THE ARITHMETIC ERROR BAR: A FIRST-PASSAGE LAW FOR THE NUMERICAL FRAGILITY OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A benchmark score carries an uncertainty nobody reports: change the arithmetic that produced it — a numerical precision, a machine — and the number moves. On a 1000-question set the aggregate accuracy is unchanged, 66.2% against 66.1%, while 53 answers differ and 29 items change verdict; the cancellation that keeps the total still is an accident, and on a second machine of a different architecture the same kind of flips cost a full point. We call the resulting band the *arithmetic error bar* — here ± 1.06 points — and we show that it is predictable in advance, item by item, from a single run, and that it can be paid down at a stated compute price. The mechanism is not the one usually assumed. Chaos predicts that reducing a perturbation makes the divergence arrive later; measurement shows that it removes it, and a state perturbation decays by an order of magnitude within ten tokens and never grows. Generation is a *hybrid* system — a contracting flow interrupted by a quantisation — and all of its sensitivity sits in the quantisation. The distance of a decision from its boundary is the *geometric margin* $\rho_j = (\ell_1 - \ell_j)/\left\|\mathbf{w}_1 - \mathbf{w}_j\right\|$; normalised by the state norm it defines a *threshold landscape* along a generation, and the probability that the text diverges is a first-passage problem whose only parameter is the amplitude of the numerical noise, measured rather than fitted. It predicts which answers change (AUROC 0.960 against 0.797 for entropy, on held-out questions), places all 29 observed correctness flips in the half of the items it had flagged, and predicts 43 changed answers against the 53 observed; carried unchanged to the second machine it predicts exact agreement in `float32`, confirmed on all 1500 questions, and 71 divergences in `bfloat16` against 66 observed. Run on two further models the law ranks well on one and not on the other. Rather than record this as a limitation we bound it: if the law holds, many questions are genuine coin tosses that no score can separate, so achievable ranking accuracy has a *ceiling* computable from the landscape and the amplitude alone, before any comparison of precisions is run. It depends on essentially one number, how often the arithmetic decides the answer, and against it all three models land at or near the best any score could achieve. Finally, re-running in `float32` only the tenth of a corpus the law flags recovers 72.4% of the correctness flips and halves the error bar, with the residual predicted at every budget.

1 INTRODUCTION

Anyone who has evaluated a language model more than once knows the experience. The same model is run on the same questions with the deterministic decoding that is supposed to guarantee repeatability; the machine changes, or the linear-algebra library, or the numerical precision, and a few answers come back different. Not many: one in twenty, one in thirty. Enough to suspect a bug, too few to find one.

That this happens is documented. Changing the evaluation batch size, the number of GPUs or the GPU model measurably changes generated responses (Yuan et al., 2025), and the dominant source has been traced to the lack of batch invariance in inference kernels, whose reduction order — and therefore whose rounding — depends on how many sequences happen to be processed together (He & Thinking Machines Lab, 2025). That line of work asks how to remove the noise. The question

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.

here is complementary and more basic: given that some numerical noise always remains, *which* outputs are vulnerable to it, and can we know in advance which ones?

The instinctive answer is chaos. A deep network is a nonlinear system with hundreds of layers; recurrent networks in the chaotic regime have positive Lyapunov exponents (Engelken et al., 2023) and signal propagation through depth is described as transient chaos (Poole et al., 2016); a perturbation at the level of the last bit, amplified exponentially, must eventually become macroscopic. The analogy is seductive and it is false, as a direct measurement shows.

Contributions.

- An uncertainty that benchmark reporting currently omits: the *arithmetic error bar* of a score, predictable in advance from a single run, together with the demonstration that the apparent stability of an aggregate is an accident of cancellation which a different machine does not reproduce (Section 4).
- A procedure that buys reproducibility back at a stated compute price, with the residual error bar predicted at every budget (Section 6).
- A ceiling on how well *any* score can rank vulnerable items, computable before any comparison of precisions, which turns a model-dependent failure into a statement about the regime (Section 5). To our knowledge this bound is not reported in comparisons of uncertainty or robustness measures; we argue it should be.
- The mechanism underlying all three: a measurement that rules out amplification and replaces it with a hybrid picture — contraction between tokens, a quantisation at each token, all the sensitivity in the quantisation — and a first-passage law on the resulting threshold landscape whose single parameter is measured on independent data rather than fitted (Sections 2–4).

2 NOT CHAOS: A HYBRID SYSTEM

The chaos and threshold hypotheses have incompatible empirical signatures. If the numerical fragility were driven by exponential amplification, reducing the perturbation amplitude ϵ would delay the onset of divergence to a later sequence position, scaling as $\lambda^{-1}\ln(1/\epsilon)$, but the divergence would still arrive. If instead the mechanism is the crossing of a discrete threshold, divergences should vanish entirely below a certain noise amplitude.

Injecting a controlled perturbation into the hidden state at a fixed position and sweeping ϵ confirms the second: below a specific amplitude, the generated text does not change. Propagating a state perturbation while holding the token sequence fixed shows that the flow contracts, decaying by an order of magnitude within about ten positions rather than amplifying. Appendix A gives the protocol and the numbers.

Language model generation is therefore not a chaotic process but a hybrid one (Goebel et al., 2012). The system alternates between a continuous, contractive evolution of the hidden state and a discrete quantisation step at each position — the step-function choice of the most probable token. The model's sensitivity to numerical noise is concentrated in those quantisation steps, and the steps can be quantified one by one, because the selection of a token corresponds to a polyhedral partition of the state space (Montúfar et al., 2014; Hanin & Rolnick, 2019): the winning token changes when the state crosses the hyperplane separating it from a competitor.

The scope of this negative claim is worth stating precisely, and Appendix A does so: it is the continuous flow that contracts. The hybrid system as a whole remains sensitive to its initial conditions, but that sensitivity is discontinuous — driven by threshold crossings — and leaves no well-defined Lyapunov exponent to measure.

3 THE THRESHOLD LANDSCAPE AND THE FIRST-PASSAGE LAW

Consider an autoregressive language model generating a sequence token by token. At each decoding step t, the prompt and the previously generated sequence pass through the network's layers to

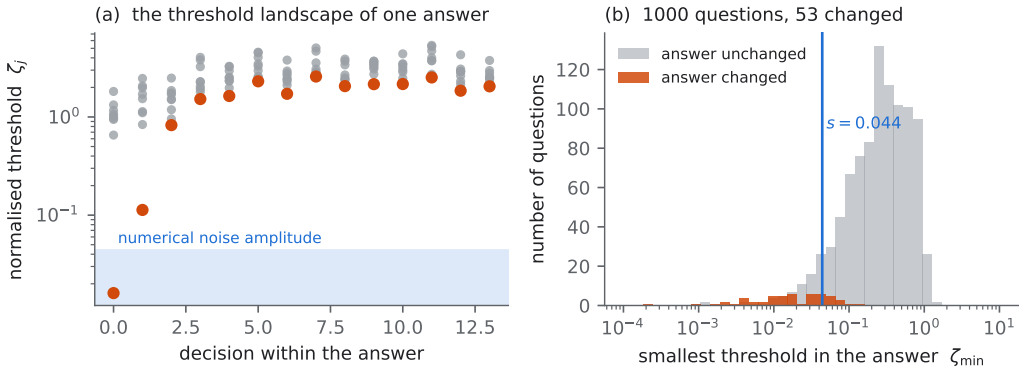

Figure 1: The threshold landscape of a generation. Most decisions sit far from their boundary; a few sit within the reach of the numerical noise, and those few decide whether the text is reproducible.

produce a final hidden state $\mathbf{h}_t \in \mathbb{R}^d$. This state is projected by an unembedding matrix W to yield a vector of logits ℓ_j, which determines the distribution over the vocabulary. Let token 1 be the most probable, that is, the winner under deterministic decoding.

For a competing token j to overtake token 1, the hidden state $\mathbf{h}_t$ must be perturbed enough to cross the decision boundary separating them. The minimum distance the state must move is

$$\rho_j(t) \;=\; \frac{\ell_1(t) - \ell_j(t)}{\|\mathbf{w}_1 - \mathbf{w}_j\|}. \tag{1}$$

This is the *geometric margin* of the decision: the classical margin of the linear classifier formed by the unembedding matrix, measured in the state space of a transformer at every decoding step.

To express the margin relative to the scale of the state, we use its normalised form

$$\zeta_j(t) \;=\; \rho_j(t)\,\frac{\sqrt{d}}{\|\mathbf{h}_t\|}. \tag{2}$$

Equation (2) states the margin in units of a relative displacement of the state: if a numerical perturbation $\delta\mathbf{h}$ with $\|\delta\mathbf{h}\| \,/\, \|\mathbf{h}_t\| = s$ is isotropic, its projection on any decision direction acts as Gaussian noise with standard deviation s, so a threshold ζ is crossed with probability $\Phi(-\zeta/s)$. The collection $\{\zeta_j(t)\}$ across all steps of a generation forms its *threshold landscape*.

Assuming these crossings are independent across competitors and steps, the probability $p_t(s)$ that a step flips and the probability $R(s)$ that the whole generation diverges are

$$p_t(s) = 1 - \prod_{j\geq 2}\Big[1 - \Phi\big(-\zeta_j(t)/s\big)\Big], \qquad R(s) = 1 - \prod_{t}\big[1 - p_t(s)\big]. \tag{3}$$

We call $R(s)$ the *fragility* of the answer. The probability that a long generation breaks exactly at step t follows a first-passage form, $P(T = t) = p_t \prod_{u<t}(1 - p_u)$, which says where a long generation breaks (Redner, 2001).

The physical regime mirrors that of an overdamped particle escaping over a potential barrier (Kramers, 1940): for $\zeta \gg s$ the Gaussian tail behaves as $\exp(-\zeta^2/2s^2)$, an Arrhenius form with barrier $\zeta^2/2$ and temperature s^2. The disorder here is *quenched* — frozen in the weights and in the prompt — so the ensemble average is taken over questions and positions rather than over time, as in the low-temperature physics of glasses (Anderson et al., 1972). Appendix B develops the analogy and states the three respects in which it is not exact.

The conceptual point of the law sits in the denominator of (1). Entropy, maximum probability and the logit gap $\ell_1 - \ell_2$ are functions of the logits alone; they discard $\|\mathbf{w}_1 - \mathbf{w}_j\|$, which encodes how similar the competing tokens are. Two tokens whose output vectors nearly coincide can swap under a minute state displacement, even when their logits are far apart.

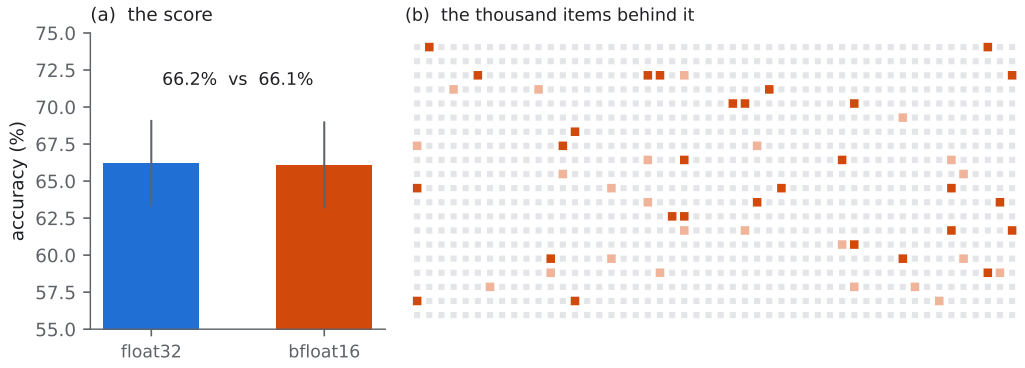

Figure 2: (a) The aggregate score does not move. (b) The thousand items that make it up: light squares are the 53 questions whose answer changes, dark squares the 29 on which correctness changes too. A reproducible score does not imply reproducible results.

4 Experiments

Setup. The reference model is Qwen2.5-0.5B-Instruct (Qwen Team, 2024), state dimension $d = 896$. The reference arithmetic is `float32` on CPU, compared against `bfloat16`. We track the nine closest competitors at each decoding step. Decoding is pure greedy with the repetition penalty disabled, which matters: with the penalty active the emitted token is not the $\arg\max$ of the logits and the geometric margin loses its meaning (Appendix E).

The task is ARC-Easy (Clark et al., 2018). Rather than prompting the model for a multiple-choice letter, we require it to write out the full text of the correct option, so that each answer spans several discrete decisions while remaining externally checkable. The 1500 questions are split by a fixed-seed permutation into 500 calibration and 1000 verification items, and three criteria with their thresholds were pre-registered before collection (Appendix C). The same protocol is then run unchanged on SmolLM2-1.7B (Allal et al., 2025) and Qwen2.5-1.5B.

4.1 Which answers change, how many, and where correctness flips

All three pre-registered criteria are confirmed. The fragility identifies which answers change with an AUROC of 0.960, against 0.729 for mean entropy and 0.797 for maximum entropy, a gap of $+0.164$ with a bootstrap interval of $[0.117, 0.215]$. All twenty-nine observed correctness flips fall, without exception, in the more fragile half of the questions. And estimating the noise amplitude on the calibration data alone predicts the scale of the churn: 54.9 changed answers against the 53 observed.

At the aggregate level the accuracy barely moves, from 66.2% in `float32` to 66.1% in `bfloat16`. Underneath that stability, 53 of 1000 answers differ and 29 items change their verdict, fourteen for the better and fifteen for the worse (Figure 2). A reader who looks only at the top-line score sees a reproducible benchmark; a reader who examines the item-by-item table does not.

The arithmetic error bar. If correctness flips are symmetric, the shift in aggregate score induced by the arithmetic has zero mean and standard deviation $\sqrt{29}/1000$, a 95% band of ± 1.06 points. We define this as the *arithmetic error bar* of a benchmark score: the uncertainty a score carries because of the hardware or the numerical precision that produced it. It is distinct from the sampling error bar, which estimates how a score generalises to unseen questions; on a fixed set of items sampling variance does not apply, and the arithmetic bar is the source of variance that remains — unreported because unsuspected. It can also be predicted without executing the alternative arithmetic: by summing the computed fragilities and rescaling by the observed fraction of changed answers that alter correctness, the law gives a band of ± 0.71 points.

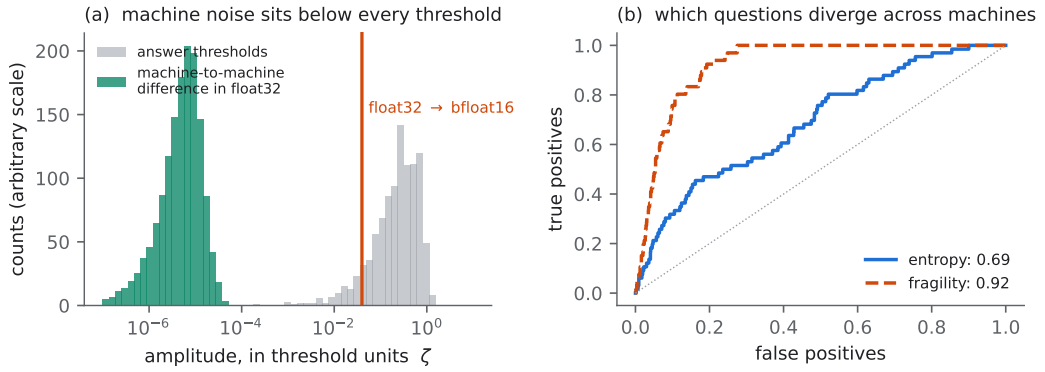

Figure 3: (a) Why `float32` is safe and `bfloat16` is not: the grey distribution is the smallest threshold of each answer, the green one the measured machine-to-machine displacement in `float32`, the orange line the amplitude of the change of precision. The machine noise lies entirely below the population of thresholds; the precision noise lies inside it. (b) In `bfloat16`, the fragility computed on one machine predicts which questions the other machine will answer differently.

4.2 The parameter is measured, not fitted

The law relies on a single parameter, and its credibility hinges on that parameter being a measurement rather than a knob. Computing the threshold landscape of the *same* generated sequence across both precisions and pairing competitors by token identity gives the dispersion of the threshold differences directly. On the reference model this yields $s = 0.044$. With that value, and nothing fitted to the outcome, the law predicts 43.4 changed answers against the 53 observed.

The measured displacement is close to Gaussian, with an excess kurtosis of 1.3; three tests exclude heavy tails, and substituting the empirical distribution for the Gaussian alters the predicted count by less than the counting noise (Appendix D).

4.3 Two machines

To test the boundaries of the framework we replicate the whole study on a second machine of a different architecture — a consumer ARM laptop against an x86 server — with identical code and question order. The contrast gives the two extremes of the same law.

In `float32` the outputs are stable: not one answer among the 1500 diverges across the two machines, and no item changes its correctness verdict, which puts a 95% upper bound of 0.20% on the divergence rate. The measured machine-to-machine displacement along the decision directions has a robust scale of 7.2×10^{-6}. Because that noise floor lies four orders of magnitude below the `bfloat16` amplitude, the law predicts fewer than 10^{-3} divergences. Zero are observed.

In `bfloat16` the two machines disagree on 66 answers and 33 correctness verdicts. Carrying over the amplitude measured for the *precision* change, $s = 0.044$, without any adjustment, the law predicts 70.6 cross-machine divergences against the 66 observed. The fragility computed on the first machine also says which questions the second will answer differently, with an AUROC of 0.922 against 0.689 for entropy.

The cancellation is an accident. On the laptop the flips cancel: of its 48 correctness flips, 25 items improve and 23 worsen, and the aggregate score does not move. On the server the balance is different — 41 flips, 13 improving and 28 worsening — and the score drops by a full point, from 66.1% to 65.1%, even though the `float32` accuracy of the two machines is identical to the last digit. The stability of an aggregate score under a change of arithmetic is therefore not a property of the phenomenon. It is what happens when local errors balance, and whether they balance is decided by the machine.

	Qwen2.5-0.5B	SmolLM2-1.7B	Qwen2.5-1.5B
state dimension d / amplitude s	896 / 0.044	2048 / 0.0185	1536 / 0.173
accuracy `float32` / `bfloat16`	66.2 / 66.1	75.3 / 75.2	86.8 / 86.2
answers that change / correctness flips	53 / 29	17 / 9	188 / 32
arithmetic error bar	± 1.06	± 0.59	± 1.11
which change: fragility / entropy	0.960 / 0.797	0.911 / 0.793	0.814 / 0.824
how many: predicted / observed	43.4 / 53	49.2 / 17	151.3 / 188

Table 1: The same protocol on three models, 1000 held-out questions each. The upper block is the phenomenon, the lower block is the law. Accuracies in per cent; *which* is AUROC.

4.4 THREE MODELS

Table 1 extends the identical protocol to two further models. The phenomenon holds on all three with the same signature: an aggregate accuracy stable to within a tenth of a point, item-level churn hidden underneath, and an arithmetic error bar on the order of a point. The ranking also transfers to SmolLM2, with an AUROC of 0.911 against 0.793 for entropy. Two things do not generalise.

The count fails on SmolLM2, for a reason outside the mechanism. There the law predicts 49.2 changed answers and 17 are observed. The discrepancy comes from where the vulnerability lies: in 77.2% of questions the thinnest threshold of the generation falls on its *last* decision, the one that terminates the answer, against 29.6% and 21.3% on the other two models. A flip at that terminal step is invisible to our measurement, because the competitors of a terminating token are other ways of ending the sequence and the first printed line is unchanged. The data confirm the blind spot: 210 questions carry a thin threshold *only* at that final decision and exactly one of them changes its observable answer, against 16 of the 120 that are thin elsewhere. Aggregating the fragility only over the decisions the observable can see brings the prediction to 15.9, against the 17 observed. We report this as the diagnosis of a failure rather than a modification of the law, because it was formulated on the data it explains and did not improve predictions on a third model where it was pre-registered.

The ranking fails on Qwen2.5-1.5B. On that model the fragility reaches an AUROC of 0.814 while entropy reaches 0.824. Three explanations can be ruled out by measurement. It is not heavy tails, nor a mis-measured amplitude — sweeping s across three orders of magnitude yields a maximum of 0.818 — nor the aggregation along the answer, since the thinnest threshold alone gives 0.813. The explanation points to a limit of the problem itself, which is the subject of the next section.

5 HOW MUCH THERE IS TO WIN

The failure on the third model prompts a question that, to our knowledge, is rarely asked: *how well could any score perform on this task?*

A ranking is evaluated by how well it separates the answers that change from those that remain stable. But if the first-passage law holds, the outcome for many questions is a genuine coin toss — the fragility is strictly between zero and one — and no score, however sophisticated, can predict the outcome of a single toss. The separability of the problem is therefore bounded, and the bound is a property of the threshold landscape and the noise amplitude rather than of the measure one chooses.

The bound is straightforward to compute. Taking the measured landscape and amplitude, we let the law itself generate the outcomes: each question changes its answer with probability equal to its own fragility. We then score those synthetic outcomes with the same fragility and measure the resulting AUROC. This is what a perfect score would achieve in a world where the law is exactly true. We call it the *ceiling*.

The ceiling collapses as crossings become common, and it is model-independent. Figure 4a sweeps the noise amplitude across each of the three measured landscapes. The three curves fall almost on top of one another: to a good approximation this bound is a function of a single variable,

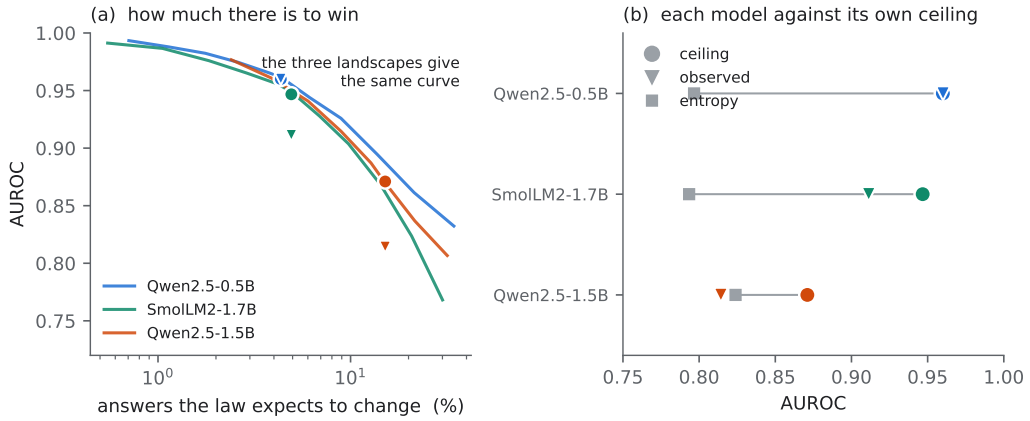

Figure 4: (a) The ceiling as a function of how often the arithmetic decides the answer, swept over each of the three measured landscapes; circles mark each model's operating point, triangles what the law actually achieves there. The three landscapes give the same curve. (b) The same three models on one axis: the distance between the square and the circle is all there is to win.

	expected to change	ceiling	fragility	entropy	headroom
Qwen2.5-0.5B	4.3%	0.960	0.960	0.797	0.163
SmolLM2-1.7B	4.9%	0.947	0.911	0.793	0.154
Qwen2.5-1.5B	15.1%	0.871	0.814	0.824	0.047

Table 2: The ceiling is what a perfect score would achieve on that landscape at that amplitude; the headroom is the distance between ceiling and entropy, that is, the most any decision measure could win over an output measure. On the third model there is almost nothing to win, and this is known before any comparison of precisions is run.

how often the arithmetic decides the answer. Near one changed answer in a hundred the ceiling sits above 0.98; at one in twenty it is about 0.95; at one in five it has fallen to 0.85. The reason traces back to the nature of the crossings. When they are rare, a question either carries a thin cell or it does not, and the fragility separates the two populations almost perfectly. When they are common, most questions carry an appreciable probability of changing and the outcome is decided by the toss.

Each model performs at or near its own ceiling. Table 2 places the three models against their bounds, and it changes the reading of Section 4.4. On the reference model the law reaches 0.960 against a ceiling of 0.960, with 0.163 of headroom over entropy. On Qwen2.5-1.5B the ceiling is 0.871 while entropy already reaches 0.824: the margin available to any decision measure, however perfect, is under five points of AUROC. What looked like a regime where geometry loses to entropy is a regime where there was almost nothing left to win, and where the law took most of what there was. SmolLM2 sits between the two, and its gap to the ceiling is the one the terminal-decision diagnosis accounts for.

Why this bound matters more than the failure it explains. The ceiling is computed from the threshold landscape and the noise amplitude, both extracted from a single reference run, before any cross-precision comparison is executed and before a single flip has been observed. A practitioner can therefore determine in advance whether applying this framework is worthwhile for their model and task. That converts a limitation into a specification: the law does not claim to outperform output-based measures universally, it works where the arithmetic rarely decides the answer, and it supplies the number that says whether a given setup is in that regime. We would suggest that comparisons of uncertainty or robustness measures report this bound. Without it, a victory won in a highly separable regime and a defeat in a regime where no victory was available look alike.

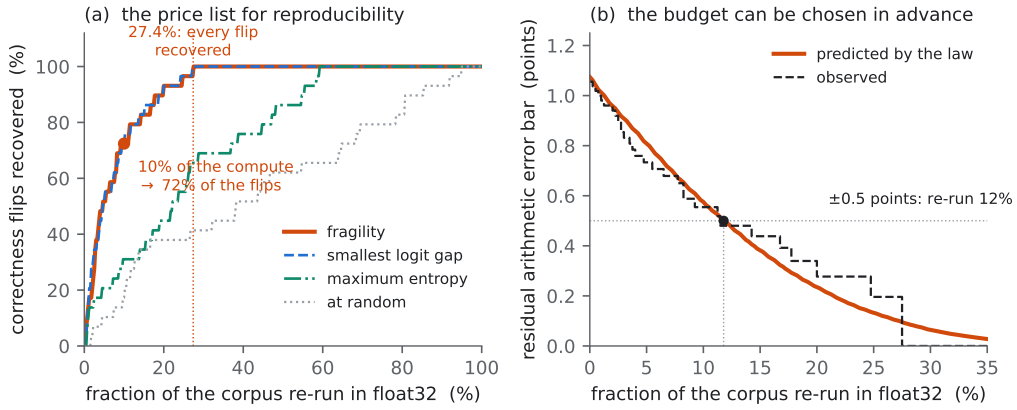

Figure 5: (a) Fraction of the corpus re-run in `float32` against correctness flips recovered. Fragility and the smallest logit gap are indistinguishable; entropy is far behind. (b) The residual arithmetic error bar, predicted by the law and observed, as a function of the budget.

6 Buying reproducibility back

Since the fragility identifies where numerical precision decides the output, there is no need to pay for high-precision arithmetic across the entire dataset. One can run the full corpus in the cheaper `bfloat16` and rank the outputs by their fragility, a ranking that costs nothing because the threshold landscape is already computed during the forward pass. Re-running only the most fragile fraction in `float32` and keeping the cheap answers for the remainder buys reproducibility back. Every re-run item agrees with the reference by construction, so the procedure is described by a single curve: the fraction of the corpus re-run against the fraction of divergences recovered. Borrowing the term from selective prediction, we call it the *deferral curve*.

On the reference model (Figure 5), re-running the top 10% of the most fragile questions recovers 72.4% of the twenty-nine correctness flips. Expanding the budget to 27.4% of the corpus recovers every flip, at which point the item table is identical to the one a full `float32` reference run would have produced. A random ranking needs 99.6% of the corpus for the same guarantee.

Two qualifications are worth stating. First, if the goal is *ranking* alone, the smallest logit gap does as well as the fragility on this setup: 72.4% recovery at the same budget, with a rank correlation of 0.957 between the two scores. The advantage of the decision-space measure is a victory over entropy, not over every conceivable alternative.

Second, a ranking has no absolute scale: it cannot say how many errors are left uncorrected in the tail that was not examined. The law can. The expected number of divergences among the items not re-run is the sum of their fragilities. At a 10% budget it predicts 8.6 residual flips against the 8 observed; at 20%, 1.4 against 2. A practitioner can therefore state the arithmetic error bar they are willing to publish — half a point, say — and the law returns the compute budget that buys it, here 12% of the corpus.

The procedure follows the ceiling. On SmolLM2, where the ceiling is high, re-running a tenth of the corpus recovers 66.7% of the correctness flips and 25.4% recovers all of them — within a tenth of the price paid on the reference model, on a different architectural family with twice the state dimension. On Qwen2.5-1.5B, where the ceiling has collapsed, the same tenth recovers 37.5% and full recovery demands 78.8% of the corpus, so the procedure is not worth running. That the law flags this in advance is the practical form of Section 5.

Finally, the ranking must be carried by the cheap `bfloat16` run itself rather than by reference thresholds. Perturbing every threshold by the measured noise amplitude and re-ranking degrades the curve without breaking it, to 59.5% recovery $[44.8, 75.9]$ at a tenth of the corpus. The figure of 72.4% is thus an upper bound, and the operating range lies between the two.

7 Related work, limitations, and what we claim

Efforts to eliminate the noise at its source — batch-invariant kernels (He & Thinking Machines Lab, 2025), higher-precision accumulation (Yuan et al., 2025) — make a controlled pipeline reproducible. The measure here predicts which outputs will fail to survive a pipeline one does not control, so the two approaches are complementary. Semantic uncertainty (Kuhn et al., 2023; Farquhar et al., 2024) refines confidence estimation in output space and remains, on this task, an impoverished projection of a state-space quantity. The geometry we use is the polyhedral partition studied in the expressivity literature (Montúfar et al., 2014; Hanin & Rolnick, 2019) and echoes the margin of adversarial robustness (Szegedy et al., 2014), evaluated here along a generated sequence and against a noise whose amplitude is measured rather than adversarial.

The limitations are specific. All models are small, 0.5B to 1.7B parameters, and were run on CPU. The perturbations are changes of numerical precision or of hardware architecture rather than of batching or GPU kernel, although the law takes the noise amplitude as its only input and Section 4 shows one amplitude transferring across two different perturbations. The task is short-answer multiple choice, so the number of decisions per answer is small. The pre-registered criterion on how many answers change failed on one of the three models; we traced the reason, but the explanation was not pre-registered and we report the criterion as failed. And the independence assumption in (3) is an approximation we test rather than derive.

What we claim is narrow. The mechanism of numerical fragility is discrete threshold crossing, not chaotic amplification. The governing quantity is a geometric distance in state space. Combined with a noise amplitude measured on independent data, that quantity predicts item-level reproducibility in regimes where the arithmetic rarely decides the answer. Benchmark scores carry an arithmetic error bar that currently goes unreported. And how much any measure can win against that uncertainty is itself measurable in advance.

Statement on the use of large language models

A large language model was used moderately in this work, as an assistant under the direction of the author. It contributed to writing and refactoring the measurement and analysis code, to running and debugging experiments and to drafting and editing this manuscript, including its English prose. The research questions, the experimental design, the pre-registered criteria and the decisions about what to report were the author's. All quantitative claims were produced by the released code from the released data and were re-derived independently of the draft text before submission; the author takes full responsibility for the content, including any errors.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When smol goes big — data-centric training of a small language model. *arXiv:2502.02737*, 2025.

P. W. Anderson, B. I. Halperin, and C. M. Varma. Anomalous low-temperature thermal properties of glasses and spin glasses. *Philosophical Magazine*, 25(1):1–9, 1972.

Riccardo Caponetto and Stefano Fazzino. A semi-analytical method for the computation of the Lyapunov exponents of fractional-order systems. *Communications in Nonlinear Science and Numerical Simulation*, 18(1):22–27, 2013.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*, 2018.

Rainer Engelken, Fred Wolf, and L. F. Abbott. Lyapunov spectra of chaotic recurrent neural networks. *Physical Review Research*, 5:043044, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Rafal Goebel, Ricardo G. Sanfelice, and Andrew R. Teel. *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press, 2012.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference, 2025. Thinking Machines Lab, Connectionism.

H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of ACL*, 2023.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

W. A. Phillips. Tunneling states in amorphous solids. *Journal of Low Temperature Physics*, 7: 351–360, 1972.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Qwen Team. Qwen2.5 technical report. *arXiv:2412.15115*, 2024.

Sidney Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. *arXiv:2506.09501*, 2025.

A Ruling out amplification

Contraction of the flow. Injecting a perturbation of relative amplitude 10^{-2} into the hidden state at one position and propagating it *with the token sequence held fixed*, the residual relative amplitude at later positions falls by an order of magnitude within about ten tokens and then settles on a small residual. There is no trace of growth: even a modest positive exponent, $\lambda = 0.15$ per token, is incompatible with the data by two orders of magnitude. Continuous injection at every step instead of an impulse gives a contraction of 3–4% per token. Holding the tokens fixed is what isolates the continuous flow from the quantisation, and answers "does the flow amplify?" without contamination from the discrete jumps.

The amplitude signature. Generating with injected isotropic noise of increasing relative amplitude and recording the position of the *first* divergence from the reference run separates the two hypotheses. Amplification predicts that divergence always occurs, at a position $t^* \simeq \lambda^{-1} \ln(\rho/\epsilon)$ growing logarithmically as ϵ decreases; crossing predicts an amplitude below which no threshold is exceeded and divergence does not occur at all. As the amplitude decreases the divergence does not slide forward, it disappears: one generation diverges at $s = 0.005$ and no longer at $s = 0.002$, another resists to $s = 0.02$ and gives way at $s = 0.05$, a third never diverges in the range explored.

Orders of magnitude. The noise internal to `float32` computation — thread count, summation grouping, reduction order — amounts to 1.2×10^{-6} in relative norm and never touches the text in our measurements. Moving to `bfloat16` amounts to $\sim 3 \times 10^{-2}$, four orders of magnitude higher, and is nearly isotropic: its mean alignment with the decision directions is 1.2–1.4 times that of a random direction in 896 dimensions. Between the two lies the difference between two machines in single precision.

The scope of the negative claim. It is worth being precise about the scope of the negative claim, because it is easily read more broadly than it should be. What contracts, and what we measured, is the *continuous flow*: with the token sequence held fixed a displacement of the state decays, and along the direction in which the generation unfolds there is no exponential amplification. The hybrid system as a whole is a different matter. Once a decision flips, the two runs emit different tokens and from that point on they are two different texts; sensitivity to the initial condition is therefore certainly present, but it is discontinuous rather than exponential, and it leaves no Lyapunov exponent to measure. Whether two runs that have diverged later re-synchronise — whether the symbolic itineraries come back together, as the contraction of the flow might suggest, or stay apart — is a separate question that our measurements do not address. The claim defended here is the narrow one: along the flow there is contraction, and what changes the text is the crossing of a boundary rather than the growth of an error.

Why the negative result is not a failure of measurement. Estimating Lyapunov exponents is delicate in systems with memory, where the naive variational approach must be adapted (Caponetto & Fazzino, 2013), and an autoregressive transformer is such a system in an extreme form: attention makes the state depend on the entire prefix. Our first attempts estimated finite-time exponents in the *layer* direction; measured over thousands of tokens they showed no predictive power for the fragility of the output. The measurement above explains why: in the sequence direction, the one in which generation unfolds, there is no positive exponent to find.

Where a long generation breaks. On 200 generations of 100 tokens each, compared with the law against a constant risk and an entropy-driven risk with parameters estimated on the same data, the mean log-likelihood difference per generation is $+1.90$ $[1.63,\ 2.17]$ against the constant risk and $+1.76$ $[1.49,\ 2.03]$ against the entropy-driven one (bootstrap over 5000 resamplings of the prompts). Two pre-declared checks, not used for the decision: the amplitude estimated from the divergence data alone comes out at $s = 0.0259$ against a directly measured ~ 0.024; and predicted crossing probability matches observed frequency over four orders of magnitude. A version of the law that considers only the *second* competitor performs as well as the full one (-2.063 against -2.059): in this regime competitors beyond the second add nothing.

A benchmark where the model is too weak. On PopQA (Mallen et al., 2023), 750 verification questions, the fragility predicts which answers change with AUROC 0.705 against 0.531 for the best entropy-based measure (difference 0.174, $[0.051,\ 0.266]$), stable across three blocks of 250 (0.732, 0.688, 0.700). There, however, the model is too weak for a changed answer to move correctness, which is why the main study uses a task on which it is competent.

B THE ACTIVATED-PROCESS READING AND ITS LIMITS

For $\zeta \gg s$ the Gaussian tail in (3) is $\Phi(-\zeta/s) \approx \exp(-\zeta^2/2s^2)$, an Arrhenius form with barrier $\Delta E = \zeta^2/2$ and temperature $k_B T = s^2$: the escape of an overdamped Brownian particle over a barrier (Kramers, 1940). The analogy is exact in its structure and inexact in three declared respects. There is no saddle and no prefactor: the boundary is a hyperplane, not a potential ridge, so the Kramers rate prefactor has no counterpart here. There is no fluctuation–dissipation relation: the noise amplitude is set by the arithmetic, not by a bath in equilibrium with the dynamics. And the boundary is absorbing rather than metastable: once crossed, the system does not return, because the emitted token changes the context.

The disorder is *quenched*. The thresholds are frozen in the weights and in the prompt; nothing trembles in time. The ensemble average implicit in p_t is taken over questions and positions, not over realisations of a noise — the same situation as in the low-temperature physics of glasses, where a

smooth law emerges from a frozen landscape sampled across many sites (Anderson et al., 1972). This is why p_t is a disorder-averaged fraction rather than a rate.

A scaling law, and why it is linear. The reading makes a quantitative prediction. The number of answers that change is $N(s) = \sum_q \int P(\zeta)\,\Phi(-\zeta/s)\,\mathrm{d}\zeta$, so if the density of thresholds behaves as $P(\zeta) \sim \zeta^\theta$ near the origin then $N(s) \propto s^{1+\theta}$. Measuring P on the verification set gives a *flat* density near zero, $\theta = -0.14$, compatible with $\theta = 0$ over a decade and a half; evaluating $N(s)$ on the recorded landscapes over a factor of thirty in amplitude gives a log–log slope of 0.97. Irreproducibility is therefore *linear* in the amplitude of the numerical error: halving the noise halves the number of non-reproducible items, no more and no less.

The finite density of thresholds at zero barrier is the same structural fact that makes glasses anomalous at low temperature, where a finite density of two-level systems at arbitrarily small energy produces a specific heat linear in T instead of the Debye T^3 (Anderson et al., 1972; Phillips, 1972). Here the two-level systems are pairs of near-degenerate tokens, and their finite density at zero is what makes a language model non-reproducible at *any* precision, only proportionally less so.

In the same language the amplitudes of Appendix A are temperatures, and their ratios are striking: $(0.044/1.2 \times 10^{-6})^2 \approx 6 \times 10^8$ between the two precisions, and $(0.044/7.2 \times 10^{-6})^2 \approx 4 \times 10^7$ between a change of precision and a change of machine in `float32`. Low precision is not slightly noisier; it is nearly nine orders of magnitude hotter. The systematic drift of Appendix D is then not a temperature but a *field*: it tilts all the barriers in the same direction instead of agitating them independently, which is exactly why treating it as independent noise over-predicts.

Two further observations are worth recording. Thresholds grow along the answer, from a median 0.358 at the first decision to 1.571 by the fourth, with the fraction of fragile positions falling from 7.5% to zero: the generation contracts onto its answer, and the decisions that matter are the early ones. And the density of thresholds near zero is a property of the *task*, not only of the model: on questions the model answers correctly it is 0.0348, on those it gets wrong 0.0864, against 0.0563 and 0.0532 for a random split of the same size.

An engine of deterministic focusing. The two observations above suggest a reading at a coarser scale than the crossing law, and it is worth stating because it inverts a familiar mechanism. There is a body of work in dynamical systems on *deterministic diffusion*: a deterministic chaotic system on a periodic lattice spreads diffusively, and its diffusion coefficient can be computed from the dynamics rather than postulated. The engine there is instability — the exponential separation of nearby orbits is what decorrelates successive steps and makes deterministic motion indistinguishable from a random walk.

The system studied here runs that mechanism backwards. It contracts, so nothing is decorrelated by stretching, and the only channel through which uncertainty survives is the discrete decision. In that reading the crossing probability p_t occupies the position the diffusion coefficient occupies there — a transport coefficient computed from the deterministic structure of the system rather than assumed — except that it measures the rate at which continuous uncertainty is converted into *symbolic* uncertainty rather than the rate at which it spreads.

Stated in one sentence: a language model under deterministic decoding is not a generator of chaotic diffusion but an engine of deterministic focusing *along the sequence*. The opening decision of an answer is the undetermined one — its state sits close to the decision boundaries, and in 7.5% of cases inside the band of the numerical noise. Once that token is emitted, the contraction carries the state towards the middle of its own cell: thresholds grow fivefold within three tokens and later decisions become progressively unassailable. Uncertainty is neither destroyed nor amplified; it is channelled into a single discrete choice and defended from then on. This is not an alternative to the activated-process reading but the same description at a different height — the Kramers form describes what happens at one boundary, focusing describes why the boundaries are arranged as they are, low at the opening decision and high afterwards, and p_t is the same number read from either side. That the variety of the text is carried in from the prompt and the weights, rather than produced by a dynamics that only contracts, is consistent with the density of thresholds near zero changing by a factor of two and a half with the questions asked.

We flag this as an interpretation of measurements rather than a result derived from them. What is measured is the growth of thresholds along the answer and the task-dependence of their density; the reading of p_t as a conversion rate is ours, and we state it because we believe it is the right frame for the numerical robustness of any network that ends in an $\arg\max$, not because the data compel it.

C Protocol, pre-registered criteria and power

Five hundred calibration and one thousand verification questions, split by a fixed-seed permutation, pilot questions excluded. Three criteria and their thresholds were fixed before collection:

which answers change the fragility predicts which answers change with an AUROC exceeding the best entropy-based measure by at least 0.10, with the bootstrap interval of the difference entirely above zero;

where the flips sit among questions with fragility above the median, the frequency of correctness flips is at least three times that below the median, with the interval of the ratio excluding 1; this requires at least ten observed flips, failing which the criterion is declared not evaluable;

how many change estimating s on the calibration questions alone, the predicted number of changed answers in the verification set falls within 25% of the observed number.

A power simulation, run before collection on resampled real landscapes from the pilot, put the probability of confirming the first criterion at 100% if the law is true and at 1% or less if the outcome were driven by a constant risk or by an entropy-driven risk. The median split for the second criterion is taken on the fragility computed before any outcome is known.

On the third model the terminal-decision rule of Section 4.4 was pre-registered in a dated document before collection, together with four numerical predictions. One of the four was confirmed — the terminal decision was predicted to matter little on a model sharing the reference tokeniser, and it changes the prediction by 11.1% — and three were not. We report this outcome as it stands.

D Calibration of the amplitude, Gaussianity and tails

The amplitude is measured, not fitted, and where it is measured matters. Computing the landscape of the *same* generated sequence under both precisions and pairing competitors *by token identity* rather than by rank — where the two orderings differ, matching by rank compares different tokens — the standard deviation of the threshold differences is $s = 0.044$ on the reference model, 0.0185 on SmolLM2-1.7B and 0.173 on Qwen2.5-1.5B.

The measured displacement is close to Gaussian: excess kurtosis 1.3 and 1.6 on the first two models, standard deviation over robust scale 1.16 and 1.10. Heavy tails are excluded by three tests, and replacing the Gaussian by the empirical distribution of displacements moves the predicted count by less than the counting noise. The measured displacement also carries a systematic drift as well as a spread; including the drift in the law makes the prediction markedly worse (86.0 against 53 observed on the reference model), which is consistent with reading it as a change of units — thresholds are normalised by $\|\mathbf{h}_t\|$, which differs slightly in reduced precision — rather than as a push: a common tilt of the whole landscape brings no decision closer to its own boundary.

One artefact of the measurement deserves a note. Read as a function of the threshold it sits on, the amplitude appears smaller in the thinnest band, but there with a *positive* drift while every other band drifts negatively. That is the signature of a sample truncated from below, which is what that band is: pairs whose decision has already flipped are discarded, and on a thin threshold those are precisely the ones the noise would have produced. The thin band is therefore not measurable this way, and the global estimate is the one the law should use.

E Numerical details, versions and the determinism control

Reference model Qwen2.5-0.5B-Instruct (Qwen Team, 2024), $d = 896$; further models SmolLM2-1.7B-Instruct (Allal et al., 2025), $d = 2048$, and Qwen2.5-1.5B-Instruct, $d = 1536$. All reference

runs in `float32` on CPU; alternative configuration `bfloat16`. Nine competitors per decision, fourteen generated tokens, answers truncated at the first newline.

Pure deterministic decoding requires one precaution that is easy to miss. The default generation configuration of many models applies a repetition penalty *even* under greedy decoding; when it is active the emitted token is not the $\arg\max$ of the logits, the boundary actually crossed is not the one computed, and the geometric margin loses its meaning. In all measurements the penalty is disabled and the generated sequence was verified token by token to coincide with the $\arg\max$.

The reference is stable. Treating `float32` on a given machine as *the* reference presumes that two runs of the same code on the same machine agree. Repeating an entire fifteen-hundred-question collection from scratch reproduces every decision: the same generated answers in both precisions, the same options, the same correctness verdicts, the same item-by-item changes. Of the 7.4×10^4 thresholds recorded, 385 differ at all, all of them in the first dozen questions of the run — a warm-up effect of the numerical libraries — with a largest discrepancy of 1.7×10^{-5} and a standard deviation of 3.2×10^{-7} over the whole archive. The largest run-to-run discrepancy is thus about a thousandth of the `bfloat16` amplitude on the same model, and of the same order as the machine-to-machine displacement of Section 4. The same check on the second machine, where ten questions happened to be recomputed after an interruption, likewise returned identical answers.

Versions. The amplitude s is a property of a model *and* of the arithmetic that executes it, so the software stack is part of the measurement. All runs use one stack per machine, held fixed across precisions and across models; exact versions are recorded in the released package alongside the landscapes, which allow every number in this paper to be recomputed without rerunning the models.