# A LANGUAGE-MODEL JUDGE'S RESOLUTION IS PREDICTED BEFORE IT IS MEASURED

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A language-model judge cannot tell every pair of answers apart. Below some difference in quality it is indifferent, and that threshold decides whether it can grade the work in front of it. We register two gates. The first is that the threshold can be predicted before it is measured, and the other is that a second budget can reverse the judge's preference outright. Judges grade worksheets whose quality is known exactly. A calibration block records the scores a judge actually uses, and from that block alone we predict, and commit with its hash, how often the judge will order fresh pairs correctly at every difference in quality, on three rating scales and six ways of reading its report. Only then is the test block's seed drawn.

**Resolution.** The prediction held on every graded judge, the largest deviation 0.078 in accuracy against a registered bar of 0.14. The scale a judge is offered is the wrong budget, and no one number drawn from its reports is the right one. The 7B judge asked for a score from 0 to 100 writes nine distinct scores and resolves 5.3 wrong answers in twenty where the scale promises one, and on every scale of every judge the channel it writes fits its accuracy better than the nominal scale does. Reading more of the report lowers the threshold by the predicted amount on five of the six graded judges.

**Reversal.** Give the worse of two worksheets a confident header and a check mark on every line. With no reasoning budget the judge prefers the decorated worse worksheet, only 0.125 of test pairs favouring the better one, and at 1024 tokens of reasoning it favours the better one in 0.95. Measuring what the evidence and the decoration are worth in separate blocks, in which the two never conflict, predicts that reversal, and a second ladder that brackets the crossing locates it, predicted at 72 tokens and observed at 57. On a second vendor's judge it is predicted at 187 and observed at 166, from that judge's own calibration.

The prediction and the comparison against the nominal scale replicate on a second vendor's models served through an API, on summaries checked against a passage, which no judge does exactly, and on readers' reviews scored against the star rating each writer gave. Read to 200 characters, to 400, and whole, the judge's resolution on those reviews coarsens in that order by the amount each budget's own calibration predicted. Every bar was fixed before the first of these runs and none was refitted. What predicts a judge's resolution is its measured channel, of which a scalar budget is one summary.

## 1 INTRODUCTION

A language model used as a judge is asked to tell two answers apart. The practice is standard (Zheng et al., 2023), and its known failures are position bias, a preference for some scores over others, and a dependence on the rating scale (Li et al., 2026). A question comes before all of these. How fine a difference in quality can the judge resolve, and can we know that before we rely on it?

An evaluator with finite resolution does not order every pair. Two answers that differ by less than it can resolve are indifferent to it, and the preference it induces is a semiorder, an ordering whose indifference is not transitive (Luce, 1956). Geometric Evaluation Theory derives that structure from what the evaluator spends on the comparison and proves that the threshold of the semiorder is set by that budget (Appendix B). For a judge that reports a score, the budget is the set of reports it actually



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

uses and how much of each report is read. It is not the scale the judge is offered, and the count of those reports is not it either. A judge asked for a score from 0 to 100 that only ever writes nine distinct scores has a report alphabet of nine symbols, and how finely it resolves depends on where those nine sit and how noisy they are as well as on how many there are. The calibration measures all three, and Appendix D scores what each of them is worth on its own.

This makes the threshold predictable. A calibration block shows the judge worksheets at every level of quality and records what it reports. Those reports fix, for every pair of levels, how often a given reading of the report orders the pair correctly, and so they fix the accuracy curve the judge will show on pairs it has never seen. The task is deliberately minimal, and that is what makes the claim testable. Worksheets of twenty single-digit multiplications have a quality that is known exactly, $1 - e/N$ for $e$ wrong answers, on a scale with twenty-one levels and no annotator disagreement. A threshold is only identifiable against a quality scale that is known rather than estimated, so a richer task would have bought realism at the cost of the thing being measured. The price is worth stating at the outset. Quality here is a count of arithmetic errors, not the contested and many-dimensional thing meant by the quality of an essay, and nothing in this paper establishes that a judge's resolution on the second can be predicted from its channel on the first. What the judge does with that scale is not minimal at all. It is free to use any of the scores it is offered, in any way, and the prediction is about what it in fact does.

We ran that prediction as a registration. Six graded judge configurations, five models across two vendors and three model generations, graded worksheets of twenty multiplications with a known number of wrong answers. Each judge's calibration was run, its predictions were written and committed with their hash, and only after that commit was the seed of the test block drawn. The order of the two commits is the evidence that the threshold was predicted before it was measured.

Four things came out of the sealed run. The calibration predicted the test block for every graded judge, at every gap of every scale and read-out, inside bars fixed beforehand from a simulation in which a faithful judge fails 3.3 percent of the time. The channel a judge actually writes fits its accuracy far better than the scale it was offered. On the 0 to 100 scale the scale predicts a threshold of one wrong answer and the judges resolve 5.3 and 2.8. Reading more of the same report lowers the threshold in the predicted order and by close to the predicted amount, by a factor of 3.5 on the 7B without changing judge, prompt or scale. Quantizing the 7B to 4-bit weights left every threshold within the registered factor of its bfloat16 value, so it produced no resolution change larger than that tolerance on this task.

A fifth thing came out of an analysis after the gates were graded, and it sharpens what is being claimed. Scoring five predictors of the same held-out curve puts an information-rate summary of the judge's scores about two fifths of the way from the nominal scale to the fully calibrated channel in the median cell (Appendix D). The theory says resolution is budgeted, and the measurement says how much of a real judge's resolution a budget buys. The two are different statements and this paper keeps them apart. What a practitioner can use is the method, a channel measured on one block predicting discrimination on another. What this paper is, then, is a preregistered demonstration of one prediction of that theory, that an evaluator's effective channel decides which differences in quality it can resolve, and that its resolution can be measured before the comparisons it governs are observed. Turning a measured channel into a predicted curve is elementary once the channel is known, and a reader may fairly say so. What is not elementary is that the prediction carried its bars before the data existed, held across six read-outs and three stimulus families with those bars never refitted, failed where the instrument ran out of resolution rather than at random, and extended to where a second budget reverses the judge outright.

The same theory says what a second budget does. A judge that may reason before answering spends tokens on the comparison, and with too few tokens it falls back on whatever is cheapest to read. We decorated the worse of two worksheets with a confident header and a check mark on every line. With no reasoning the judge preferred the decorated worse sheet, and with enough reasoning it preferred the better one. That is the theory's three regimes on one axis, reversal, indifference, and distinction. Where it turns over was predicted from blocks in which the decoration never sat on the worse sheet, at 174 tokens of reasoning against 128 observed on a ladder too coarse to place it more finely than a factor of two, then to 57 tokens on a ladder built to bracket it and to 166 on a second vendor's judge, and the additive account of the two influences fit the sealed test cells seven times better than ignoring the decoration and twenty times better than ignoring the evidence.

Four rules for anyone building a judge follow from the measurements, in Appendix A. An earlier gate measured the same judge on a numeric task and found its threshold equal to the grid on the numbers it reads and writes, which calibrates the instrument rather than testing the theory, since the judge does that arithmetic exactly and the grid is then the whole story (Appendix C). Worksheet grading is a judgement the model does not do exactly, which is where a prediction can fail.

## 2 RESOLUTION AND THE BUDGETS OF A JUDGE

A judge maps an answer to a report, and its preference between two answers is the order of their reports. Two answers whose reports are equal are indifferent to it, and a tie counts as an error. Quality here is known exactly. A worksheet has $N = 20$ items, $e$ of them wrong, and its quality is $1 - e/N$. The judge's resolution is read from its accuracy curve, the share of pairs it orders correctly as a function of the gap in wrong answers between them, and its threshold is the gap at which that accuracy reaches $0.75$.

Five quantities have to be kept apart, because which of them is claimed to predict which is the whole question. The *capacity* is what the scale offers, 101 levels on a 0 to 100 scale, whatever the judge does with them, and it is set by the experimenter. The *channel* is what the judge does with them, the conditional distribution of its score given the true quality, measured on the calibration block. Its support is the codebook, and the mutual information between score and error count is a rate in bits. The *read-out* is how much of the report is read, one symbol for the greedy score, the whole distribution for the expected score, $n$ symbols for the mean of $n$ scores drawn at temperature one, and for a judge that may reason, the tokens it spends before answering. The *resolution* is the threshold, and it is what everything here predicts. The *reliability* is the lapse, error that does not fall as the gap grows. Capacity is given, the channel and the read-out are the judge's and the reader's, and resolution and reliability are what come out.

The prediction follows from the channel and the read-out. For a read-out $r$ and two levels $e < e'$, the calibration block gives the share of cross pairs, one worksheet at each level, that $r$ orders correctly. Averaging those shares over the levels a test pair can occupy predicts the accuracy at every gap, and the predicted threshold is where that curve reaches $0.75$. The theory's content is in what the prediction depends on. The channel sets the curve, not the capacity. A rival judge that uses every level of the nominal scale, exactly, predicts a threshold of one wrong answer on the 0 to 100 scale, and the claim is that this rival fits worse than the judge's own channel on every scale. The registered hypothesis is that reading more of the report lowers the threshold, in the order and by the amounts the calibration predicts. Direct pairwise comparison has no report grid at all, and if the judge's internal resolution is what limits it, its threshold sits at the floor of the pointwise read-outs.

This paper claims two separable things. The first is a theory. A length budget makes the induced preference a semiorder whose threshold is the budget, which is Theorem 2(b). A grid of step $h$ is not that budget, since it coarsens the distance and so induces on a set of options the weak order of Theorem 2(a) on the quantized distance, which no single threshold represents against the true distance. What Proposition 1 gives is a rule for one comparison, that two options are distinguished when their difference reaches the distance from the lower of them up to the next boundary. Where that alignment is unknown the tolerance is uniform on $(0, h]$, so the accuracy curve is its distribution function and a threshold read at accuracy $0.75$ is a quantile rather than a number the judge carries. The second is a method. A channel measured on one block predicts discrimination on another the judge has not seen. What the theory does not say is how much of a real judge's resolution a scalar budget accounts for, because a judge's channel is not a grid, its codebook is not evenly spaced and its scores are noisy. That is a measurement rather than a deduction, and Appendix D makes it by scoring predictors that know less than the whole channel against the same held-out curves. The mutual information the judge's score carries about quality, read as a count of levels, closes about two fifths of that distance in the median cell. It is a proxy for the theory's rate budget rather than the budget, which carries a workload and a distortion that this summary does not.

## 3 PROTOCOL

**Judges.** Qwen2.5 at 1.5B, 7B, and 14B parameters (Qwen Team, 2025), the 7B also with 4-bit NF4 weights (Dettmers et al., 2023), and Gemma 3 at 4B (Gemma Team, 2025), each at a recorded revision, in bfloat16 unless stated.

**Stimuli and elicitation.** A worksheet lists twenty single-digit multiplications with a student's answers. Wrong answers differ from the product by one of the operands, by ten, or by one or two. The judge rates a worksheet on a scale of 1 to 5, 0 to 9, or 0 to 100, and in a pairwise elicitation it is shown two worksheets in both orders and asked which student got more right, read from the logits of the answer letters. For each rating we record the greedy score and the judge's exact probability over the scale. On the two single-token scales that probability is the first-token distribution. On the 0 to 100 scale a score takes up to three tokens, and we compute its distribution exactly with a digit tree, extending every digit prefix whose probability exceeds $10^{-5}$ by one cached decoding step with the model's repetition penalty applied as sampling applies it. Samples are drawn from that distribution, so they are exact samples of the judge at temperature one and cost no generation. Real sampling matched the tree within multinomial noise on every worksheet checked (Appendix H).

**Blocks.** The calibration block holds 40 worksheets at each of the 21 error counts. The test block holds 200 fresh pairs at each gap of 1, 2, 3, 4, 6, 8, and 12 wrong answers, the better sheet's error count uniform on what the gap allows.

**Prediction before measurement.** The registration, with its claims, estimator, bars, and grading code, was committed and its hash recorded before any sealed run. Each judge's calibration block was then run, and its predictions were written and committed with their hash. Only after that commit was pushed was the test seed drawn, with a system random source, and the test block run and graded by the registered script.

**What this discipline has cost.** A registration is only worth what it refuses, so here is what ours refused, all of it visible in this paper. A theorem was false as first stated. The uniqueness result of Theorem 3(c) omitted a hypothesis and fails in rank one, which an external reader found and which is corrected here, with the machine-checked version being the whole-space case. Two of the five judges of the first registration could not order worksheets twelve errors apart and were declared vacuous from their calibration blocks, before their test blocks were drawn, and are reported and not graded. The read-out ordering claim then failed outright on one judge of the replication, and because the registration holds that claim only if every graded judge passes it, the claim fails and its gate fails with it. The registered threshold estimator turned out to report $0.87h$ where the quantity it names is $0.9h$, an artifact of interpolating on a coarse ladder rather than a property of any judge, so the fitted step is reported beside it. The budget at which the preference reverses could not be located to better than a factor of two on the ladder first registered, and a second ladder was registered to locate it. And the framing this paper began with, that a scalar budget predicts resolution, survives only in a weakened form. The predictor ladder of Appendix D puts a rate in bits about two fifths of the way from the nominal scale to the calibrated channel, so what predicts is the channel. None of these was discovered by inspection after the fact. Each is a bar, fixed before a seed was drawn, doing the thing bars are for. The wider programme this belongs to keeps a ledger in which a claim's class is raised only by a sealed pass and lowered by a sealed miss. It currently carries refutations and one retraction alongside its confirmations, and a gate in it has already refuted a scaling law this theory had posited.

**Bars.** A pilot on separate seeds fixed the bars by simulation. Each simulated replicate draws a calibration block and a test block from disjoint halves of the pilot's worksheets and grades them with the registered code, so a replicate is the experiment with nothing wrong. All eighteen read-outs share the same test pairs, so their deviations move together, and the bars are quantiles of the largest deviation over all read-outs and gaps of one judge at once. A faithful judge fails the prediction check in 1.4 percent of replicates, the ordering check in 2.0 percent, and either in 3.3. A judge that cannot order worksheets twelve errors apart with accuracy 0.9 on some scale is declared vacuous from its calibration block before its test block is drawn, and it is reported and not graded.

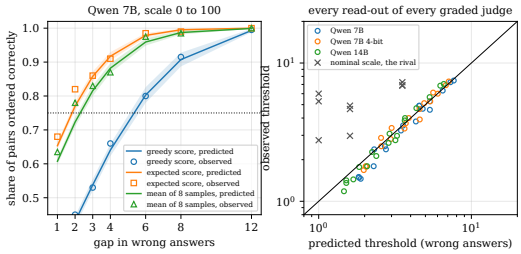



Figure 1: Left, the 7B judge on the 0 to 100 scale. Lines are what the calibration block predicts, bands two standard deviations of its bootstrap, points the sealed test block on worksheets the judge had not seen. The dotted line is the 0.75 level that defines the threshold. Right, every read-out of every graded judge, 54 points, predicted threshold against observed. Points on the diagonal were predicted before the test seed existed. The crosses are the rival that reads the scale at face value, and it predicts a finer threshold than the judge has, by up to a factor of five.

Table 1: Sealed thresholds in wrong answers out of twenty, predicted from each judge's calibration block and observed on its test block. The scale a judge is offered is not its budget. On 0 to 100 the 7B writes nine distinct scores and the 14B thirteen, and a rival that used all 101 levels would predict a threshold of one. Gemma 3 4B and Qwen 1.5B fail anti-vacuity from their calibration blocks and are reported, not graded.

| | | greedy score | | expected score | | mean of 8 samples | |
|---|---|---|---|---|---|---|---|
| Judge | scale (levels used) | pred. | obs. | pred. | obs. | pred. | obs. |
| Qwen 7B | 1 to 5 (5) | 7.29 | 7.26 | 1.84 | 1.48 | 4.33 | 3.78 |
| | 0 to 9 (8) | 4.56 | 4.91 | 1.88 | 1.45 | 2.31 | 2.38 |
| | 0 to 100 (9) | 5.35 | 5.29 | 1.83 | 1.50 | 2.29 | 1.79 |
| Qwen 7B 4-bit | 1 to 5 (5) | 6.81 | 6.91 | 2.03 | 1.79 | 3.65 | 3.36 |
| | 0 to 9 (8) | 4.58 | 4.65 | 2.05 | 1.78 | 2.57 | 2.50 |
| | 0 to 100 (9) | 6.02 | 6.00 | 1.97 | 1.68 | 2.56 | 2.88 |
| Qwen 14B | 1 to 5 (5) | 6.26 | 6.86 | 1.52 | 1.36 | 3.68 | 4.00 |
| | 0 to 9 (9) | 3.26 | 2.98 | 1.52 | 1.42 | 2.08 | 1.79 |
| | 0 to 100 (13) | 3.18 | 2.76 | 1.46 | 1.18 | 1.69 | 1.44 |

## 4 THE THRESHOLD IS PREDICTED BEFORE IT IS MEASURED

**The calibration block predicts the test block.** For every graded judge, on every scale and every read-out, the observed accuracy stayed inside the registered bars at every gap. The largest deviation was 0.078 for the 7B, 0.055 at 4-bit and 0.072 for the 14B, against a bar of 0.14, and the largest deviation in noise units was 2.17, 1.62 and 2.20 against a bar of 4.55. Those bars came from a simulation on pilot data in which a faithful judge fails 3.3 percent of the time, and they were fixed before the run. Figure 1 shows the curves and the 54 thresholds. The commit history supplied with this paper records each judge's predictions with their sha256 in an ancestor of the commit that draws the test seed, so the order of the two is checkable in the artifact. It is the repository's own history and not a third-party timestamp.

**What the judge does with the scale, and not the scale, predicts its resolution.** On the 0 to 100 scale the 7B judge writes nine distinct scores and the 14B thirteen. A rival judge that used all 101 levels exactly would order two worksheets one wrong answer apart, and the sealed thresholds are 5.29 and 2.76. The registered comparison puts that noiseless rival against the judge's whole measured channel, the conditional distribution of its score given the true quality. On the 0 to 100 scale the rival's squared error against the observed curve is 1.169 and the channel's is 0.001, and the channel fits better on every scale of every graded judge. Which part of the channel does the predicting is a separate question, answered in Appendix D by scoring predictors that know only how many scores the judge used, only where those scores sit, or only the rate in bits. In the median of the eighteen graded cells the squared error is 0.419 for the rival, 0.278 knowing only how many

scores the judge wrote, 0.240 knowing only the rate in bits, 0.220 knowing only where its scores sit, and 0.0026 for the channel, so the best budget-only predictor closes 39 percent of the distance from the rival to the channel and the channel is better than all of them in every cell. The scale is not the resolution. What the judge writes is.

**Reading more of the report lowers the threshold, by the predicted amount.** The greedy score is one symbol from the codebook, the expected score reads the whole distribution over it, and the mean of eight samples reads eight. Their thresholds fall in that order, and the calibration predicts each one. On the 7B's 0 to 100 scale the threshold falls from 5.29 to 1.50 wrong answers, a factor of 3.5, without changing the judge, the prompt or the scale. Kendall's tau between predicted and observed thresholds across read-outs was 0.876, 0.908 and 0.948, against a registered floor of 0.86.

**Quantization moved no threshold further than the registered tolerance.** The 7B at 4-bit weights passed on its own terms, and every one of its eighteen read-out thresholds sits within the registered factor of 1.75 of the bfloat16 value, with individual ratios from 0.85 to 1.60. That is a bound on how far resolution moved rather than a demonstration that it held fixed, since a ratio of 1.60 is a change this ladder cannot separate from none. The earlier numeric gate, on arithmetic the judge does exactly, separated the two effects further and found the step unchanged with a lapse of 0.031 at 4-bit (Appendix C).

**Pairwise comparison resolves near the floor of the pointwise read-outs.** Shown both worksheets and asked which student did better, with no reasoning, the 7B reached 0.75 accuracy at 2.88 wrong answers against a floor of 1.83 predicted from its finest pointwise read-out, and the 14B at 1.96 against 1.46, both inside the registered factor. The 4-bit judge was coarser, 3.60 against 1.97. The judge's first-position preference was 0.85 to 0.86, which the two-order average removes by construction.

**Two judges are declared vacuous before their test blocks are drawn.** Gemma 3 4B and Qwen 1.5B order worksheets twelve wrong answers apart with their greedy score at best 0.86 and 0.72 of the time, below the registered 0.9, and carry at best 0.98 and 0.45 bits about quality. The registration calls such a judge vacuous from its calibration block, and its curves are reported rather than graded. Their predictions were nonetheless accurate, with largest deviations 0.097 and 0.083, which says the method describes a judge that cannot do the task just as it describes one that can.

**The same two results on a second vendor and a newer generation, read through an API.** The judges above share a tokenizer, a vendor and a loader, and the two that did not were vacuous. A third registration puts the same worksheets, blocks, read-outs, estimator and bars in front of three judges that share none of those, namely Gemma 4 at 31B and at 12B, both at weights the operator quantized to 4 bits, and Qwen3 at 27B, all served by a public research API that returns only the twenty most probable tokens per position. Nothing of the method changed but the judge. The bars are the ones fixed from the 7B judge's pilot, and using another gate's bars is the point, since a bar moved for a new judge is not a bar. All three met anti-vacuity and were graded. The prediction held on all three, the largest deviation 0.077 against the same bar of 0.14, and the measured channel beat the nominal scale on every scale of every judge. On the 0 to 100 scale the three write 21, 15 and 17 distinct scores of the 101 offered, and both Gemma judges use only multiples of five, a heaping the Qwen2.5 judges do not share. The read-out ordering failed, on one judge of the three. Gemma 4 at 31B returned Kendall's tau 0.804 against the bar of 0.86, and the registration holds the claim only if every graded judge passes it, so the claim fails and that gate fails with it. The thresholds were not the problem, since all eighteen of its read-outs landed within the registered factor of their predictions. What failed is the rank statistic, and seven of its eighteen read-outs have a predicted threshold of exactly one wrong answer in twenty, the smallest gap the ladder contains. A judge finer than the instrument cannot be placed on it, and read-outs tied at the floor cannot be ranked. Across the six graded configurations of the two registrations the count of read-outs at that floor is 7 for the Gemma 31B, 1 for the Qwen3 and 0 for the Gemma 12B, and 0 for every earlier judge, and the only one with seven is the only one to fail. That is an association over six judges and not a demonstration. What would settle it is a ladder whose steps go below one wrong answer, which needs worksheets longer than twenty items, and a probe reported in the supplement shows that buys the greedy read-out but not the expected score, whose threshold on that judge lies below one part in a hundred of quality.

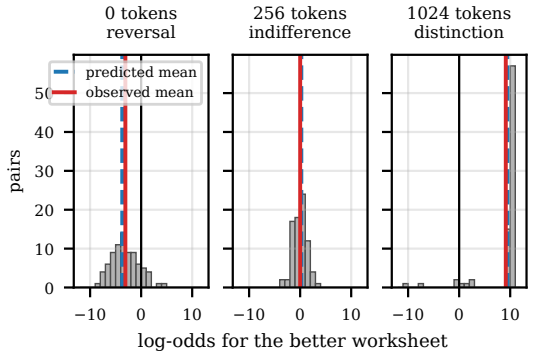



Figure 2: The sealed test block of the second gate, the cell where the worse worksheet carries the decoration. Each panel is the distribution over pairs of the judge's log-odds for the better worksheet at one budget, with the predicted and observed means. With no reasoning the judge prefers the decorated worse sheet, with reasoning it cannot finish it is near indifferent, and with reasoning it can finish it prefers the better sheet. Every prediction comes from blocks in which the decoration never sat on the worse sheet.

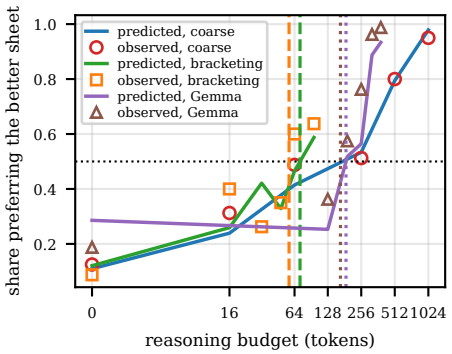



Figure 3: Where the preference turns over. The share of test pairs preferring the better worksheet against the reasoning budget, predicted from calibration blocks (lines) and observed on sealed test blocks (points). Vertical lines mark the predicted and observed crossings on the ladder that brackets it (dashed) and on Gemma 3 12B (dotted). The numbers are in the text.

## 5 Reversal, indifference and distinction on one budget

A judge that may reason before it answers spends a second budget, and the theory says the same thing about it. Below the resolution that budget buys, the judge is indifferent to the quality difference and decides on whatever else is in front of it. That much is the theory's. Two further things are not, and this gate tests them as empirical hypotheses rather than deriving them. The theory does not say that the thing decided upon is a decoration, nor which cue a judge will take when it is indifferent. That is a fact about judges, and the one used here was chosen because it carries no information about quality. Nor does the theory give the additive form. That what the evidence is worth and what the decoration is worth combine by adding on the log-odds scale is a modelling assumption, tested here against two rivals that drop one term each, and it could fail while the indifference the theory predicts held. We gave the worse of two worksheets a confident header and a check mark on every line, and measured the two influences apart. One calibration cell shows plain worksheets three wrong answers apart, which measures what the evidence is worth at each budget. The other shows worksheets of equal quality with one of them decorated, which measures what the decoration is worth. Neither cell contains a pair in which the two conflict.

Their sum on the log-odds scale predicts both test cells at every budget and the budget at which the judge stops preferring the decorated worse sheet, and the registration fixed the bars for that prediction before any pilot. On the sealed test block the prediction held, with deviations of at most 2.51 and 3.11 noise units against a bar of 3.341. The crossing was predicted at 174 tokens of reasoning and observed at 128, a deviation of 0.37 against a bar of 2.576 (Figure 3). That agreement did not locate the crossing. The rungs on either side, 64 and 256, gave 39 and 41 of 80 pairs to the better sheet, and the bootstrap spread was 1.16 in $\log_2$ budget, a factor of 2.2 either way. The additive account fit the two test cells with a squared error of 10.1, against 70.7 for an account that ignores the decoration and 228.7 for one that ignores the evidence. A second registration says where. Its first draft put the rungs at 64 to 256 tokens, and a pilot found the reversal already at one half at 64, so the sealed ladder is 0, 16, 32, 48, 64 and 96 tokens, on the same judge at fresh seeds, and the pilot with its failed rival check is in the registration. The calibration cells predicted the crossing at 72 tokens and the sealed test block observed it at 57, 0.62 noise units against 2.576, with the bootstrap spread at 0.52 in $\log_2$ budget, a factor of 1.4 where the first ladder gave 2.2. Additivity held at every rung, at most 3.00 noise units against 3.341, and the additive account fit the test cells with a squared error of 1.7 against 64.4 for evidence alone and 16.4 for the cue alone (Figure 3). A third registration asks whether that

Table 2: The second gate, sealed. The two calibration cells measure the evidence and the decoration apart, in log-odds for the sheet of interest. Their difference and sum predict the two test cells, where the decoration sits on the worse and on the better worksheet. The last columns are the share of test pairs preferring the better worksheet, predicted and observed, whose crossing of one half is the flip.

| | calibration | | decoration on worse | | decoration on better | | share better | |
|---|---|---|---|---|---|---|---|---|
| tokens | evidence | cue | pred. | obs. | pred. | obs. | pred. | obs. |
| 0 | 2.45 | 6.15 | -3.70 | -3.09 | 8.59 | 7.66 | 0.11 | 0.12 |
| 16 | 0.67 | 1.54 | -0.87 | -0.41 | 2.21 | 2.28 | 0.24 | 0.31 |
| 64 | 0.33 | 0.63 | -0.30 | -0.03 | 0.96 | 1.27 | 0.41 | 0.49 |
| 256 | 0.67 | 0.44 | 0.23 | 0.01 | 1.11 | 0.81 | 0.53 | 0.51 |
| 512 | 5.53 | 0.77 | 4.76 | 3.60 | 6.31 | 4.09 | 0.79 | 0.80 |
| 1024 | 10.44 | 1.00 | 9.44 | 9.13 | 11.44 | 10.02 | 0.98 | 0.95 |

is a fact about one judge. Google's Gemma 3 12B, on the same harness, was first registered on a ladder of 0 to 192 tokens and came out vacuous by the registered rule, its calibration predicting the decorated worse sheet preferred at every rung, so the ladder was moved to 0 to 384 tokens where that calibration placed the crossing, and both records ship. On the sealed test block additivity held at every rung, at most 2.83 noise units against 3.341, the crossing was predicted at 187 tokens and observed at 166, 0.91 noise units against 2.576 with a bootstrap spread of 0.18 in $\log_2$ budget, and the additive account fit with a squared error of 3.9 against 78.4 and 95.5 for the two rivals. Gemma holds the cue about three times longer than Qwen, and each judge's own calibration predicted its own crossing.

Figure 2 is the theory's three regimes on one axis. With no reasoning the judge prefers the decorated worse worksheet, at log-odds of $-3.09$ where the calibration predicted $-3.70$. At 256 tokens it is near indifferent, $+0.01$ against a predicted $+0.23$. At 1024 tokens it prefers the better worksheet, $+9.13$ against a predicted $+9.44$.

One feature of the middle regime was not anticipated and is worth stating plainly. The evidence term is not monotone in the budget. It is $+2.45$ with no reasoning, falls to $+0.33$ at 64 tokens and reaches $+10.44$ at 1024. Reasoning that stops mid-count leaves the judge worse than answering at once, so the indifference in the middle of the ladder is made by truncation as much as by any balance of evidence against cue. The prediction does not depend on monotonicity, since each budget is calibrated on its own, and the registration records the effect as a weakness of the ladder rather than a finding about judges.

## 6 Related work

Finite resolution producing intransitive indifference is Luce's semiorder (Luce, 1956), with the representation of Scott & Suppes (1958) and the evidence of Tversky (1969). Here the threshold is the budget, so it is a property of the evaluator that can be predicted rather than a constant fitted afterwards. Xu et al. (2025) measure a different intransitivity in the same setting, cycles in a judge's strict preferences over models, and repair the rankings with round-robin tournaments and a Bradley-Terry model. A semiorder keeps strict preference transitive and gives up transitivity of indifference, so the two are separate failures and a judge can show either without the other. What is measured here is the width of that indifference.

On language-model judges, Zheng et al. (2023) established the practice and its position and verbosity biases, and Li et al. (2026) found human alignment highest on a 0 to 5 scale, with finer scales worse. That finding is about the scale a judge is asked to use, and ours explains it as a budget the judge does not spend. Reading a judge's expected score from its probabilities over the scale is the read-out of G-Eval (Liu et al., 2023), which introduced it to make scores finer and better aligned with human ratings. We use it as one read-out among six and predict its threshold before measuring it.

Estimating a judge's conditional behaviour from gold-labelled data is a standard move put to other ends. Lee et al. (2025) use it to correct the bias in a reported evaluation score and to give that score a confidence interval on a test set the calibration did not contain. They bound the error of an

aggregate, where the quantity here is a curve, how often a judge orders a pair correctly against the difference in quality, committed with its hash before the test seed is drawn. Choi et al. (2026) fit an item response model and read discrimination and threshold parameters off it, treating the judge as a measurement instrument as this paper does, and those parameters are fitted to the responses they explain rather than written down for responses nobody has drawn. Vossler et al. (2026) place a judge's confusion channel in a probability simplex and ask when rankings from judges alone are sound with no human labels, and they find a binary score can rank better than a multi-level one, which agrees with the finding here that the levels a scale offers are not the levels a judge uses. Usami et al. (2026) report a judge as a measurement instrument with a datasheet, its response to empty input, to surface variation at equal quality, to position and to a controlled quality ladder, and find that a tie instruction moves the criterion and not the resolution, and Zhao et al. (2026) show that telling a judge to ignore presentation cues can suppress the bias while destroying its resolution on validated quality gaps. Both characterise the instrument on the data they collect, where here its resolution is predicted, with its bars, before the data it will be scored on exists, and the budget at which a cue stops moving the judge is predicted from blocks in which cue and evidence never conflict. Judges are known to favour answers that carry marks of authority, such as citations that are fake (Ye et al., 2025), and our decoration is a cue of that kind. What is new here is not that a cue sways a judge, or that reasoning can weaken it, but that where it stops swaying the judge is predicted from measurements in which evidence and cue never conflict. Reasoning does not remove such cues. Wang et al. (2025) find that large reasoning models used as judges remain susceptible to authority, bandwagon, position and distraction biases, are more robust than other models only on fact-related data, and are swayed by text that merely imitates reflection. Their comparison is across models. Ours is along one judge's reasoning budget, with the turning point predicted.

The nearest neighbour with a budget in economics is rational inattention (Sims, 2003; Caplin & Dean, 2015; Matějka & McKay, 2015), in which a cost of information about the state produces stochastic choice. The budget here limits the resolution of consequences rather than the acquisition of information about states.

## 7 Scope

The stimuli are three families, and only the third is prose, in one register. Worksheets give an exact quality scale and so do summaries scored by how many statements a passage does not support, the second being the one no judge does exactly. Neither is the contested and many-dimensional thing meant by the quality of an essay. The prediction assumes that a worksheet's score depends on how many items are wrong and not on which, and the record reports the score variance within an error count so that a reader can see how much noise that assumption absorbs. Matching predicted to observed accuracy is itself a test that the judge scores the two blocks alike, and the nominal-scale comparison and the read-out ordering can fail separately. The bars were fixed from one judge's pilot, the 7B judge at bfloat16, by the simulation of Section 3. Two constants were fixed after pilot values they bear on had been seen, the factor that labels the pairwise threshold and the information bar for anti-vacuity, and the registration says so for each. Two of the first registration's five judges, Gemma 3 4B and Qwen 1.5B, failed the anti-vacuity rule on calibration and are reported and not graded, so its verdicts rest on the 7B at two precisions and the 14B. The replication adds two sizes of Gemma 4 and one Qwen3, so on worksheets the prediction and the comparison against the nominal scale rest on six graded configurations of five models across two vendors and three generations, and the read-out ordering on five, having failed on the sixth for the reason given there. The second family regrades two of those configurations and adds no judge, so it buys a task and not breadth (Appendix F). The third scores readers' reviews against each writer's own rating at three input lengths (Appendix G), its budget claim resting on one judge and its ordering claim ungradable on five levels. No judge here was trained or tuned by us, and three of the six were served by a third party at weights we cannot inspect. The reversal is measured on two judges of two vendors, at one gap of three errors in ten, with one decoration, and a third judge, Llama 3.1 8B, is recorded as one the cue cannot move. Additivity of evidence and cue at that gap need not hold at others. The budget is a ceiling on reasoning tokens, not the reasoning spent, which the record keeps, and that axis is not monotone (Section 5). Reasoning is decoded greedily, so the noise in those estimates is over worksheets and not over samples of the judge. Nothing here is a claim about human evaluators.

## AI USE STATEMENT

Generative AI tools were used in this work, and the uses that the ICLR policy lists as requiring disclosure were the following. They assisted in the writing of proofs of the theorems in Appendix B, in refining the predictions into registered hypotheses with bars, in designing and giving feedback on the experimental protocols, in implementing the methods (the stimulus generators, the judge harness, the exact score distributions, the estimators, and the grading scripts), in generating the synthetic judges on which every gate's self-test was run, and in interpreting results into the written record. They were not used to generate the theory's definitions or its central claims, which were the authors' and precede the tools' involvement, and they were not used to produce any reported number, which come from executing the registered code on the recorded seeds. The language-model judges are the objects measured and had no part in the writing. Among the uses the policy lists as recommended for disclosure, the tools drafted and edited sections of this paper, searched and summarized related literature, produced the figures from the graded records, wrote and edited code, and formatted references. Verification was as follows. Every experiment was registered and sealed with its hash before its run seed was drawn, its self-test and probe were run first, its calibration predictions were committed with their hash before the test block was scored, and its verdict was produced by a grading script against the sealed bars and not by any model. The exact score distributions that replace sampling were checked against real samples and against brute force (Appendix H). Parts of the theory are machine-checked in Lean against Mathlib, in ten files with no unproven step and no axiom beyond `propext`, `Classical.choice` and `Quot.sound`. The checked statements are the tolerance condition of Theorem 1(c), the semiorder of Theorem 2(b), the hull law of Theorem 3(b), the whole-space case of the uniqueness in Theorem 3(c), the reversal example, part (a) of Proposition 1 together with its two quantizers and the bound on the tolerance, and the statements on shared-representation regret in Appendix K, with Ky Fan's principle taken as a named hypothesis. Two further files carry results of the wider development that the main text does not use, on distinguishability under a read operator and on the horizon of a budgeted reader. The remaining statements were checked by hand, the open-set form of Theorem 3(c), which was corrected after an external reader found a defect in it, and parts (b) and (c) of Proposition 1, which are measure-theoretic. Every citation was checked against the publisher's record. All AI-assisted text, code, and proofs were reviewed by the authors, who take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

The supplementary material contains, for every gate, the sealed registration with its hash, the configuration with model revisions and library versions, the seeds, the commits that order prediction before measurement, the graded records, and the scripts that produce every figure and table from them. A notebook in the package runs inside it with no network, no model and no GPU. It regenerates every figure from the graded records, checks the principal numerical claims made here against those records and reports any that disagree, re-runs the grader over the score records of the sealed gates and compares the verdicts it derives with the ones shipped, and checks by hash that the predictions graded are the ones each test seed was drawn against.

## REFERENCES

Laura Blow, Martin Browning, and Ian Crawford. Revealed preference analysis of characteristics models. *Review of Economic Studies*, 75(2):371–389, 2008. doi: 10.1111/j.1467-937X.2008.00485.x.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015. doi: 10.1257/aer.20140117.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of LLM-as-a-judge via item response theory. *arXiv preprint arXiv:2602.00521*, 2026.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems 35*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36*, 2023.

Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949. doi: 10.1073/pnas.35.11.652.

Peter C. Fishburn. Intransitive indifference with unequal indifference intervals. *Journal of Mathematical Psychology*, 7(1):144–149, 1970. doi: 10.1016/0022-2496(70)90062-3.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. *arXiv preprint arXiv:2511.21140*, 2025.

Weiyue Li, Minda Zhao, Weixuan Dong, Jiahui Cai, Yuze Wei, Michael Pocress, Yi Li, Wanyan Yuan, Xiaoyue Wang, Ruoyu Hou, Kaiyuan Lou, Wenqi Zeng, Yutong Yang, and Yilun Du. Grading scale impact on LLM-as-a-judge: Human-LLM alignment is highest on 0–5 grading scale. *arXiv preprint arXiv:2601.03444*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

R. Duncan Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24(2):178–191, 1956. doi: 10.2307/1905751.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015. doi: 10.1257/aer.20130047.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Dana Scott and Patrick Suppes. Foundational aspects of theories of measurement. *Journal of Symbolic Logic*, 23(2):113–128, 1958. doi: 10.2307/2964389.

Herbert A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1): 99–118, 1955. doi: 10.2307/1884852.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003. doi: 10.1016/S0304-3932(03)00029-1.

Amos Tversky. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969. doi: 10.1037/h0026750.

Hiroyasu Usami, Keisuke Hara, Ayato Tsuboi, and Naohiko Matsuda. LLM judges have dark current: A psychometric datasheet for LLM-as-a-judge evaluation. *arXiv preprint arXiv:2606.15610*, 2026.

Patrick Vossler, Fan Xia, Yifan Mai, Adarsh Subbaswamy, and Jean Feng. LLMs judging LLMs: A simplex perspective. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Qian Wang, Zhanzhi Lou, Zhenheng Tang, Nuo Chen, Xuandong Zhao, Wenxuan Zhang, Dawn Song, and Bingsheng He. Assessing judging bias in large reasoning models: An empirical study. *arXiv preprint arXiv:2504.09946*, 2025.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in LLM-as-a-judge. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *International Conference on Learning Representations*, 2025.

Liang Zhao, Yong Wang, and Jiangzhe Chen. Debiasing as a measurement intervention: Calibrated ties and resolution loss in LLM-as-a-judge evaluation. *arXiv preprint arXiv:2609.12439*, 2026. To appear at EMNLP 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.

## A PRACTITIONER'S SUMMARY

Four rules for building or trusting a language-model judge, each one a consequence of a measurement in this paper rather than of the theory alone.

**1. Measure the channel the judge writes, and do not trust the scale you offered.** Asked for a score from 0 to 100, the judges here wrote between 9 and 21 distinct scores. A rival that assumes all 101 levels are used predicts a threshold of one wrong answer in twenty, and the judges resolve between 1.5 and 8.5. On every scale of every judge the measured channel, the conditional distribution of the score given the quality, fits the observed accuracy better than the nominal scale does. The measurement is cheap, one forward pass per item on a calibration set whose quality you know, then the conditional score distribution at each quality level. The count of distinct scores is the cheapest summary of that distribution and it is not the part that does the predicting, which Appendix D measures. Reporting a judge's agreement without reporting its channel leaves out the quantity that decides what the agreement could have been.

**2. Read the expected score from the probability vector, not the token the judge wrote.** The greedy score is one symbol, and the expected score under the judge's own distribution over the codebook uses the whole vector. On a single-token scale such as 1 to 5 or 0 to 9 that vector is the first-token distribution and costs nothing beyond the forward pass you already ran (Liu et al., 2023). On a 0 to 100 scale a score spans up to three tokens, so the vector has to be built with a digit tree, one cached decoding step for every prefix worth extending, which is far cheaper than sampling and is not free. On the 7B judge and the 0 to 100 scale this moved the threshold from 5.29 wrong answers to 1.50, a factor of 3.5, without changing the model, the prompt or the scale. Averaging several sampled scores helps too and costs a forward pass each, so on the single-token scales the expected score dominates it, the same information in one pass. The ordering across read-outs was predicted before it was measured, so the gain is not a lucky artifact of one judge.

**3. A hard token cap on a reasoning judge can leave it worse than no reasoning at all.** On the one judge, one cue and one forced-answer protocol of Section 5 the budget axis is not monotone. With no budget the judge preferred the decorated worse worksheet in 7 pairs of 8, and at the two smallest non-zero budgets it sat near indifference, which is a harder failure to detect than a confident wrong answer. One judge does not establish that every reasoning judge behaves this way, and the mechanism belongs to a protocol that forces an answer when the cap is reached. A cap chosen for cost can still turn a wrong answer into a coin flip without announcing it, so bound cost with a budget the judge can plan against, or accept the answer it gives when it stops on its own, and measure what your cap does to the judge you have.

**4. Check the judge can do the task at all, and check it before you look at the results.** Two of the five judges in the first registration could not order worksheets twelve errors apart in twenty with accuracy 0.9 on any scale, and carried under one bit about quality. Their predictions were still accurate, which is the trap, since a method that describes a judge faithfully will describe a judge that cannot do the job, faithfully. Fix the bar, apply it to a calibration block, and decide whether to grade the judge before its test data is drawn.

# B THE THEORY

This appendix states the evaluation object and the results the main text uses. Proofs are in Appendix I.

## B.1 THE EVALUATION OBJECT

**Definition 1** (Evaluation object)**.** An evaluation object is a six-tuple $\mathcal{E} = (X, A, C, G, B, \mathcal{K})$ in which $X$ is a set of states carrying a probability law $\mu$, the evaluator's belief. $A$ is a set of actions. $C : A \times X \to \mathcal{Y}$ is the consequence map, into a consequence space $\mathcal{Y}$ that is a $C^1$ manifold. $G$ is the metric of evaluation, a continuous field of positive semidefinite quadratic forms on the tangent spaces of $\mathcal{Y}$, together with an ideal point $t \in \mathcal{Y}$. $B$ is the resolution budget, one of a rank cap, a length $\varepsilon \geq 0$, or a rate triple $(R, \nu, \delta)$ with $R$ in bits, with $B = \infty$ for the unbudgeted case. $\mathcal{K} \subseteq A$ is the set of admissible actions. Two restrictions belong in the definition rather than in the results that use them. $G$ is allowed to be degenerate, so it is a pseudo-metric and not a Riemannian metric in the usual sense. Theorem 3 chooses a positive semidefinite $G$, and its uniqueness clause is stated for $G_1 \neq 0$ rather than for $G_1$ definite, so calling it Riemannian here would contradict the results that use it. A rank cap is only defined when $\mathcal{Y}$ carries the linear structure its operations need, so a rank budget is a pair $(k, \nu)$ in which $\mathcal{Y} = \mathbb{R}^m$ with its Euclidean structure and $\nu$ is a workload, a probability measure on $A$ under which the second moment of consequences is taken. A rate budget needs a source and a cost for the same reason, so it is a triple $(R, \nu, \delta)$ with $\nu$ again a workload on $A$ and $\delta$ a distortion on $\mathcal{Y}$, since a number of bits says nothing without something to encode and a price for encoding it badly. Without $\nu$ the phrase *second moment of the evaluator's consequences* has no referent, since $\mu$ is a law on states and not on actions. The length budget needs neither restriction and is defined on any $\mathcal{Y}$ as above.

The world supplies $(X, A, C)$ and the evaluator supplies $(G, B, \mathcal{K})$. Two evaluators of the same world differ only in the second triple. The split is the source of the theory's content. If the evaluator could also choose $C$, every weak order on a finite $A$ would be an evaluation object, since a one-dimensional consequence $C(a) = t - d(a)$ represents any real function $d$. With $C$ fixed by the world, which preferences an evaluator can hold becomes a representation theorem and what its choices reveal becomes a uniqueness theorem.

The evaluation distance of an action is the root of the expected squared geodesic distance of its consequence from the ideal, $d(a) = (\mathbb{E}_\mu[d_G(C(a, X), t)^2])^{1/2}$, and its cost is $c(a) = d(a)^2$. It is a root mean square and not an expected distance, and the two differ by Jensen's inequality. It is the squared form that makes the cost additive over independent coordinates and gives Theorem 3 its quadratic $Q_W$. Under a rank cap $(k, \nu)$ the evaluator retains the top-$k$ eigenspace of $\mathbb{E}_{a\sim\nu}\mathbb{E}_\mu[C(a, X)C(a, X)^\top]$, the second moment of the consequences its workload actually presents, and evaluates the projection. Under a length $\varepsilon$ the distance is $d$ itself and the budget acts on comparison. Under a rate $(R, \nu, \delta)$ the evaluator holds an $R$-bit description of the consequence and evaluates the reconstruction, and the budget is treated as the length budget given by the least distortion attainable at that rate under $\nu$ and $\delta$. That last step is a modelling convention of this paper and is not proved here. The length budget is the one this paper measures.

## B.2 WHAT THE OBJECT INDUCES

**Theorem 1** (Induced distinctions)**.** *Let $\mathcal{E}$ be an evaluation object. (a) The relation $a \sim b$ if and only if $C(a, x) = C(b, x)$ for $\mu$-almost every $x$ is an equivalence relation on $A$. (b) Under a rank cap $k$ the relation $a \sim_k b$ if and only if the retained projections of $C(a, \cdot)$ and $C(b, \cdot)$ agree almost everywhere is an equivalence relation each of whose classes is a union of classes of $\sim$. (c) Under a length $\varepsilon$ the relation $a \approx_\varepsilon b$ if and only if $|d(a) - d(b)| \leq \varepsilon$ is reflexive and symmetric, and it is transitive on $A$ if and only if $\varepsilon = 0$ or the set $\{d(a)\}$ contains no three points $r_1 < r_2 < r_3$ with $r_2 - r_1 \leq \varepsilon$, $r_3 - r_2 \leq \varepsilon$, and $r_3 - r_1 > \varepsilon$.*

Part (c) is the whole difference between a resolution budget and a rank budget. A rank budget coarsens the evaluator's world into classes. A length budget produces a tolerance, and the failure of transitivity is the signature of finite resolution rather than a defect of the evaluator.

**Definition 2** (Induced preference)**.** Write $d_B$ for the budgeted distance and $\varepsilon_B$ for the length budget, with $\varepsilon_B = 0$ under a rank cap or no budget. Then $a \succ_B b$ if and only if $d_B(a) + \varepsilon_B < d_B(b)$, and $a \sim_B b$ if and only if $|d_B(a) - d_B(b)| \le \varepsilon_B$.

**Theorem 2** (Preference from geometry)**.** *Let $A$ be finite. (a) For $\varepsilon_B = 0$ the relation $\succsim_B$ is a weak order represented by the utility $u_B = -d_B$, unique up to strictly increasing transformation. (b) For $\varepsilon_B > 0$ the pair $(\succ_B, \sim_B)$ is a semiorder in the sense of Luce, and $u_B = -d_B$ with threshold $\varepsilon_B$ is a threshold representation of it. (c) The semiorder of (b) is a weak order if and only if the condition of Theorem 1(c) holds, and as $\varepsilon_B \to 0$ it converges to the weak order of (a).*

Proofs are in Appendix I. Luce (1956) derived semiorders from just-noticeable differences in utility with the threshold as a primitive, Scott & Suppes (1958) gave the representation, and Fishburn (1970) and Tversky (1969) the extensions and the evidence. Here the threshold is the budget and the utility is the distance to the ideal, so both are fixed by the evaluation object. A semiorder with intransitive indifference is what a finite-resolution evaluator must exhibit, and its threshold is a property of the evaluator rather than a fitted constant. Choice from a menu is the set of admissible actions no admissible action strictly beats, nonempty on a finite menu because a semiorder is acyclic, and under a length budget every admissible action within $\varepsilon$ of the least distance is chosen, which is satisficing (Simon, 1955) with the aspiration level as the ideal and the tolerance as the budget.

**Theorem 3** (Representation, hull law, uniqueness)**.** *Let $X$ be a point, $\mathcal{Y} = \mathbb{R}^m$, and the consequences $y_a = C(a)$ fixed by the world, with the evaluator choosing a positive semidefinite $G$ and an ideal $t$, so that $c(a) = q(y_a)$ with $q(y) = (y - t)^\top G(y - t)$.*

- *(a) A weak order on finite $A$ is representable by some $(G, t)$ if and only if there is a positive semidefinite $(m + 1) \times (m + 1)$ matrix $W$ such that $Q_W(y) = (y, 1)W(y, 1)^\top$ satisfies $Q_W(y_a) < Q_W(y_b)$ whenever $a \succ b$ and $Q_W(y_a) = Q_W(y_b)$ whenever $a \sim b$. These are linear constraints on $W$ over the semidefinite cone. The strict inequalities make the feasible set open, so the program solved in practice fixes a margin $\delta > 0$ and asks for $Q_W(y_b) - Q_W(y_a) \ge \delta$ whenever $a \succ b$, which is a closed and standard semidefinite feasibility problem, and a solution at any $\delta > 0$ scales to a solution of the strict system and conversely, since $W$ may be scaled freely.*

- *(b) If a preference is representable, then for every action $a$ and finite set $S$ with $y_a$ in the convex hull of $\{y_s : s \in S\}$ there is $s \in S$ with $a \precsim s$. No action is strictly worse than every action whose consequences surround it, and this holds for every convex evaluation cost.*

- *(c) Let $(G_1, t_1)$ and $(G_2, t_2)$ with $G_1 \ne 0$, $G_2 \ne 0$ induce the same weak order on a nonempty open connected set $U$ of consequences. Then $G_2 = \lambda G_1$ for some $\lambda > 0$. If in addition $\operatorname{rank} G_1 \ge 2$, or $U$ contains a point at which $q_1$ attains its minimum, then also $G_1(t_2 - t_1) = 0$.*

Part (b) is the first law of the theory that a scalar utility does not state, and it is checkable on data whenever the consequence map is known. Part (c) says what choices identify. The metric is identified up to scale from any open set of consequences. The ideal is identified up to the metric's kernel when the metric has rank at least two or the data reach the ideal, and not otherwise. With $\mathcal{Y} = \mathbb{R}$, $U = (1, 2)$, $q_1(x) = x^2$ and $q_2(x) = (x+10)^2$ the two orders agree on $U$ and $G_1(t_2 - t_1) = -10$. An earlier statement of this theorem omitted the rank or coverage hypothesis and was false in rank one, which an external reader found. The machine-checked version of the result is the whole-space case, which is the case the identification gate of Section C.3 uses. A rank budget also produces a reversal that no single utility records. There is an object and two actions with $b \succ a$ at rank cap 1 and $a \succ b$ at rank cap 2, with the same metric and the same ideal at both budgets (Appendix I).

### B.3 A GRID ON THE SYMBOLS IS A LENGTH BUDGET THE EVALUATOR CANNOT SEE PAST

The experiment does not hand the judge a tolerance $\varepsilon$. It places the consequences the judge reads, or the report it writes, on a grid of step $h$, and the relation this induces on the underlying distances is not the deterministic band of Definition 2. Two distances $0.6h$ apart round to the same value or to different values depending on where they sit relative to the grid. The next proposition says what a grid does to an exact evaluator, and it is the quantity the experiment measures.

**Proposition 1** (Grid channel)**.** *Let an evaluator order two options by comparing $\rho(d_1)$ and $\rho(d_2)$, where $d_1 < d_2$ are their true distances, $\rho$ is nondecreasing and constant between consecutive boundaries spaced $h$ apart, and a tie counts as an error. Rounding and truncation to a grid of step $h$ are both such a $\rho$. Write $\Delta = d_2 - d_1$ and let $\varepsilon$ be the distance from $d_1$ up to the next boundary.*

*(a) $\varepsilon \in (0, h]$, and the pair is ordered correctly if and only if $\Delta \geq \varepsilon$. This is the rule of Definition 2 applied to that one comparison with tolerance $\varepsilon$, the two differing only in the boundary case $\Delta = \varepsilon$. The tolerance depends on $d_1$, so the relation a grid induces on a set of options is not in general the fixed-threshold semiorder of Definition 2.*

*(b) If that position is uniform on a cell, $\varepsilon$ is uniform on $(0, h]$.*

*(c) Hence $P(\text{correct} \mid \Delta) = P(\varepsilon \leq \Delta) = \min(1, \Delta/h)$, and the gap at which the evaluator reaches accuracy $\alpha$ is the $\alpha$ quantile of its tolerance, $\alpha h$. If the evaluator further errs with probability $\lambda$ independently of the pair, the probability is $(1 - \lambda)\min(1, \Delta/h)$.*

*Proof.* (a) $\rho$ is nondecreasing, so $\rho(d_1) \leq \rho(d_2)$ always and the pair is ordered correctly exactly when $\rho(d_1) < \rho(d_2)$, which happens exactly when a boundary lies in $(d_1, d_2]$. The least boundary above $d_1$ is $d_1 + \varepsilon$, so that is $\Delta \geq \varepsilon$, and $\varepsilon \leq h$ because consecutive boundaries are $h$ apart. (b) The boundaries are a lattice of spacing $h$, and the distance from a point to the next lattice point above it carries the uniform law on a cell to the uniform law on $(0, h]$. (c) The first equality is (a) and the second is (b). The $\alpha$ quantile of the uniform law on $(0, h]$ is $\alpha h$, and the lapse multiplies by $1 - \lambda$. □

Part (a), and the bound $\varepsilon \in (0, h]$, are machine-checked in Lean for both quantizers. Parts (b) and (c) are not, and the proof above is the one to read for them.

*Remark.* The dependence on $d_1$ in part (a) is not slack in the statement and cannot be removed. Take the unit grid $\rho(d) = \lfloor d \rfloor$ and the three distances $0.1$, $0.9$ and $1.1$. The first two share a report and the last two do not, so a single tolerance would have to be at least $0.8$ and less than $0.2$ at once. What a grid induces on a whole set of options is the weak order of Theorem 2(a) on the quantized distance, whose indifference is transitive, and not the semiorder of Theorem 2(b) on the true distance. The semiorder with one threshold for every pair is what a length budget gives. Part (a) is the bridge between the two and it holds one comparison at a time.

Part (a) is what connects a grid to a length budget at all. One comparison at a time the grid is the rule of Definition 2 exactly, with the distance to the next boundary as its tolerance. What a grid takes away is knowledge of that alignment, and part (b) says the tolerance is then uniform on $(0, h]$. The accuracy curve of part (c) is therefore the distribution function of the tolerance, so an experiment that orders pairs identifies the whole law of the threshold rather than one number, and $h$ is its largest value. Three things follow. The gap reported at accuracy $\alpha$ is the $\alpha$ quantile of the tolerance, so the level is part of the definition of that statistic and not a property of the evaluator, which is why the step is reported beside it here. A constant threshold of Definition 2 represents the relation on a finite set of options exactly when some $\varepsilon$ is at least the largest separation between two options sharing a bin and below the smallest separation between two options in different bins. A grid does not arrange that in general. And the curve is a ramp because the alignment is unknown rather than because the evaluator is noisy. It has one parameter, the step, and no shape left to fit, which is what makes the fit of Appendix C a test of the proposition and not a description of the data. The proposition also fixes what the registered threshold statistic measures. The registration read the threshold as the gap at accuracy 0.9, interpolated linearly in log gap between ladder points. On a ladder with points at $h/2$ and $h$, where the proposition puts accuracy at $0.5$ and $1$, that interpolation returns $h \cdot 2^{-0.2} = 0.87h$. The number $0.87$ that appears throughout the record is therefore a property of the ladder and the interpolation, not of the judge, and the quantity to report is the fitted step $h$ with its interval. The proposition also says what that number is a biased estimate of. The gap at accuracy $0.9$ is $0.9h$, so interpolating on this ladder returns $0.967$ of it. The self-test on a synthetic rounding scorer had already returned $0.87h$ at every $h$, which is the same fact seen from the other side.

# C A NUMERIC TASK AS A CALIBRATION OF THE INSTRUMENT

## C.1 THE TEST IS A JUDGE WITH A KNOWN CONSEQUENCE MAP

The prediction is that for a fixed world and evaluator the indifference threshold tracks the resolution budget. Coarsening the resolution at which the evaluator resolves consequences enlarges the threshold, and the ordering of pairs separated by more than the largest threshold does not change. The test needs an evaluator whose consequence map is known and whose budget can be set from outside. A language-model judge on a numeric task is such an evaluator, and it is one the theory did not build.

**Evaluator.** Qwen2.5-7B-Instruct (Qwen Team, 2025) at a recorded revision, loaded in bfloat16, in 8-bit (Dettmers et al., 2022), and in 4-bit NF4 (Dettmers et al., 2023). The judge is a scorer. Shown, in its chat template, a target value and one value, it is asked how far the value is from the target and answers with a number, read as the first number in its greedy generation. The prompt is "Target: {target}. Value: {value}. How far is the value from the target? Answer with a single number and nothing else." The induced preference between two options compares their reported distances. Equal reports, and any unparsable report, are indifference. This is the theory's evaluator, a map from a consequence to a scalar cost whose order is the preference, and the model's own output resolution and arithmetic are its floor.

Two earlier instruments were tried on a probe seed and are kept in the record. A pairwise chooser asked which of two options is closer, read from the next-token logits of the answer letters with both orders shown. With options on either side of the target it was vacuous, choosing the larger number in every straddling pair. With options on one side it reached accuracy 0.99 at a gap of 10 but stayed between 0.43 and 0.63 from gap 0.01 to gap 2, because its first-position preference held the two-order average near one half until the gap was large. A pairwise chooser measures its position bias before it measures a threshold, so it was replaced by the scorer before any pilot.

**Consequences.** The target is 100. An option is a number whose distance from the target lies between 10 and 40, on either side with equal probability, so that options fall in $[60, 90]$ or $[110, 140]$. Its consequence is that distance, which has two digits before the decimal point so that a report of fixed length has fixed resolution. A pair is two options whose distances differ by a gap on the ladder 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, each option on either side of the target with equal probability, 200 pairs per gap drawn by seed. The options are drawn continuously, so their positions relative to any grid on the ladder are uniform, which is the hypothesis of Proposition 1.

**Budgets.** Three, set independently. Rendering precision is the first. The numbers are shown with 0, 1, 2, or 3 decimals, the target at the same precision, so the judge reads consequences rounded to a grid of step $10^{-d}$. Report length is the second. The scorer's answer is limited to $k$ generated tokens, $k$ in 1, 2, 3, 4, 5, 6, 12. This tokenizer emits one character per token for digits and the decimal point (Appendix J), so with two-digit distances a report of $k$ tokens is the true distance truncated to a grid of step 10 at $k = 1$, 1 at $k = 2$ and $k = 3$ (the third character is the decimal point), 0.1 at $k = 4$, 0.01 at $k = 5$, and 0.001 from $k = 6$. This is the length budget taken literally. The evaluator may spend $k$ symbols on its cost, and the grid is the resolution those symbols afford. Weight precision is the third, at bfloat16, 8-bit and 4-bit. The cells are rendering by weight precision at full report length, twelve cells, and the report ladder at three decimals and full weights, six more.

**Registered estimator and bars.** For each cell and gap, accuracy is the share of pairs on which the preference for the closer option strictly exceeds one half. A tie and an unparsable report are scored as failures, so indifference counts against the evaluator rather than being discarded. The registered threshold of a cell is the gap at which accuracy first reaches 0.9, interpolated linearly in log gap between ladder points, infinite if never reached and the smallest gap if reached there. The scorer's threshold in a cell is predicted to be the larger of the budget it was set and a floor, the scorer's own resolution at full weights, full rendering, and full report length. P1 says that along the rendering ladder at fixed weight precision the threshold is non-increasing in decimals within a tolerance factor, and at every decimal count whose step exceeds the floor by a factor of at least 3 it is within that factor of the step. P1b says the same along the report ladder for $k \geq 2$, with the $k = 1$ cell reported and not graded because its resolution of 10 sits at the edge of the gap ladder. P2 says that along the weight

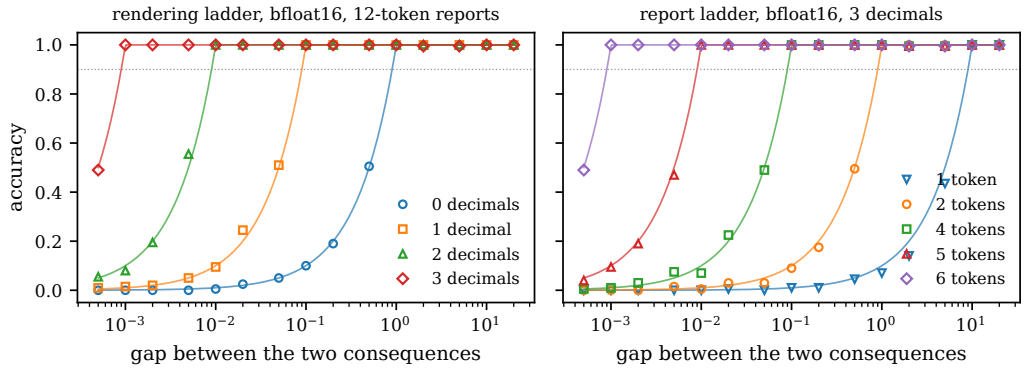



Figure 4: Accuracy against gap, bfloat16 weights, with the grid-channel law $\min(1, \Delta/h)$ drawn through each cell at the step that was set. Left, the rendering ladder at full report length. Right, the report ladder at three decimals. Every point sits on its line. The accuracy at a gap is the gap divided by the grid step, and the line moves by one decade with each step of the budget. The dotted line is the 0.9 level of the registered threshold.

ladder at three decimals and full report length the threshold is non-decreasing as bits are removed, within the factor. P3 says that for every gap at least twice the largest threshold measured in any cell, accuracy is at least 0.95 in every cell. An anti-vacuity rule required accuracy of at least 0.95 at gap 20 and a floor of at most 0.3 at full precision. The tolerance factor was fixed at 1.5 from a pilot on a separate seed. The registration records that at full weights the rendering and report ladders test only that the scorer's arithmetic resolves what it is shown and allowed to say, which a competent model makes nearly a tautology, and that the weight ladder is where the prediction can fail.

**Channel fit.** Proposition 1 was not part of the registration, and the fit that follows is an analysis of the sealed record rather than a registered test. The two parameters are fitted in a fixed order, the lapse first and the step second. The lapse $\hat{\lambda}$ is one minus the mean accuracy at gaps of at least twice the set step, where the law is saturated and any shortfall is lapse. Accuracy is then divided by $1 - \hat{\lambda}$ to remove it, and the step is fitted to the corrected accuracies $y_i = a_i/(1 - \hat{\lambda})$ at the ladder gaps $x_i$ below 0.9 by least squares through the origin. Under the law of Proposition 1 the corrected accuracy is $x/h$, so the fitted slope estimates $1/h$ and the step is its reciprocal, $\hat{h} = \sum x_i^2 / \sum x_i y_i$. Reading $\hat{h}$ as the slope itself would invert the quantity, and the values in Table 3 are the reciprocal. Intervals are a parametric bootstrap that redraws each gap's 200 outcomes as binomial and refits, 4000 draws. A paired bootstrap across cells would be preferable and the record does not support it, since it holds per-gap aggregates and not per-pair outcomes.

**Discipline.** The registration, with its hypothesis, world, statistic, bars, exclusions and null, was committed and its hash recorded before the run seed was drawn. The self-test, the probes, and the pilot were run and recorded before sealing, and the run was graded by a script against the sealed bars.

### C.2 Results

The sealed verdict is a pass on P1, P1b, P2 and P3. Figure 4 shows the accuracy curves, Table 3 the fitted steps, and Figure 5 the registered thresholds. No report was unparsable in any cell. The registered thresholds, at bfloat16, were 0.869, 0.0868, 0.00856, and 0.00087 at 0 to 3 decimals and 8.85, 0.872, 0.872, 0.0873, 0.00877, and 0.00087 at 1 to 6 tokens, each with a bootstrap interval of width below three percent of its value, and the 8-bit and 4-bit thresholds equal the bfloat16 values to the digits shown except at three decimals under 4-bit weights, 0.00093, a ratio of 1.06 inside the registered tolerance of 1.5.

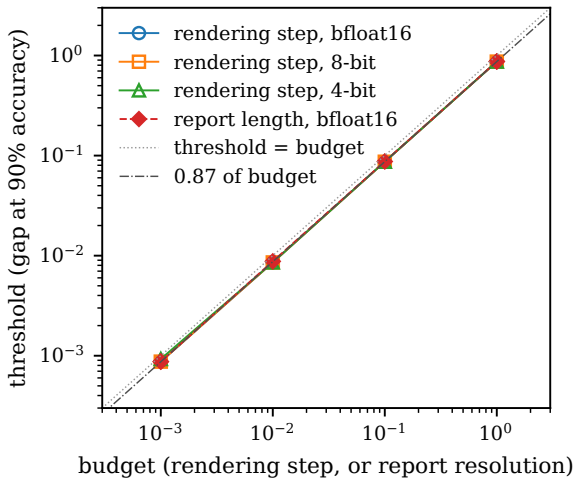



Figure 5: The registered threshold against the budget that was set, log axes. Rendering cells at three weight precisions, and the report ladder. Every point lies on the line at 0.87 of the budget, which is where the registered estimator, log-gap interpolation to the 0.9 level on this ladder, places the threshold of a grid channel. The three weight precisions coincide. The fitted steps in Table 3 are the quantity to read, and this figure shows that the sealed statistic behaved as the proposition says it must.

Table 3: The step fitted from each accuracy curve against the step that was set, with 95 percent bootstrap intervals, and the lapse above resolution. Rendering cells at 12-token reports, report cells at 3 decimals and bfloat16. Every graded interval contains the set step. The one-token cell is reported and not graded. An interval is given for the lapse only where it is not pinned at zero, and in every other cell the point estimate is zero and the bootstrap does not move it. The registered threshold is not a column here, since it is given in the text and plotted in Figure 5.

| Cell | weights | set $h$ | $\hat{h}$ | interval | $\hat{h}/h$ | lapse $\hat{\lambda}$ |
|---|---|---|---|---|---|---|
| 0 decimals | bf16, 8-bit, 4-bit | 1 | 0.998 | [0.89, 1.14] | 1.00 | 0.000 |
| 1 decimal | bf16, 8-bit, 4-bit | 0.1 | 0.0957 | [0.086, 0.109] | 0.96 | 0.000 |
| 2 decimals | bf16, 8-bit, 4-bit | 0.01 | 0.00924 | [0.0083, 0.0104] | 0.92 | 0.000 |
| 3 decimals | bf16 | 0.001 | 0.00102 | [0.00089, 0.00119] | 1.02 | 0.001 |
| 3 decimals | 8-bit | 0.001 | 0.00102 | [0.00089, 0.00118] | 1.02 | 0.001 |
| 3 decimals | 4-bit | 0.001 | 0.00099 | [0.00087, 0.00115] | 0.99 | 0.031 [0.024, 0.038] |
| 1 token | bf16 | 10 | 11.9 | [10.4, 13.8] | 1.19 | 0.000 |
| 2 tokens | bf16 | 1 | 1.03 | [0.92, 1.17] | 1.03 | 0.000 |
| 3 tokens | bf16 | 1 | 1.03 | [0.92, 1.17] | 1.03 | 0.002 |
| 4 tokens | bf16 | 0.1 | 0.100 | [0.090, 0.113] | 1.00 | 0.001 |
| 5 tokens | bf16 | 0.01 | 0.0106 | [0.0094, 0.0122] | 1.06 | 0.001 |
| 6 tokens | bf16 | 0.001 | 0.00102 | [0.00089, 0.00119] | 1.02 | 0.001 |

**The accuracy curve is the grid-channel law.** In every cell the accuracy at a gap below the set step equals the gap divided by the step, and above the step it is one, to within the binomial noise of 200 pairs. The largest deviation from $\min(1, \Delta/h)$ at any ladder point in any graded cell is 0.055, at a point where the law predicts 0.5 and the standard error is 0.035. The step fitted from the curve is within eight percent of the step that was set in every graded cell, and every interval contains the set step. That is the content of Figure 4: the curves are one line, $\Delta/h$, drawn at four values of $h$ on the left and five on the right.

**Rendering and report length are the same budget.** A grid on what the judge reads and a grid on what it writes produce the same accuracy curve at the same step. At three decimals and two tokens the fitted step is 1.03 against 1, at one decimal and full report it is 0.096 against 0.1, and at

two decimals and five tokens the two fitted steps are 0.0092 and 0.0106 against 0.01. Two and three tokens give the same step because the third character is the decimal point, and six and twelve tokens give the same step because the three-decimal rendering is the floor. This is the length budget of Definition 1 on both sides of the evaluator. The symbols the judge is given and the symbols it may spend set its resolution, and nothing between them does.

**Weight precision changes reliability and not resolution.** At 8-bit weights every accuracy curve equals its bfloat16 curve to the digits shown. At 4-bit weights the curves at 0, 1, and 2 decimals equal the bfloat16 curves, and at three decimals the fitted step is 0.00099 against 0.00102 at bfloat16, an interval of $[0.00087, 0.00115]$ containing the set step, while the accuracy above resolution falls from the 0.99 to 1.00 range to a plateau of 0.95 to 0.98. The two-parameter law reads this as a step unchanged and a lapse of 0.031, interval $[0.024, 0.038]$, where the bfloat16 and 8-bit judges have a lapse of 0.001. The registration named the weight ladder as the place where the threshold prediction could fail, and it did not. What the weights did move is a second property, the judge's reliability once a difference is resolved. On this task, then, a judge has two numbers and not one. It has a resolution, set by the symbol grid on either side of it, and a reliability above that resolution, which its weight precision can lower without moving the resolution.

**Well-separated pairs keep their order, on one gap.** The largest registered threshold in any cell is 8.85, in the ungraded one-token cell, so the only ladder gap at least twice that is 20, and P3 as registered is a single-gap test. At gap 20 the worst accuracy over all eighteen cells is 0.985. A budget change enlarged the threshold by four orders of magnitude and reordered nothing at that gap. A registration with gaps beyond 20, and with P3 defined on the largest graded threshold, is the natural next gate and is listed in Section 7.

### C.3 Identification on synthetic evaluators

Theorem 3(c) says what choices can recover, and the gate that tests it uses the whole-space case that is machine-checked. Forty evaluators per cell, in two to four dimensions with random metrics including singular ones, were given batteries of consequences of increasing size, and the feasibility program of Theorem 3(a) with a margin was solved from their weak orders alone. Errors were scored against the chance level of guesses from the evaluator prior. Theorem 3(c) fixes the metric only up to a positive scale and the ideal only up to the metric's kernel, so recovery is scored by quantities invariant to both. The metric is scored by the chordal distance between the recovered and true forms after normalising each to unit Frobenius norm, and the ideal by the distance between its recovered and true components in the range of the metric, divided by the diameter of the consequence battery. Chance is the same statistic under a metric and ideal drawn at random from the same generator. In all twelve cells on a fresh seed, under a sealed registration, the metric and the ideal's range component were recovered at 64 consequences above the dimension, to a few percent of chance, with the error falling at every step as the battery grew, while at or below the design bound, off the span of a battery confined to a subspace, and in the kernel of a singular metric, nothing was recovered. A semiorder that withholds the fifth of pairs closest in distance identified more slowly, to between 5 and 25 percent of chance at the same battery. Further sealed gates, a hull law on a public survey that held as a population tendency, the same law on natural decisions where the violation rate fell with the rank at which consequences were read, as registered, and a shared-representation bound on attention heads that returned indeterminate, are reported in Appendix K at the size of their verdicts.

## D Where the predictive power enters

The comparison the registration graded sets one predictor built from a judge's calibration block against one rival that assumes the judge uses every level of the nominal scale. That rival knows nothing and the calibrated predictor knows the judge's whole conditional distribution, so the win says little about which part of that knowledge did the work. This appendix puts five predictors of the same quantity on one ladder and scores them all against the same observed test curve. It was run after both gates were graded, it is an analysis of sealed records and not a registered test, and it changes no verdict.

Four of the five form a chain in which each knows more than the one before. The fifth, a rate in bits, is the theory's own budget form and is not a superset of the count, so it sits beside rung 2 rather than above it. Each predicts the accuracy curve of the greedy read-out, so the comparison is like for like. Rung 1 is the capacity, every level of the nominal scale, evenly spaced. Rung 2 knows only how many distinct scores the judge wrote, assumed evenly spaced, which is a budget as a count of symbols. Rung 2b knows the mutual information in bits between the judge's score and the error count instead, read as $2^I$ evenly spaced levels. This is a proxy for the theory's rate budget and not that budget, which carries a workload and a distortion that this rung has neither of, and $2^I$ evenly spaced levels summarises the measured channel rather than naming a code chosen to minimise a distortion. Rung 3 knows how many scores and where they sit, mapping quality to the nearest of the judge's own scores, still without noise. Rung 4 is the calibrated channel of the registration. The statistic is the squared error against the observed curve, summed over the gap ladder, which is the statistic the registration already uses for its own rival.

Table 4: Squared error of each predictor against the same observed greedy-accuracy curve, over the graded configurations of both gates. Lower is better. Reading left to right within a row is the price of knowing less than the channel. The column order is not a ranking. By median squared error the positions are the best of the three budget-only predictors and the rate is second, and by the share of the distance to the channel that each closes in the median cell the order of those two is reversed, which the text gives in full. The channel is better than all four in every one of the eighteen cells.

| | | levels | | | squared error | | | | |
|---|---|---|---|---|---|---|---|---|---|
| judge | scale | scale | used | $2^I$ | capacity | count | bits | positions | channel |
| qwen14b__full | 0-100 | 101 | 13 | 4 | 0.491 | 0.239 | 0.302 | 0.170 | 0.0056 |
| qwen14b__full | 0-9 | 10 | 9 | 3 | 0.258 | 0.188 | 0.540 | 0.164 | 0.0029 |
| qwen14b__full | 1-5 | 5 | 5 | 3 | 0.286 | 0.286 | 0.097 | 0.286 | 0.0145 |
| qwen7b__full | 0-100 | 101 | 9 | 3 | 1.169 | 0.547 | 0.174 | 0.282 | 0.0014 |
| qwen7b__full | 0-9 | 10 | 8 | 3 | 0.521 | 0.366 | 0.252 | 0.323 | 0.0023 |
| qwen7b__full | 1-5 | 5 | 5 | 2 | 0.449 | 0.449 | 0.227 | 0.449 | 0.0054 |
| qwen7b__int4 | 0-100 | 101 | 9 | 2 | 1.450 | 0.776 | 0.527 | 0.455 | 0.0014 |
| qwen7b__int4 | 0-9 | 10 | 8 | 3 | 0.472 | 0.333 | 0.283 | 0.163 | 0.0046 |
| qwen7b__int4 | 1-5 | 5 | 5 | 2 | 0.318 | 0.318 | 0.330 | 0.318 | 0.0022 |
| gemma12b | 0-100 | 101 | 15 | 4 | 0.468 | 0.270 | 0.318 | 0.077 | 0.0064 |
| gemma12b | 0-9 | 10 | 9 | 3 | 0.393 | 0.310 | 0.419 | 0.310 | 0.0102 |
| gemma12b | 1-5 | 5 | 5 | 2 | 0.489 | 0.489 | 0.210 | 0.489 | 0.0071 |
| gemma31b | 0-100 | 101 | 21 | 10 | 0.061 | 0.061 | 0.094 | 0.061 | 0.0001 |
| gemma31b | 0-9 | 10 | 10 | 7 | 0.051 | 0.051 | 0.035 | 0.051 | 0.0006 |
| gemma31b | 1-5 | 5 | 5 | 4 | 0.021 | 0.021 | 0.014 | 0.021 | 0.0006 |
| qwen3_27b | 0-100 | 101 | 17 | 5 | 0.444 | 0.270 | 0.161 | 0.270 | 0.0032 |
| qwen3_27b | 0-9 | 10 | 10 | 4 | 0.133 | 0.133 | 0.340 | 0.133 | 0.0020 |
| qwen3_27b | 1-5 | 5 | 5 | 3 | 0.059 | 0.059 | 0.160 | 0.059 | 0.0017 |
| median | | | | | 0.419 | 0.278 | 0.240 | 0.220 | 0.0026 |

Two orderings of the three budget-only rungs come out of this table and they do not agree, so both are given. By median squared error the positions are best at 0.220, the information rate next at 0.240 and the count last at 0.278, against 0.419 for the capacity and 0.0026 for the channel. By the share of the distance from the capacity to the channel that a rung closes in the median cell, the information rate is best at 39 percent and the count and the positions are level at 22 percent. The two disagree because a median of ratios is not a ratio of medians, and the share is computed only on the 17 cells of 18 in which the capacity and the channel differ by more than 0.05 in squared error, a rule the script fixes and which drops a judge the capacity already predicts well. The information rate beats the capacity in 12 of 18 cells and the channel is the best of the five in all 18. What none of the three budget-only rungs does is close most of the distance, and they are not nested, so what is left over cannot be divided between the noise of the channel and the placement of its scores. It needs the conditional distribution itself.

The conclusion is the one stated in Section 2. Resolution is budgeted, and a scalar budget accounts for a substantial minority of it. The quantity that predicts a judge's held-out discrimination is its measured channel, of which a budget is one summary.

## E DOES A FINER LADDER REMOVE THE TIES?

The read-out ordering failed on one judge because seven of its eighteen read-outs had a predicted threshold at one wrong answer in twenty, the smallest gap the registered ladder contains, and read-outs tied at the floor cannot be ranked. The obvious repair is a finer ladder, which on this stimulus means a longer worksheet, since one wrong answer in a hundred is a quality gap of 0.01 rather than 0.05. This appendix reports a probe of that repair, run after the gates were graded, on 180 worksheets of 100 items drawn from a probe seed, with error counts in $[0, 20]$ so that the calibration levels stay contiguous, and 30 pairs at each gap. Thirty pairs put a binomial standard error of about 0.08 on every accuracy below, which is why the reading here is coarse and the numbers are not a measurement of any threshold. It is exploratory and grades nothing.

The table gives the share of pairs one item apart that each read-out orders correctly. Below 0.75 the threshold would be measured rather than censored.

Table 5: Accuracy at the finest gap a 100-item worksheet allows, one wrong answer in a hundred. A longer worksheet uncensors most read-outs but not all. The expected score of the finer judge is still past the crossing at the smallest gap this stimulus can express.

| judge | scale | greedy | expected | mean of 8 | scores used |
|---|---|---|---|---|---|
| Gemma 4 31B | 0–100 | 0.700 | **0.900** | **0.833** | 23 |
| Gemma 4 31B | 0–9 | 0.033 | 0.700 | 0.367 | 3 |
| Gemma 4 31B | 1–5 | 0.067 | **0.800** | 0.167 | 3 |
| Gemma 4 12B | 0–100 | 0.400 | 0.667 | 0.633 | 14 |
| Gemma 4 12B | 0–9 | 0.167 | 0.533 | 0.400 | 4 |
| Gemma 4 12B | 1–5 | 0.133 | 0.500 | 0.267 | 4 |

Two things follow, and only one of them is the hoped-for answer. The repair mostly works. Every read-out of the 12B judge becomes measurable, and the greedy read-out of the 31B judge does too, at 0.700. On the evidence here a ladder built on 100-item worksheets would cut that judge's censored read-outs from seven of eighteen to about three. It does not work completely. The expected score of the 31B judge is at 0.900 on the 0 to 100 scale and 0.800 on the 1 to 5 scale, so its threshold lies below one part in a hundred of quality, and measuring it would need worksheets of several hundred items, at which point the stimulus no longer resembles anything a judge would be asked to grade and the result would be about that oddity.

A second effect appears that the repair did not anticipate, and it is the theory's own claim turned back on the experiment. Narrowing the quality band to the top fifth collapses the codebook on the coarse scales. The 31B judge uses 3 distinct scores of the 5 and 10 offered, so its greedy read-out orders one-item pairs correctly 3 to 7 percent of the time, which is not error but ties. A judge spends its codebook across the range of quality it expects, not across the slice it is shown, so resolution in quality units is not bought simply by showing it finer differences.

The statement that survives is therefore narrower than the repair promised. A finer ladder would remove most of the ties and turn most of the failure into a measurement, and a registered gate on 100-item worksheets is the natural next step. It would not remove all of them, and on this stimulus family the finest read-out of the finest judge stays out of reach.

## F A SECOND STIMULUS FAMILY, WHERE THE JUDGE IS NOT EXACT

Every gate above grades one family of stimuli. Worksheets were chosen because their quality scale is exact, and that choice has a cost no number of further judges pays off, since the judge does the arithmetic exactly and the grid on what it reads and writes is the whole story. This appendix reports a registration that changes the stimuli and nothing else.

**The stimuli.** A passage of twenty numbered facts and a summary restating them, of which $e$ are unsupported. Quality is $1 - e/20$, the scale the worksheets used, so the gap ladder and every estimator carry over unchanged. Neither text is written. Both are rendered from one record of twenty facts, each a template with one slot value, and an unsupported statement is one whose slot

was replaced by a different value of the same kind. A statement is therefore unsupported exactly when it was perturbed, and the count is known without an annotator. Two conditions make that unambiguous and both are checked for every stimulus rather than assumed. The replacement differs from the truth, and it appears nowhere in the passage. The second is checked against the rendered text, which the self-test showed is not the same as checking the values, since a wrong metal can sit inside another word of the passage.

**What did not change.** The blocks, the three scales, the six read-outs, the estimator, the grading script and the bars are those of the worksheet gates, taken by reference. The bars matter most. They were fixed by a null simulation on a judge of a third family against worksheets, and they were not refitted here. A bar moved for a new task is not a bar.

**Order.** The registration was sealed and published with its hash while no test seed existed. Each judge's calibration block was then run, its predictions written from that block alone and published with their sha256. Only then was the test seed drawn from a system random source, and the graded record pins each judge's predictions by hash at the moment of drawing. The supplied commit graph records that order, since the prediction commits are ancestors of the commit that draws the seed, and the sequence is checkable against the public repository once the authors are known.

Table 6: The second family, sealed. Bars are the worksheet gates', unchanged. Every threshold of both judges landed within the registered factor. The rightmost column is the squared error of the judge's measured channel against its observed curve, and of a rival reading the nominal scale at face value, ranged over the three scales. The channel won all six judge and scale cells.

| judge | deviation (bar 0.14) | noise units (bar 4.55) | Kendall's $\tau$ (bar 0.86) | at the floor | squared error | |
|---|---|---|---|---|---|---|
| | | | | | channel | nominal |
| Gemma 4 31B | 0.066 | 2.97 | 0.869 | 2 of 18 | 0.0007 to 0.0038 | 0.15 to 0.85 |
| Gemma 4 12B | 0.124 | 3.55 | 0.908 | 0 of 18 | 0.0031 to 0.0212 | 0.14 to 0.32 |

**The read-out ordering that failed on worksheets passes here.** Appendix E asked whether a finer ladder would remove the ties that made the 31B judge fail that claim, and answered that it mostly would. A harder task does the same work by another route. Seven of that judge's eighteen read-outs sat at the smallest gap the worksheets could express. On summaries two do, and Kendall's $\tau$ rises from 0.804 to 0.869. The judge resolves summaries less finely than arithmetic, so its finest read-outs sit above the floor instead of piling onto it. That mechanism was read off the calibration blocks and recorded before the test seed was drawn, so the repair was predicted and not observed after the fact.

**What this does not establish.** Two judges, one vendor. This gate adds a task and not breadth, and the two judges are two of the six already graded on worksheets. Two of its margins are thin. The 31B judge clears the rank bar by 0.009, so one read-out ordering differently would fail the claim, and the 12B judge's largest deviation is 0.124 against a bar of 0.14 where the worksheet judges sat at 0.047 and 0.066. The prediction held on the harder family and held less comfortably. Two judges of the registration were not graded, one dropped before sealing and one withdrawn after it, both because the serving latency of a shared research API made their blocks unaffordable rather than for anything measured about them, and the registration records both with the measurements. And the family is still not prose.

## G A THIRD FAMILY, WHERE THE LABEL IS A HUMAN JUDGEMENT, AT THREE INPUT BUDGETS

The two families above were both written by us, so a reader can say the method has only been shown where quality is a thing the experimenter built. This registration scores book reviews written by readers, and the label that stands in for quality is the star rating the review's own writer gave the book, that writer's ordered evaluation of the book and not any merit of the review. That label is a human judgement, rendered in the writer's own prose, and it is still exactly known, because it is what the person did rather than an estimate of what a panel would say. There is no annotator

disagreement because there is no second annotator. Asking a model what a reader thought is also a deployed use of a judge rather than a proxy for one.

**The stimuli.** One review of between 400 and 3000 characters, drawn from a balanced sample of 200,000 Goodreads reviews with 40,000 at each star rating. The label enters as $1 - e/4$ for $e = 5 -$ rating, so the scale has five levels and the gap ladder runs from one to four. A text appearing under two ratings has no label and is dropped, a text repeated under one rating is kept once, and each level is split in two by position so that no review can sit in both a calibration and a test block. The median review is 786 characters long and two in a thousand are 400 or shorter.

**A budget on the input.** No earlier gate limited what the judge could read. Here each review is shown to the judge in three cells, its first 200 characters, its first 400, and the whole of it, cut back to a word boundary and never cutting the label. One calibration seed and one test seed serve all three cells, so the cells score the same reviews and the same test pairs at different lengths and differ in nothing else. The theory's length budget says resolution coarsens as that budget shrinks, and the registration states it as a claim to be graded on held-out data. On every scale the observed greedy threshold at 200 characters must exceed the one at full length, and the thresholds predicted from calibration must fall along the ladder. A fourth cell at 100 characters was piloted and dropped before sealing, because the judge there orders reviews four levels apart only 0.79 of the time, below the a priori bar of 0.9, and the registered rule would have declared it vacuous from its calibration block.

**What did not change, and the one bar that did.** The three scales, the six read-outs, the pairwise elicitation, the estimator, the grading script and five of the six bars are the worksheet gates', taken by reference and not refitted. The sixth is the information bar for anti-vacuity, which the first registration set at one bit after reading its own pilot and disclosed. A five-level scale can carry at most $\log_2 5$ bits about quality where the worksheet scale could carry $\log_2 21$, so the bar here is the same fraction of capacity, 0.53 bits, rounded down to 0.5. It was set before sealing and after the pilot, as the original was, and on every pilot cell it agrees with the a priori accuracy bar, which one bit would not have done, refusing two cells in which the judge orders the extremes at 0.90 to 0.95.

**What the design-matched null decided before the seal.** A null simulation for this design, by the procedure and code that fixed the worksheet bars, was run on each pilot cell and reported beside the transferred bars without grading anything. It found two things. A faithful judge fails the transferred rank bar of 0.86 in 31 to 96 percent of replicates depending on the cell, because eighteen read-outs on a five-level scale sit within a level of each other and cannot be ranked, so the read-out ordering claim is reported in every cell and graded in none. And the transferred bar of 4.55 noise units is stricter here than in the gate that set it, with a faithful judge exceeding it in 3.8 to 5.8 percent of replicates per cell against 1.4 there, so across two judges and three cells a pair of faithful judges fails the prediction claim somewhere in about one run in four. The bars were not moved. The record reports each cell against both the transferred bar and the design-matched quantile, so a miss inside the latter is reported as a miss against the worksheet bar and nothing more.

**Order.** The registration was sealed and published with its hash while no test seed existed. The six calibration blocks were run, one per judge and cell, each cell's predictions written from its own calibration block alone and published with their sha256, and only then was one test seed drawn from a system random source for every cell at once.

**The prediction held in every graded cell.** Five cells were graded and the calibration block predicted the test block in each, with deviations from 0.035 to 0.084 against the bar of 0.14 and noise units from 2.45 to 3.08 against 4.55. The design-matched null had put the transferred noise-unit bar at a false-alarm rate of 3.8 to 5.8 percent per cell, so the margins here are real and not the null's. The measured channel beat the nominal scale in every cell, squared errors of 0.0006 to 0.0092 against 0.05 to 0.33, and the pairwise elicitation read at the floor of the pointwise read-outs in every cell. The read-out ordering statistic is reported and was not graded, as the registration decided from the null before sealing, and its values show why. Kendall's $\tau$ runs from 0.64 to 0.90 with up to ten of eighteen read-outs at the floor of a four-step ladder, which cannot rank them. Gemma 4 12B at 200 characters is the one vacuous cell. Its greedy read-out orders reviews four levels apart 0.76 to 0.80

Table 7: The third family, sealed, test seed 2660660289 for every cell. Each cell is graded on its own against the worksheet bars, 0.14 and 4.55, and holds on every graded cell. The rank column is reported and not graded, for the reason given in the text. The last column is the greedy threshold in levels of the five-level scale, predicted from that cell's calibration block and observed on its test block, on the three rating scales. † marks the one cell declared vacuous from its calibration block by the registered rule, reported and not graded.

| judge | chars read | deviation (bar 0.14) | noise (bar 4.55) | at the floor | $\tau$ (reported) | greedy threshold, predicted / observed 1 to 5 | 0 to 9 | 0 to 100 |
|---|---|---|---|---|---|---|---|---|
| Gemma 4 31B | 200 | 0.083 | 2.81 | 0 of 18 | 0.752 | 1.84 / 1.82 | 1.68 / 1.82 | 1.60 / 1.65 |
| | 400 | 0.059 | 2.72 | 4 of 18 | 0.895 | 1.58 / 1.60 | 1.36 / 1.48 | 1.21 / 1.25 |
| | all | 0.036 | 2.45 | 10 of 18 | 0.641 | 1.35 / 1.43 | 1.13 / 1.27 | 1.00 / 1.00 |
| Gemma 4 12B | 200† | 0.084 | 2.78 | 0 of 14 | 0.791 | 2.92 / 3.71 | – | – |
| | 400 | 0.079 | 2.83 | 1 of 18 | 0.758 | 1.94 / 2.22 | 1.94 / 2.35 | 1.59 / 2.09 |
| | all | 0.035 | 3.08 | 8 of 18 | 0.706 | 1.48 / 1.55 | 1.25 / 1.22 | 1.00 / 1.00 |

of the time against the a priori 0.9, decided from its calibration block before the seed was drawn, and its curves are reported rather than graded, so the budget claim below rests on one judge.

**Resolution is ordered by the input budget.** On Gemma 4 31B, the one judge graded at all three budgets, the greedy threshold observed at 200 characters exceeds the one at full length on every scale, 1.82 against 1.43 on the 1 to 5 scale, 1.82 against 1.27 on 0 to 9, and 1.65 against 1.00 on 0 to 100, and the thresholds predicted from each budget's own calibration fall along the whole ladder on every scale, 1.84, 1.58, 1.35 and 1.68, 1.36, 1.13 and 1.60, 1.21, 1.00 at 200, 400 and all characters. Both halves are the registered claim and both held. The same shape appears on Gemma 4 12B on the two cells where it is graded, 2.22, 2.35 and 2.09 at 400 characters against 1.55, 1.22 and 1.00 at full length, and is not graded because the 200-character cell is vacuous. The theory's length budget, applied to prose that a person wrote, coarsens the judge's resolution in the order it predicts, and the amount by which it coarsens was predicted from the calibration block at each budget before the test pairs existed. The expected-score read-out sits at the ladder floor at every budget on the 31B judge except 200 characters, where it predicts 1.15 and observes 1.00, so the budget's effect is visible on the greedy read-out and censored on the finer ones, which is the same instrument limit met on worksheets from the other side.

**What this does not establish.** Two judges, one vendor, and one register of English prose about books. The label is one reader's rating, which is exact and idiosyncratic, and nothing here says a panel would agree with it. A judge that recognises a famous book could recall its reception rather than read the review, which cannot supply this reader's rating but could pull a judge toward the crowd. Three rungs can show that resolution is ordered by the input budget and cannot show the shape of the dependence. The budget claim is graded on one judge, since the other is vacuous at the shortest budget. The read-out ordering claim was not gradable on this family, and a family with more levels would be needed to grade it.

## H VERIFICATION OF THE EXACT DISTRIBUTIONS

The read-outs that use samples do not generate them. They draw from a score distribution computed exactly, which removes one full pass of the prompt per sample. This appendix reports the checks that the computed distribution is the one the judge samples from. All were run before the registration was sealed, on the 7B judge at bfloat16 unless stated, and the records are in the supplementary material.

**Single-token scales.** On the 1 to 5 and 0 to 9 scales a score is one token, and the distribution is the first-token distribution restricted to the codebook. On a block of 42 worksheets no report was unparsable on any scale and all first-token mass fell on the codebook. Real samples at temperature one were compared with the recorded distribution on six worksheets of the 0 to 9 scale, at 0, 4, 8, 12, 16 and 20 errors, with 256 draws each. The statistic is the total variation distance, and the bar is the 95th percentile of that distance under multinomial sampling from the recorded distribution. Five worksheets fell inside the bar. The worksheet with 16 errors did not, at 0.105 against 0.073. Two

follow-ups placed the miss as chance. The distribution did not depend on batch shape. Computed alone, in 32 copies, and inside a padded batch of mixed prompts, it was identical, which rules out bfloat16 batch numerics. Redrawn with 2,048 samples the same worksheet sat at 0.015 against a bar of 0.026 ($p = 0.36$), and the worksheet with 12 errors at 0.014 against 0.027 ($p = 0.59$).

**The digit tree.** On the 0 to 100 scale a score takes up to three digit tokens. One pass gives the first digit. Every digit prefix with probability above $10^{-5}$ is extended by one cached single-token step that gives each next digit and the probability of stopping, with the model's repetition penalty applied as sampling applies it. The pruned mass, the mass on values outside the scale, and the mass on a first token that is not a digit are recorded for every worksheet. On six worksheets, 1,024 real samples each matched the tree's distribution within multinomial noise in all six, with $p$ from 0.21 to 0.95, no unparsable or out-of-range draw, and at most $6.2 \times 10^{-5}$ of the mass pruned. Over the 42-worksheet block the pruned mass was at most $1.2 \times 10^{-4}$, no mass fell outside the scale, and the first token was always a digit. Recomputing the tree's 37 largest probabilities by one uncached forward pass each gave relative differences of up to 24 percent. That is one bfloat16 rounding step in the logits, about 0.125 at the logit sizes seen here, between the cached decoding path that sampling uses and an uncached full pass. A judge's score distribution is therefore defined only up to one bfloat16 step in its logits, depending on the kernel path, and the tree reproduces the path that sampling takes, which is what the agreement with real samples shows. The tree scores a worksheet in 1.16 seconds, against 4.5 for generating eight samples.

**The chunked tree for the 14B judge.** The 14B judge ran out of memory at the second digit of the tree, because every live two-digit prefix carried its own copy of the prompt's cache beside 27.5 GiB of weights on a 31.7 GiB card. For that judge alone the tree runs the same single-token step on at most eight prefixes at a time and keeps the cache only of prefixes that have children. Every other judge uses the whole-frontier tree unchanged. On 22 worksheets the 7B judge's chunked and whole-frontier trees differed by at most $5.2 \times 10^{-5}$ in any probability, which is batch composition in bfloat16. The 14B judge's chunked tree matched brute force on eight scores to a relative difference of at most 0.21, the largest on a score of probability 0.07, against 0.24 for the 7B judge's tree. It scores a worksheet in 3.7 seconds at a peak of 29.5 GiB.

# I PROOFS

*Proof of Theorem 1.* (a) Equality almost everywhere is an equivalence relation. (b) The relation is equality of a function of $a$, hence an equivalence, and $C(a, x) = C(b, x)$ implies equality of the retained projections, so each $\sim_k$ class contains the $\sim$ classes of its members. (c) Reflexivity and symmetry are immediate. If three such points exist then $r_1 \approx_\varepsilon r_2$ and $r_2 \approx_\varepsilon r_3$ but not $r_1 \approx_\varepsilon r_3$. If none exist, take $a \approx_\varepsilon b \approx_\varepsilon c$ with values $r_a, r_b, r_c$. If $r_a \le r_b \le r_c$ or the reverse, the absence of such a triple gives $|r_a - r_c| \le \varepsilon$, and in every other ordering $|r_a - r_c| \le \max(|r_a - r_b|, |r_b - r_c|) \le \varepsilon$. □

*Proof of Theorem 2.* (a) A relation defined by comparing a real function is complete and transitive, and the function represents it. Uniqueness up to monotone transformation is the standard ordinal uniqueness. (b) A relation defined by $u(a) > u(b) + \varepsilon$ with fixed $\varepsilon > 0$ is irreflexive and satisfies the two semiorder axioms. If $a \succ b$ and $c \succ e$, then $u(a) > u(b) + \varepsilon$ and $u(c) > u(e) + \varepsilon$, so either $u(a) > u(e) + \varepsilon$ or $u(c) > u(b) + \varepsilon$, since if both failed we would have $u(a) \le u(e) + \varepsilon < u(c)$ and $u(c) \le u(b) + \varepsilon < u(a)$, a contradiction. If $a \succ b \succ c$ and $e$ is any action, then $u(a) > u(c) + 2\varepsilon$, so either $u(a) > u(e) + \varepsilon$ or $u(e) \ge u(a) - \varepsilon > u(c) + \varepsilon$. This is the easy direction of the Scott and Suppes representation of semiorders, and the representation is by construction. (c) The semiorder is a weak order exactly when indifference is transitive, which is Theorem 1(c), and for finite $A$ the condition holds for all sufficiently small $\varepsilon$. □

*Proof of Theorem 3.* (a) Given $(G, t)$, put $M = [\, I \quad -t \,]$, an $m \times (m+1)$ matrix, and $W = M^\top G M$. Then $W$ is positive semidefinite and $Q_W(y) = (y - t)^\top G (y - t) = q(y)$, so the order constraints hold for $W$. Conversely let $W$ be positive semidefinite with blocks $G$, $b/2$, $b^\top/2$, $\gamma$, so that $Q_W(y) = y^\top G y + b^\top y + \gamma$. Then $G$ is positive semidefinite, and $b$ lies in the range of $G$: if $Gz = 0$ then $(z, s) W (z, s)^\top = s\, b^\top z + \gamma s^2 \ge 0$ for every real $s$, which forces $b^\top z = 0$. Choose $t$

with $Gt = -b/2$. Then $Q_W(y) = (y-t)^\top G(y-t)+\gamma-t^\top Gt$ differs from $q$ by a constant, so $(G, t)$ orders $A$ as $Q_W$ does. The order constraints are linear in the entries of $W$ for fixed consequences, and the cone is convex. (b) $q$ is convex because $G$ is positive semidefinite. If $y_a = \sum_s \lambda_s y_s$ with $\lambda_s \geq 0$ summing to one, then $q(y_a) \leq \sum_s \lambda_s q(y_s) \leq \max_s q(y_s)$, so $a \precsim s^*$ for a maximizer $s^*$. The argument used only convexity. (c) Same weak order on $U$ means $q_2 = \varphi \circ q_1$ on $U$ for a strictly increasing $\varphi$ on $q_1(U)$, an interval because $U$ is connected and $q_1$ is continuous.

*Metric.* Fix $y \in U$ and a direction $v$ with $v^\top G_1 v > 0$, and write $g_i = v^\top G_i v$, $\ell_i = v^\top G_i(y - t_i)$. Along $y + sv$, $q_1 = q_1(y) + 2\ell_1 s + g_1 s^2$ and $q_2 = q_2(y) + 2\ell_2 s + g_2 s^2$, both quadratics in $s$, and $q_2 = \varphi(q_1)$ on an interval of $s$ around 0. Two quadratics in $s$ related by a strictly increasing map on an interval have the same critical point if the interval contains it, and in general the level sets of $q_1$ along the segment, which are pairs $\{s, s'\}$ symmetric about $s_1^* = -\ell_1/g_1$, must be level sets of $q_2$. If the segment contains two distinct such pairs, $q_2$ is symmetric about the same point, so $s_2^* = s_1^*$ and $\varphi$ is affine on the traced interval with slope $g_2/g_1$. When $\operatorname{rank} G_1 \geq 2$ such a segment exists through every $y \in U$: choose $v$ with $v^\top G_1 v > 0$ and $v^\top G_1(y - t_1) = 0$, which exists because a positive semidefinite form of rank at least two does not vanish on the hyperplane $\{v : v^\top G_1(y - t_1) = 0\}$, and then $s_1^* = 0$ is interior to the segment. When $\operatorname{rank} G_1 = 1$, write $G_1 = g\, ww^\top$. The level sets of $q_1$ in $U$ are pieces of the hyperplanes $w^\top y = \text{const}$, $q_2$ is constant on each, so $q_2$ restricted to $U$ is a function of $w^\top y$ alone, hence a quadratic in $w^\top y$, and its quadratic part is $g'\, ww^\top$ with $g' > 0$ because $q_2$ is nonconstant on $U$. In both cases $G_2 = \lambda G_1$ with $\lambda > 0$: in the rank-one case directly, and in the higher-rank case because $\varphi$ is affine with the same slope $\lambda$ on overlapping intervals along every chain of segments through the connected set $U$, so $q_2 = \lambda q_1 + \mu$ on the open set $U$, two polynomials that agree on an open set agree identically, and the quadratic parts match.

*Ideal.* In the case $\operatorname{rank} G_1 \geq 2$ the identity $q_2 = \lambda q_1 + \mu$ also matches linear parts, $G_2 t_2 = \lambda G_1 t_1$, which with $G_2 = \lambda G_1$ is $G_1(t_2 - t_1) = 0$. In the case that $U$ contains a minimizer $y^*$ of $q_1$, so $G_1(y^* - t_1) = 0$, every segment through $y^*$ in a direction with $v^\top G_1 v > 0$ has $s_1^* = 0$ interior, and the argument above gives $v^\top G_2(y^* - t_2) = 0$ for every such $v$, hence for a spanning set of the range of $G_1$, and with $G_2 = \lambda G_1$ this is $G_1(y^* - t_2) = 0$ and so $G_1(t_2 - t_1) = 0$. When neither hypothesis holds the ideal need not be identified, as the one-dimensional example after the theorem shows. □

*Resolution reversal.* Take $\mathcal{Y} = \mathbb{R}^2$, $G = I$, $t = 0$, $X$ a point, and a workload whose consequence second moment has top eigenvector $e_1$, for example $\operatorname{diag}(4, 1)$. Let $C(a) = (1, 0)$ and $C(b) = (0.9, 2)$. Then $d_1(a) = 1$, $d_1(b) = 0.9$, $d_2(a) = 1$, and $d_2(b) = \sqrt{0.81 + 4} = 2.19$, so $b \succ a$ at rank cap 1 and $a \succ b$ at rank cap 2. A function representing both would satisfy $u(b) > u(a)$ and $u(a) > u(b)$. □

## J TOKENIZATION AND PARSING OF REPORTS

The report-length ladder rests on the claim that the judge emits one character per generated token for the digits and the decimal point. Table 8 shows the token identifiers the recorded tokenizer revision assigns to representative reports. Each digit and the decimal point is a single token, so a report limited to $k$ tokens is the decimal expansion of the judge's answer truncated to $k$ characters. A report is parsed as the first number in the decoded string, a trailing decimal point is dropped, and a string with no number counts as unparsable and as indifference. No report was unparsable in any cell of the run. The probes at full report length, whose raw generations are in the record, show the judge answering with exactly three decimals and rounding the last, for example a true distance of 11.6876 reported as 11.688 and 10.2039 as 10.204. The record does not hold per-pair generations for the run or the count of reports that stopped before the budget, which a future gate will log under the programme's rule that every stage persists what it saw.

## K TWO FURTHER SEALED GATES

**Hull law on a public survey, pass as a tendency.** Theorem 3(b) needs only a known consequence map and observed choices, so it can be read off data nobody collected for it. In the 1972 American National Election Study respondents placed candidates and parties on seven-point issue scales and rated them on feeling thermometers, which makes the consequence of an object a point the respon-

Table 8: Token identifiers for representative reports under the recorded tokenizer revision. Every digit and the decimal point is one token, so a report of $k$ tokens is the answer truncated to $k$ characters and the token budget is a grid on the reported distance, which is what makes report length a resolution budget rather than a proxy for one.

| report | token identifiers | decoded tokens |
|---|---|---|
| `10` | 16, 15 | 1, 0 |
| `10.` | 16, 15, 13 | 1, 0, . |
| `10.1` | 16, 15, 13, 16 | 1, 0, ., 1 |
| `10.12` | 16, 15, 13, 16, 17 | 1, 0, ., 1, 2 |
| `10.123` | 16, 15, 13, 16, 17, 18 | 1, 0, ., 1, 2, 3 |
| `27.375` | 17, 22, 13, 18, 22, 20 | 2, 7, ., 3, 7, 5 |
| `0.001` | 15, 13, 15, 15, 16 | 0, ., 0, 0, 1 |

dent supplied and the evaluation a number the respondent gave. On the registered pair of scales 489 of 557 respondents had a menu in which the law could bind, and 246 of them, 50.3 percent, violated it at least once, against 66.4 percent when each respondent's ratings were shuffled, and five other pairs gave the same ordering, 54 to 63 percent observed against 67 to 74 percent shuffled. The registration recorded both halves. The law holds as a population tendency, and about half of the respondents violate it. One qualification was found after that gate was graded and is kept here rather than repaired. A pair of scales is a rank-two reading of whatever space a respondent evaluates in, and a convex combination survives any linear map, so a point inside a hull stays inside under projection while a point outside can fall inside. Reading fewer coordinates can therefore add violations and cannot remove them, and the observed rate bounds the rate in the respondent's own space from above rather than estimating it. That omitted attributes can make rational choice fail a revealed-preference test is established for characteristics models (Blow et al., 2008), and the next gate measures the size of that effect.

**The same law on natural decisions, read at increasing rank, pass.** The cap of Theorem 1(b) is a cap on the rank at which consequences are read, and this gate applies it to choices nobody elicited. In ten regular seasons of professional football play calls, 303,414 decisions, the consequence of a call is a five-coordinate vector estimated from outcomes and pooled over every team, and a menu can bind only on a plane when four calls are available. On the registered pair of coordinates 8 of 402 testable units violated the law, 2.0 percent against 22.9 percent shuffled, with none of 200 shuffles reaching the observed rate. One replication pair was violated above chance, and that gate graded indeterminate under its own sensitivity clause. A follow-on gate with thirteen calls then registered, before any rate in that menu had been computed, that the violation rate would not rise as more of the five coordinates are read. Over every graded subset the mean rate was 0.914 at two coordinates, 0.444 at three and 0.139 at four, and all 26 subsets sat below their shuffled rate, including the pair that had been violated above chance. Perturbing each consequence by its own standard error moved that pair's rate by 0.089 against a margin of 0.079 over its null, and the registered pair's by 0.006 against a margin of 0.209. Reading more of the consequence space lowers the violation rate, which is the ordering Section 2 predicts for a judge's read-outs, in a world with no language model in it. A second arm of the same registration tried to recover a metric from pooled choices and recovered none. One metric and one ideal served about eight game states while a stratum held 189, so that arm is a miss of the measurement and is reported as one.

**A third gate, indeterminate, summarized.** A third registration read the query heads of one grouped-query attention group as evaluators sharing a rank-$k$ code of a key cache, under Ky Fan's maximum principle (Fan, 1949), which gives both a zero-regret condition and a weighted bound. At ranks 2, 4 and 8 in 128 dimensions the measured attention loss was 20 to 330 times the second-order prediction, so the regret formula could not be tested and the verdict is indeterminate. The ordering claim survived on the measured losses alone, the leading eigenspace of the summed geometries beating every random code and the key-covariance code in 32 of 32 cases. The registration of that run also compared a one-key prediction to an all-keys measurement, a defect recorded rather than repaired. The loss measurements, the perturbation run that matched the second-order prediction at the median, and the full record are in the supplementary material.

## L PER-CELL ACCURACY

Table 9 gives the accuracy at every ladder gap in every cell of the run, from which the registered thresholds and the fits of Table 3 were computed.

Table 9: Accuracy by gap for all eighteen cells, the record from which every threshold and fitted step in this appendix is computed. Columns are the gap ladder and rows are weight precision, decimals and report tokens. Reading down a block, accuracy at a fixed gap falls as the budget coarsens. Reading across, each row climbs at the gap that equals its grid step, to one except where a lapse holds it below.

| cell | .0005 | .001 | .002 | .005 | .01 | .02 | .05 | .1 | .2 | .5 | 1 | 2 | 5 | 10 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bf16, 0, 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.05 | 0.10 | 0.19 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| bf16, 1, 12 | 0.01 | 0.01 | 0.02 | 0.05 | 0.10 | 0.24 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| bf16, 2, 12 | 0.06 | 0.08 | 0.20 | 0.56 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| bf16, 3, 12 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| bf16, 3, 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.04 | 0.07 | 0.14 | 0.43 | 1.00 | 1.00 |
| bf16, 3, 2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.03 | 0.03 | 0.09 | 0.17 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| bf16, 3, 3 | 0.00 | 0.00 | 0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.09 | 0.17 | 0.49 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| bf16, 3, 4 | 0.01 | 0.01 | 0.03 | 0.07 | 0.07 | 0.23 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| bf16, 3, 5 | 0.04 | 0.10 | 0.19 | 0.47 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| bf16, 3, 6 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| 8-bit, 0, 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.05 | 0.10 | 0.19 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 8-bit, 1, 12 | 0.01 | 0.01 | 0.02 | 0.05 | 0.10 | 0.24 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 8-bit, 2, 12 | 0.06 | 0.08 | 0.20 | 0.56 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 8-bit, 3, 12 | 0.49 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 |
| 4-bit, 0, 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.05 | 0.10 | 0.19 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 4-bit, 1, 12 | 0.01 | 0.01 | 0.02 | 0.05 | 0.10 | 0.24 | 0.51 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 4-bit, 2, 12 | 0.06 | 0.08 | 0.20 | 0.56 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 4-bit, 3, 12 | 0.49 | 0.95 | 0.97 | 0.97 | 0.97 | 0.95 | 0.96 | 0.96 | 0.98 | 0.96 | 0.95 | 0.98 | 0.98 | 0.97 | 0.98 |