WOMEN WHO CODE LEADERSHIP SUMMIT · NEW YORK · 8 OCTOBER 2026

Score It Yourself

How reliable is the AI that grades your AI?

Mason Ulrika Borchard
Senior Software Engineer, Stride, Inc. · Doctoral Student, Valdosta State University

4.6 / 5

Story quality

94%

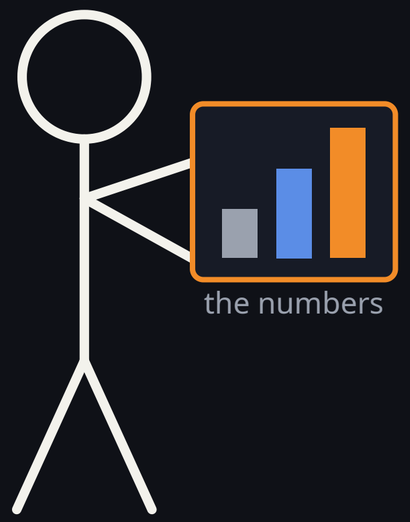
Pass rate

What would you want to know before you approved this?

WHY THE TESTING HARNESS EXISTS

Can we upgrade Story Studio's content generator from GPT-3.5?

Even GPT-4.1 mini wins. Here's the research.



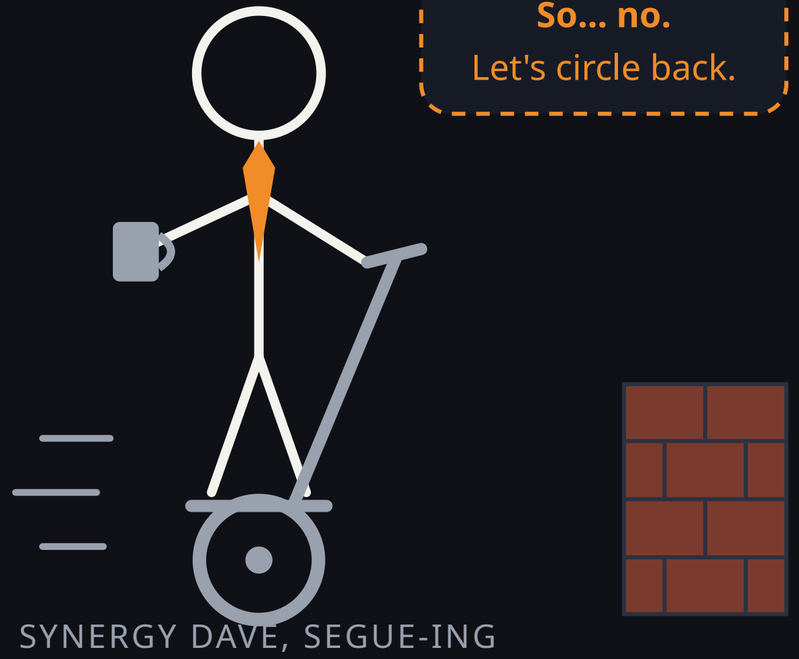
Great evidence!

Very thorough research!

I agree!

So... no.

Let's circle back.



SYNERGY DAVE, SEGUE-ING

Circling back all day without getting anything done.

STORY STUDIO

An AI reading app for grades K to 8: a story, its pictures and a quiz, generated for one child.

Began as the winning entry in the company's 2023 hackathon. The CEO asked for it to be built.

THE TEAM

One frontend engineer. One backend engineer. One QA engineer on two teams. One UX designer. A curriculum lead who reports to the CEO. Me, and since the budget cuts, the reading specialist too.

HOW IT IS TESTED

Automated checks that count. BQTH, the Borchard Qualitative Testing Harness: AI evaluators from four LLM families. A QA engineer, testing every release. A former teacher, reading.

WHAT IT COST

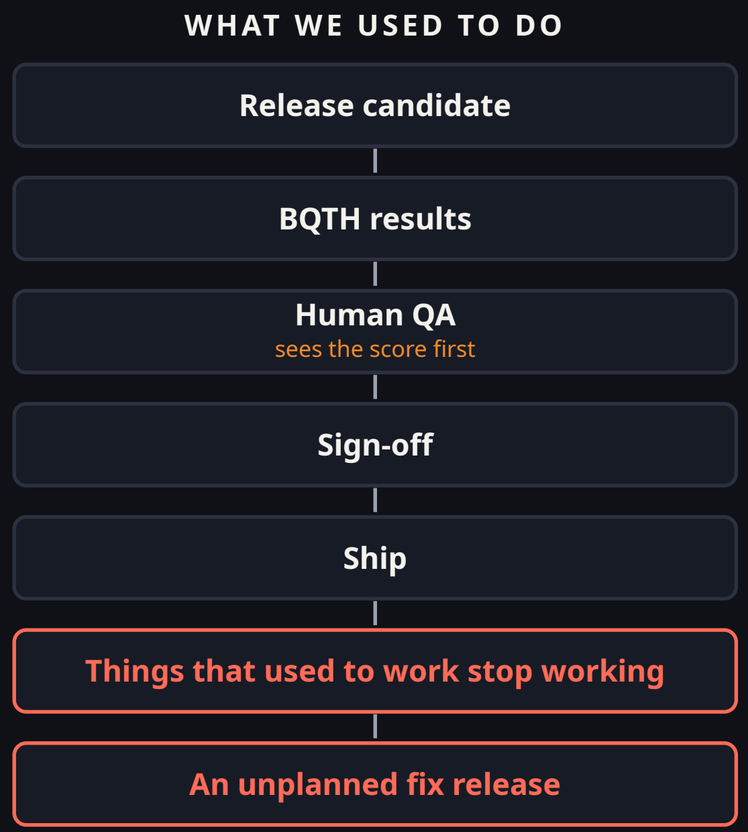
The first AI evaluator runs came out of my own pocket: a few hundred dollars. The experiment that stopped a multi-week build cost $2.50.

1,900+
stories

1,800+
images

1,500+
quizzes, in about five months

HOW WE USED TO GET TO THE SIGN-OFF



Most of what follows was found after this process. Today: you take a test → how the number is made → five ways it lies → how people learn to read it → three questions

Based on specific details in the story, explain why Bubbles and Cat could not have saved the alien ship if only one of them had tried.

A question from a children’s reading app, 695L, about grade 4. The story is removed. Pick an answer anyway.

A The alien ship's signal was too weak for one frequency to boost.

B The transmitter was too heavy for one animal to hold steady.

C One animal would have gotten too tired before finishing.

D The device required two different frequencies adjusted simultaneously, and each animal had different paws suited to different tasks.

What did Brad see in the sky over the ice?

The same app, 1227L, about grade 8. The story is removed. Pick an answer anyway.

A A spiral of green aurora light

B A ring of falling winter stars

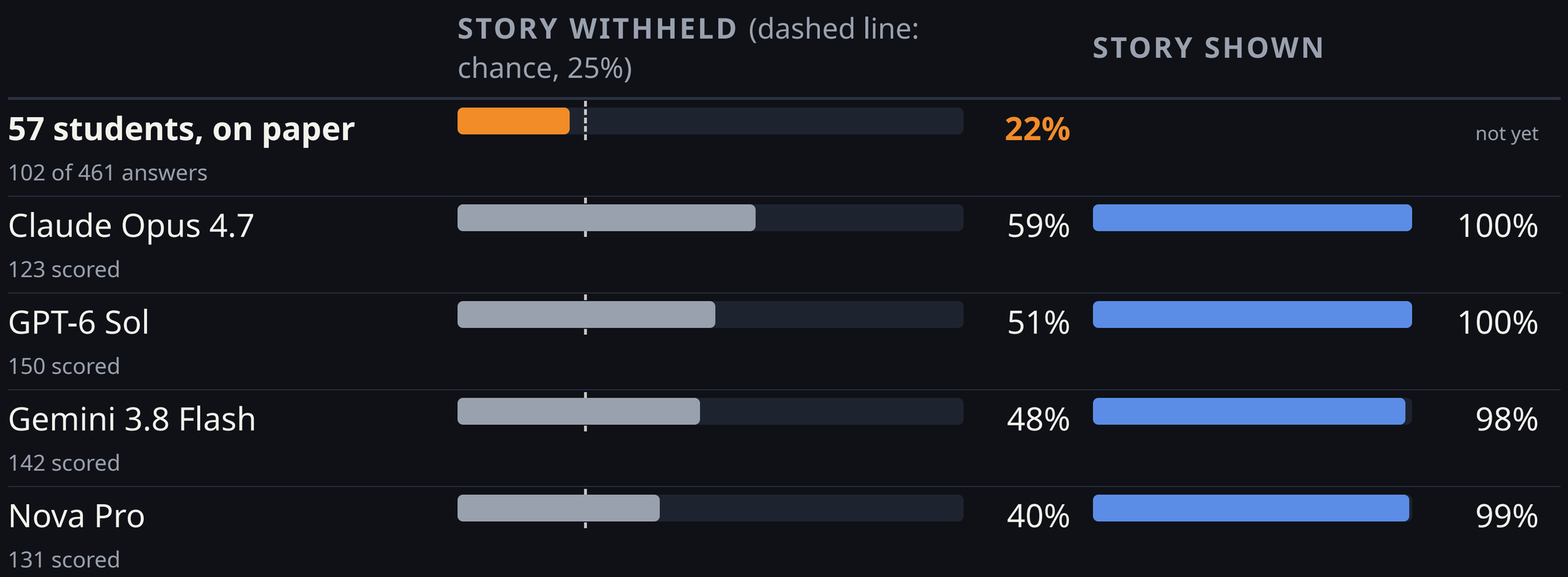
C A wall of black storm cloud

D A fissure like a cracked horizon

EXERCISE 1 · FIVE WRITTEN UNDER THE OLD RULES, ONE UNDER THE NEW · FOUR AI EVALUATORS, NO STORY, THREE RUNS EACH

#	LEVEL	KEY	RIGHT, EVERY RUN	WHY
1	695L	D	4 of 4	Longest choice by a mile, the only one with detail
2	331L	D	4 of 4	The kind answer
3	1090L	B	4 of 4	The shortest choice. Still the one a person would feel
4	908L	B	4 of 4	24 words against 9. "Immediately." "Ever."
5	331L	B	3 of 4	Question 2 gives it away
6	1227L	D	0 of 4	Written under the new rules. With the story: 4 of 4. Students: 1 of 6

Fifty items under the new rules: evaluators 40 to 59% without the story, 98 to 100% with it. Fifty-seven children on paper, no story: at the chance line.



Guessing at random on four choices: **25%**. Every wrong answer was written to be the natural guess from the question alone. Unreadable evaluator answers count as missing, not wrong.

At the chance line without the story. A guessability check, not a reading score.

PASSAGE DEPENDENCE · 95 QUESTIONS · 24 STORIES · GRADES 1 TO 8 · 3 RUNS EACH

	WITH THE STORY	WITHOUT
AI Evaluator #1	98.9%	97.5%
AI Evaluator #2	99.3%	96.8%
AI Evaluator #3	99.6%	96.5%
AI Evaluator #4	99.6%	89.5%

25%

Random guessing

Calibrating a judge against human labels would pass every one of these.
Calibration checks the judge. It doesn't check the question.

Tuinman (1973); Keenan & Betjemann (2006); Säuberli & Clematide (2024). Full references on the last slide.

THE EVALUATION PANEL, ON THE LATEST BATCH

GENERATE

Stories, quizzes and picture prompts
the product's own prompts; Haiku 4.5 wrote the batch

Pictures
Stable Image Core, gpt-image-2, Gemini image

EVALUATE

Anthropic
Claude Opus 4.7, Bedrock
text, pictures, quiz

OpenAI
GPT-6 Sol
text, pictures, quiz

Amazon
Nova Pro, Bedrock
text, pictures, quiz

Google
Gemini 3.8 Flash
text, pictures, quiz

Four LLM families, four prompt roles: teacher, reviewer, item writer, editor. Every model played every role once.
8 new stories, a pilot: switching the model moved the text score 0.7 points; switching the role, 0.2.

four runs, logged



COUNT

Not AI: words, failures, cost
free, and the same answer every time

SHEET

One row per item, per judge, per role, per run
quoted evidence beside each score; 120 clean judgments of 128 on 8 stories, and every story's verdict split across the evaluators

ARGUE

Research first
it reads before it argues

Every citation checked
fetched and read, not trusted

Another family argues
against the conclusions

A person reads the splits
wherever evaluators disagree

The second layer exists only to attack the first: it audits what BQTH concludes, not the stories.

Panel of families: Verga et al. (2024). Self-preference in judges: Panickssery, Bowman & Feng (2024).

THE HARNESS, THE VERSION BEING BUILT

GENERATE

Stories, quizzes and picture prompts
Haiku 4.5 writes the batch; gpt-3.5-turbo stays as the unchanged comparison

Pictures
Stable Image Core, gpt-image-2, Gemini image

EVALUATE

Anthropic
Claude Opus 4.7
literacy teacher

OpenAI
GPT-6 Sol
ed-software reviewer

Amazon
Nova Pro
assessment item writer

Google
Gemini 3.8 Flash
children's book editor

Four LLM families. A role is a prompt condition, not an expert. Every model plays every role once.
Rubric: story text, pictures, quiz. Gates, checks, 1-to-4 scales with written anchors.

four runs, logged



COUNT

Not AI: words, failures, cost
free, and the same answer every time

SHEET

One row per item, per judge, per run, per role
20 percent of items scored twice by people; every split between a evaluator and a person read by a person

ARGUE

Research first
it reads before it argues

Every citation checked
fetched and read, not trusted

Another family argues
against the conclusions

A person reads the splits
wherever evaluators disagree

[human comparison: pending, arrives after this talk]

Panel of families: Verga et al. (2024). Self-preference in judges: Panickssery, Bowman & Feng (2024).

THE DISCIPLINE

Four rules, none of them technical

What makes anything the machinery produces mean something.

01

The same test items, every run

The 95-question set has not changed since it was written, so a difference cannot come from the items.

02

An unchanged version beside every change

The four-version experiment under sign 5 ran each one against what production already did.

03

Scoring rules written before scoring

The rubric was fixed before the 300-story test. People cannot see the AI scores until theirs are in.

04

Everything runs more than once

Three runs each. One evaluator changed its own answer on 10 of 24 stories.

THE DISCIPLINE

Four rules, none of them technical

What makes anything the machinery produces mean something.

01

The same test items, every run

The 95-question set has not changed since it was written, so a difference cannot come from the items.

02

An unchanged version beside every change

The four-version experiment under sign 5 ran each one against what production already did.

03

Scoring rules written before scoring

The rubric was fixed before the 300-story test. People cannot see the AI scores until theirs are in.

04

Everything runs more than once

Three runs each. One evaluator changed its own answer on 10 of 24 stories.

THE DISCIPLINE

Four rules, none of them technical

What makes anything the machinery produces mean something.

01

The same test items, every run

The 95-question set has not changed since it was written, so a difference cannot come from the items.

02

An unchanged version beside every change

The four-version experiment under sign 5 ran each one against what production already did.

03

Scoring rules written before scoring

The rubric was fixed before the 300-story test. People cannot see the AI scores until theirs are in.

04

Everything runs more than once

Three runs each. One evaluator changed its own answer on 10 of 24 stories.

THE DISCIPLINE

Four rules, none of them technical

What makes anything the machinery produces mean something.

01

The same test items, every run

The 95-question set has not changed since it was written, so a difference cannot come from the items.

02

An unchanged version beside every change

The four-version experiment under sign 5 ran each one against what production already did.

03

Scoring rules written before scoring

The rubric was fixed before the 300-story test. People cannot see the AI scores until theirs are in.

04

Everything runs more than once

Three runs each. One evaluator changed its own answer on 10 of 24 stories.

Counting

How many words is the right answer?
How often did it come back broken?
What did it cost?

Almost free. Same answer every time.

Opinion

Is this a good question?
Is the marked answer actually right?

Costs real money. Moves.

Five versions of the instructions, 80 questions each, counting only: about 75 cents. 53% gave the answer away by length; the best alternative, 22%.

FIVE WAYS A QUALITY SCORE LIES TO YOU

1. The score rewards sameness
2. The score can't move
3. The evaluator is the variable
4. The failures aren't in the average
5. Nothing was compared

WARNING SIGN 1 OF 5





page 1

Prompt 1 reading in the beanbag by the round window, daytime



page 2

Prompt 2 kneeling at the bookshelf, searching for a brass key, low angle



page 3

Prompt 3 standing with Mr Penguin by the chalkboard, wide shot



page 4

Prompt 4 night: the room dark, lit by one small lamp

Four prompts, one pinned seed. Consistency score: 5 out of 5, every page. Night palette: 1 out of 5. Three of the four files are byte-for-byte identical.

> **If the cheapest way to pass is also a bad outcome, something will find it.**

"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

Campbell (1979), *Evaluation and Program Planning*, 2(1), 67 to 90.
Goodhart's law as usually quoted: Strathern (1997).

What is the cheapest way to score well on this?

A support team judged on tickets closed closes tickets without solving them. A summary scored on words shared with the document copies the document.

WARNING SIGN 1 OF 5

The score rewards sameness

Same failure, a different product, no images anywhere near it.

Grade 6 answer choices over the length limit

55%

AUTOMATED CHECK

Clean.

Not one flagged.

THE CHECK ASKED

Is the right answer the single longest choice? If so, the question gives itself away.

THE HOLE

If every choice runs long, none stands out as the longest. The cheapest way to pass: make every answer long.

Nobody intended it.

WARNING SIGN 2 OF 5

The score cannot move

32 of 32 test cells

100%

After harder test material

91%

WHY IT'S A PROBLEM

A score at the top of its range can't go down. It would catch a collapse, not ordinary decline.

THE FIX

Harder material until it moves again. A score that sits at the ceiling for months retires.

At the ceiling, a score has gone blind to the decline you need to catch.

WARNING SIGN 3 OF 5

The evaluator is the variable

The question this session is named after.

The same 12 images, scored out of 5 by two AI evaluators

4.92
AI Evaluator #1

3.92
AI Evaluator #2

Same pictures, a full point apart. If only one number reached you, you would never know the other existed.

Same evaluator, same 95 questions, three runs

	STORIES WHERE THE ANSWER CHANGED, OF 24	SINGLE QUESTIONS THAT FLIPPED, OF 95
AI Evaluator #4 (GPT-4o)	10	10
AI Evaluator #2 (Claude Opus)	3	4
AI Evaluator #3 (gpt-5.4)	3	3
AI Evaluator #1 (Gemini)	1	1

AI Evaluator #4 is the oldest of the four. Its answers moved between runs; the choices were reshuffled every run, for every evaluator.

None of the four is a child. Use them, and check them.

Run-to-run inconsistency: Haldar & Hockenmaier (2025). Position and length bias: Zheng et al. (2023). Self-preference: Panickssery et al. (2024).

1. Two evaluators agree on the answer key in 9 of 23. I withdraw "the answer keys are sound."
2. A third AI evaluator agrees with the first in 31 of 35.

Of the 23 all three evaluated: a majority says valid in 21. The second AI evaluator is the lone holdout in 13.

> **The vote tells you where to look, not what's true.**

The run summary recommended Model A.

	STORIES REQUESTED	FAILED BEFORE SCORING	STORIES SCORED
Model A	80	11	69
Model B	80	0	80

How many started, and what happened to the ones that aren't here?

WARNING SIGN 5 OF 5

Nothing was compared

THE RESULT

A 20-point gain

In how well pictures matched their story.

THE PLAN

Weeks of building

A new stage in the product, on the strength of it.

THE TEST

Four versions

8 images each, every one against what production already did.

$2.50

The whole gain was already in the step production ran. We didn't build it.

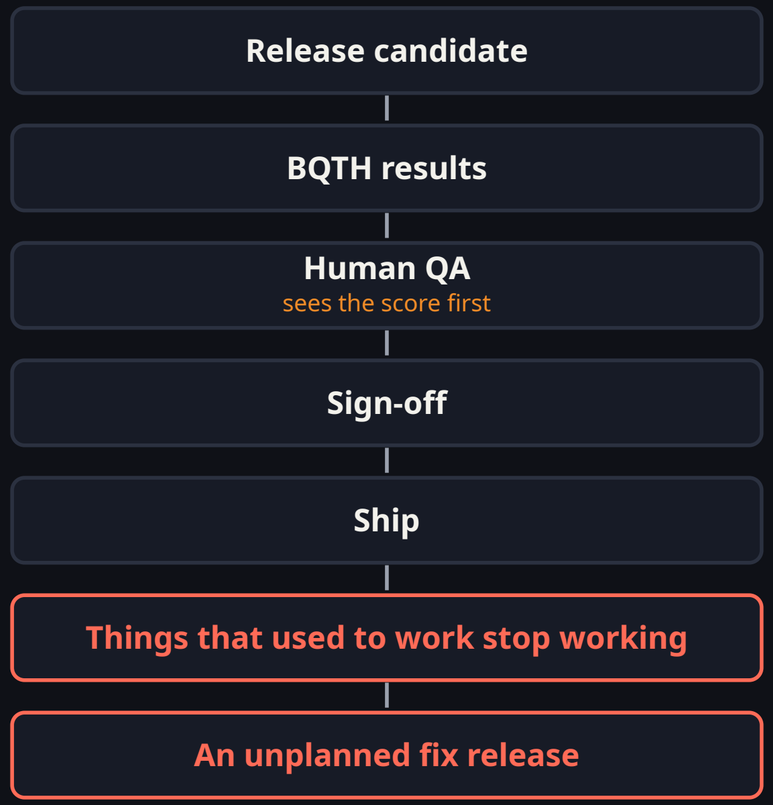
Compare every change against what already runs.

FIVE SIGNS, FIVE FIXES

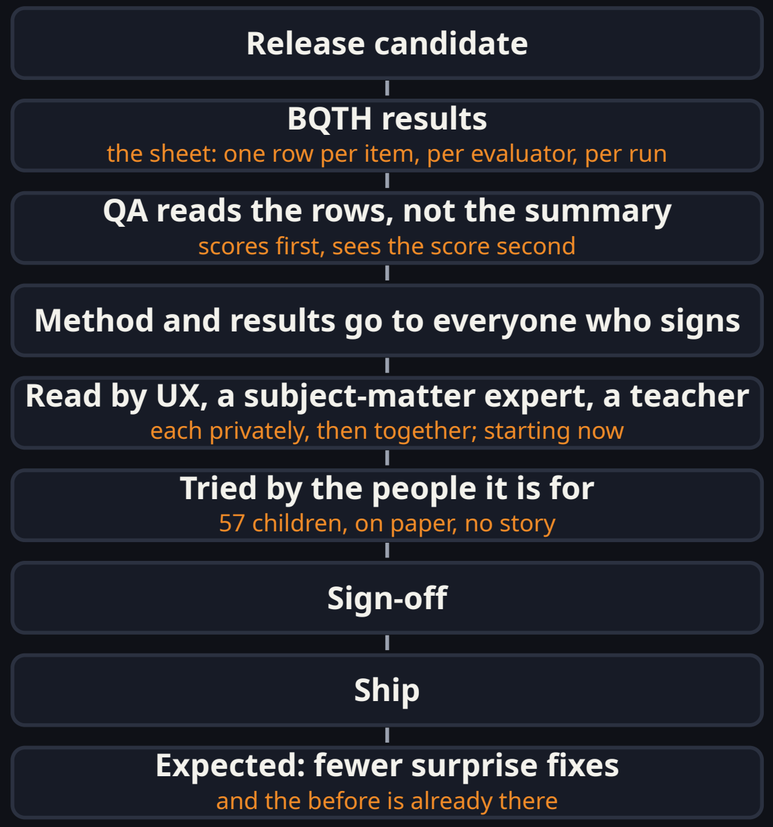
THE SIGN	WHAT WE DID ABOUT IT
The score rewards sameness	Ask what the cheapest way to score well is, and put that case in the test set
The score can't move	Make the test harder until it moves. A score at the ceiling retires
The evaluator is the variable	A second AI evaluator from another LLM family, a second run, and a person reads where they split
The failures aren't in the average	Every summary reports requested, returned, scored
Nothing was compared	An unchanged version beside every change, on the same items

TWO WAYS TO GET TO THE SIGN-OFF

WHAT WE USED TO DO



WHAT I'M PROPOSING



What shipped because of the measuring: a length fix that holds on a second AI evaluator · a gate that says no to its own rewrites · a model pick reversed before the cutover · a build that was never built · a fallback that announces itself

ADULT LEARNING AND DEVELOPMENT

How adults learn to read evaluation results

Most people who sign off on an AI feature were never taught how to read the evidence behind it.

01 · THEIR OWN PROBLEMS

Teach on your own items

Adults learn from problems in their own work, and want to know why before how.

02 · WORKED EXAMPLES, THEN FADING

Show three, then let them score one

Study an expert's worked example, then take over more of the steps while explaining each one.

03 · A BRIEFING IS NOT TRAINING

Practice with feedback

Untrained readers judged AI-written stories at chance. Instructions and examples raised that to 55% at best.

04 · THE EXPERTISE REVERSAL EFFECT

One page, in layers

Help a newcomer needs gets in an expert's way. Verdict on top, evidence a click below.

Knowles (1980); Atkinson, Derry, Renkl & Wortham (2000); Atkinson, Renkl & Merrill (2003); Clark et al. (2021); Kalyuga, Ayres, Chandler & Sweller (2003).

FOR THE ENGINEERS

How to present qualitative results so others can use them

Four habits, each grounded in research on how people read numbers and evidence.

01 · NATURAL FREQUENCIES

Counts, not percentages

“3 of 50 failed” beats “94%”. Untrained readers reason correctly far more often with counts.

02 · A BOUNDARY OBJECT

The item beside its score

A bare number transfers. The item, its score and the quoted evidence let other roles translate it and push back.

03 · COGNITIVE LOAD

Define the terms first

Evaluator, rubric, passage dependence: one line each, before the first number.

04 · IMPLEMENTATION INTENTIONS

End with an if-then

“If one evaluator flags it and the other does not, a person reads it before release.”

Gigerenzer & Hoffrage (1995); Carlile (2004); Mayer & Moreno (2003); Gollwitzer & Sheeran (2006); Roediger & Karpicke (2006).

WHAT BQTH WRITES WHEN IT RUNS · ONE REVIEW WORKBOOK PER GRADE

Read me first
What the sheet is, how to work a tab, what every column means

Roll-up
One row per story: the evaluators' verdicts filled in, the reviewer's left blank

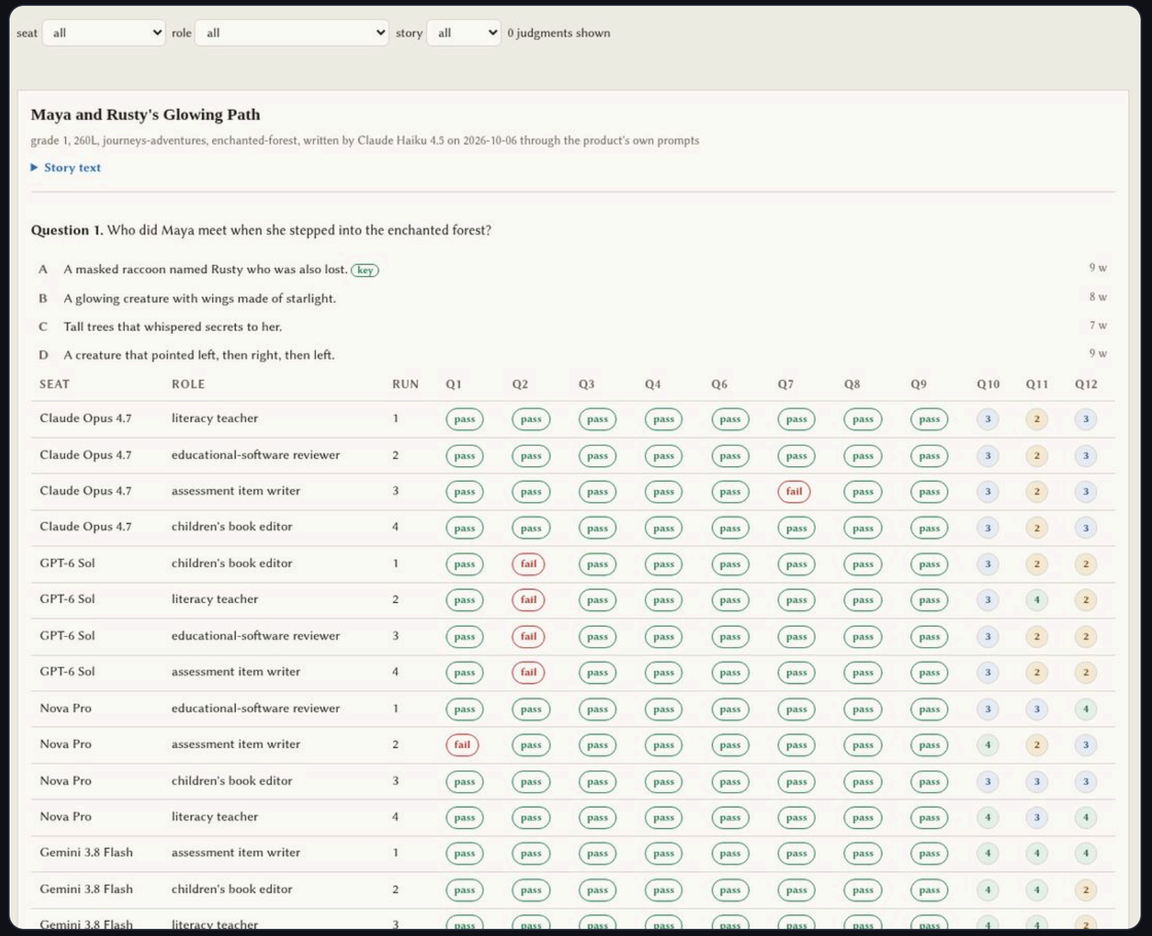
One tab per story
Four blocks, always in this order →

1 · THE STORY
Selections, a link to read it in the app, every paragraph with its picture

2 · THE QUIZ
Questions, four choices with the key marked, explanations, hints

3 · YOUR ASSESSMENT
One row per rubric item: your score, the evidence (quote it), notes
“Fill this in before expanding the block below.”

4 · THE EVALUATORS' ASSESSMENT
The same rows, a score and the quoted reasoning per evaluator
Collapsed when the tab is created. Open it after your scores are in.



SEAT	ROLE	RUN	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Claude Opus 4.7	literacy teacher	1	pass	pass	pass	pass	pass	pass	pass	pass	3	2	3
Claude Opus 4.7	educational-software reviewer	2	pass	pass	pass	pass	pass	pass	pass	pass	3	2	3
Claude Opus 4.7	assessment item writer	3	pass	pass	pass	pass	pass	fail	pass	pass	3	2	3
Claude Opus 4.7	children's book editor	4	pass	pass	pass	pass	pass	pass	pass	pass	3	2	3
GPT-6 Sol	children's book editor	1	pass	fail	pass	pass	pass	pass	pass	pass	3	2	2
GPT-6 Sol	literacy teacher	2	pass	fail	pass	pass	pass	pass	pass	pass	3	4	2
GPT-6 Sol	educational-software reviewer	3	pass	fail	pass	pass	pass	pass	pass	pass	3	2	2
GPT-6 Sol	assessment item writer	4	pass	fail	pass	pass	pass	pass	pass	pass	3	2	2
Nova Pro	educational-software reviewer	1	pass	pass	pass	pass	pass	pass	pass	pass	3	3	4
Nova Pro	assessment item writer	2	fail	pass	pass	pass	pass	pass	pass	pass	4	2	3
Nova Pro	children's book editor	3	pass	pass	pass	pass	pass	pass	pass	pass	3	3	3
Nova Pro	literacy teacher	4	pass	pass	pass	pass	pass	pass	pass	pass	4	3	4
Gemini 3.8 Flash	assessment item writer	1	pass	pass	pass	pass	pass	pass	pass	pass	4	4	4
Gemini 3.8 Flash	children's book editor	2	pass	pass	pass	pass	pass	pass	pass	pass	4	4	2
Gemini 3.8 Flash	literacy teacher	3	pass	pass	pass	pass	pass	pass	pass	pass	4	4	2

OVERVIEW, FILTER, DETAILS ON DEMAND

Roll-up first, one item when asked

Filter by evaluator and role before opening any item.

OUTCOMES PEOPLE CAN COUNT

A row per run, not an error bar

Non-experts read discrete outcomes more accurately than intervals.

EXPLORABLE REPORTS

Let the reader change the view

Switching the analysis shows how solid a finding is.

COMMIT FIRST

Their judgment before ours

The blind block in the sheet: score first, then open the evaluators.

Shneiderman (1996); Hullman, Resnick & Adar (2015); Kay, Kola, Hullman & Munson (2016); Dragicevic, Jansen, Sarma, Kay & Chevalier (2019).

WHAT USUALLY REACHES THE SIGN-OFF

Quiz quality: 94% pass. Approved.

WHAT A READER CAN WORK WITH

47 of 50 quizzes passed, scored by two AI evaluators from different LLM families.

3 failed. Here they are, with the evidence each evaluator quoted.

The evaluators disagreed on 5. A QA engineer read those; 2 need rewriting.

The same 50 on the current model: 41 passed.

Rubric fixed before scoring. Every row is in the sheet.

EXERCISE 2 · 3 MINUTES

A. Which model should we ship?

	REQUESTED	SCORED	SUMMARY SAYS
Model A	80	69	Ship this one
Model B	80	80	

B. Are the answer keys sound?

EVALUATOR	SCORED	AGREES WITH 1
1	35	–
2	23	9 of 23
3	35	31 of 35

On the 23 all three evaluated, a majority says valid on 21. AI Evaluator #2 is the only dissenter on 13.

C. Did the sentence fix work?

	STORIES	SENTENCES	OVER LEVEL
Before	3	147	28%
After	12	800	12%

Same model both rows. Only the instructions changed.

With the person next to you, for each one:
act on it, don't, or ask one specific question first.
If you'd ask, write the question down.

Same prompt. Same grade. Two different stories.

15% 67%

The story mattered more than the thing I was testing.

> **Before and after on three items measures the three items. Ask: how many, and how different were they?**

The decline effect: Schooler (2011), *Nature*, 470, 437.

WHAT HELD

	WITH THE STORY	WITHOUT	CHILDREN, WITHOUT
The 95 items from the start	98.9 to 99.6%	89.5 to 97.5%	–
50 items under the new rules	98 to 100%	27 to 55%	22%

Chance is 25%. Fifty-seven children, K to 8, on paper.

- Still the longest choice 53% of the time after a word budget. One rule about the set of four: 28%. Counted, not judged.
- In production the rule still needed the arithmetic gate: 183 rewrites on 132 items. 173 kept, 10 declined.

- **Count before you judge.** Arithmetic is free, repeatable, and has no opinion. Arushee Garg's rule of thumb, from the summit: don't ask an LLM to judge what code can verify.
- **Distrust the ruler once.** Ours counted characters and fired on a two-letter difference. Everything before the fix is a different instrument.
- **Trust the file, not the version tag.** Three deploys, three version numbers, one package compiled four days earlier.

MY CURRENT QUALITATIVE TESTING SYSTEM



Every score carries a quoted piece of evidence. Word counts, key position and the length cue are counted, never judged.

Results: still being scored. The people's scores, and how they compare with the AI evaluators, will be posted at masonborchard.com/wwc2026.

WHERE PEOPLE COME IN

The checks that caught the checkers

A rule, applied every time.

THE REVIEWER AI

Cited two papers that don't exist

Misdescribed two that do. Conceded all four when challenged.

HOW IT WAS CAUGHT

Every citation fetched and checked

A review skill reads the research first, then checks each source before anything leans on it.

MY OWN TESTS

56 passed, 0 failed

They never ran. They sat below a line that ends the program.

BEFORE THE CITATION RULE

Ten bad citations in one review

Four in the turn meant to correct the first six. None reached the page.

Ask any AI review process: what has it retracted?

1 Who did the judging, and who checked the judge?

2 What would move this number?

3 What did the same inputs produce before the change?

Ask for the methods. Ask for the full dataset. Not the summary.

Before you sign, send for three things: the rows, the funnel, the before.

Training and instructions do not prevent over-reliance on automation, in experts as well as novices (Parasuraman & Manzey, 2010). Fifteen researchers watched using AI tools: the confident tone hid which outputs needed scrutiny, trust was fragile, and the ones who coped checked by hand and built trust by auditing step by step (Gautam, Liu, Choi, & Lease, 2026).



Thank you.

The full deck and the resources from the talk:

masonborchard.com/wwc2026

linkedin.com/in/masonborchard
masonulrikaborchard@gmail.com

Connect on LinkedIn

REFERENCES

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*(4), 774–783. https://doi.org/10.1037/0022-0663.95.4.774

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Carlile, P. R. (2004). Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization Science, 15*(5), 555–568. https://doi.org/10.1287/orsc.1040.0094

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text. *Proceedings of ACL-IJCNLP 2021*, 7282–7296. https://doi.org/10.18653/v1/2021.acl-long.565

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300295

Gautam, S., Liu, H., Choi, Y., & Lease, M. (2026). How researchers navigate accountability, transparency, and trust when using AI tools in early-stage research: A think-aloud study. *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*. ACM. https://doi.org/10.1145/3805689.3812262

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in Experimental Social Psychology, 38*, 69–119. https://doi.org/10.1016/S0065-2601(06)38002-1

Haldar, R., & Hockenmaier, J. (2025). Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 24986–25004. arXiv:2510.27106

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS ONE, 10*(11), e0142444. https://doi.org/10.1371/journal.pone.0142444

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5092–5103. https://doi.org/10.1145/2858036.2858558

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10*(4), 363–380. https://doi.org/10.1207/s1532799xssr1004_2

Knowles, M. S. (1980). *The modern practice of adult education: From pedagogy to andragogy* (Rev. ed.). Cambridge.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems, 37*. arXiv:2404.13076

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437. https://doi.org/10.1038/470437a

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the 1996 IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Strathern, M. (1997). "Improving ratings": Audit in the British university system. *European Review, 5*(3), 305–321.

Säuberli, A., & Clematide, S. (2024). Automatic generation and evaluation of reading comprehension test items with large language models. *Proceedings of the 3rd Workshop on Tools and Resources for People with Reading Difficulties (READI), LREC-COLING 2024*. ACL Anthology 2024.readi-1.3

Tuinman, J. J. (1973–1974). Determining the passage dependency of comprehension questions in 5 major tests. *Reading Research Quarterly, 9*(2), 206–223. https://doi.org/10.2307/747135

Verga, P., Hofstätter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., & Lewis, P. (2024). Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv:2404.18796

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*. arXiv:2306.05685