Noisy and Biased Judges - and Subjective Tasks

Problem, impact, reporting, management.

Fabio Casati

Two different questions: the job, and the judge

The task and its output

A task, and the judgment over theoutput, can be **objective** or **subjective**.

Is this invoice total correct? **Objective.**
Is this reply friendly enough? **Subjective.**
Is this summary complete, 1–5? **In between.**

Where it sits depends partly on how well we define the criteria.

The judge

A judge can be **noisy** or **biased**, even when the job is objective.

Noisy: errors scattered across cases, in no particular direction.
Biased: errors that lean the same way.

An important distinction: We assume judges (even with noisy with respect to the truth) are consistent. Same task, same output, is assessed in the same way.

Inconsistency adds noise, and this is noise that is easy to mesure and detect when we have AI judges. Humans are notoriously inconsistent too, and that's harder to measure.

The judge is part of the system

The AI-powered service

Retrieval, prompt, model and tools produce an answer.

The judge

Rubric, evidence, judge prompt and model produce a score. We build and iterate on it too.

One change at a time

Change the service or the judge, not both, so you know what moved the score.

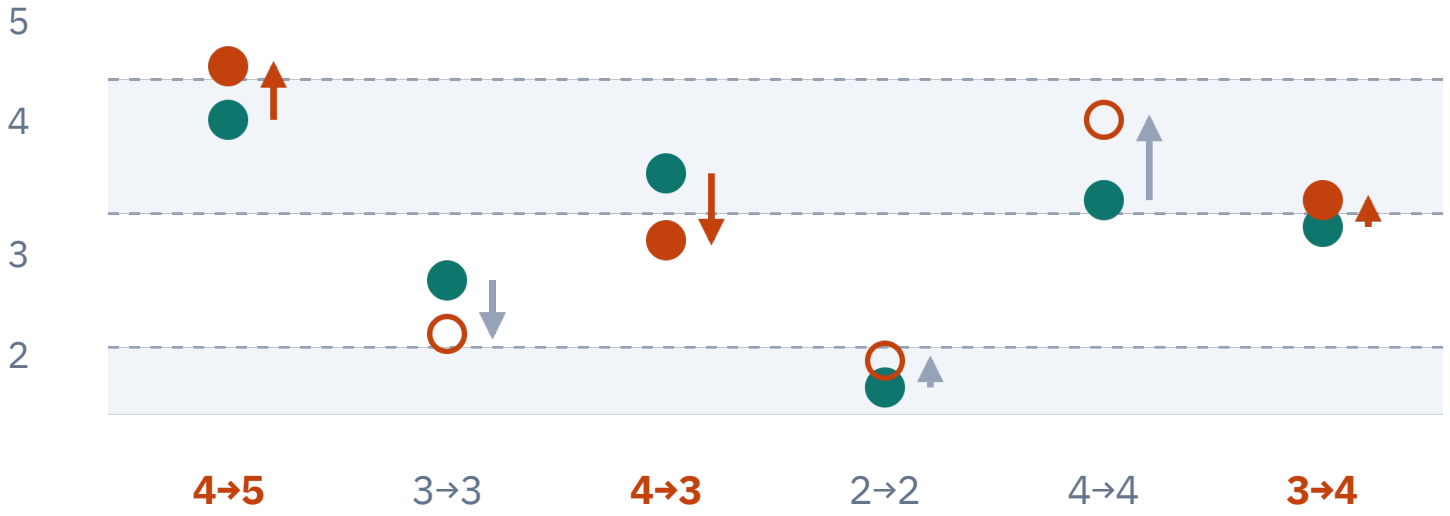
Version it, test it, improve it. Just not at the same time as the service.

Noise alone

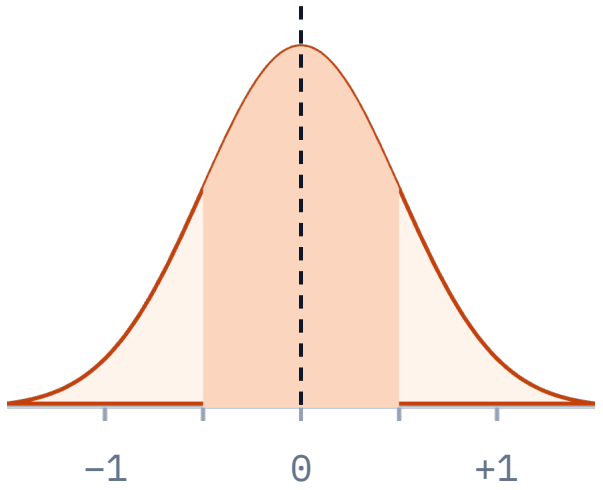
Errors that scatter, in no particular direction. The judge is consistent and has no bias.

How we model a noisy judge

We assume we have a Perfect Oracle that can give us the correct answer for some test dataset



● Oracle's score ● judge's score, rating flips ○ judge's score, rating stays

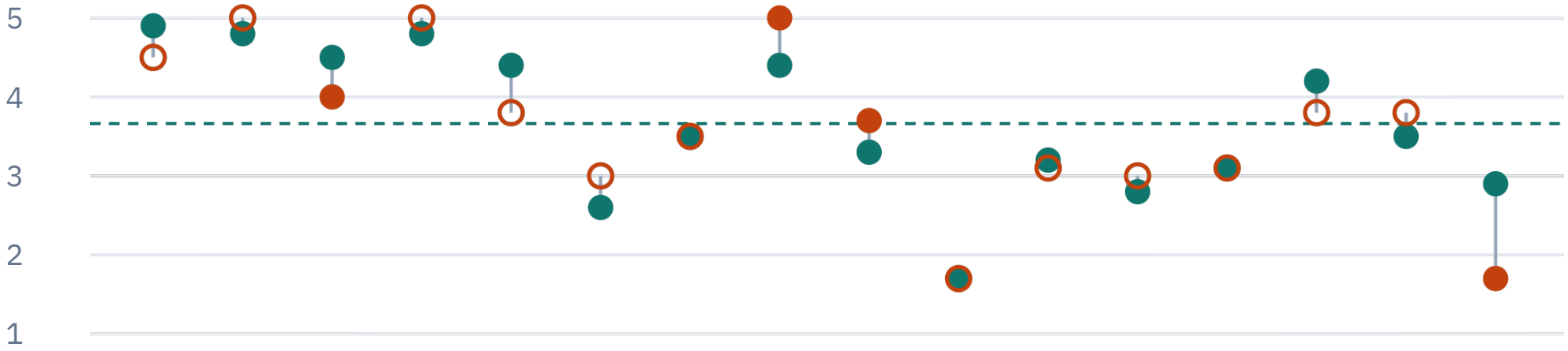


Noise per answer: centered on 0, **σ = 0.5** (shaded: ±σ)

Judge's score = Oracle's score + "random noise" error. Noise: the errors scatter around 0.

Noise on a 1–5 scale

Same model as before, σ = 0.5, now on sixteen answers.



● Oracle's score ● judge, rating flips (4 of 16) ○ judge, rating stays

Dashes: Oracle's mean 3.66. σ = 0.5, clamped to 1–5. Judged mean 3.61. RMSE 0.45.

Noise averages out of the mean, not out of each score: the mean stays near the Oracle's, while the typical miss (RMSE) grows with σ.

Playground: play with noise

1. Bias at 0, noise at 0.5: the picture from the last slides. Flips go both ways.

2. Press New judge: other answers flip. The mean barely moves.

3. Raise noise: more flips, a wider error histogram, and RMSE climbs.

Open the bias and noise tab

Noise on a binary scale

Oracle ↓ Judge →	AIS answer is correct	AIS answer is wrong
AIS answer is correct (80)	72	8
AIS answer is wrong (20)	2	18

AIS = the AI system under test. Same model on a 0–1 scale: the noise flips 10% of verdicts, in either direction. Illustrative counts.

Oracle: AIS correct

80%

Judge: AIS correct

74%

No lean, yet the judge's number is pulled toward 50%, and more cases do not close the gap.

Implications of noise

Single scores are unreliable

Any verdict can be off, borderline answers most of all

Means survive

Errors cancel over many answers: the average stays near the Oracle's

Rates do not

Share correct, or rated 4+, is pulled toward 50%, however many cases

Trust the average, not the single score, and be careful with thresholds.

How to measure noise in a consistent judge

Same as measuring any system, with the Oracle as ground truth.

1

Collect data points

Answers the judge has scored

2

Get the Oracle

The correct score for each one

3

Compare

Errors centered on 0; their spread is σ.
Binary: flips both ways.

The more data points, the better we can estimate σ.

How to mitigate noise

Lever	What it fixes
More cases	Averages noise out of the mean; no single verdict gets better
Clearer rubric, anchors and examples	Less case-by-case scatter
Several independent judges, averaged	Cancels errors only if they are independent
Correct a pass rate when q is known	true = (measured − q) / (1 − 2q)

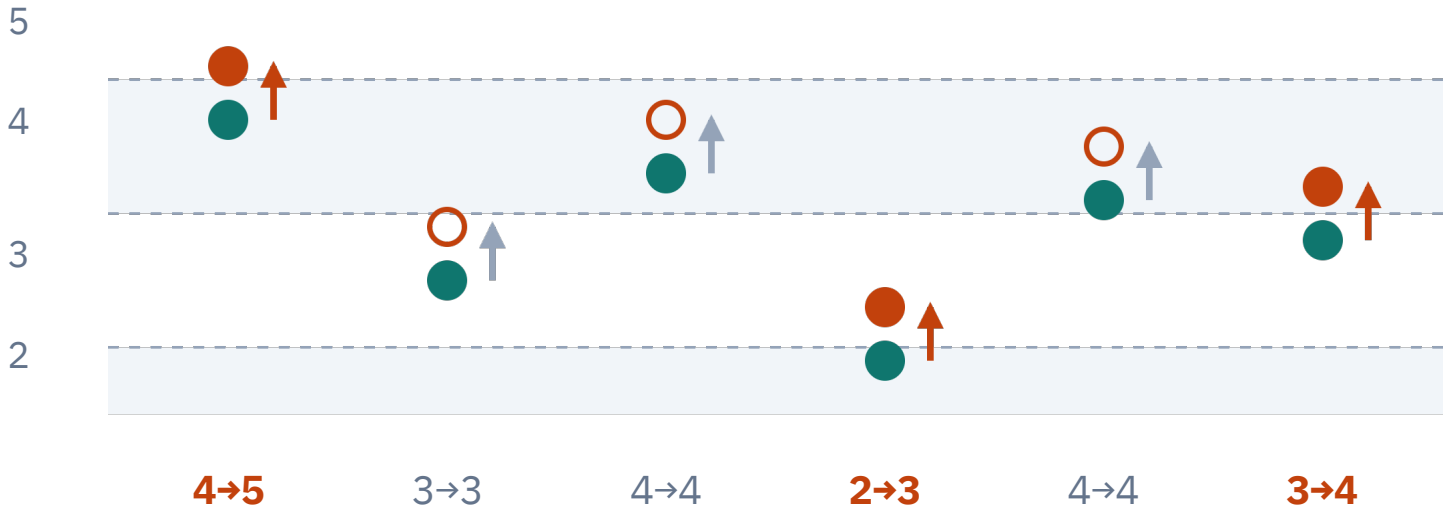
On a pass/fail or "share rated 4+" metric, noise does not average away.

Bias alone

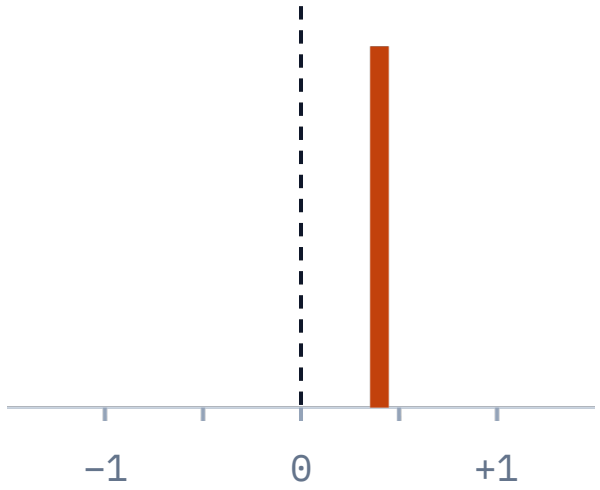
Errors that lean the same way. The judge is consistent and has no noise.

How we model a biased judge

We assume we have a Perfect Oracle that can give us the correct answer for some test dataset



● Oracle's score ● judge's score, rating flips ○ judge's score, rating stays

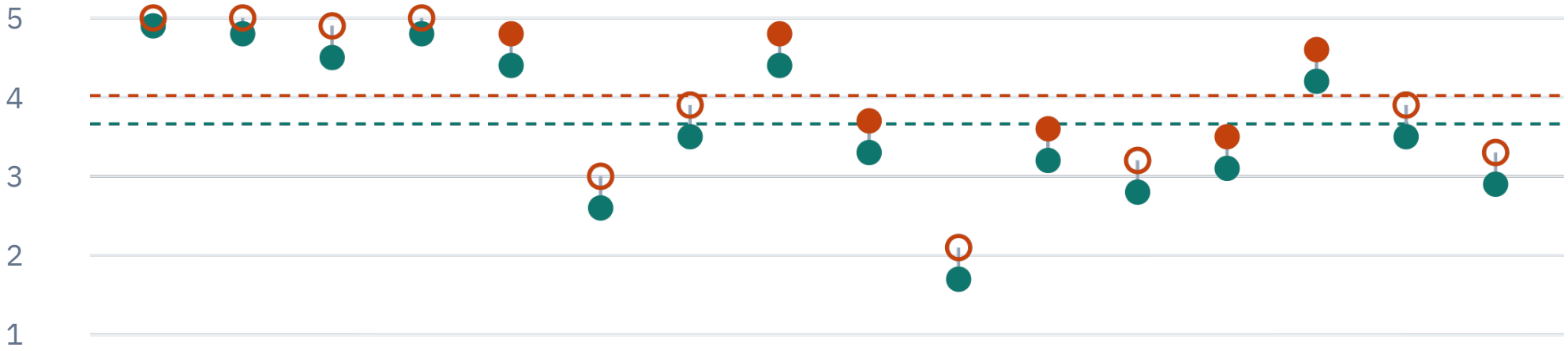


Error per answer: always **b = +0.4**, never centered on 0

Judge's score = Oracle's score + a constant error *b*. Bias: every error leans the same way.

Bias on a 1–5 scale

Same model as before, b = +0.4, now on sixteen answers.



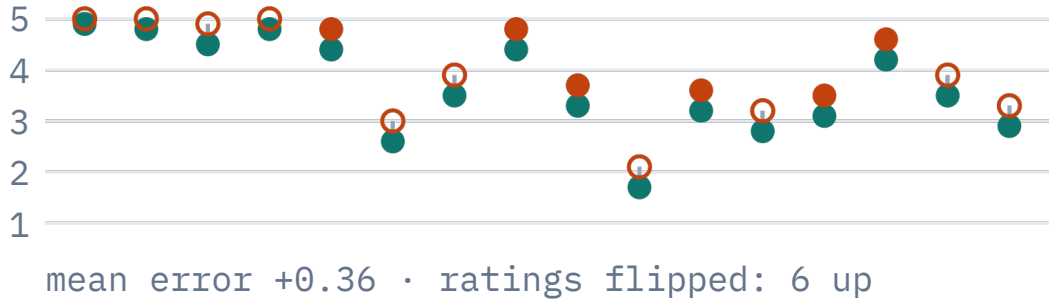
● Oracle's score ● judge, rating flips (6 of 16) ○ judge, rating stays

Dashes: Oracle's mean 3.66, judge's mean 4.02. b = +0.4, clamped at 5. RMSE 0.37.

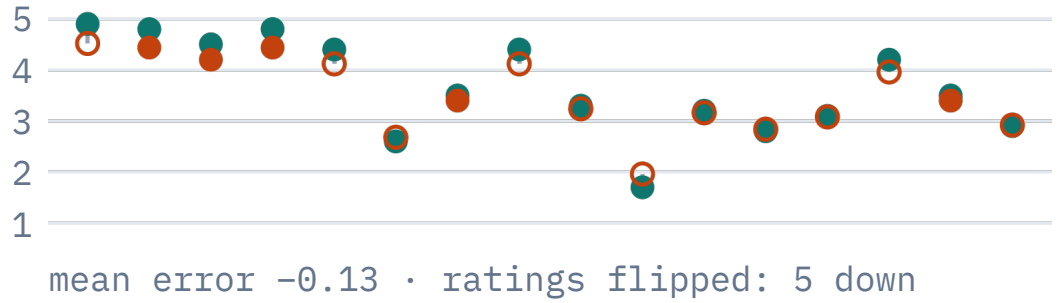
The whole picture slides up: the mean is off by about b, and more cases do not help.

Four shapes of bias on a 1–5 scale

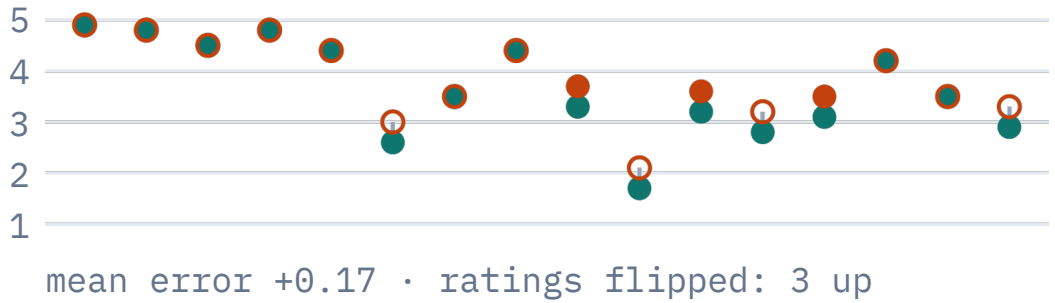
Constant: every answer +0.4



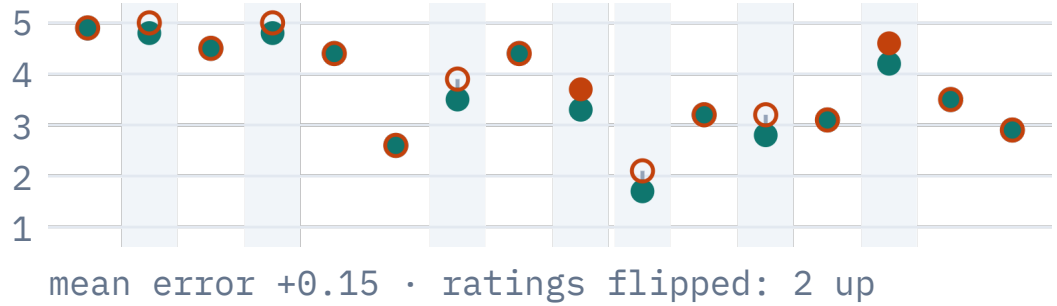
Toward the middle: weak up, strong down



Lenient on weak: only answers below 3.5



Rewards length: only long answers (shaded)



● Oracle's score ● judge, rating flips ○ judge, rating stays strength $b = 0.4$

Bias can depend on the score or on the answer, and some shapes hide from the mean error.

Playground: play with bias

1. Noise at 0, bias at +0.4: every dot moves up, flips only go up. The histogram is a spike at +0.4.

2. Press New judge: nothing changes.

3. Try the other bias shapes. With Toward the middle, the mean error is near 0, yet most answers are wrong.

Open the bias and noise tab

Bias on a binary scale

Oracle ↓ Judge →	AIS answer is correct	AIS answer is wrong
AIS answer is correct (80)	80	0
AIS answer is wrong (20)	8	12

AIS = the AI system under test. Same model on a 0–1 scale: the lean flips 40% of wrong answers to correct, and no correct ones. Illustrative counts.

Oracle: AIS correct

80%

Judge: AIS correct

88%

Every mistake pushes the same way. More cases only make you surer of 88%.

How a judge becomes systematically wrong

Failure source	What the score can reward
Missing evidence	Plausibility instead of support
Unclear rubric	A different definition of "good"
Length, order or model family	Incidental features of the answer
Leniency or harshness	A shifted acceptance standard

Bias is broader than the familiar position and verbosity effects.

Implications of bias

Every number leans

The headline score is off by the lean, in one direction

More data does not help

The interval narrows around the wrong value

Invisible without the Oracle

A consistent biased judge looks perfectly reliable

Bias moves the headline number, and can change which system wins.

How to measure bias in a consistent judge

Same as measuring any system, with the Oracle as ground truth.

1

Collect data points

Answers the judge has scored

2

Get the Oracle

The correct score for each one

3

Compare

Errors centered away from 0: their mean is *b*. Binary: flips go one way.

The more data points, the better we can estimate *b*, also per slice.

How to mitigate bias

Lever	What it fixes
More cases	Nothing: the interval tightens around the wrong value
Evidence and a calibrated rubric	Removes a known lean; validate on held-out cases
Correct with audited error rates	Pass/fail: true = (m − FPR) / (1 − FPR − FNR)

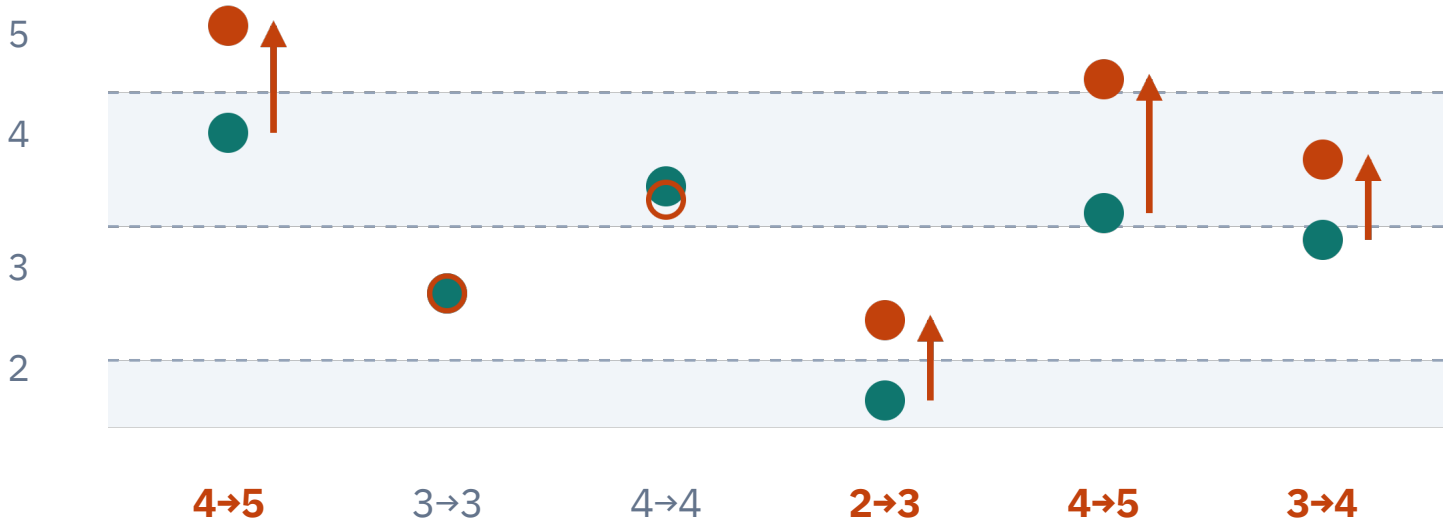
Every fix is a hypothesis: re-audit on cases the fix never saw.

Bias and noise together

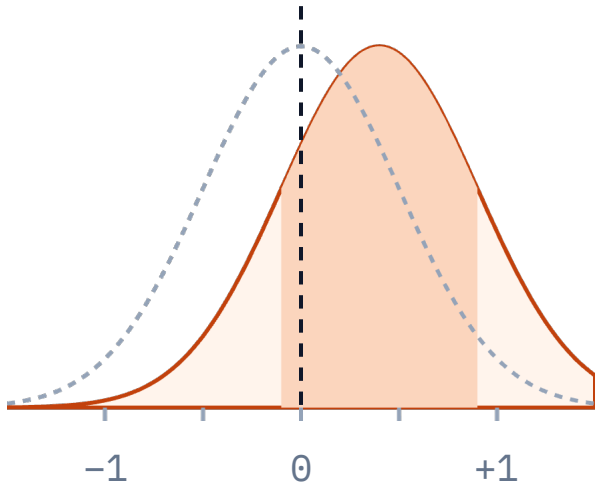
Real judges have both. What does more data fix, and what does it leave alone?

How we model a biased and noisy judge

We assume we have a Perfect Oracle that can give us the correct answer for some test dataset



● Oracle's score ● judge's score, rating flips ○ judge's score, rating stays



Error: centered on **b = +0.4**, σ = 0.5
(grey: noise alone)

Judge's score = Oracle's score + *b* + noise. The errors scatter around *b*, not around 0.

Playground: bias and noise together

1. Bias at +0.4, noise at 0.5: the histogram is the noise curve, shifted to +0.4.

2. Press New judge: which answers flip changes; the mean error stays near +0.4.

3. Raise the test-set size to 200: the histogram fills in around +0.4, not 0.

Open the bias and noise tab

One biased judge, one noisy judge

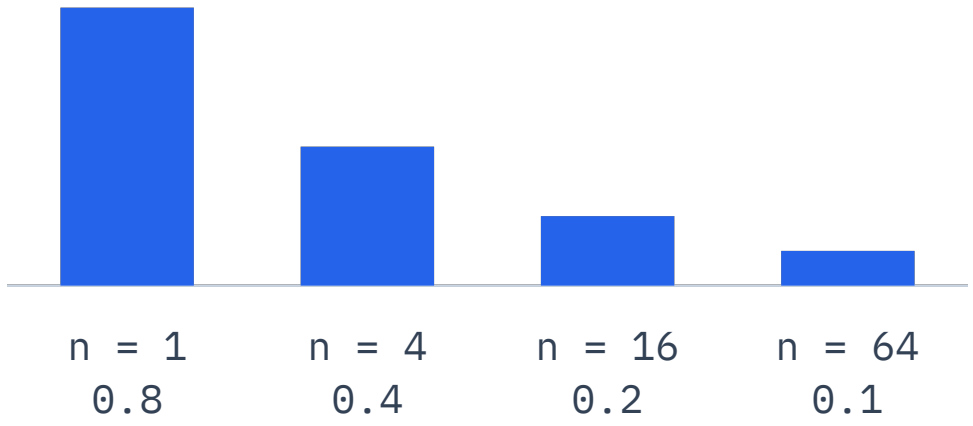
Answer	1	2	3	4	5
Reference	3	4	2	5	3
Judge A (biased)	4	5	3	5	4
Judge B (noisy)	2	5	3	4	3

Illustrative ratings on a 1–5 scale. Ask either judge again: the same scores come back.

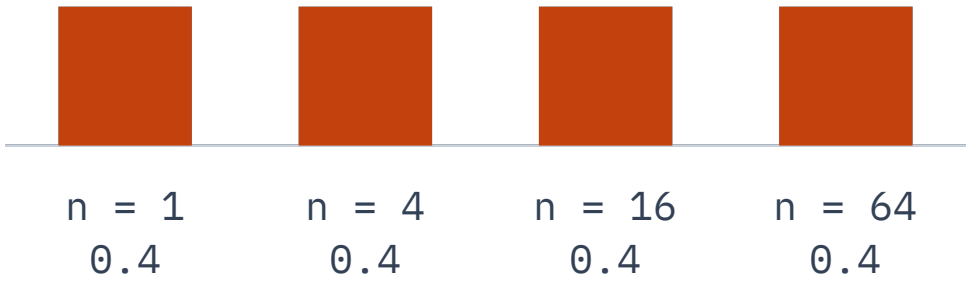
Both are perfectly repeatable. A is biased: it leans high. B is noisy: its errors scatter.

More cases reduce some errors

Random error of the mean



Shared systematic offset



n = independent cases scored by a consistent judge; values in score points. Illustrative additive-error model: per-case error with mean 0 and σ = 0.8, plus a shared offset of 0.4.

A narrow interval can still miss the target

Sampling uncertainty

Which examples and judgments happened to be observed?

Measurement bias

Where does this judge systematically depart from the standard?

Reporting limit

An interval around judge scores does not establish their validity.

The interval describes sampling behavior under a protocol. It does not cover the gap between that protocol and the intended standard.

Implications of bias and noise together

One number hides the mix

RMSE mixes the lean and the scatter

Confident and wrong

More data removes the noise, keeps the bias

A narrow interval is not enough

It speaks to noise, not to bias

Precision is not accuracy.

How to measure bias and noise together

Measure	What it tells you
Mean error and spread, reported separately	$RMSE^2 = \text{mean error}^2 + \text{spread}^2$; one number hides which you have
A reference-labeled audit set	The only way to tell the two apart
The interval next to the judge audit	Interval width says nothing about bias
Uncertainty of the audit itself	Ten bad cases give a very wide false-pass rate

A narrow interval is a statement about noise. Bias needs its own evidence.

How to mitigate bias and noise together

Step	Why
Fix or correct bias first	More data cannot remove it
Then size the sample for the noise	Noise in the mean shrinks as $1/\sqrt{n}$
Carry the judge's error into the claim	Uncertainty about bias limits what you can conclude
Re-audit when the judge changes	A new prompt or model is a new instrument

Order matters: calibrate, then collect, then report.

Subjective outputs

When reasonable people disagree, there is no single reference to be noisy or biased against.

Noise, bias and subjectivity

Source	What changes?	What to examine
Noise	Errors scatter across cases	Spread of errors vs reference
Bias	Systematic error vs reference	Error direction and affected slices
Subjectivity	Legitimate standards differ	Whose preferences define quality?

Define the reference and the population before calling an error bias.

Human disagreement is evidence too

Factual disagreement

Check the source and adjudicate the error.

Ambiguous instructions

Clarify the rubric and re-label affected cases.

Different preferences

Report whose standard the judge represents.

A majority label may serve a declared purpose. It does not erase the minority.

Implications of subjectivity

No single Oracle

Correct depends on whose standard

Disagreement is not always error

Judge vs human can be taste, not mistake

One score can hide trade-offs

Groups can want different things

Before measuring the judge, decide whose quality counts.

How to measure subjectivity

Measure	What it tells you
Several human raters per case	One label hides the disagreement
Agreement statistics, such as kappa	How much raters differ, not how wrong the judge is
The distribution of ratings, by rater group	Whose standard differs, and how
Re-label after clarifying the rubric	Disagreement that disappears was ambiguity, not taste

Separate disagreement about facts, about instructions, and about preferences.

How to mitigate subjectivity

Lever	What it does
Sharper criteria	Moves the job toward objective where a fact exists
Declare the audience standard	The judge represents someone; say who
Separate scores per criterion or group	One blended score hides real trade-offs
Report the distribution, not only the majority	The minority view is still evidence

Do not average away legitimate disagreement. Decide whose quality counts.

The HR-policy Judge Auditor

`10 answers × 5 prompts × 3 models × 5 repeats`

ATTEMPTED JUDGMENTS

750

UNIQUE TEST CASES

10

Fixed

Fictional policy, questions, generated answers and reference judgments

Varied

Judge prompt, model family/size and repeated calls

Open the Judge Auditor

Playground: fifteen judges, ten answers

1. Open S07, the justified refusal. Compare the small model under P3 and P4.

2. Open S05, the padded answer. Compare conciseness under P4 and P5.

3. Open the scorecard. Find the prompt that agrees most on faithfulness. Is it a good judge?

Open the Judge Auditor at S07 · Open the scorecard

Five prompts, different judging conditions

Prompt	What the judge receives
P1 Bare	Question and answer
P2 Rubric	Criteria, but no policy excerpts
P3 Evidence	Rubric and policy excerpts
P4 Examples	Sharper criteria, excerpts and worked examples
P5 Helpfulness	A deliberately different, more permissive standard

Read these as diagnostic contrasts, not causal estimates of any single factor.

Repeat noise on one fixed answer

S07 · CAN I TAKE A 3-MONTH UNPAID SABBATICAL?

"The policy excerpts I have do not cover sabbaticals, so I cannot confirm whether they are offered. Please ask HR directly."

P3, middle model	Run 1	Run 2	Run 3	Run 4	Run 5
Completeness	3	4	5	4	5

Saved course run, 16 September 2026. Reference completeness: 5.

Same output and rubric. Completeness varies from 3 to 5.

Calibration can repair a known failure

Small model, S07	Five completeness scores
P3: rubric + policy	1 1 1 1 1
P4: clarified rubric + examples	5 5 5 5 5

Saved course run. Reference completeness: 5.

P4 explicitly recognizes a justified refusal as a complete answer.

Validate the revised judge on new, held-out examples.

A changed rubric changes the target

S05 · IS PARENTAL LEAVE PAID IN FULL FOR THE FIRST 3 MONTHS?

"Thank you for reaching out about parental leave, which is understandably an important question for many colleagues at this stage of life. Subject to the usual eligibility rules, the first 12 weeks are paid at 100%…"

Middle model, same verbose answer	Conciseness ratings
P4: nothing unnecessary	2 2 2 2 2
P5: well organized and easy to read	4 4 4 4 4

Saved course run. Reference conciseness: 2.

Both judges are repeatable here. They apply different standards.

A true answer can miss the question

QUESTION

"I joined Northwind 4 months ago. Am I eligible for parental leave?"

ANSWER

"Parental leave is 26 weeks: 12 weeks at 100% of base salary and 14 weeks at 60%. Submit your request in the HR portal at least 4 weeks ahead."

POLICY

Eligibility requires at least six months of service.

Define faithfulness and completeness before scoring.

A judge audit is an experiment

Hold fixed

Cases, evidence, product outputs and scoring definitions

Measure

Reference disagreement, within-case variation and failure rates

Keep separate

Calibration examples and the final judge audit

Record unique cases, repeats, judge version and reference provenance.

Agreement can hide the costly error

Reference judgment	Judge passes	Judge fails
Acceptable	88	2
Unacceptable	7	3

Illustrative counts. Report the denominator and uncertainty for each error rate.

91%

agreement

7 of 10

bad answers pass

Playground: audit the release gate

1. Score the judge with the defaults: 91% agreement, 7 of 10 bad answers pass.
2. Raise "catches bad answers" to 80%. How much does agreement move?
3. Back to the defaults. Switch the audit set to 1,000. What happens to the interval?

Open the "audit the gate" lab

A report needs more than one number

What to report	Why it matters
Error direction and affected slices	Bias can cancel in the overall average
Within-case repeat distribution	A mean can hide unstable verdicts
Unparseable or failed calls	Missing judgments change the evidence
Intervals and sampling unit	Repeats are not new test cases

Agreement statistics can supplement this description. They cannot replace it.

Controlled probes reveal specific biases

Suspected bias	Controlled probe
Position	Swap answer order and repeat each order
Verbosity	Vary length while preserving substance
Identity or family preference	Blind identities and compare conditions
Leniency / harshness	Audit false passes and false failures

A single flip may be ordinary stochastic noise, not position bias.

Different problems need different fixes

Observed problem	Intervention to test
Unstable ratings	Clear anchors, repeat calls, fixed settings
Systematic disagreement	Better evidence and rubric calibration
Shared errors across judges	Diverse judges and independent validation
Legitimate subjectivity	Explicit audience standard or separate scores

Every intervention is a hypothesis. Measure whether it worked on held-out cases.

Judge changes need versioned comparisons

Version the rubric, prompt, model, settings and score mapping.

When the judge changes, re-score baseline and candidate.

A lower score can reflect a better measurement.

The same judge can still favor one system

Illustrative comparison	System A	System B
Reference quality	3.8	3.6
Judge bias on these outputs	0.0	+0.5
Measured quality	3.8	4.1

Paired testing helps with sampling noise. Differential bias can reverse the ranking.

Playground: compare two systems with one judge

1. Run with the defaults: bias 0 on A, +0.5 on B. Which system looks better?
2. Switch to 1,000 cases. Read the belief that B is better.
3. Set both biases to +0.5. What happens to the ranking?

Open the "compare two systems" lab

A decision report people can use

FINDING	Candidate lift [value], interval [limits], on [N] unique cases
JUDGE AUDIT	False passes [count / reference-bad cases], repeats [K], failures [count]
DECISION	Ship, hold or collect more evidence against a pre-agreed rule

Attach reference provenance, versions, affected slices and unresolved disagreements.

Discussion: would you ship?

Judge 1: candidate improves by 0.2 points.

Judge 2: candidate falls by 0.1 points.

The candidate's answers are longer.

Both judges are consistent across repeated calls.

What evidence would resolve the disagreement?

What a trustworthy result must survive

Understand

Separate repeat noise, systematic bias and legitimate disagreement.

Report

Show errors, affected slices, failures and uncertainty.

Manage

Calibrate, validate independently and version the judge.

Next: calibration can overfit too. Protect the evidence used to validate it.

Agreement statistics: useful, with limits

$$\kappa = (\text{observed agreement} - \text{expected agreement}) / (1 - \text{expected agreement})$$

Useful when raw agreement hides class imbalance.

Sensitive to class prevalence, margins and weighting.

Read alongside counts, error types and reference quality.

Never let a single agreement statistic replace error analysis.

Bias and error spread add in squared units

$$\text{mean squared error} = (\text{mean error})^2 + \text{variance of error}$$

RMSE measures typical error on the chosen numeric scale.

Repeat runs separate rejudging noise from item differences.

Normality is not required for this identity.

RMSE is a typical error, not a confidence interval or a standard error of the mean.

Score mappings can change a ranking

Five illustrative ratings	System A	System B
Ratings	4, 4, 4, 4, 4	1, 5, 5, 5, 5
Mean linear score (0–100)	75	80
Share with rating ≥ 4	100%	80%

A mapped 0–100 score is not automatically a percentage.

Playgrounds and further reading

LLM as a Judge: bias, noise and uncertainty

Judge Auditor: prompts, models and repeated judgments

Noisy and Biased Judges lesson page: the gate-audit and two-system labs

MT-Bench: judging LLM judges and their biases

Replacing Judges with Juries: evidence on diverse panels

Full course: Experiment Design and Evaluation of AI Systems