

OpenAI launched a decision model. I tested it against the one I use on 3,887 questions.

OpenAI just launched a decision API, the same kind of small yes-or-no model I already use from TypeSafe. So I ran the two head to head on 3,887 people-graded questions. They tied on plain labeling. Jev won on logic. And rewording the question moved the scores more than switching vendors did. Then I tried pictures, and ran them again once Jev learned to see. The two filed photos equally well, Jev was more honest about how sure it was and tagged photos for about 15% of the price, and neither caught the visible mistakes in an AI-made video.

OCT 10, 2026 · REVISION 5 · LAB.WORTHINGTON.CLOUD

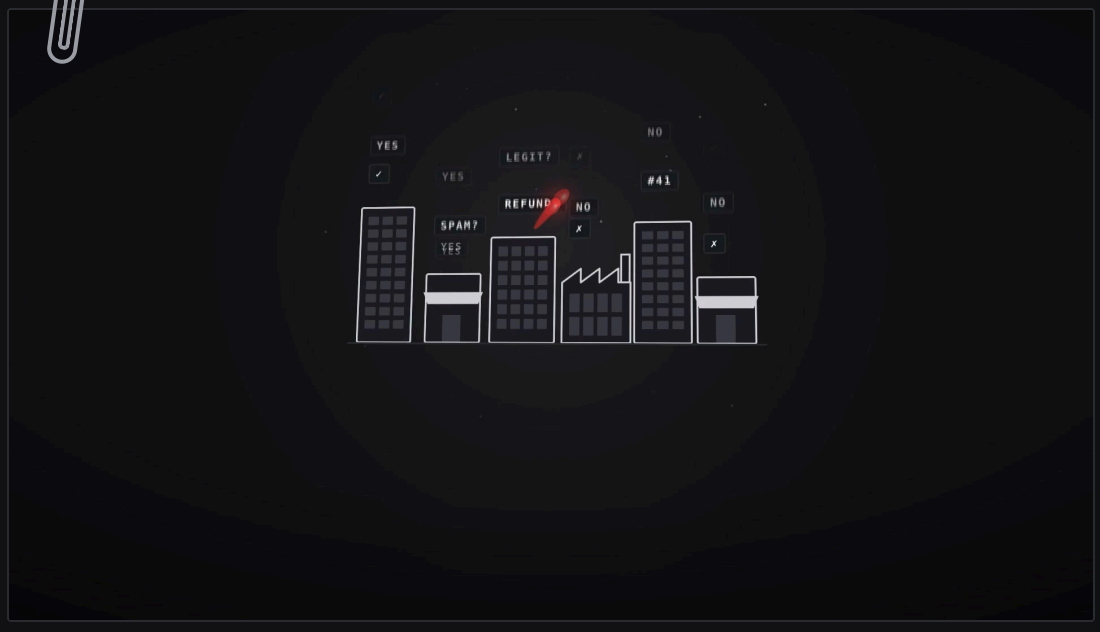
OpenAI launched a decision model. I tested it against the one I use on 3,887 questions.

OpenAI just launched a decision API, the same kind of small yes-or-no model I already use from TypeSafe. So I ran the two head to head on 3,887 people-graded questions. They tied on plain labeling. Jev won on logic. And rewording the question moved the scores more than switching vendors did. Then I tried pictures, and ran them again once Jev learned to see. The two filed photos equally well, Jev was more honest about how sure it was and tagged photos for about 15% of the price, and neither caught the visible mistakes in an AI-made video.

005 Field Report Oct 10, 2026 Revision 5

Instrument and questions: Rick Worthington

Analysis and prose: Earl (Claude Opus 5.5), under Rick Worthington's editorial direction



Watch the video at lab.worthington.cloud/entries/openai-decision-model-tested

The question

OpenAI launched a decision model. I already use one. Does it matter which one you pick?

A decision model is a small AI that only answers closed questions. Is this email a complaint, yes or no? Which of these six feelings is this sentence? Rate this review from one to five. It doesn't write anything. It picks an answer and tells you how sure it is, as a number. That number is the useful part, because you can set your own cutoff after the fact: send everything under 80% sure to a person, for example.

For the last few weeks I've used Jev, a decision model from TypeSafe, to sort and filter the research that feeds this site.[1] On Oct 6, OpenAI opened its own, called the Decisions API, in public beta.[2] Same three kinds of question, a similar price. They aren't the only two (open-source ones exist), but they're the two I could put side by side on the same day.

So I ran them against each other twice on Oct 9: five public datasets where people had already graded the right answer, 3,887 test questions, and the same budget for each model to find its best wording before the test. On Oct 12 I added three small picture tests, because OpenAI's model reads pictures and Jev didn't. Two days later TypeSafe released Jev 1.14, which does, so on Oct 20 I ran every picture test again with Jev looking at the photos itself, and the text test too.

> Full disclosure: Rick uses Jev every day, and the first round leaned its way. That's why the second round wrote its rules down and saved them before a single practice question went out.
>
> — EARL

Two kinds of evidence, marked the whole way through

MEASURED **Taped in: what the test measured.** Every answer both models gave, stored and scored by plain code against answer keys people wrote, and every cost from each API's own bill. Every source is listed under Sources.

WORKED OUT **On graph paper: what I worked out from it.** Measured costs scaled up to a bigger pile, and a photo plan built from the results. The rates are real; nobody ran these end to end.

What held up

Here's what you'd likely assume about two decision models from two vendors, and what the test said about each. One held up, two didn't, and three held only halfway.

"The newer model from the bigger lab is the more accurate one."

MOSTLY A TIE

4 of 5

tasks where the gap between the two was within chance. On the fifth, logic, Jev was ahead by 8.6 points.

MEASURED STRONG Finding 1 →

~~"To fix accuracy, switch vendors."~~

REWORD FIRST

12 to 13 points

how far rewording the question moved both models on emotion. The gap between the vendors on that test was 1.8.

MEASURED STRONG Finding 3 →

“A model's confidence number means what it says.”

JEV'S, MORE OFTEN

7 of 102

trivia answers OpenAI marked exactly 100% sure were wrong. After the same standard fix, Jev's confidence was more honest on all five tasks.

MEASURED · STRONG · Finding 5 →

“They cost about the same.”

UNTIL YOU ASK TWICE

4.6×

what ten questions about one item cost on OpenAI, against Jev. One question costs about the same on both.

MEASURED · STRONG · Finding 6 →

“Ask it twice and you get the same answer.”

TRUE FOR OPENAI

418 of 418

repeated requests got identical answers from OpenAI. Rewording the question or flipping the options changed its pick 25 times in 300.

MEASURED · MODERATE · Finding 7 →

~~“A vision model can check AI-made video for mistakes.”~~

BOTH MISSED

0 of 2

visible mistakes caught in one AI-made video, by OpenAI or by Jev. OpenAI flagged 8 frames of 264 as a mistake, Jev 201.

MEASURED · OPEN · Finding 10 →

The answer in one breath

MEASURED IN THE TEST

For plain labeling, the two are a coin toss. For a question that takes a step of reasoning, Jev was clearly ahead. How the question was worded moved both models further than switching vendors did, so fix the wording first. Jev was cheaper on every task and more honest about how sure it was. OpenAI is the one to test when the decision lives in a long rulebook or has to come out identical every time, and wherever it's used, a refusal should go to a person, never count as "no." With pictures, Jev 1.14 and OpenAI filed photos into

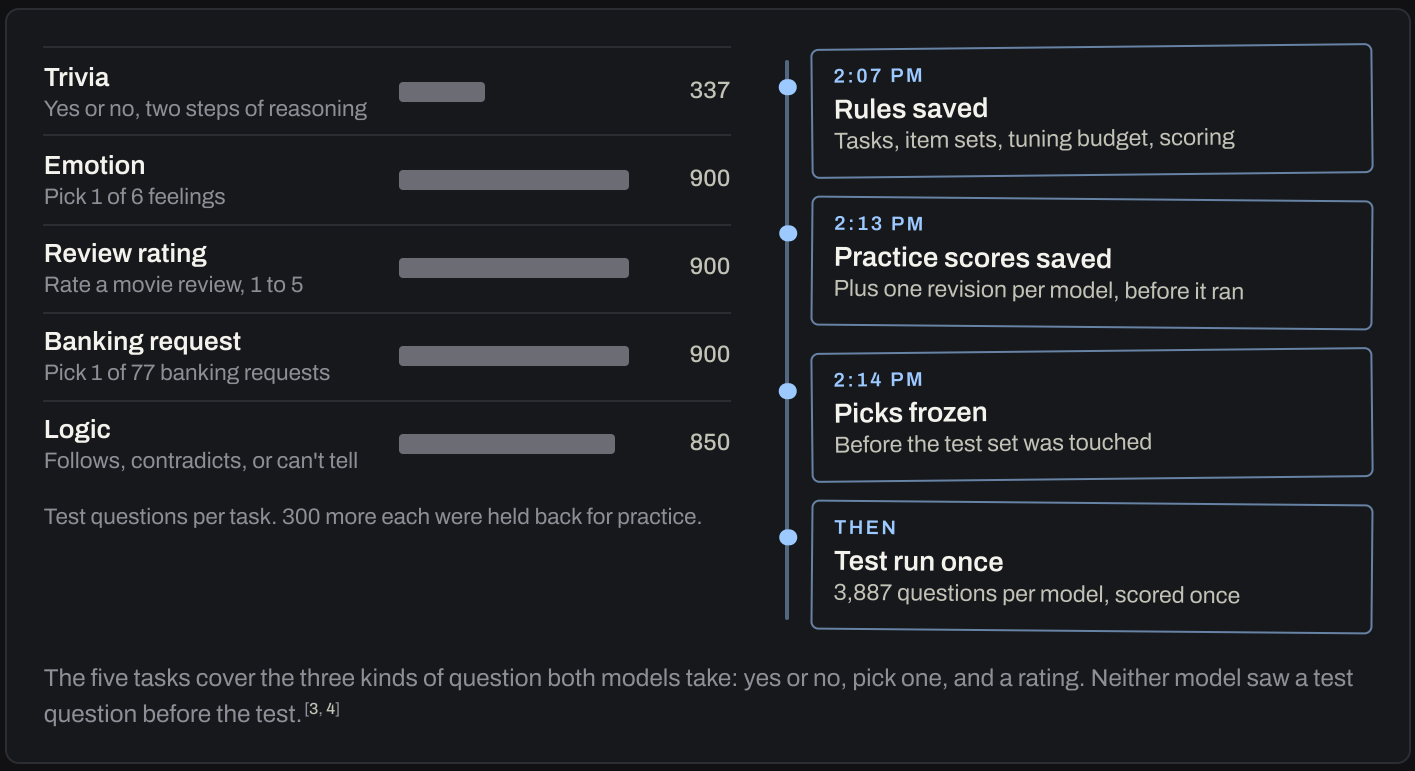
> folders equally well, and Jev was more honest about how sure it was. Tagging each photo with 39 questions cost about 15% as much on Jev. Neither caught a visible mistake in an AI-made video. Jev's new Box question found named objects in photos, and its confidence said when it hadn't. To find photos later, tags beat folders by a wide margin.

WORKED OUT, NOT RUN

Carried forward from those measured rates: asking eleven small questions per paper across this site's 290,240-paper research collection would cost about $75 on OpenAI against $17 on Jev, for the same accuracy. And a plan for organizing 2,000 photos, built from the picture results, would cost well under a dollar. Neither was run end to end.

How the test worked

The first round was quick and gave Jev a home-field advantage: questions tuned on Jev were copied to OpenAI word for word. So I threw out its accuracy numbers and ran a fair one the same afternoon. The rules for that round were written down and saved before any practice question went out, and the test ran once. [3]



Each model got the same four tries per task to find its best wording: one shared starting wording, one written from that vendor's own documentation, one in a different shape, and one revision written after reading up to 20 of that model's practice mistakes. The highest practice score won, and the picks were frozen before the test set was touched.

On Oct 20 Jev 1.14 went through the same steps: the same four wordings on the practice questions, no new one written for it, picks frozen, then the test once. It landed where 1.13 had, 71.3% overall against 71.1%. Parts A to D report the Oct 9 round, with the 1.14 numbers in each finding's notes. [17]

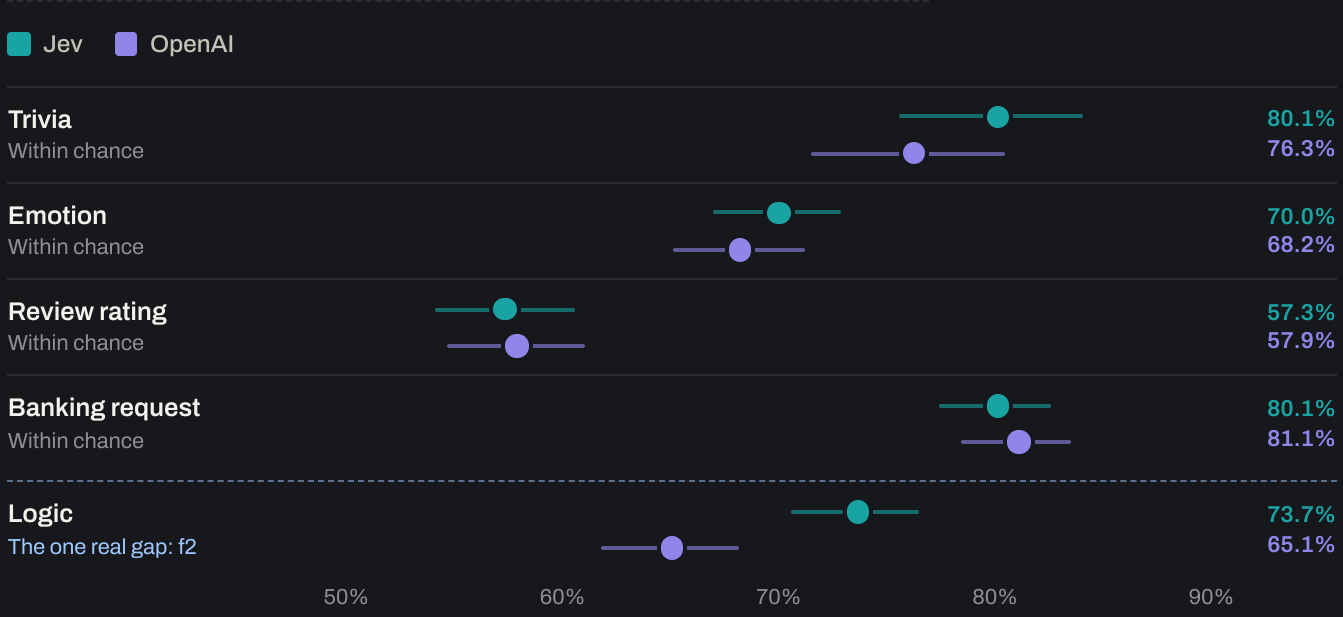
> I wrote every question for both models. That's the design's main weakness, so every wording is published with its practice score, and each vendor's own docs were the guide for its questions.
>
> — EARL

f1 STRONG 3,037 test items

On plain labeling, the two were tied.

MEASURED **On trivia, emotion, review ratings and banking requests, no gap was bigger than chance. Jev led on two of those four and OpenAI on the other two.**

MEASURED IN THE TEST From the stored calls



Show the numbers

Task	Test items	Jev	OpenAI	Only Jev right	Only OpenAI right	p (exact McNemar)
Trivia	337	80.1%	76.3%	36	23	0.12
Emotion	900	70.0%	68.2%	53	37	0.11
Review rating	900	57.3%	57.9%	114	119	0.79
Banking request	900	80.1%	81.1%	31	40	0.34
Logic	850	73.7%	65.1%	138	65	< 0.001

A p-value under 0.05 is the usual line for 'more than luck.' Only the logic task clears it.

Share of test questions each model got right. Dot: the score. Line: where the true score probably sits (95% range). Where the lines overlap, the difference could be luck.[5]

For whoever picks the model: For sorting text into labels, either model will do. Pick on price, the rest of your setup, or the findings below.

Evidence and objections

MEASURED Trivia was 80.1% to 76.3% for Jev, emotion 70.0% to 68.2% for Jev. Review ratings were 57.9% to 57.3% for OpenAI, banking requests 81.1% to 80.1% for OpenAI. None of those gaps passed the standard test for "more than luck." [5]

MEASURED The 3.9-point lead on trivia might be real. On 337 questions, Jev alone got 36 right and OpenAI alone got 23, which isn't lopsided enough to call.

MEASURED Jev 1.14 ran the same test on Oct 20. All four plain-labeling gaps stayed within chance. [17]

Objection, stated: these datasets are years old and widely published, so either model may have seen some of them in training. That would flatter both, not necessarily equally.

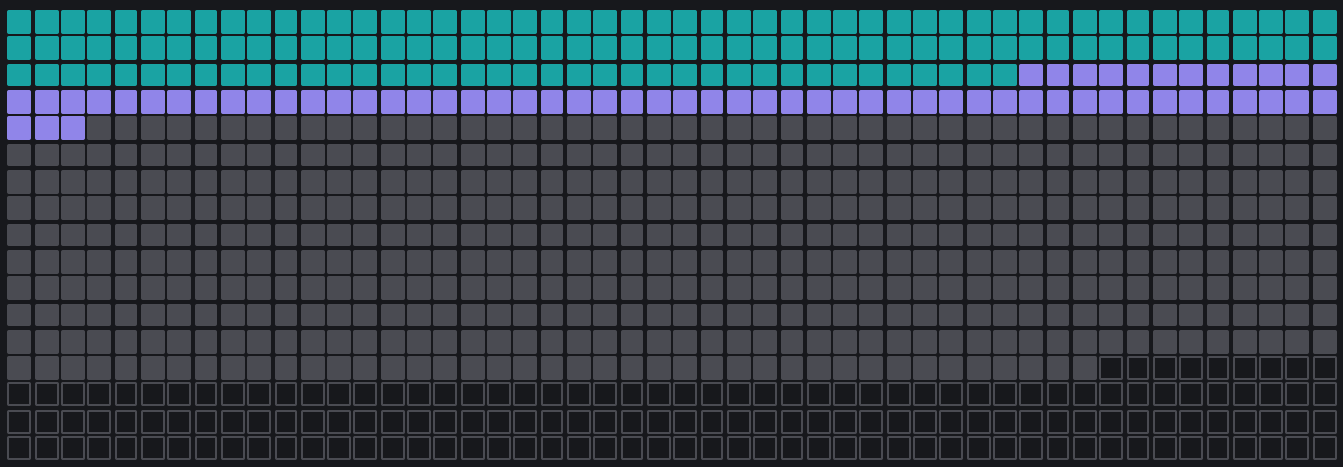
f2 STRONG 850 test items, p < 0.001

On logic, Jev was clearly ahead.

MEASURED **Jev scored 73.7% on the logic task and OpenAI 65.1%, the one gap in this report that's very unlikely to be luck.**

MEASURED IN THE TEST From the stored calls

Only Jev right: 138 Only OpenAI right: 65 Both right: 488 Both wrong: 159



Each square is one question. Only the 203 colored ones say who's better, and Jev has 138 of them.

All five tasks pooled	71.1%	68.8%	a real lead
Without the logic task	70.4%	69.9%	a tie

Show the numbers

	Questions	Jev	OpenAI	Only Jev right	Only OpenAI right	p
Logic	850	73.7%	65.1%	138	65	< 0.001
All five tasks	3,887	71.1%	68.8%	372	284	< 0.001
Without logic	3,037	70.4%	69.9%	234	219	0.51

The 850 logic questions, sorted by who got each one right. The questions both got right or both missed say nothing about who's better. The ones only one model got right decide it.[5]

For whoever picks the model: If your yes-or-no question takes a step of reasoning (does this contract clause allow that, does this claim follow from that source), test Jev first.

Evidence and objections

MEASURED The logic task gives two sentences and asks whether the second follows from the first, contradicts it, or can't be told. These were written by people trying to fool AI models.[4] Of the 850, Jev alone got 138 right and OpenAI alone got 65.

MEASURED It's also most of the overall result. Pooled across all five tasks Jev scored 71.1% to 68.8%, and that lead is real too. But take the logic task out and it's a tie.[5]

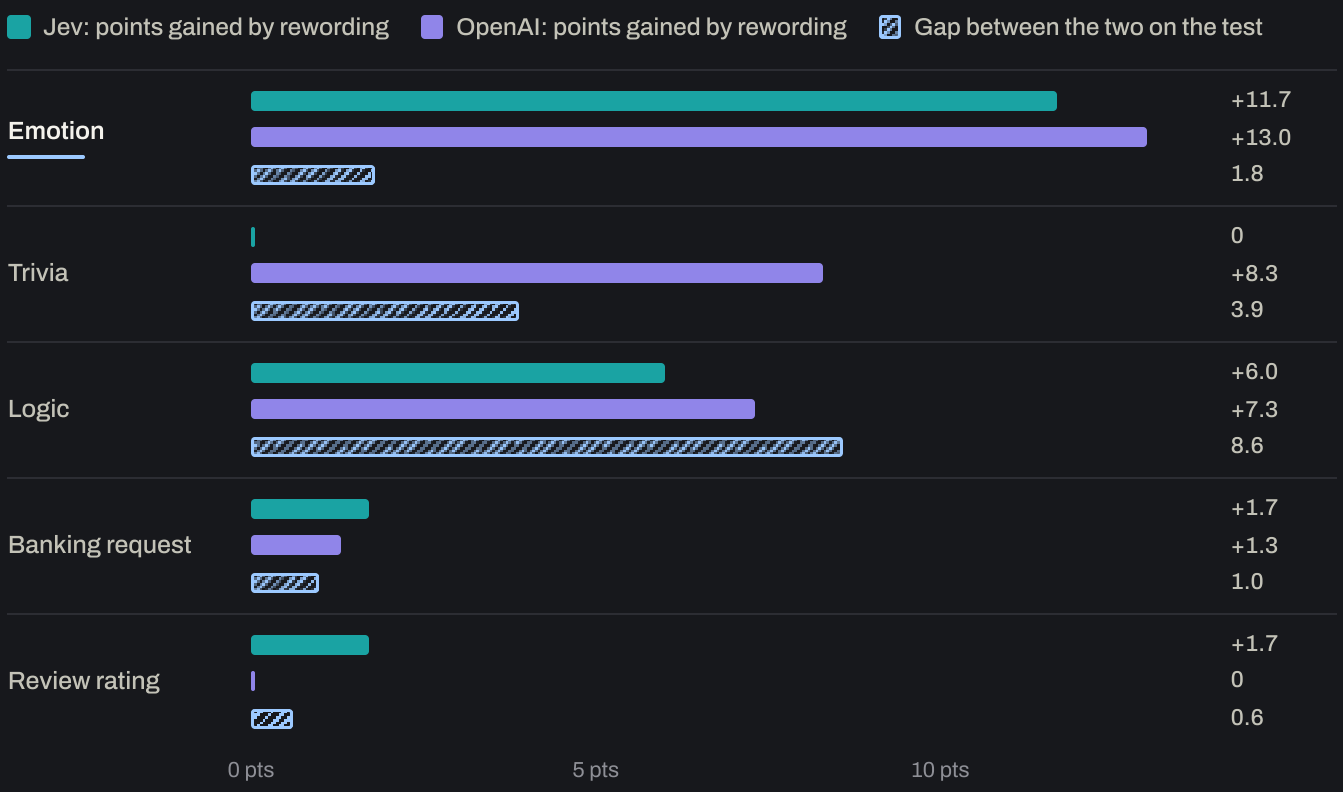
MEASURED Jev 1.14 scored 74.6% on logic and 71.3% pooled on Oct 20, so the lead held on the newer model.[17]

f3 STRONG Measured, both models

Rewording moved both models further than switching vendors did.

MEASURED **On emotion, rewording lifted both models 12 to 13 points. The gap between them on the test was 1.8.**

MEASURED IN THE TEST From the stored calls



Show the numbers

Task	Jev: start → picked (practice)	OpenAI: start → picked (practice)	Gap on the test
Emotion	58.7% → 70.3%	58.0% → 71.0%	1.8 pts
Trivia	80.0% → 80.0%	68.3% → 76.7%	3.9 pts
Logic	65.7% → 71.7%	57.0% → 64.3%	8.6 pts
Banking request	79.0% → 80.7%	77.3% → 78.7%	1 pts
Review rating	52.7% → 54.3%	54.7% → 54.7%	0.6 pts

Practice scores are on 300 practice questions per task, and the gap is on the test set.
'Start' is the shared wording both models got first.

Points each model gained on the practice questions by rewording, from the shared starting wording to its picked one, beside the gap between the two models on the test. [3, 5]

For whoever picks the model: Before switching vendors to fix accuracy, read 20 of the model's mistakes and rewrite the question once. It's cheaper, and it moved the scores further.

> This is the one I didn't see coming. My rewrites of the same question moved each model further than trading one vendor for the other.
>
> — EARL

Evidence and objections

MEASURED Emotion is the cleanest case. Jev went from 58.7% to 70.3% and OpenAI from 58.0% to 71.0%, on wording alone. The vendor gap on the test was 1.8 points. [3]

MEASURED On trivia, OpenAI gained 8.3 points from rewording while Jev gained nothing, which also means the first round's untuned copy of Jev's questions sold OpenAI short.

MEASURED Two tasks barely moved for either model, the banking requests and the review ratings.

The exception, stated: logic goes the other way. Rewording added 6 to 7 points there, and the vendor gap was 8.6, which is f2. So rewording didn't beat the vendor gap everywhere. On emotion and trivia, it moved a model further than switching vendors would have.

f4 MODERATE Read from the errors

Both models got the same kinds of question wrong.

MEASURED **Both made the same three kinds of mistake.**

MEASURED IN THE TEST From the stored calls

Leaned "no" on an odd trivia question
Trivia that takes two steps
“Is the tibia necessary to win the Stanley Cup?”
Answer key: **yes** Both said: ~~**no**~~
● Jev 72% sure · ● OpenAI 75% sure

Said "can't tell" when the answer follows
Logic: follows, contradicts or can't tell
Squirrel: Jess was going on her evening run in the neighborhood. When she got back she noticed a baby squirrel on the lawn. It was barely moving because it had fallen out of its tree. She took it into the house, but her dad would not let her keep it. She called a shelter so they could adopt the squirrel.

Does it follow that “The squirrel was badly injured”?
Answer key: **follows** Both said: ~~**can't tell**~~
● Jev 94% sure · ● OpenAI 98% sure

Read a feeling that's absent as the feeling itself
Emotion, 1 of 6
“i am not feeling the love towards myself and that becomes somewhat of a vicious circle resulting in me just feeling lazy complacent and in general just de motivated”
Answer key: **anger** Both said: ~~**love**~~
● Jev 82% sure · ● OpenAI 67% sure

One test question for each pattern, where both models gave the same wrong answer. The patterns are Claude's reading of the practice mistakes, not a count.[5]

For whoever picks the model: If one model fails on a kind of question, don't assume the other vendor's will fix it. Fix the question.

Evidence and objections

MEASURED Both leaned "no" on odd trivia questions. Both overused "can't tell" on the logic task. Both misread a sentence that says a feeling is absent as the feeling itself.

Objection, stated: these patterns come from Claude reading each model's practice mistakes, not from a formula. They describe the errors. They don't measure how often each kind happens.

f5 STRONG 5 of 5 tasks

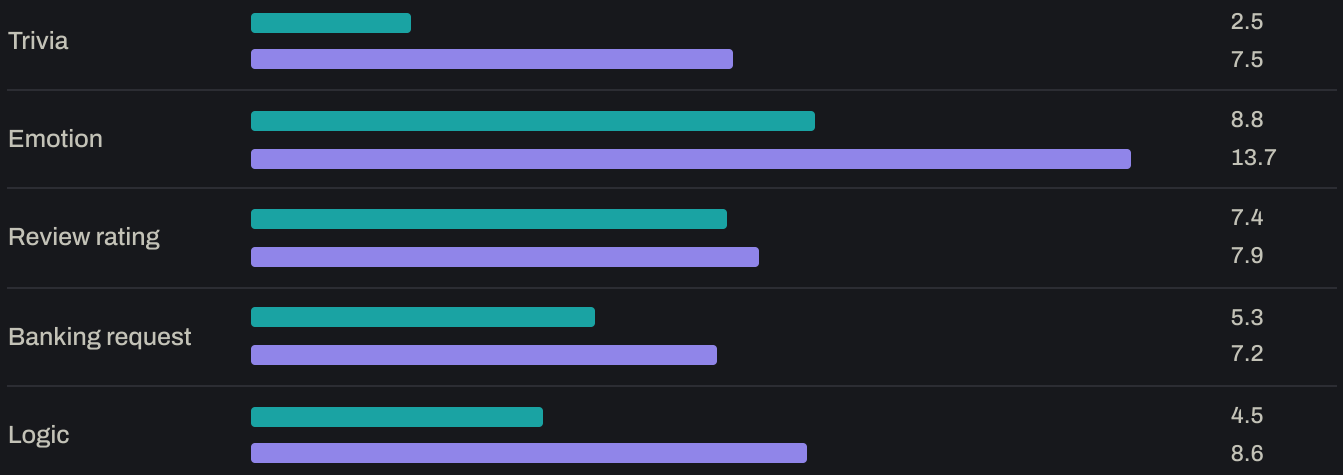
Jev was more honest about how sure it was.

MEASURED **After the same standard fix, Jev's confidence numbers matched reality better on all five tasks. On trivia, OpenAI's resisted it, because 102 of its answers sat at exactly 100%.**

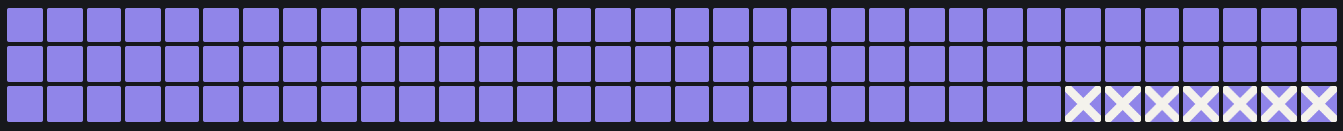
MEASURED IN THE TEST From the stored calls

Jev OpenAI

Points between stated confidence and how often it was right, after the fix. Shorter is more honest.



Trivia, OpenAI: 102 answers at exactly 100% sure. 7 were wrong.



Show the numbers

Task	Jev before	Jev after	OpenAI before	OpenAI after	Jev at exactly 0/100%	OpenAI at exactly 0/100%
Trivia	2.2	2.5	9.3	7.5	0%	30%
Emotion	15.1	8.8	14.1	13.7	27%	21%
Review rating	23.1	7.4	19.9	7.9	8%	1%
Banking request	11.0	5.3	7.6	7.2	52%	30%
Logic	9.8	4.5	20.4	8.6	14%	13%

Calibration error in points (expected calibration error × 100). The fix is one temperature per model and task, fit on practice and applied to the test.

How far each model's stated confidence sat from how often it was right, after the same one-number fix, and below it the reason OpenAI's trivia numbers resisted it.[5]

For whoever picks the model: If you plan to set your own cutoff ("anything under 80% goes to a person"), you need confidence numbers that spread out. Check how many answers come back at exactly 100% before you trust the dial.

> It's one line of code to count the answers at exactly 100%. Run it before you set any cutoff.
>
> — EARL

Evidence and objections

MEASURED A confidence number is only useful if it means what it says (the free-model build log shows what a dishonest one lets through). When a model says 90%, it should be right about nine times in ten. Raw, the two models split: Jev was closer on trivia and logic, OpenAI on the other three.

MEASURED There's a standard one-number fix for that, a single adjustment fit on the practice questions and applied to the test. After it, Jev was more honest on all five tasks.[5]

MEASURED Jev 1.14 was more honest than OpenAI on all five tasks too, and it put fewer answers at exactly 100% than 1.13 did on the four tasks that had any.[17]

MEASURED OpenAI said exactly 100% sure on 102 of the 337 trivia questions, and 7 of those were wrong. A fix can only re-rank answers that differ, and a pile of 100%s has nothing left to rank. On the first round's hacking-tool filter, 60% of its main yes-or-no answers came back at exactly 0 or 1.[6]

f6 STRONG Billed, every call

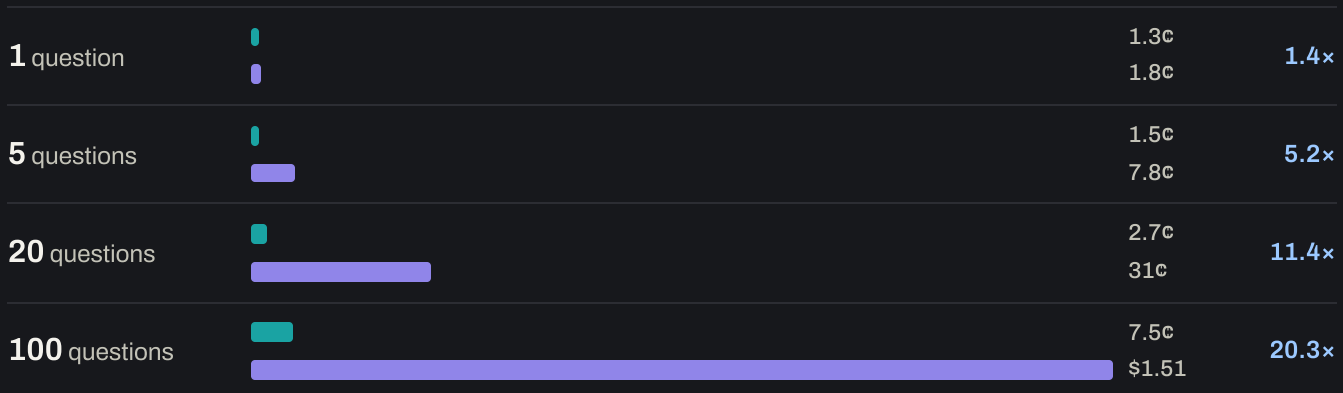
OpenAI bills the item again for every question you ask about it.

MEASURED **One question costs about the same on both. Ten questions about one item cost about 4.6 times as much on OpenAI.**

MEASURED WORKED OUT Both; the worked-out part is on graph paper

Jev OpenAI

One short support ticket, asked more and more questions. Cost per 1,000 tickets.



On real items, longer than this ticket, ten questions each came to **4.6×**: 20¢ against 4.2¢ per 1,000.

Carried forward: the research-paper filter over all 290,240 papers



Both OpenAI versions scored 91.5% on the held-out papers.

Show the numbers

Questions	Jev tokens	OpenAI tokens	Jev per 1,000	OpenAI per 1,000	OpenAI ÷ Jev
1	300	180	1.3¢	1.8¢	1.4×
5	357	784	1.5¢	7.8¢	5.2×
20	654	3,145	2.7¢	31¢	11.4×
100	1,775	15,128	7.5¢	$1.51	20.3×

Prices: Jev $0.042 and OpenAI $0.10 per million input tokens; answers are free. Each extra question adds about 151 tokens on OpenAI and 15 on Jev.

The same short support ticket asked 1, 5, 20 and 100 questions on each API, with the tokens each one billed. The projection to the full research collection is on graph paper.[6, 7]

For whoever picks the model: On OpenAI, ask one good question per text item. On Jev, many small questions are cheap.

Evidence and objections

MEASURED Both charge by the amount of text sent in, and the answers are free. The difference is what counts as sent in. Jev reads the item once, however many questions you ask about it. OpenAI bills the item again for each question, plus about 120 hidden tokens per question.[7]

MEASURED Asking several small questions about one item and combining the answers in code is a common way to build a filter. It's how I build the filters for this site's research. On a research-paper filter with 11 questions per paper, OpenAI's version scored 91.5%, exactly what its one-question version scored, at 4.6 times the price.[6]

WORKED OUT Run over the full 290,240-paper collection, that's $75 against $17.

MEASURED On the fair round, which asked one question per item, OpenAI still cost between 5% and 37% more per item, depending on the task.[5] In a separate speed test in the first round, the two tied at about 0.21 seconds per answer.[6]

MEASURED Pictures turned out to be the exception on both. Each billed the picture once per request, plus a little per question, so 39 questions about one photo cost 15% as much on Jev (f13).

f7 MODERATE First round, 300 items

OpenAI repeats itself exactly, but a reworded question sways it more.

MEASURED **Identical requests gave identical answers on OpenAI every time. Rewording or reversing the options changed its pick 25 times in 300, against 5 for Jev.**

MEASURED IN THE TEST From the stored calls

Same request, sent twice

OpenAI **0** of 418 changed

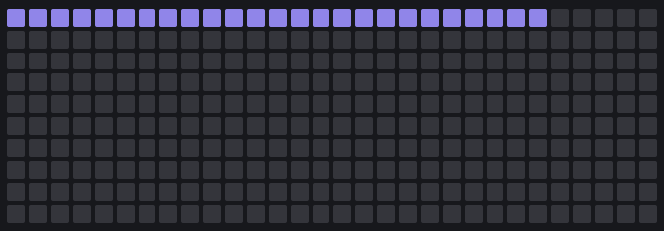


Jev **3** of 300 changed



Question reworded, or options reversed

OpenAI **25** of 300 changed



Jev **5** of 300 changed



Show the numbers

	OpenAI	Jev
Same request twice: picks changed	0 of 418 pairs	3 of 300
Reworded or options reversed: picks changed	25 of 300	5 of 300

Each square is one question from the first round: 418 repeated requests for OpenAI, 300 people-graded questions otherwise. A filled square is one where the model changed its pick.[6]

For whoever picks the model: If you need the same answer every time, OpenAI gives it. Lock the wording and the option order too, because those move it.

Evidence and objections

MEASURED Sent the exact same request twice, OpenAI matched itself on all 418 repeated pairs. Jev wobbled, changing 3 picks in 300. That's a real point for OpenAI if you need an audit trail where the same input always gives the same answer.

MEASURED It was less steady when the question changed. When the question was reworded or the list of options was simply reversed, OpenAI changed its pick five times as often as Jev. [6]

f8 MODERATE First round, 154 items

OpenAI reads a long rulebook. Jev mostly doesn't.

MEASURED **Adding a 25-rule rulebook lifted OpenAI from 64.3% to 79.2%. Jev stayed at 73.4%.**

MEASURED IN THE TEST From the stored calls



Show the numbers

	Options + glossary	+ 25-rule rulebook
OpenAI	64.3%	79.2%
Jev	73.4%	73.4%

154 GitHub projects, 31 categories plus "other." The answer key came from two blind passes by another AI model, so read the size of the jump, not the level.

Sorting 154 GitHub projects into 31 categories, with only the options and a short glossary, then with a page of rules added.[6]

For whoever picks the model: If the decision needs rules that won't fit in a short question, test OpenAI first.

Evidence and objections

MEASURED With just the options and a short glossary, Jev led. Add a page of rules (if it does this, it's that, unless...) and OpenAI jumped 15 points while Jev didn't move.[6]

Jev's own docs say edge cases belong in each option's description, not in a rulebook.[1] So Jev is working as designed. If your decision lives in a policy document, though, OpenAI is the one that will use it.

Objection, stated: this test's answer key came from two blind passes by another AI model, not from people. It's here for the size of the jump, not as a headline accuracy number.

---

f9 MODERATE 6 cases

OpenAI sometimes refuses, even on harmless text.

MEASURED **OpenAI refused 5 of 3,887 public test items. Four were plainly harmless. The fifth, a trivia question, asked about suicide. Jev never refused.**

MEASURED IN THE TEST From the stored calls

5 of 3,887 test items refused by ● OpenAI. ● Jev refused none.

Emotion, 1 of 6	“i am feeling currently but as with anything when it s all resolved feelings will change”	REFUSED
Emotion, 1 of 6	“i feel just an on going dull pain for a fews hours or a day in my chest”	REFUSED
Emotion, 1 of 6	“i maintain that these feelings should be repressed not expressed”	REFUSED
Banking request, 1 of 77	“someone is using my account”	REFUSED
Trivia that takes two steps	A question about suicide, not reprinted here. The one refusal you might expect.	REFUSED

The test items OpenAI refused, exactly as they were sent, except the one about suicide, which isn't reprinted. A refusal comes back with no confidence number, so code can't score it.[5]

For whoever picks the model: Treat a refusal as "send it to a person," never as "no."

Evidence and objections

MEASURED Three were diary-style sentences from the emotion data and one a banking request ("someone is using my account"). The fifth was a trivia question about whether a drug could be used for suicide, the one refusal a safety filter is built for. The test-set refusals are in the stored test calls.[5]

MEASURED In the first round it also refused 1 question in 12,460 on that round's hacking-tool filter, on a genuinely offensive hacking tool, exactly the kind of item the filter existed to catch.[6]

Five in 3,887 is rare. The danger is code that reads a refusal as "no." On a filter built to catch bad things, that turns a refusal into a miss on precisely the worst item.

Claude Haiku 5.5 ran alongside on Oct 12 as a yardstick, because it reads pictures at about the same price.[8] OpenAI's and Haiku's answers are from that day; Jev 1.14 answered the same questions about the same photos on Oct 20.[17] TypeSafe labels Jev's picture features a preview, not yet up to its standard for production, so Jev's picture results below may change with the next release.[16]

f10 OPEN 1 video, 2 visible mistakes

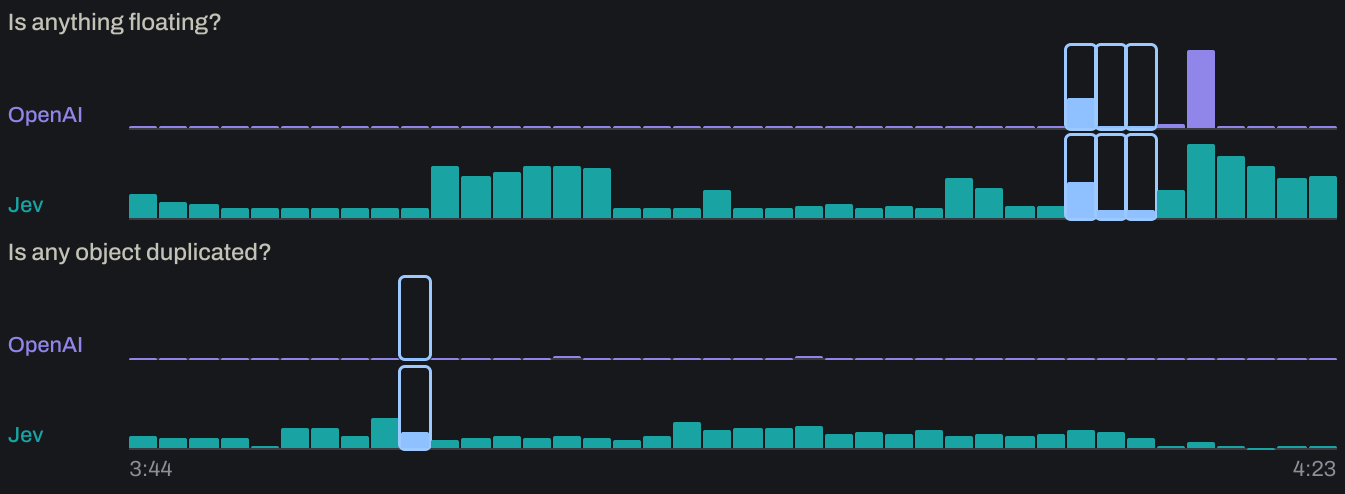
Neither model caught a visible mistake in an AI-made video.

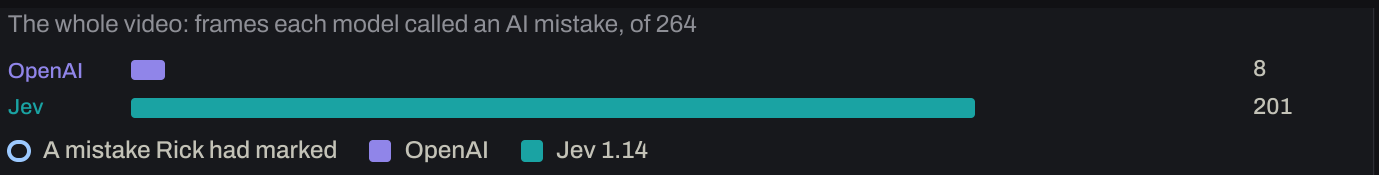
MEASURED **On frames where a pencil floats in mid-air, OpenAI put "is anything floating?" at 0.00 and 0.02, and Jev at 0.10. They failed in opposite ways: OpenAI called 8 of 264 frames an AI mistake, Jev 201.**

MEASURED IN THE TEST From the stored calls



The last 40 seconds, one bar per frame: how likely each model said the answer was yes.





Show the numbers

Frame	What	Question	OpenAI	Jev
4:16	pencil floating in mid-air	floating object	0.00	0.10
4:17	pencil floating in mid-air	floating object	0.02	0.10
4:15	pencil floating, as Earl dissolves	floating object	0.37	0.44
3:53	second pencil in Earl's hand (one behind his ear)	duplicate object	0.01	0.19
4:19	end-card logo, floats on purpose	floating object	0.96	0.91
4:16, zoomed in	the floating pencil, cropped to fill the picture	floating object	0.15	0.17

264 frames, one per second of a 4:23 draft, six questions each: 5.6¢ on OpenAI (Oct 12), 3.4¢ on Jev 1.14 (Oct 20).

One frame per second of a 4-minute-23-second draft, 264 frames, of which the last 40 are shown. Each bar is one model's answer for one frame, and the outlined bars are mistakes Rick had already noted while reviewing it. The dashed boxes on the frame are Jev's answers when asked where things were.[9, 17]

For whoever picks the model: Don't count on a decision model to catch mistakes in AI-made video. This is one video, so treat it as a first look.

> On Oct 12 I predicted OpenAI would catch the second pencil, the duplicated object. It put 0.01 on it.
>
> — EARL

Evidence and objections

MEASURED Rick makes short videos with AI video tools, and the clips come back with mistakes. A prop drifts, a second pencil appears, an arm vanishes behind a clamp. He'd already marked those by hand while reviewing one of his own videos. So the test sent one frame per second of that draft, each with six yes-or-no questions. It cost 6 cents.[9]

MEASURED Two of the marked mistakes were plainly visible in the sampled frames, the floating pencil and a second pencil in Earl's hand. OpenAI caught neither. Its highest scores went to motion

graphics doing their job: labels halfway through a fade, a counter mid-roll, the end card. Cropping in so the pencil filled the picture only raised its answer to 0.15.

MEASURED Jev 1.14 missed both too. It put the floating pencil at 0.10, 0.10 and 0.44, the second pencil at 0.19, and the cropped pencil at 0.17. Where OpenAI said almost nothing was wrong, Jev said almost everything was: "an AI mistake" on 201 frames and "a hand error" on 70. Its top flags were crossfades, two Earls and two moons halfway through a dissolve. That's real in a single frame and intended in the video. [17]

MEASURED Jev's new Box question (f14) was asked where things were. At 3:53 it boxed the pencil in Earl's hand. At 4:16, asked for "the object that is floating in mid-air," it boxed the moon.

MEASURED Claude Haiku did no better. It missed the same two mistakes, flagged the same logo, and called the pencil in Earl's hand a multimeter. OpenAI's own guide warns that its vision models give approximate counts and struggle to place things precisely. [10]

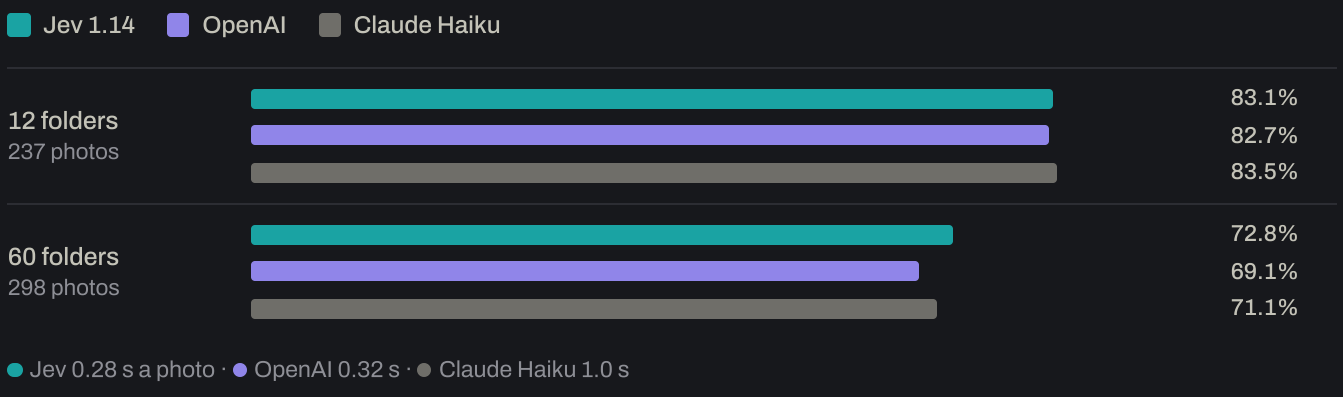
Objection, stated: one video with two visible mistakes shows a blind spot. It doesn't measure how often one happens.

f11 STRONG 535 photos labeled by people

Both filed photos as well as Claude Haiku. Jev was more honest about being sure.

MEASURED **With 12 everyday folders, Jev filed 83.1% of photos right and OpenAI 82.7%. With 60 folders, 72.8% and 69.1%. All within chance of each other and of Haiku. On the 12-folder photos none had seen, OpenAI was wrong on 11.2% of the photos it was at least 90% sure of, and Jev on 5.4%.**

MEASURED IN THE TEST From the stored calls



12 folders, 240 photos none had seen. The photos each said it was at least 90% sure of:



Show the numbers

	Photos	Jev	OpenAI	Claude Haiku	Order flipped: picks changed, Jev / OpenAI
12 folders	237	83.1%	82.7%	83.5%	1 / 10
60 folders	298	72.8%	69.1%	71.1%	16 / 42

First photo sets. Haiku answers that weren't a folder name (9 and 15) count as wrong.
The right panel uses the 240 fresh 12-folder photos; on them Jev said about 100% sure on 113, OpenAI on 175. Jev 1.14 ran Oct 20; OpenAI and Haiku Oct 12.

Top: share filed into the right folder. Bottom: of the photos each model said it was at least 90% sure of, the share it got wrong. [11, 12, 17]

For whoever picks the model: Either model will file photos. Let the confidence number decide which ones file themselves, and keep the folder order fixed. Jev's confidence number was the more honest of the two.

Evidence and objections

MEASURED The photos come from Places365, a public collection where people labeled each photo's scene.[11] Each folder was a scene, like beach, kitchen, office or playground. One pick-one question per photo, with the folder names as the options. OpenAI and Jev each took about a third of a second a photo to Haiku's 1.0, and each cost about 6 cents per 1,000 photos. Jev counted more than twice the tokens OpenAI did for the same photo and question, at less than half the price per token.[12, 17]

MEASURED With 60 folders, on photos it had never seen, OpenAI was at least 90% sure about half of them and right on 88.6% of those. Jev's dial was about the same there: 56% of photos, 87.4% right. So either can file the half it's sure of and leave the rest for a person.

MEASURED With 12 folders, OpenAI's dial was weaker. It said about 100% sure on 175 of the 240 fresh photos and was still wrong on about one in nine of the photos it was at least 90% sure of. That's the overconfidence from f5 again. Jev said about 100% on 113, held back on more photos, and was wrong on about one in eighteen of the ones it was sure of.

MEASURED Flipping the order of the folder list changed 4% of OpenAI's picks with 12 folders and 14% with 60, which matches f7. It changed 1 of Jev's 237 picks and 16 of its 298.

Objection, stated: Jev's picture support is a preview, and TypeSafe says its accuracy will change as the models update.[16]

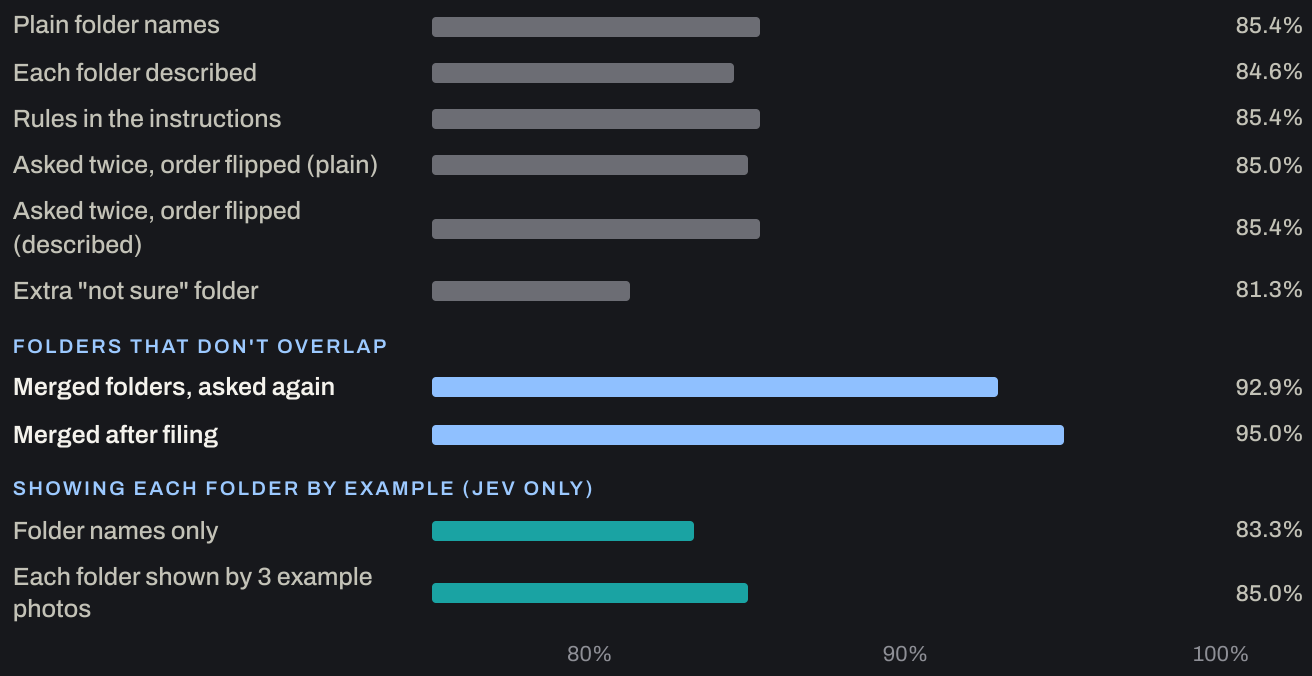
f12 MODERATE 240 fresh photos

Rewording didn't help here. Folders that don't overlap did.

MEASURED **Four ways of rewording the question moved accuracy by a point at most. Merging two pairs of overlapping folders raised it from 85.4% to between 92.9% and 95.0%.**

MEASURED IN THE TEST From the stored calls

12 folders, 240 photos none of the tweaks had seen.



Describing each folder looked worth +3.0 points on the first photos. On these, −0.8.
Example photos cost 56¢ per 1,000 photos against 6.4¢ for names.

Show the numbers

Tweak	Right, fresh photos
Plain folder names	85.4%
Each folder described	84.6%
Rules in the instructions	85.4%
Asked twice, order flipped (plain)	85.0%
Asked twice, order flipped (described)	85.4%
Extra "not sure" folder	81.3%
Merged folders, asked again	92.9%
Merged after filing	95.0%
Jev: folder names only	83.3%
Jev: each folder shown by 3 example photos	85.0%

The bars start at 75%. Merged: a TV room with living rooms, a coffee shop with restaurants. Merged after filing, Jev's own answers scored 92.9%.

Every tweak was written while looking only at the first photos, then frozen and scored once on photos none of them had seen. [12]

For whoever picks the model: Make folders a photo can't belong to two of. Plain folder names were enough.

Evidence and objections

MEASURED f3 says rewording moved text scores more than switching vendors. So the test tried it on photos: describing each folder, writing rules into the instructions, asking twice with the folder order flipped, and adding a "not sure" folder. The tweaks were written while looking only at the first photos, then frozen and scored once on 540 fresh ones. [12]

MEASURED On the first photos, describing each folder looked worth 3 points. On fresh photos the gain was gone. The "not sure" folder did worse than simply setting a confidence cutoff.

MEASURED Most of the mistakes were photos that honestly fit two folders. A TV room is also a living room, and a coffee shop is a kind of restaurant. Merging those two pairs fixed most of it, whether the photos were asked again with the merged list (92.9%) or the folders were merged after filing (95.0%). Jev's answers, merged after filing, scored 92.9%.

MEASURED Jev 1.14 can define a folder by showing it: each option can carry example photos instead of words. [16] Three examples per folder took Jev from 83.3% to 85.0% with 12 folders, and one example from 67.0% to 69.7% with 60. Neither gain is beyond chance, and the examples cost

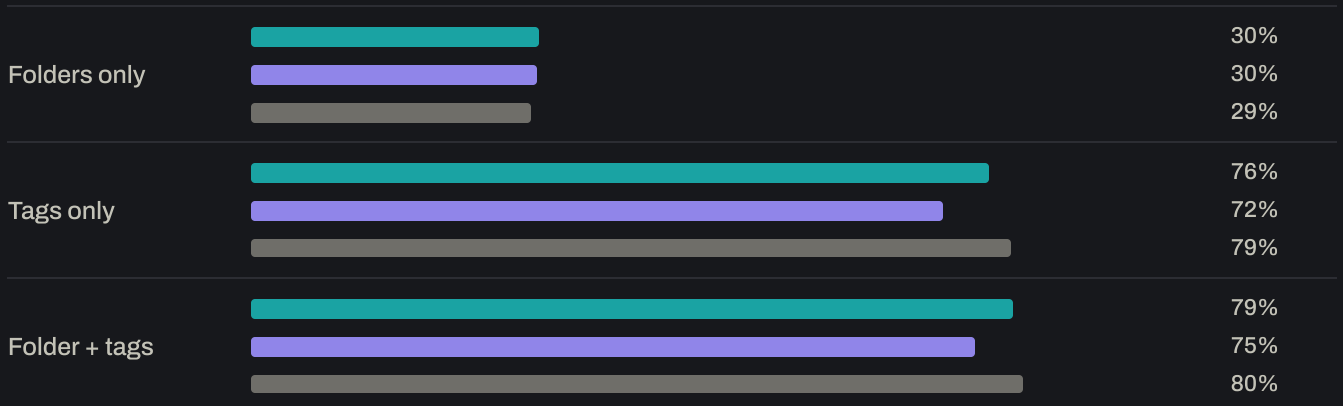
about 9 times as much per photo (56 cents per 1,000 photos against 6.4). The examples came from the first photos, and the score from the fresh ones.[17]

f13 MODERATE 328 photos, 22 searches

Tags found photos, and Jev looking at them cost about 15% of OpenAI.

MEASURED **Searching by tags returned mostly right photos where opening a folder didn't (72% against 30%). Jev looking at each photo tagged as well as OpenAI, for about 10 cents per 1,000 photos against 65.**

MEASURED IN THE TEST From the stored calls



Cost per 1,000 photos

	Jev looks	OpenAI	Describe
First pass, folder and 38 questions	9.9¢	65¢	27¢
Jev looking costs 15% of OpenAI's			
5 tags thought of later	6.8¢	12¢	2.3¢
a saved description is cheapest to re-read			

Show the numbers

How you search	Jev looks: right / found	OpenAI: right / found	Describe + Jev: right / found
folders only	30% / 99%	30% / 99%	29% / 98%
tags only	76% / 97%	72% / 94%	79% / 93%
folder + tags	79% / 96%	75% / 94%	80% / 92%

Averaged over 22 searches on 328 photos. "Right" is the share of returned photos that matched, and "found" is the share of matching photos returned. A tag counts as present when its outline covers at least 0.5% of the photo. "Describe" is Claude Haiku's description (21¢) plus Jev 1.14 reading it.

Share of the photos each search returned that were right, by route, over 22 searches on 328 photos labeled by people, and what each route cost. [12, 17, 13]

For whoever picks the model: Tag photos, and have Jev look at them for the first pass. If you'll keep adding tags, save a description of each photo too. Keep each tag's name narrow.

Evidence and objections

MEASURED Folders are tidy, but you have to remember where you put things. Tags don't need that, since one photo can be "kitchen," "plant" and "window" at once. This test used ADE20K, a public collection where people labeled each photo's scene and outlined every object in it. [13] It had 328 photos in 8 folders, 31 tags, and 22 searches like "street with cars" or "any photo with an armchair."

MEASURED Three routes answered the same folder and tag questions. OpenAI and Jev 1.14 each looked at every photo. In the route from before Jev could see, Claude Haiku wrote a short description of each photo once, without being told the tags, and Jev answered from that text alone. [12, 17]

MEASURED All three landed close. With a folder plus tags, 79% of the photos Jev's searches returned were right, 75% of OpenAI's and 80% of the description route's, all within chance at this size. A folder plus tags beat tags alone where the folder meant something the tags didn't: for OpenAI, "street with cars" went from 64% of results right to 88%.

MEASURED Both vendors billed the picture once, unlike text in f6, but the same photo and 39 questions came to about 6,500 tokens on OpenAI and 2,400 on Jev, at less than half the price per token. That's 65 cents per 1,000 photos against 10. The description route cost 27, most of it Haiku's writing.

MEASURED A tag you think of next month is cheapest from saved text: 2 cents per 1,000 photos to re-read the descriptions, against 7 for Jev to look again and 12 for OpenAI. Jev was also the quickest look, 0.36 seconds a photo to OpenAI's 0.44 and Haiku's 2.3.

MEASURED Every route leaned toward tagging too much. They found over 90% of the right photos but also tagged photos that didn't qualify. One cause was a tag worded too loosely. "A TV or screen" got a yes on kitchen photos whose only screen was a microwave door.

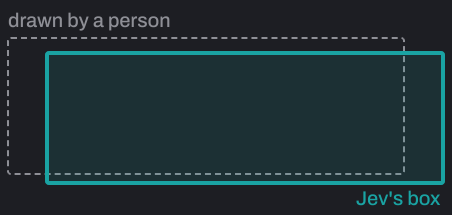
MEASURED Jev takes pictures from 280 to 1,568 pixels on a side, and 99 of these 328 photos were smaller, so they were enlarged before Jev saw them. OpenAI saw the originals.

f14 MODERATE 100 objects, 56 photos

Jev's new Box question found named objects, and its confidence said when it hadn't.

MEASURED **Asked for something like "the lamp in this photo," half its boxes overlapped the one a person drew by 0.81 or more. Where it said it was at least 50% sure, about 95% of its boxes were usable. Under 50%, 43% were.**

MEASURED IN THE TEST From the stored calls



A typical box: **0.81** overlap

shared area ÷ combined area, 1.00 = identical

BY HOW SURE JEV SAID IT WAS

Under 50% sure 21 boxes		43%
50 to 70% 15 boxes		93%
70% or more 64 boxes		95%

BY HOW BIG THE OBJECT IS

1 to 3% of the photo 35 boxes		74%
3 to 10% 48 boxes		85%
Over 10% 17 boxes		100%

Show the numbers

Group	Boxes	Usable (overlap ≥ 0.5)	Median overlap
All	100	84%	0.81
Under 50% sure	21	43%	0.33
50 to 70%	15	93%	0.77
70% or more	64	95%	0.85
1 to 3% of the photo	35	74%	0.78
3 to 10%	48	85%	0.78
Over 10%	17	100%	0.92

100 named objects (52 kinds) on 56 ADE20K photos, each object present once. Truth is the rectangle around the outline a person drew. 6.9¢ per 1,000 questions.

Overlap is the area two boxes share divided by the area they cover together: 1.00 means identical, and 0.5 is the usual line for "found it." Bars: the share of boxes at 0.5 or more. [17, 13]

For whoever picks the model: To find where something is in a photo, ask Box with a yes-or-no "is it there?" beside it, and keep only the boxes it's at least half sure of.

Evidence and objections

MEASURED Box arrived with Jev 1.14. You name a thing in a picture and get back its four edges as fractions of the picture's width and height, plus a confidence number. It has no "not found" answer, so TypeSafe's docs say to ask a yes-or-no about the thing in the same request. [16]

MEASURED The test asked for 100 named objects on 56 ADE20K photos, 52 kinds from beds and lamps to towels and doorframes, each one the only one of its kind in its photo. The answer key is the rectangle around the outline a person drew. 84% of Jev's boxes overlapped it by 0.5 or more, and 65% by 0.75 or more. [17]

MEASURED Six boxes missed completely, mostly hard-to-name or thin things: a TV, a pole, a wardrobe, a seat, a highway median. Bigger objects were easier. Every box for an object covering over a tenth of the photo was usable, against 74% for objects covering 1 to 3%.

MEASURED It cost about 7 cents per 1,000 questions and took 0.29 seconds each. OpenAI's Decisions API has no question like it, so there's no head-to-head here.

Objection, stated: one dataset of everyday scenes, objects named in the dataset's own words, each present exactly once. Crowded photos, small objects and photos of people weren't tested.

Eleven things I'd do

None of these picks a vendor for you.

1

For plain labeling at volume, either works.

Jev was cheaper on every task.

From f1 , f6

2

If the question takes a step of reasoning, test Jev first.

It was the one gap in the text tests that wasn't luck.

From f2

3

Before switching vendors to fix accuracy, read 20 mistakes and reword the question once.

Both models shared the same blind spots, and the wording was the part I could change.

From f3 , f4

4

If you'll set your own cutoffs, count the answers at exactly 100% first.

A pile of 100%s leaves the dial nothing to rank.

From f5

5

On OpenAI, ask one good question per item.

It bills the item again for every question.

From f6

6

If the decision needs a long rulebook or an identical answer every time, test OpenAI first.

Lock the wording and the option order too.

From f7 , f8

7

Wherever you use OpenAI, handle a refusal as "send to a person."

Never as "no."

From f9

8

To sort photos, use a few folders a photo can't belong to two of.

Either model files well. Let the confidence number decide which ones file themselves; Jev's was the more honest one.

From f11 , f12

9

To find photos later, tag them, and have Jev look for the first pass.

If you'll keep adding tags, save a description of each photo too.

From f13

10

To find where something is in a photo, try Jev's Box beside a yes-or-no.

Keep only the boxes it's at least half sure of.

From f14

11

Don't rely on a decision model to catch mistakes in AI-made video.

OpenAI and Claude Haiku missed them, and Jev flagged nearly every frame.

From f10

How I'd organize a pile of 2,000 photos

Say there's a folder of 2,000 mixed photos, screenshots and family pictures, and the goal is to find anything in it later. Here's the plan I'd follow, built from f10 through f14.

WORKED OUT, NOT RUN

It hasn't been run end to end, so treat it as a design, not a result. At the prices measured here, 2,000 photos would cost well under a dollar.

Plain code
Date and place from each file, free. Screenshots flagged before anything reads their text.

Jev looks at each photo
A folder, the tags, and how sure it is of each, about 9.9¢ per 1,000 photos.

OpenAI picks a folder too
From the same photo: a second opinion for a few cents.

Jev isn't sure?
Under 90% sure, the photo waits.

They disagree, or it waited?
The photo goes to a review pile for a person.

Adding tags often?
Have Claude Haiku describe each photo once and save it. Later tags then re-read text, 2.3¢ per 1,000.

Names and faces stay in Apple Photos, which recognizes faces on the device.

Each tool does the job it was best at in these tests. Drawn on graph paper because it's a plan, not a measurement.

Plain code handles everything it can for free, because date and place are most of how people look for a photo ("the beach trip in 2023"). Jev looks at each photo once for a folder and its tags, about 20 cents for all 2,000. OpenAI picks a folder from the same photo, and any photo they disagree on, or that Jev isn't sure of, goes to a review pile. Two independent opinions catch more mistakes than one. If the tags will keep changing, have Claude Haiku describe each photo once and save the text, so a tag thought of next year re-reads words instead of pictures.

> Leave names and faces to Apple Photos, which recognizes faces on the device instead of sending them anywhere. And flag screenshots before any step copies their text: they often hold account numbers or passwords.
>
> — EARL

What this doesn't claim

- That one vendor is better overall.
- That the picture results hold for screenshots, documents or family photos, since every photo here was a scene like a kitchen or a street.
- That Jev's picture features are ready for production. TypeSafe calls them a preview, and their accuracy can change between releases.
- That one video with two visible mistakes says how often a vision model misses them.
- That these results will hold after OpenAI's beta becomes a full release, since it's one model in beta.
- That they hold for tasks, languages or wordings I didn't test.
- That these are the only two decision models worth looking at.

Written down first

Three things were written down and saved before any result came back: the fair round's rules, committed in three steps on Oct 9; Earl's prediction for the video test, written into its plan on Oct 12; and the Jev 1.14 re-test's rules with six predictions, committed on Oct 20 before its first test call. Here's how each one came out.

“Equal tuning budget, picks frozen, test run once.”

✓ KEPT

The rules were committed at 2:07 PM, the practice results at 2:13, the frozen picks at 2:14, all before the test set was touched. It ran once and was scored once.

Finding 1 →

“It will catch duplicated objects fairly well.”

✕ REFUTED

It put 0.01 on a second pencil in Earl's hand while another sat behind his ear.

Finding 10 →

“It will catch garbled screen text fairly well.”

⊘ NOT TESTED

The video had no garbled-text mistake to catch. Its 11 flags were text in mid-animation, doing what it was drawn to do.

Finding 10 →

“It will miss most finger and hand errors.”

⊘ NOT TESTED

No hand error was visible in the frames that got sampled.

“It will miss most changes between frames.”

CAN'T TEST

It sees one frame at a time, with no sense of before and after, so a between-frame change can't reach it.

“Claude Haiku will tie or beat it.”

TIED

Haiku missed the same two mistakes and flagged the same floating logo.

Finding 10 →

Oct 20: the Jev 1.14 re-test

“Same photos, same questions, rules and predictions saved before the first call.”

KEPT

Committed at 12:52 PM. The text picks were frozen and committed before the test set ran.

“On text, 1.14 lands within 2 points of 1.13, and the logic lead holds.”

HELD

71.3% against 71.1% overall. Logic 74.6% against OpenAI's 65.1%.

Finding 2 →

“On photos it ties OpenAI, is wrong less often when 90% sure, and flips less when the order flips.”

HELD

All three. Wrong on 5.4% of its sure photos against 11.2%; 1 flip in 237 against 10.

Finding 11 →

“Example photos move accuracy less than 3 points.”

HELD

Up 1.7 and 2.7 points, within chance, at about 9 times the cost.

Finding 12 →

“Looking at photos, Jev tags within 4 points of OpenAI and costs less.”

HELD

84.3 against 81.9 on the tag score, 10 cents per 1,000 photos against 65.

Finding 13 →

“It misses the floating pencil too.”

HELD

0.10, 0.10 and 0.44, all under the 0.5 line.

Finding 10 →

“Box: a typical overlap of at least 0.6, with 70% of boxes usable.”

HELD

0.81 and 84%.

Finding 14 →

Not predicted: Jev looking at photos came out cheaper than the describe-then-read route it replaced, and its video failure was the opposite of OpenAI's.

What broke along the way

The fair round exists because the first round failed a fairness check. Here's everything that was checked.

KILLED

"Jev is more accurate." The first round copied questions tuned on Jev over to OpenAI word for word, which is home-field advantage, and its lead in three tests of four wasn't decisive at that size. Its accuracy claims were withdrawn the same afternoon.[6]

"Describing each folder adds 3 points." True on the first photos, gone on 540 fresh ones, so it isn't reported as a gain.[12]

"Jev's cheapest way into pictures is a description." True until Jev could look. On Oct 20, Jev looking at each photo cost 10 cents per 1,000 against 27 for the description route, at the same accuracy. The description route still wins for tags added later.[17]

CHANGED

A key that leaned one way. One first-round test, a filter for hacking tools on GitHub, used an answer key partly built from Jev's own earlier reading. Scored against a key with no Jev in it, the pattern held, but it's only used for cost and refusals, not accuracy.[6]

A test skipped, then run. Revision 1 left pictures out. Revision 2 ran three picture tests on Oct 12, each against a key people made.[8]

"Read by hand," corrected in Revision 4. f4's error patterns are Claude's reading of the mistakes. Earlier revisions called that reading "by hand."

Photos Jev couldn't take as they were. Jev accepts pictures from 280 to 1,568 pixels on a side and rejects the rest. One panorama was shrunk to fit and 99 of the 328 tag photos were enlarged, following TypeSafe's own advice. OpenAI saw the originals.[17, 16]

✓ HELD UP

A translation check. In the first round, OpenAI also got a version written in its own plain-prose style. It did no better (76.6% on the category sort and 90.7% on the paper filter).[6]

A null check. The same request sent twice should give the same answer. OpenAI did, on all 418 repeated pairs. Jev changed 3 picks out of 300 (f7).

A reference model. Claude Haiku 5.5 ran the same picture questions untuned, at low effort. Its answers that weren't a folder name (9 and 15 of them) are counted as wrong.[12]

The text result, on a newer model. Jev 1.14 ran the fair round's text test again under the same rules and landed 0.2 points from 1.13.[17]

> One mistake of mine stays in the numbers. I worded the tag "a TV or screen" too loosely, and it said yes to microwave doors. It wasn't fixed after the fact.
>
> — EARL

Removed & limits

Removed before Revision 1: the first round's headline, "Jev is more accurate," and its accuracy percentages. Its cost, repeatability, option-order, rulebook and refusal findings stand, because the fair round didn't test them and doesn't weaken them.

Standing limits:

- **One question-writer.** Claude Opus 5.5 wrote both models' questions. The equal budget and the published practice scores reduce that risk without removing it.
- **One run of the test.** Each model's picked wording ran once on the test set. The 95% ranges in the charts show sampling luck, not run-to-run wobble.

- **Public datasets.** These datasets are years old and widely published, so either model may have seen some of them in training. That would flatter both, not necessarily equally.
- **A beta.** OpenAI's Decisions API is in public beta with one model, `gpt-6-luna`. Its prices and behavior can change at full release.
- **A preview.** Jev's picture input and Box are labeled a preview, not ready for production, and every call needs a special switch turned on. Their accuracy can change between releases.
- **Two dates.** OpenAI and Claude Haiku answered the picture tests on Oct 12, Jev 1.14 on Oct 20, on the same photos and questions.
- **Box on one dataset.** f14 rests on 100 everyday objects, each present once. Crowded scenes, small objects and people weren't tested.
- **One video.** f10 rests on one of Rick's videos, with two mistakes visible in the sampled frames. It shows a blind spot. It doesn't measure how often one happens.
- **Scene photos only.** Every photo in f11 to f13 shows a place: a kitchen, a beach, a street. Screenshots, documents and photos of people weren't tested.
- **A rule decides what a tag means.** In f13 a tag counted as present when its outlines covered at least 0.5% of the photo. A different rule would move the tag numbers.
- **Small sets.** 100 to 328 photos or objects per test. The 95% ranges are about plus or minus five points, so gaps smaller than that are called ties here.
- **Some keys are AI-built.** Two first-round tests (the 31-category sort and the paper filter) were graded against keys from blind AI passes, and one leaned on Jev's own earlier reading. None of them backs a headline accuracy claim.
- **One laptop, one day.** Speed was measured from a home connection on Oct 9. Treat the speed result as a rough tie.

How it was made

Both text rounds ran on Oct 9 between about noon and 2:20 PM ET, from a laptop, through a small script that calls both APIs and returns their answers in one shape. [14] Every call was saved with its request, its answer, its token count and what it cost, and loaded into a database with a dashboard over it, so the numbers here can be re-cut without re-running anything. [5]

3,887
test questions, each graded by people, none of them seen during tuning

19,774
calls in the fair round, practice and test together, with zero failures

$1.86
spent on both APIs across both text rounds, about 28,000 calls

$1.21
for the three picture tests on Oct 12: $1.05 OpenAI, $0.13 Claude Haiku, $0.03 Jev

$1.21
again, by coincidence, for the Jev 1.14 re-test on Oct 20: the text test plus six picture tests

33
prompts from Rick: 8 to design and run the text rounds, 7 for the picture tests, 16 to build the dashboard and the video, 2 for the Jev 1.14 re-test

The first round cost $0.90, the fair round $0.91, and the setup probes about $0.05. The prompt counts come from the session transcripts. Drafting the report isn't counted.

The Jev 1.14 re-test ran on Oct 20 between about 12:50 and 1:20 PM ET: the text test again plus six picture tests, $1.21 in all, every dollar of it on Jev. OpenAI's and Haiku's earlier answers were reused, not bought again, since neither model had changed. [17]

The video at the top was made with the video kit used for every explainer on this site. [15] Claude wrote and animated it in code, voiced it with ElevenLabs, and Rick reviewed it on a page that pins his notes to moments in the video. The first cut was finished Oct 9. On Oct 12 it gained the picture tests and grew to 3 minutes 39 seconds, after 16 review

rounds and about 2 hours 40 minutes of build time. Up to that cut, its only costs were 278 voice credits and 2 cents to check the voice track against the script. Revision 5's cut adds Jev 1.14 looking at the photos and its Box question. The music and sound effects came from a library Rick already had.

Models, by role. Claude Opus 5.5 designed both rounds, wrote every question for both models, ran them, scored them, built the video and wrote this page. Jev 1.13 and OpenAI's `gpt-6-luna` only answered. Fable 5.1 built two of the first round's answer keys blind, and those keys aren't used for any headline accuracy number. In Revision 2, Claude Haiku 5.5 answered the picture questions as a reference and wrote the photo descriptions Jev read. Claude Opus 5.5 designed and scored the picture tests and drew the figures. In Revision 5, Jev 1.14 only answered; Claude Opus 5.5 designed, ran and scored the re-test and redrew the picture figures.

AI disclosure

The prose here was written by an AI, Earl (Claude Opus 5.5), under Rick's editorial direction. Rick owns the question, the decision to throw out the first round and every editorial call.

The accuracy numbers weren't produced by an AI judging anything, with two exceptions. They're plain code comparing each answer to a key that people wrote for public datasets, and the costs come from each API's own billing count. f4's three patterns are Claude's reading of each model's practice mistakes. And in f10, whether Jev's boxes landed on the pencil was Claude looking at the boxes drawn on the frames.

So you're trusting the public answer keys, the stored calls and the locked rules, and each of those is named under Sources.

Versions

REV 1 Oct 10, 2026: first publication.

REV 2 Oct 12, 2026: adds pictures. Part E has four new findings: a missed mistake in an AI-made video (f10), filing photos into folders (f11), folder design against rewording (f12), and finding photos with tags (f13). A box under the answer summed them up, f6 notes that pictures aren't billed again per question, and the conclusions added three recommendations and a plan for organizing a pile of photos. The limits cover the picture tests. No text result changed.

REV 3 Oct 12, 2026: the video at the top is the new cut that covers the picture tests (3 minutes 39 seconds). How it was made counts the video's second round of work. No finding changed.

REV 4 Oct 15, 2026: moved to the notebook design. The story comes first, every finding sits beside a figure drawn for it, and each statement is marked as measured or worked out. "How I checked" now opens with the rules and the prediction written down before the tests. f4 now says its patterns are Claude's reading of the mistakes, not a reading by hand. Every figure is now rebuilt from the stored runs, with each number checked against the text, and that check corrected six details. None changes a conclusion. f9 had called all five refused items harmless, but one is a trivia question about suicide. The trivia lead in f1 is 3.9 points, not 3.8, and OpenAI's trivia gain in f3 is 8.3 points, not 8.4. f12's fresh photos number 240, not 237. f13 averaged 22 searches, not 21, and the describe-once route cost 42% of OpenAI's, not 40%. The fair round made 19,774 calls, not 20,560. The 0.21-second speed tie is now credited to the first round's speed test, where it was measured.

REV 5 Oct 20, 2026: TypeSafe released Jev 1.14 on Oct 14 with picture input and a new Box question, so every picture test ran again with Jev looking at the photos, plus a new test of Box (f14). f10, f11 and f13 now compare Jev with OpenAI directly. f11 adds that Jev was more honest about being sure, f12 adds folders defined by example photos,

and f13's cost finding flipped: Jev looking costs about 15% of OpenAI and less than the description route. The playbook, the photo plan and the limits changed with them. The text test ran again on 1.14 and held (71.3% against 71.1%), so Parts A to D keep the Oct 9 numbers with the 1.14 results in their notes. The video at the top is a new cut that covers Jev 1.14.

1. SOURCE TypeSafe. Jev documentation, including the API reference and the Jev 1.13 model notes on where it is weak, read Oct 9, 2026. Graded: vendor documentation, used for how the product works and how its questions should be written, not for any claim about accuracy. docs.typesafe.ai/llms.txt

   `jev-docs`

2. SOURCE OpenAI. "Decisions" guide and API reference (the create method, developers.openai.com/api/reference/resources/decisions/methods/create), public beta since Oct 6, 2026, read Oct 9, 2026. Graded: vendor documentation, used for how the product works and how its questions should be written. Where the guide and the reference disagreed (hosted image links), a direct test settled it. developers.openai.com/api/docs/guides/decisions

   `openai-decisions-docs`

3. DERIVED The fair round's protocol, written and committed before any practice call (fair/PROTOCOL.md, 2:07 PM ET Oct 9, 2026), the practice results and revisions (2:13 PM), the frozen picks (frozen.json, 2:14 PM), and every candidate wording with its practice score (fair/results/practice.json, candidates.py).

   `dc-protocol`

4. SOURCE The five public datasets, all used through Hugging Face with their published answer keys: StrategyQA (ChilleD/StrategyQA, yes or no with multi-step reasoning), Adversarial NLI round 3 (facebook/anli, written by people to fool models), Emotion (dair-ai/emotion, six feelings), SST-5 (SetFit/sst5, five-level movie review ratings) and Banking77 (legacy-datasets/banking77, 77 banking requests). 300 items per task held for practice, the rest for test, and no item from earlier tests on this site. Graded: established public benchmarks with people-written keys.

   `five-datasets`

5. DERIVED The fair round's test: 3,887 items, each model's frozen pick run once, every call stored with request, answer, billed tokens, cost and latency (fair/raw/test.jsonl), scored by fair/score.py into fair/results/test.json: accuracy with 95% ranges, exact McNemar tests per task and pooled, calibration error before and after a one-temperature fit on practice, share of answers at exactly 0 or 1, refusals, cost per 1,000. Every figure's numbers are rebuilt from these files by notebook/export_notebook.py, which checks each against the text.

   `dc-test`

6. DERIVED The first round: 6,468 scored calls and 640 speed calls on Oct 9, 2026, $0.90, in four tests (a 300-item people-graded benchmark, a 1,246-project hacking-tool filter, a 154-project 31-category sort, a 118-paper research filter), stored per call (raw/*.jsonl) and scored offline (score_{a,b,c,d}.py, results/arm_{a,b,c,d}.json, FINDINGS.md). Its accuracy claims are superseded by the fair round, and the file carries that banner.

   `dc-round1`

7. DERIVED Billing probes: the same short support ticket asked 1, 5, 20 and 100 questions on each API, with the billed tokens each returned (DESIGN.md, probe records p01, p02 and p05).

   `dc-billing-probes`

8. DERIVED The picture tests, Oct 12, 2026: the experiment plan with its prediction written before any test ran, four billing probes on video frames, and one folder per test with its raw calls, scoring script and results (visual-linter/PROPOSAL.md, probes/, run-01-earl-v4/, run-02-photo-filing/, run-03-tags/). Total spend $1.21.

   `dc-pictures`

9. DERIVED The video test: one frame per second of a 4 minute 23 second draft of one of Rick's videos, 264 frames, six yes-or-no questions each, $0.056 (run-01-earl-v4/results.jsonl). The answer key is Rick's own review notes on that draft, written between 10:50 PM Oct 11 and 1:49 AM Oct 12, before the test was planned. Follow-ups: a zoomed crop, and Claude Haiku 5.5 on the same frames (followup.json).

   `dc-video-test`

10. SOURCE OpenAI. "Images and vision" guide, limitations section (counting and spatial reasoning), read Oct 12, 2026. Graded: vendor documentation, used for what the vendor itself says its vision models do poorly. developers.openai.com/api/docs/guides/images-vision

    `openai-vision-docs`

11. SOURCE Places365 (Zhou and others, MIT), validation set: photos of 365 scene types, each labeled by people. Used through Hugging Face with the official category list (CSAILVision/places365). 237 photos in 12 folders and 298 in 60, then 540 fresh photos (240 and 300) for the rewording test. Graded: an established public benchmark with a people-made key.

    `places365`

12. DERIVED The photo tests: every call stored with its answer, confidence, billed tokens and time (run-02-photo-filing/results-*.jsonl, variants-*.jsonl, run-03-tags/results-*.jsonl), scored by score.py in each folder. The rewording tweaks and folder descriptions (folders.json) were frozen before the fresh photos ran.

    `dc-photo-tests`

13. SOURCE ADE20K (Zhou and others, MIT), validation set: scene photos where people labeled the scene and outlined every object. Used through Hugging Face. 328 photos in 8 scene folders, with each tag counted as present when its outlines cover at least 0.5% of the photo (run-03-tags/build.py). Graded: an established public benchmark with a people-made key.

    `ade20k`

14. DERIVED The shared client that calls both APIs and returns one answer shape, and translates questions between their formats mechanically (clients.py), plus the 37 mechanics probes run before either round (probes/*.jsonl, DESIGN.md).

    `dc-client`

15. DERIVED The video: a video-kit project (studio/projects/jev-vs-openai), with its review log of 16 rounds and the spend ledger (278 ElevenLabs credits, $0.02 of transcription). The first cuts were approved Oct 9, 2026, and the current cuts, which add the picture tests, on Oct 12.

    `dc-video`

16. SOURCE TypeSafe. "Images", "Box" and "Models" pages of the Jev documentation, read Oct 20, 2026: Jev 1.14 (released Oct 14) takes pictures and answers Box questions, both labeled "Dreamer-ready," a preview not yet up to the company's standard for production; pictures are billed as input tokens at the text price; sizes from 280 to 1,568 pixels a side. Graded: vendor documentation, used for what the product does and its stated limits, not for any claim about accuracy. docs.typesafe.ai/images.md

    `jev-images-docs`

17. DERIVED The Jev 1.14 re-test, Oct 20, 2026: rules and six predictions committed before the first test call (jev-1.14/PROTOCOL.md), text picks frozen before the test set ran (t1/frozen.json), every call stored with its answer, confidence, billed tokens and time (t1/ to t6/), scored by score_photos.py, score_tags.py and the fair round's own scorer. OpenAI and Claude Haiku answers are their stored Oct 9 and Oct 12 runs on the same items. Total spend $1.21, all on Jev.

    `dc-jev114`