thresholds

Break even acceptance is the acceptance rate at which AI alone costs exactly what a person costs; below it the AI lane loses on every unit. The supervision ceiling is the most review time per unit the AI lane can carry before the saving is gone. Payback is how many months of saving it takes to earn back the build, on a 12 month horizon.

How far you sit from each line tells you how fragile the verdict is. Two points of headroom on acceptance is not headroom.

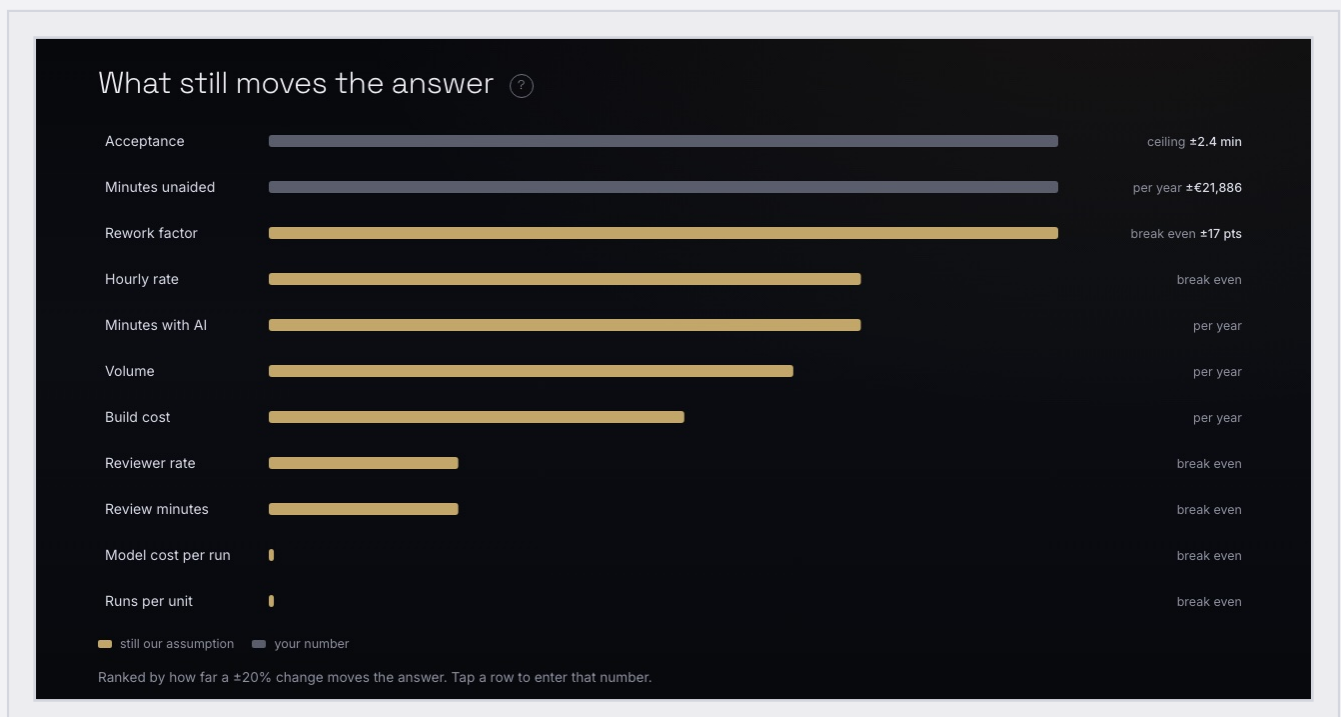


Try it: set acceptance to the break even value and watch the verdict change.

6. What still moves the answer

The tornado ranks every input by how far a 20 % change moves the answer at your current numbers. Yours are marked; ours are the ones worth measuring next. The top row is the number to check first: if it is still ours, the verdict is still ours too.

Tap a row to enter that number. The ? beside each input in All numbers shows the same swing as a sentence, so you can see what a change would do before you make it.



Try it: enter the top row's number and watch the ranking reshuffle.

7. Next number and the honesty gate

The Next number card asks one question at a time, starting with the two inputs frozen for the role. Each has a slider, a field, and Use this value for when our figure is close enough to yours.

The Yield Card unlocks only when those two are yours, whatever else you enter. That is the honesty gate. It is not there to collect emails; it is there so nobody shares a card built entirely on our assumptions. The pips count down 2, 1, 0.

Try it: enter your acceptance rate and the minutes a person takes unaided.

8. The five judgement calls

Arithmetic cannot answer these, so they override it. Does the AI output meet the same bar? Can a wrong output be undone? Must a named person answer for it? Is the know-how written down? How often does the process change? Each answer can cap a lane, so a person stays in the loop, or block the build entirely.

Filled disc: clear. Half disc: a warning. Ring with a bar: caps the lane or blocks the build. Answer them honestly first. A cheap lane that is not permitted is not an option.

Five judgement calls ?

- Does the AI output meet the person's standard? ?
 - Yes, the same bar
 - Not yet
- ⊖ If the output is wrong, can it be undone? ? caps at: person using AI
 - Yes, cheaply
 - Yes, at a cost
 - No
- ⊖ Must a named person answer for it? ? caps at: person using AI
 - No
 - Informally
 - By regulation or contract
- ⚠ Is the know-how written down? ? warning: caps at person using AI
 - Yes, reachable
 - Partly
 - No, in people's heads
- How often does the process change? ?
 - Rarely
 - A few times a year
 - Constantly

Filled disc: clear. Half disc: a warning. Ring with a bar: caps the lane or blocks the build.

Try it: for Recruiter, change Can it be undone to Yes, cheaply. The verdict flips, and what was left on the table becomes a saving.

9. All numbers, the model you use, and currencies

Open All numbers to edit any figure the model reads. The summary counts which are yours and which are ours. Pick the model you use from a price list checked against the vendors' pricing pages, with the source link and the date we checked. The cost per run is derived from your token counts and the list price, and the row shows the arithmetic.

Currency: your amounts stay exactly as typed. Only the model price converts, at the ECB reference rate shown on the page, or at your own rate. Nothing is converted behind your back.

All numbers (edit any)
16 inputs · 2 yours · 14 ours · model Claude Sonnet 5
CLOSE

Every number the model reads. ? Each one is yours or ours. ?

VOLUME AND PEOPLE

AI BEHAVIOUR

Volume
1200 a month ASSUMPTION ?
Hourly rate
38 € per hour ASSUMPTION ?
Reviewer rate
38 € per hour ASSUMPTION ?
Minutes unaided
7 min YOUR INPUT ?
Minutes with AI
4 min ASSUMPTION ?
Review minutes
1 min ASSUMPTION ?

Acceptance
good enough first time 62 % YOUR INPUT ?
Rework factor
1.4 × the fresh cost ASSUMPTION ?
Runs per unit
retries and re-runs 1.2 runs ASSUMPTION ?

PROJECT

Build cost
12000 € ASSUMPTION ?
Maintenance
not in the ranking: small effect, tracks the build 500 € a month ASSUMPTION ?
Subscriptions
seats and tools, added to maintenance 0 € a month ASSUMPTION ?
Bar to clear
the larger of €5,000 and the build €12,000 a year ASSUMPTION, THE BUILD COST ?

MODEL

ALL ANTHROPIC OPENAI GOOGLE MISTRAL

MODEL	USD IN / OUT PER 1M	USD PER RUN	EUR PER RUN	CHECKED
ANTHROPIC				
● Claude Fable 5.1 Anthropic · standard tier	10.00 / 50.00	0.0800	0.0701	23 Sep 2026
● Claude Opus 5.5 Anthropic · standard tier	4.00 / 20.00	0.0320	0.0280	23 Sep 2026
● Claude Sonnet 5 Anthropic · standard tier	2.00 / 10.00	0.0160	0.0140	23 Sep 2026
● Claude Haiku 4.5 Anthropic · standard tier	1.00 / 5.00	0.0080	0.0070	23 Sep 2026
OPENAI				

Tokens in per run
4000 tokens ASSUMPTION ?
Tokens out per run
800 tokens ASSUMPTION ?

Picking a model with our token counts does not count toward the card. Type your own tokens and it does.

Model price, Claude Sonnet 5
Anthropic · standard tier
USD 2.00 / 10.00 per 1M
PUBLISHED PRICE, CHECKED 23 SEP 2026 ?

Model cost per run
= 4,000 in × USD 2.00 + 800 out × USD 10.00 per 1M = USD 0.0160 × 0.8763 = EUR 0.0140 per run
€0.014 ASSUMPTION, FROM A PUBLISHED PRICE ?

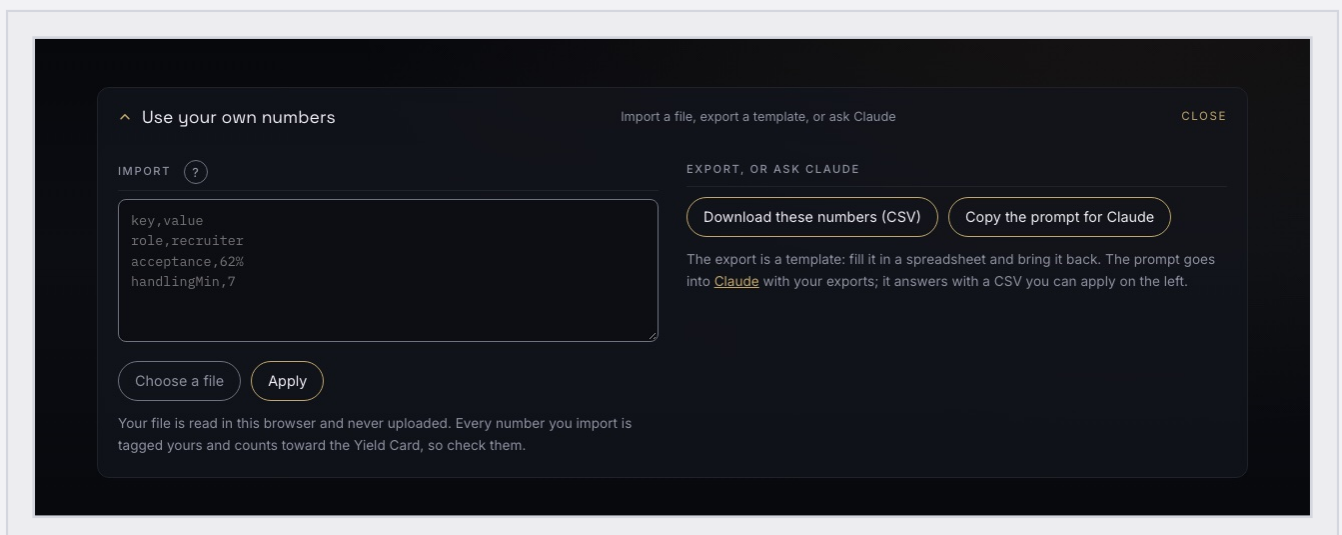
1 USD in EUR
ECB reference rate, 23 Sep 2026, inverted from 1.1411 USD per EUR
0.8763 EUR PUBLISHED RATE ?

Try it: pick a model, then type your tokens per run. The cost per run becomes yours.

10. Bringing your own numbers

Open Use your own numbers. Download these numbers gives you a CSV with every input, its unit and a one line description. Fill in what you know in a spreadsheet and Apply it back: every number you import is tagged yours and counts toward the Yield Card. Your file is read in the browser and never uploaded.

Copy the prompt for Claude gives you a prompt with every input's meaning, unit and range. Paste it into Claude with your CRM, ticket or timesheet exports; it answers with a CSV you can apply. Check anything an assistant estimated.

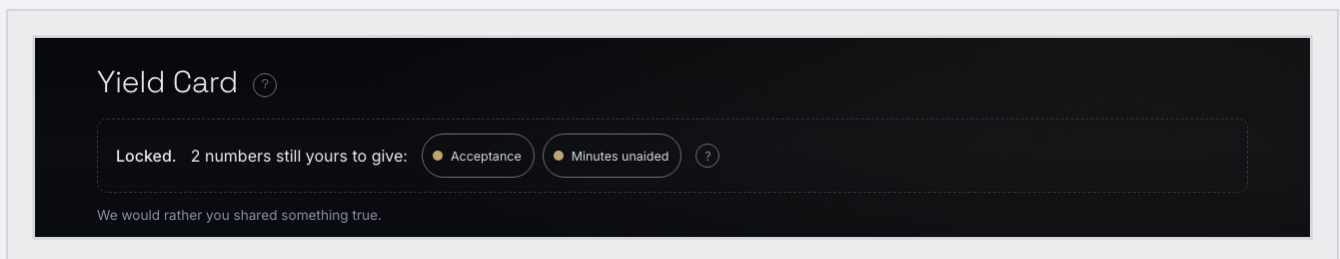


Try it: export, change two values, import.

11. The four verdicts, and what to do next

Automate it: AI alone is cheapest, the calls are clear, the build pays back. Rarer than the market implies; the marginal gates are the real risk. Augment the person: a person using AI wins. Budget the review time; it erodes first. Keep it human: not permitted, or too small to matter. Not an AI problem; spend the effort elsewhere. Not yet: the economics work and one named thing blocks the build. Fix that first.

Worked example, Recruiter: AI alone is cheapest, but a screening cannot be undone for the candidate and a person must answer for it, so the page keeps it human and shows about EUR 15,000 a year left on the table.



Try it: share the link. It carries every figure and where it came from.

12. Before you trust the number

- ☐ The two frozen inputs are yours, measured, not remembered.
- ☐ Acceptance was measured on real units against the real standard.
- ☐ Review minutes include the checking that actually happens, not the checking that is planned.
- ☐ The reviewer's rate is the reviewer's, not the replaced person's.
- ☐ Runs per unit came from logs.
- ☐ The judgement calls were answered by someone who would sign the output.
- ☐ The model price is published with a date, or it is your own contract price.
- ☐ The saving clears the bar and the payback sits inside the horizon.

If two of these fail, the verdict is ours, not yours. If you want a second pair of eyes on a case, say what you are looking at: aios.institute/contact.