

Quantitative study of sustained focus in large language models via repetitive deterministic prediction tasks

Wanda Hou^{1,*}, Leon Zhou^{2,‡}, Hong-Ye Hu^{3,§}, Yubei Chen^{4,||}, Yi-Zhuang You^{1,†} and Xiao-Liang Qi^{5,¶}

¹EdenCode Inc., San Jose, California, USA

²Stevenson School, Pebble Beach, California 93953, USA

³Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

⁴Department of Electrical and Computer Engineering, University of California, Davis, Davis, California 95616, USA

⁵Path Integral Technology, Inc., Cupertino, California 95014, USA



(Received 22 March 2026; accepted 22 September 2026; published 13 October 2026)

We investigate the performance of large language models (LLMs) on repetitive deterministic prediction tasks and study how the sequence accuracy rate (SAR) scales with output length. Each such task involves the repetition of the same operation N times. Examples of such tasks include letter replacement in letter strings following a given rule, integer addition, and multiplication of string operators in many-body quantum mechanics. If the LLM performs the task by a simple repetition algorithm, the success rate would follow an exponential decay with sequence length. In contrast, our experiments on leading LLMs reveal a crossover that is sharper than exponential: $-\log \text{SAR}$ grows superlinearly with N , and accuracy collapses around a characteristic length N_* , the accuracy cliff that separates reliable from unreliable generation. The hypothesis of independent per-step errors is rejected for every model and task we studied. The crossover is well described by a double-exponential accumulation law, $\text{SAR} = \exp(-\beta_0 N \alpha^{N-1})$, whose crossover scale N_* does not depend on the functional form chosen to fit it. To interpret this behavior, we introduce a minimal effective model in which step-correctness variables interact through dense random couplings and compete with an external field set by the prompt. Solved by direct enumeration, the model reproduces the superlinear error accumulation and the accuracy cliff qualitatively, and it assigns to each model-task pair two interpretable parameters, an intrinsic error rate and an error-accumulation factor.

DOI: [10.1103/fn1m-wn2b](https://doi.org/10.1103/fn1m-wn2b)

I. INTRODUCTION

Accuracy rate is among the most fundamental and widely adopted evaluation metrics in machine learning, especially for classification and other discriminative tasks. Its appeal lies in the clarity of its definition: Each input has a well-defined ground-truth label, and the model's prediction can be judged unambiguously as correct or incorrect. This binary criterion makes accuracy a natural measure of performance, enabling direct comparison across models and datasets.

However, this paradigm does not directly extend to large language models (LLMs). Many LLM applications, such as machine translation, summarization, or dialogue, are inherently sequence-to-sequence generation problems. In these settings, the model generates text probabilistically, token by

token, sampling from a distribution of plausible continuations. Crucially, there is rarely a single “correct” output: Multiple valid translations, summaries, or conversational responses may exist. Consequently, simple exact-match accuracy is insufficient for evaluation, and researchers rely instead on a suite of task-specific metrics (e.g., BLEU, ROUGE, perplexity, or human judgment) [1–3] that capture fluency, coherence, and relevance.

Yet, this limitation is context dependent. In scientific domains—including mathematics, physics, and computer science—the nature of the task often changes. Here, many problems admit a unique correct solution: a definite numerical value, a closed-form expression, or a precise code output. For instance, the question “what is $2\,342\,342 + 3\,442\,342$ ” has only one acceptable answer, and the model's response must exactly match the correct sequence of digits to be considered correct [4–6]. In such applications, the probabilistic flexibility of LLMs does not diminish the importance of exact correctness. The evaluation once again reduces to a binary criterion: Either the model outputs the correct sequence or it does not.

This observation motivates the introduction of a new evaluation concept tailored for LLMs in scientific applications: the sequence accuracy rate (SAR). SAR measures the proportion of problem instances for which the LLM produces the exact correct sequence as output. By aligning the evaluation of LLMs with the well-established standard of accuracy in classification, SAR provides a rigorous

* Contact author: hwanda@edencode.ai

† Contact author: yzyou@edencode.ai

‡ Contact author: lzhou26@stevensonschool.org

§ Contact author: hongyehu@fas.harvard.edu

|| Contact author: ybchen@ucdavis.edu

¶ Contact author: phynics@pathintegral.xyz

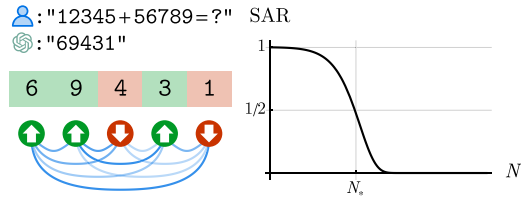


FIG. 1. Illustration of the experimental setup and theoretical framework. The left panel shows examples of deterministic sequence prediction tasks (e.g., arithmetic problems) used to evaluate LLMs, where each token's correctness is represented by a binary variable (Ising spin). Errors at different positions are modeled as Ising spins interacting through dense random couplings, an effective description of how correlations and noise propagate during sequence generation. The right panel displays the typical behavior of the SAR (i.e., the probability for the whole output sequence to be correct) as a function of the sequence length N : SAR remains high for short sequences and then drops beyond a characteristic crossover scale N_* , forming the accuracy cliff.

and interpretable metric for assessing LLMs in domains where precision and correctness are paramount, as illustrated in Fig. 1.

LLMs have achieved remarkable progress across a wide range of tasks, yet their reliability on structured, multistep reasoning remains poorly understood [7,8]. While in generative modeling the challenge has often been framed as a trade-off between diversity and realism [9], multistep reasoning tasks demand consistent correctness across sequential subtasks. In the simplest case, success in these settings is approximately *factorizable*: For tasks whose output positions are computed independently (such as cyclic letter replacement), the probability of solving the overall problem is the product of per-step accuracies, making performance highly sensitive to individual step error rates. Other tasks introduce explicit dependencies between steps—integer addition couples digits through long-range carry propagation, and Pauli string multiplication couples sites through a shared global phase—so that the strict factorization holds only as a leading-order approximation. This range of dependency structures is exactly what allows us to probe whether models exploit compositional structure and execute each step independently, rather than treating complex problems monolithically [10].

To probe this challenge, we introduce a minimal and fully deterministic evaluation class where each problem instance admits a unique ground-truth solution and complexity can be tuned by a single parameter N . We present three complementary benchmarks: (1) cyclic-letter-replacement task, a purely symbolic problem over a finite alphabet that isolates step-wise error propagation without memory; (2) integer addition, which tests the handling of local carry rules and long-range carry propagation across digits; and (3) Pauli string multiplication, which requires tracking both local symbol rules and a global phase factor, thereby testing the model's ability to maintain a lightweight memory across many steps. All three benchmarks share the key property that overall complexity decays multiplicatively with sequence length N , providing a clean probe of cumulative error accumulation in deterministic reasoning [11–14].

Beyond empirical evaluation, we propose a statistical-physics interpretation of the observed behavior of SAR, drawing an analogy to an Ising model in which the correctness variables of the individual task positions influence each other through effective random couplings. Solved by direct enumeration, this minimal model reproduces the superlinear error accumulation and the accuracy cliff qualitatively and provides a two-parameter description of each model-task pair. The model is an *effective*, phenomenological description and should not be read as a microscopic theory of the network; what the data establish is the superlinear accumulation of errors, and the mechanism behind it remains open. Together, our benchmarks and theoretical framework establish a controlled setting for studying how LLMs manage compositional reasoning, memory, and robustness in deterministic problem families.

II. EXPERIMENT

We prompted a set of LLMs—gpt-5, gemini-2.5-pro, gemini-2.5-flash, grok-4, and claude-4-sonnet—to perform a series of deterministic sequence prediction tasks that each admit a unique correct solution. These tasks provide a well-defined setting for evaluating the SAR, enabling a quantitative analysis of how the accuracy of generated sequences scales with the problem size, measured by the output sequence length. During evaluation, all models were tested in a closed setting with external tools such as code execution and web access explicitly disabled. Concretely, every query was issued through a plain text completion/chat call with no tool-use, function-calling, or code-interpreter interfaces enabled (see the source code); consequently, the models cannot invoke external calculators or solvers, and all results reflect the models' internal, in-context reasoning capabilities only.

Estimation of SAR. For each task and each sequence length N , we pose n independent problem instances in batches of ten freshly sampled problems per prompt, typically ten batches per length ($n = 100$; a subset of grok-4 runs use fewer, $n \leq 30$). Throughout the paper, the empirical SAR at length N is the fraction of instances whose complete answer is correct, $\bar{\text{SAR}}(N) = k/n$; the same estimator is used for every task and every figure. A batch whose reply could not be parsed into answers at all (an empty or malformed response) is treated as a protocol failure and excluded from n . Such batches make up between 0% and 9% of the total depending on the model and occur mostly at large N ; counting their instances as wrong instead shifts the fitted crossover scales by at most 5%. Because the ten instances of a batch are generated in one reply, their outcomes are positively correlated. The between-batch variance of the batch accuracy exceeds the binomial value by a design effect $\hat{c}$, which we estimate for every model-task pair: $\hat{c} = 1.0$ – 1.9 for letter replacement and addition, 2.1 – 4.2 for the reasoning models, and 4.8 – 5.9 for Pauli multiplication. All error bars are therefore 95% Wilson intervals computed with the effective sample size $n/\hat{c}$, all fits maximize the binomial likelihood of the instance counts, and the information criteria are rescaled accordingly (quasi-Akaike information criterion, $\text{QAIC} = 2P - 2 \log L/\hat{c}$, with P the number of fitted parameters; see Appendix C). Here and throughout, $\log$ denotes the natural logarithm. Sampling temperature was set to 0

wherever the API exposes it; the reasoning models (gpt-5, grok-4) use their fixed internal sampling and do not expose a temperature parameter. While the quantitative results may vary slightly with prompt phrasing or instruction style, the qualitative behaviors reported here are robust across formulations. The prompt templates, per-model settings, raw records, and the complete analysis pipeline are available at [15].

A. Cyclic letter replacement

We start with a simple cyclic-letter-replacement task defined over an alphabet

$$\mathcal{A} = \{A, B, C, \dots\}. \quad (1)$$

$$T = ((x_1 + 1) \bmod |\mathcal{A}|, (x_2 + 1) \bmod |\mathcal{A}|, \dots, (x_N + 1) \bmod |\mathcal{A}|). \quad (4)$$

For example, for $\mathcal{A} = A, B, C, D$,

Input: *ADBAA* $\Rightarrow$ Output: *BACBB*,

Input: *DABDC* $\Rightarrow$ Output: *ABCAD*.

For this task, the results are presented in Figs. 2 and 3. Figure 2 shows results from gemini-2.5-pro and gemini-2.5-flash with alphabet sizes $|\mathcal{A}| = 4, 9, 13, 26$. Figure 3 compares gpt-5, gemini-2.5-pro, gemini-2.5-flash, grok-4, and claude-4-sonnet at $|\mathcal{A}| = 13$.

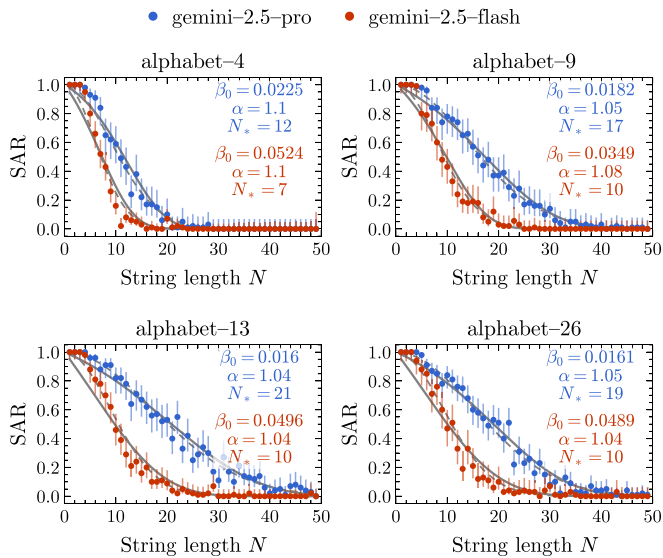


FIG. 2. Cyclic-letter-replacement benchmark on gemini-2.5-pro and gemini-2.5-flash with alphabet size $|\mathcal{A}| = 4, 9, 13$, and 26. Points: fraction of correct instances (up to $n = 100$ per length); error bars: 95% Wilson intervals with the batch design effect. Solid curves: maximum-likelihood fit of the accumulation law [Eq. (14)], with the fitted β_0 , α , and crossover scale N_* ; dashed: the quadratic law $e^{-aN - bN^2}$ (Sec. III A).

Each letter is mapped to an integer index $x \in \{0, 1, \dots, |\mathcal{A}| - 1\}$. The transformation rule is a cyclic shift by one modulo the alphabet size:

$$y = (x + 1) \bmod |\mathcal{A}|. \quad (2)$$

The output is then mapped back to a letter in $\mathcal{A}$. Each position in the input string is transformed independently, so for a string of length N the task is simply to apply the shift rule to each site.

Formally, given an input string

$$S = (x_1, x_2, \dots, x_N), \quad x_j \in \{0, \dots, |\mathcal{A}| - 1\}, \quad (3)$$

the transformed output string is

B. Integer addition

Next, we consider *integer addition* with a fixed number of digits. The task probes arithmetic consistency and multidigit carry propagation—requiring sequential reasoning rather than purely single-site transformations.

Each query presents an addition problem of the form

$$A + B, \quad (5)$$

where $A, B \in \mathbb{N}$ are positive integers with the same digit lengths. The model is asked to compute the sum and return the result as a plain numeric string.

For example, at $N = 4$,

Input: *1234 + 5678* $\Rightarrow$ Output: *6912*,

Input: *9999 + 0001* $\Rightarrow$ Output: *10 000*.

The corresponding results are shown in Fig. 4, evaluating gemini-2.5-flash, gemini-2.5-pro, grok-4, and claude-4-sonnet.

C. Pauli algebra

As a third deterministic benchmark, we consider *Pauli string multiplication* task with tunable string length. The Pauli operators X, Y , and Z are 2×2 matrices

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (6)$$

They originate as generators of the Lie algebra $\mathfrak{su}(2)$, up to constants. For our purposes, we treat them algebraically, using only their multiplication rules and phase factors. The identity operator is denoted I . The single-qubit multiplication

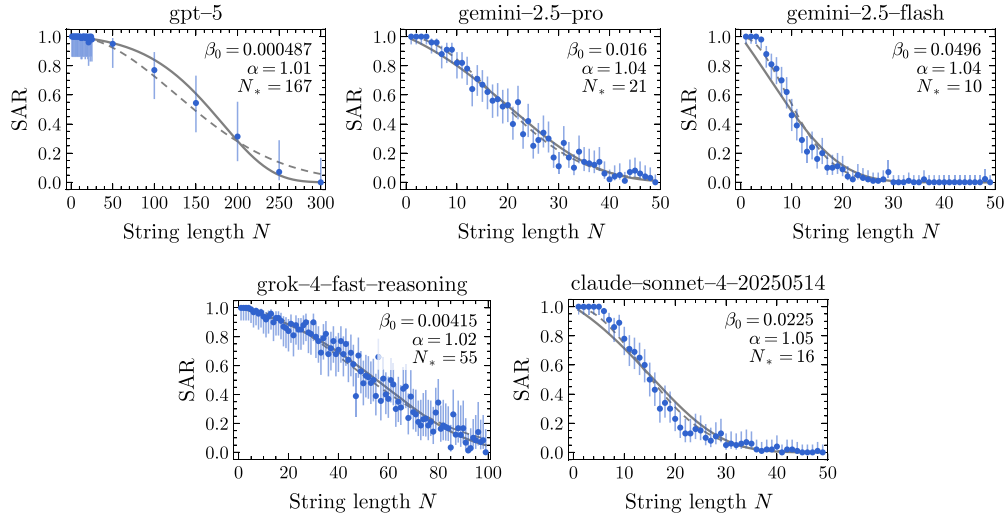


FIG. 3. Cyclic letter replacement with alphabet size $|A| = 13$ across different models. Points: fraction of correct instances (up to $n = 100$ per length; $n \leq 30$ for some grok-4 lengths); error bars: 95% Wilson intervals with the batch design effect. Solid: fit of Eq. (14); dashed: quadratic law.

rules are

$$X \times X = Y \times Y = Z \times Z = I, \quad (7)$$

$$X \times Y = iZ, \quad Y \times X = -iZ, \quad (8)$$

$$Y \times Z = iX, \quad Z \times Y = -iX, \quad (9)$$

$$Z \times X = iY, \quad X \times Z = -iY, \quad (10)$$

together with $I \times P = P$ for $P \in \{I, X, Y, Z\}$.

A Pauli string on N sites is

$$S = \alpha \bigotimes_{\ell=1}^N P_{\ell}, \quad (11)$$

where $\alpha \in \{\pm 1, \pm i\}$ is a global phase and $P_{\ell} \in \{I, X, Y, Z\}$. Because Pauli operators acting on different sites commute, the product of two strings reduces to independent local multiplications together with accumulation of the global phase:

$$\left(\alpha_1 \bigotimes_{\ell=1}^N P_{\ell} \right) \left(\alpha_2 \bigotimes_{\ell=1}^N Q_{\ell} \right) = \left(\alpha_1 \alpha_2 \prod_{\ell=1}^N \phi_{\ell} \right) \bigotimes_{\ell=1}^N R_{\ell}, \quad (12)$$

where $P_{\ell} \times Q_{\ell} = \phi_{\ell} R_{\ell}$, with $\phi_{\ell} \in \{\pm 1, \pm i\}$ and $R_{\ell} \in \{I, X, Y, Z\}$.

This structure makes the task explicitly factorizable: Each site requires a local Pauli multiplication and a phase update, while the overall result depends on consistent accumulation of all local outcomes. See Appendix A for a theoretical complexity analysis.

The need to consistently maintain and update the global phase highlights the role of *memory handling* in multistep reasoning. Unlike purely sitewise problems, errors here can accumulate not only locally but also through the global phase register.

For this task, the results are presented in Fig. 5, evaluating gemini-2.5-pro and gemini-2.5-flash under two evaluation criteria: one requiring both the operator sequence and accumulated phase to be correct, and a relaxed criterion considering only the operator sequence while ignoring phase errors.

III. ANALYSIS

A. Empirical formula

In our experiments, we observed that the SAR depends strongly on the output sequence length N . Assuming independent-error events at the individual steps of the task, the baseline expectation is that the sequence accuracy rate decays exponentially with the sequence length N :

$$\text{SAR}(N) = e^{-\beta N}, \quad (13)$$

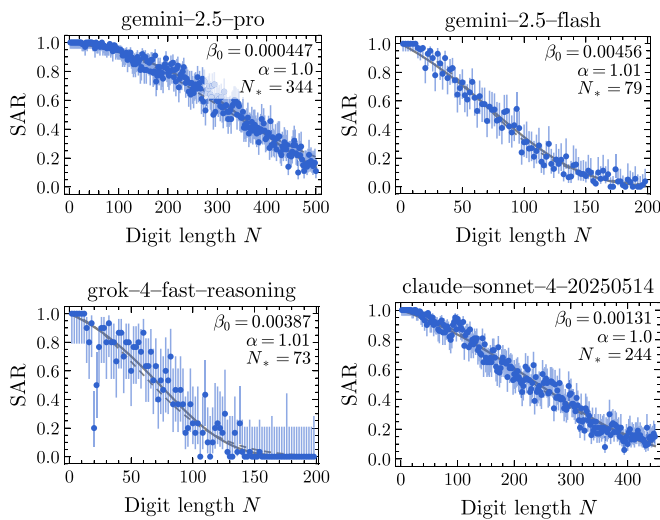


FIG. 4. Integer-addition benchmark across different models. Points: fraction of correct instances (up to $n = 100$ per length; $n \leq 30$ for grok-4); error bars: 95% Wilson intervals with the batch design effect. Solid: fit of Eq. (14); dashed: quadratic law.

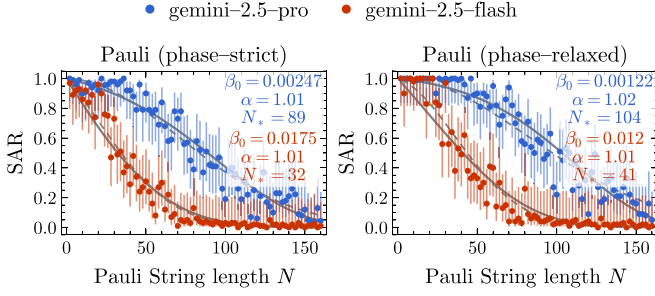


FIG. 5. Pauli string multiplication benchmark on gemini-2.5-pro and gemini-2.5-flash evaluated under strict (phase-matched) and relaxed (phase-ignored) criteria. Points: fraction of correct instances (up to $n = 100$ per length); error bars: 95% Wilson intervals with the batch design effect ($\hat{c} \approx 5$ for this task). Solid: fit of Eq. (14); dashed: quadratic law.

where $\beta > 0$ denotes a constant per-step error rate. However, as we have seen in the above experiments, the empirical results deviate significantly from this simple picture. Instead of a smooth exponential decay, SAR exhibits a crossover: It remains nearly perfect for short sequences and then drops to zero around a characteristic length scale N_* , more rapidly than an exponential would.

This observation indicates that the errors at different steps are not independent: Earlier steps affect the error rate of later ones. To capture this history dependence, we treat the per-step error rate as a function of the sequence length N rather than a constant and adopt the following empirical fit:

$$\text{SAR}(N) = e^{-\beta(N)N}, \quad \beta(N) = \beta_0 \alpha^{N-1}, \quad (14)$$

where $\beta_0 > 0$ quantifies the *intrinsic* error rate and α represents the error-accumulation factor, which characterizes how local errors compound as the sequence length increases. When $\alpha = 1$, the model reduces to the standard exponential decay expected from independent errors. For $\alpha > 1$, however, the effective error rate $\beta(N)$ grows exponentially with N even when β_0 is small, leading to a rapid drop in accuracy once $\beta(N)$ becomes of order one. The onset of this accuracy cliff occurs at a characteristic sequence length

$$N_* \simeq 1 + \frac{\log(1/\beta_0)}{\log \alpha}, \quad (15)$$

beyond which scale the model's reliability collapses.

Is the functional form warranted? Two questions must be kept apart: Whether the data reject the independent-error law [Eq. (13)], and whether they single out the particular N dependence of Eq. (14). To answer both we fitted, by binomial maximum likelihood, Eq. (14) together with its one-parameter special case $\alpha = 1$ and with three alternative two-parameter crossover laws: the quadratic exponent $\exp(-aN - bN^2)$, which is the leading-order prediction of the model of Sec. III B and the natural competitor suggested by the perturbative expansion; a stretched exponential $\exp[-(N/N_0)^\gamma]$; and a logistic crossover $[1 + e^{(N-N_0)/w}]^{-1}$. The laws are compared by quasi-AIC and by the quasi-Bayesian information criterion (quasi-BIC), both rescaled by the design effect $\hat{c}$, and by leave-one-length-out predictive likelihood; the full table is given in Appendix C. On the first question, the answer is

clear. The exponential law is rejected on every one of the 19 model-task curves, by $\Delta\text{QAIC} \geq 35$ and up to 1.6×10^3 , so the superlinear growth of $-\log \text{SAR}$ with N is an established feature of the data. On the second question, the answer is negative. The four crossover laws describe the data comparably: The stretched exponential has the lowest QAIC on 14 of the 19 curves, Eq. (14) on 3, and the logistic on 2, and the differences, $|\Delta\text{QAIC}| \leq 1.6 \times 10^2$, are one to two orders of magnitude smaller than the margin against the exponential. The data alone therefore do not discriminate between crossover forms of comparable flexibility. We adopt Eq. (14) as the working parametrization for two reasons. It is the exponentiation of the leading-order correction of the model in Sec. III B, and the quantity we use to compare models and tasks does not depend on the choice: The crossover scales N_* obtained from the four laws agree to within 6% (median over curves) and 15% at most. The fitted accumulation factors are modest, $\alpha = 1.003\text{--}1.10$, i.e., $\log \alpha$ between 3×10^{-3} and 0.1; the crossover is nevertheless far from exponential because α^{N-1} is evaluated at $N \sim N_*$, which ranges from 7 to 344 across our curves.

B. Hypothesis and statistical model

Empirically, the observed crossover length N_* is typically on the order of $10 \sim 300$ —well within the attention window ($\sim 10^5$ or larger) of modern LLMs. This rules out limited context length as the primary cause of the accuracy cliff. We therefore propose an *effective* model in which the superlinear accumulation of errors is captured by dense couplings among position-correctness variables. This choice is motivated by—though not derived from—the all-to-all structure of the self-attention mechanism. While the attention weights in next-token prediction are structured and typically sparse rather than uniformly distributed over all past tokens, the architecture nonetheless allows fluctuations and errors to spread through the network in an effectively all-to-all fashion, propagating globally across the entire generated sequence. As a result, small local errors can be amplified and spread through these dense interactions, effectively inducing collective fluctuations among the accuracies at different positions.

To analyze this mechanism, we abstract away the content of the output and focus solely on whether each position of the task is answered correctly. We introduce binary task-position correctness variables $s_i = \pm 1$, $i = 1, \dots, N$, with $s_i = +1$ if the i th position of the task (a letter, a digit, or a Pauli site, as opposed to a tokenizer token) is correct and $s_i = -1$ otherwise. We model the joint distribution of position errors $s = \{s_i\}$ using an effective Ising-like energy function,

$$E_J[s] = - \sum_{i < j} J_{ij} s_i s_j - h \sum_i s_i, \quad (16)$$

where h denotes an external field reflecting the baseline bias from the prompt toward correct generation and J_{ij} represent effective couplings between position-correctness variables. In the independent-position limit, $J_{ij} = 0$ and the model predicts the exponential decay of Eq. (13). The rejection of that law in Sec. III A implies correlations among the outcomes at different positions.

We model these correlations phenomenologically as random couplings drawn independently from a Gaussian ensemble, a form inspired by (but not derived from) the all-to-all structure of attention,

$$\mathbb{E}_J[J_{ij}] = 0, \quad \mathbb{E}_J[J_{ij}^2] = J_0^2, \quad (17)$$

where J_0 characterizes the strength of internal interference or “noise coupling.” The random couplings capture, in an effective way, both constructive and destructive interference among position-correctness variables. Together with the external field h , which sets the intrinsic per-position accuracy, this formulation defines a statistical-physics model for the collective behavior of errors. Its energy function is that of the Sherrington-Kirkpatrick (SK) spin glass, with two important differences of interpretation that we make explicit.

Scaling of the couplings. The coupling variance J_0^2 is independent of N . This differs from the conventional SK normalization J_0^2/N , which is chosen to make the free energy extensive in the thermodynamic limit. No such limit is taken here, because N is precisely the quantity varied in the experiments. The N -independent variance encodes the assumption that each additional output position couples to all earlier positions with a fixed characteristic strength, so that the total disorder energy grows as N^2 , while the field energy grows as N . We stress that this growth is an assumption built into the effective model, motivated by the all-to-all structure of attention but not derived from it, and that it should not be read as a prediction of spin-glass physics; with the standard $1/\sqrt{N}$ normalization, the same model produces no cliff (Appendix B 5). In this sense, the model is a dense random-coupling Ising model of SK type rather than the SK model proper, and we use the SK terminology only for the mathematical structure.

Symmetric couplings and position labels. The couplings are symmetric, whereas autoregressive generation is directed: Later outputs depend on earlier ones and not conversely. The energy model therefore describes the joint distribution of correctness outcomes and makes no statement about the generation dynamics; a causal variant with $J_{ij} \neq 0$ only for $j < i$ is a natural extension that we leave to future work. Likewise, the index i labels a task position and not a tokenizer token or an attention head, and nothing in the model identifies J_{ij} with attention weights.

We emphasize that a genuinely disordered coupling is essential to the observed phenomenology, and that simpler alternatives are insufficient. A model containing only a field (uniform, or even a site-dependent random field h_i) but no pairwise coupling factorizes over sites; its $\log \text{SAR}$ is then strictly linear in N , reproducing the single exponential of Eq. (13) with no crossover. Uniform *ferromagnetic* couplings ($J_{ij} = J_0 > 0$) would instead drive a smooth mean-field transition, not the sharp, sample-to-sample crossover seen here. The random signs of J_{ij} introduce *frustration*—competing pairwise constraints that cannot all be satisfied simultaneously—while the field h favors the all-correct configuration. It is the competition between this disorder (internal interference among generated positions) and the field (external conditioning from the prompt) that the model captures; the disorder contributes a super-linear-in- N term to $-\log \text{SAR}$, and hence an accuracy cliff. We note that the field explicitly breaks the $s \rightarrow -s$ symmetry, so there is no $\mathbb{Z}_2$ degeneracy

between the all-correct and all-incorrect states: At small N , the field selects the all-correct configuration (paramagnetic regime), whereas at large N the disorder energy ($\propto N^2$) overtakes the field, so the all-correct configuration is no longer favored—and, in the strong-disorder/weak-field regime, an incorrect configuration can become energetically competitive. This asymmetry is precisely what the field term is designed to encode.

For a given realization of the couplings J_{ij} , the probability of observing a configuration of position-correctness variables s is

$$p_J[s] = \frac{1}{Z_J} e^{-E_J[s]}, \quad (18)$$

with the partition function

$$Z_J = \sum_s e^{-E_J[s]}. \quad (19)$$

In particular, the SAR is associated with the probability of the fully correct configuration $s = (1, 1, \dots, 1)$, averaged over disorder realizations of the couplings, which play the role of problem instances. The empirical estimator (fraction of correct instances) corresponds to the arithmetic disorder average $\mathbb{E}_J p_J[s = \mathbf{1}]$; the replica calculation of Appendix B is most naturally carried out for the geometric average,

$$\text{SAR}_{\text{geo}} := \exp(\mathbb{E}_J[\log p_J[s = \mathbf{1}]]), \quad (20)$$

and the two coincide to leading order in the disorder and whenever SAR is close to one (Appendix B 1); both are evaluated in the numerics of Appendix B 5. To leading order in the coupling strength (Appendix B 4),

$$\begin{aligned} \log \text{SAR}_{\text{geo}} &= -N \log(1 + e^{-2h}) \\ &\quad - \frac{J_0^2}{4} N(N-1)(1 - \tanh^4 h) + \mathcal{O}(J_0^4) \end{aligned} \quad (21)$$

(we have verified this coefficient against exact enumeration, Appendix B 4). The disorder thus contributes a term quadratic in N with a fixed negative sign, which is precisely the quadratic law $\exp(-aN - bN^2)$ compared in Sec. III A, with $a = \log(1 + e^{-2h})$ and $b \simeq 2J_0^2 e^{-2h}$ for $h \gg 1$. The empirical form [Eq. (14)] corresponds to exponentiating the relative correction, $1 + 2J_0^2(N-1) \rightarrow e^{2J_0^2(N-1)}$, which suggests the identification $\alpha \simeq e^{2J_0^2}$, $\beta_0 \simeq e^{-2h}$. This exponentiation is a phenomenological closed form rather than a controlled resummation, and we do not use it for quantitative statements. Instead, we evaluate the model directly by exhaustive enumeration of the 2^N configurations for $N \leq 19$ and 200 disorder realizations (Appendix B 5, Fig. 11). The exact curves display the qualitative phenomenology of the data. A plateau near $\text{SAR} = 1$ is followed by a crossover whose location is controlled by J_0 and h and which is much sharper than the exponential law with the same h . The crossover of the exact curves is, however, broader than the exponentiated form predicts, and the leading-order quadratic law is accurate only well before the crossover. We therefore regard the model as a minimal effective model that qualitatively produces superlinear error accumulation and an accuracy cliff.

From the statistical-physics perspective, the accuracy cliff corresponds to a finite-size crossover in the underlying spin

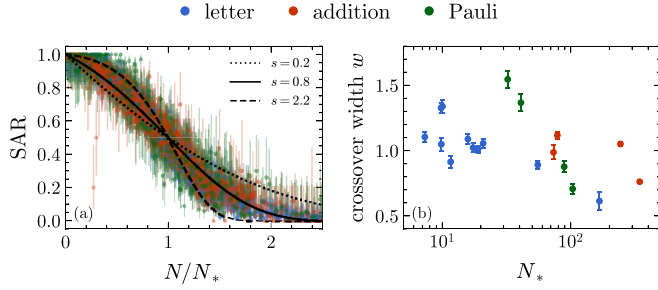


FIG. 6. Scaling collapse of the sequence accuracy rate. (a) SAR against the rescaled length N/N_* [with N_* the fitted SAR = 0.5 crossover scale of Eq. (14)] for all 19 model-task curves; colors denote task families; error bars are 95% Wilson intervals with the batch design effect. The black curves are the rescaled form of Eq. (14), $\exp[-(\ln 2)x e^{s(x-1)}]$, for shape parameters $s = N_* \log \alpha = 0.2, 0.8$, and 2.2 spanning the fitted range. (b) The law-independent crossover width $w = [N(\text{SAR} = 0.2) - N(\text{SAR} = 0.8)]/N_*$ of every curve against its N_* , with standard errors; the widths differ significantly between model-task pairs.

system. For small system size N , the external field h dominates, and the system behaves as a paramagnet, aligning most spins in the correct direction ($s_i = +1$). As N increases, the disorder energy of the random couplings, which by construction scales as N^2 , eventually exceeds the field and entropic contributions that grow only linearly with N . Beyond a characteristic size N_* , the fully aligned configuration is no longer favored, leading to the accuracy degradation observed as the accuracy cliff.

Scaling collapse. A crossover governed by a single scale suggests plotting SAR against N/N_* . Figure 6(a) shows that the data of all model-task pairs then fall within a common band. Its scatter about a nonparametric master curve (RMS 0.068) equals the scatter about each curve's own fit (0.072) and the typical statistical error (0.071). This collapse should not, however, be read as evidence of a universal scaling function. Because N_* is fitted for each curve, every curve passes through $(1, \frac{1}{2})$ by construction, and the rescaled form of Eq. (14), $\text{SAR}(xN_*) = \exp[-(\ln 2)x \alpha^{N_*(x-1)}]$, retains one shape parameter $s = N_* \log \alpha$. Across our 19 curves, s ranges from 0.21 to 2.15, and the law-independent crossover width $w = [N(\text{SAR} = 0.2) - N(\text{SAR} = 0.8)]/N_*$ from 0.61 to 1.55 with standard errors of about 0.04 [Fig. 6(b)], the variation of w is highly significant ($\chi^2 = 9 \times 10^2$ for 18 degrees of freedom). Rescaling instead with N_* of the quadratic or of the stretched-exponential law gives the same scatter about the corresponding master curve (0.068), so the collapse does not discriminate between crossover laws either. The collapse therefore holds at the level of the statistical noise, but the fitted widths differ significantly between model-task pairs, so the master curve is common only approximately.

Correlation-error map. Using the fitted parameters of Eq. (14), we pin each model-task pair onto a correlation-error map. Each point corresponds to a pair $(\log \alpha, \log \beta_0)$, with the crossover scale N_* [the solution of $\text{SAR}(N_*) = 0.5$] shown as background contours. Within this parametrization, a smaller $\log \alpha$ means that errors accumulate less strongly along the output, and a smaller $\log \beta_0$ means a lower error rate for

executing a single unit task; since the two parameters are correlated in the fit and α is form dependent (Sec. III A), the robust content of the map is the position of each pair relative to the N_* contours. The mapping results grouped by model are shown in Fig. 7, and those grouped by task are shown in Fig. 8.

IV. IMPLICATION: DIVIDE-AND-CONQUER STRATEGY

The empirical scaling law of Eq. (14) not only provides a qualitative description of how the accuracy inevitably degrades with the output sequence length N for LLMs, but also offers a *constructive implication*: When $\alpha > 1$, the reliability of long-sequence generation can be improved by adopting a divide-and-conquer strategy. We emphasize that the prediction below follows directly from the empirical accumulation law [Eq. (14)]; it does not depend on the effective model of Sec. III B, and the experiment in this section is presented as a practical implication of the accumulation law, not as a test of the statistical-mechanical picture.

In this context, divide and conquer refers to partitioning a sequence-generation task of total length N into k subtasks of length N/k , prompting the model separately for each subtask, and then merging the partial outputs to form the complete sequence. The intuition is that the accuracy cliff originates from correlated error propagation along the output: When $\alpha > 1$, the effective per-position error rate grows exponentially with N . Dividing the sequence into shorter fragments resets this accumulation.

We introduce a multiplicative overhead factor $\theta_{N,k} \in (0, 1]$ to account for the reduction of SAR due to the additional operations involved in applying the divide-and-conquer strategy, such as segmentation, formatting, boundary handling, and content copying. The overall success probability under divide and conquer is then

$$\begin{aligned} \text{SAR}_{\text{DC}}(N, k) &= \theta_{N,k} [\text{SAR}(N/k)]^k \\ &= \theta_{N,k} \exp[-\beta_0 N \alpha^{\frac{N}{k}-1}]. \end{aligned} \quad (22)$$

Define the logarithmic gain of divide and conquer relative to the single-shot strategy as

$$\begin{aligned} \Delta(N, k) &:= \log \frac{\text{SAR}_{\text{DC}}(N, k)}{\text{SAR}(N)} \\ &= \log \theta_{N,k} + \beta_0 N (\alpha^{N-1} - \alpha^{\frac{N}{k}-1}). \end{aligned} \quad (23)$$

This formulation shows that for $\alpha > 1$, the gain term increases rapidly with N , indicating that the larger α is, the more advantageous segmentation becomes.

Theorem 1 (Advantage of divide and conquer). Let the intrinsic error rate $\beta_0 > 0$ and the correlation amplification factor $\alpha > 1$, and fix the integer $k \geq 2$. Under the empirical scaling law $\text{SAR}(N) = \exp[-\beta_0 N \alpha^{N-1}]$, for the segmentation into k equal subtasks to yield a positive gain $\Delta(N, k) > 0$, it would be sufficient if the sequence length exceeds the following bound:

$$N \geq N_{\text{DC}} = 1 + \frac{1}{\log \alpha} \left(\log \left(1 - \frac{2 \log \theta_{N,k}}{\beta_0} \right) + \frac{\log 2}{1 - 1/k} \right), \quad (24)$$

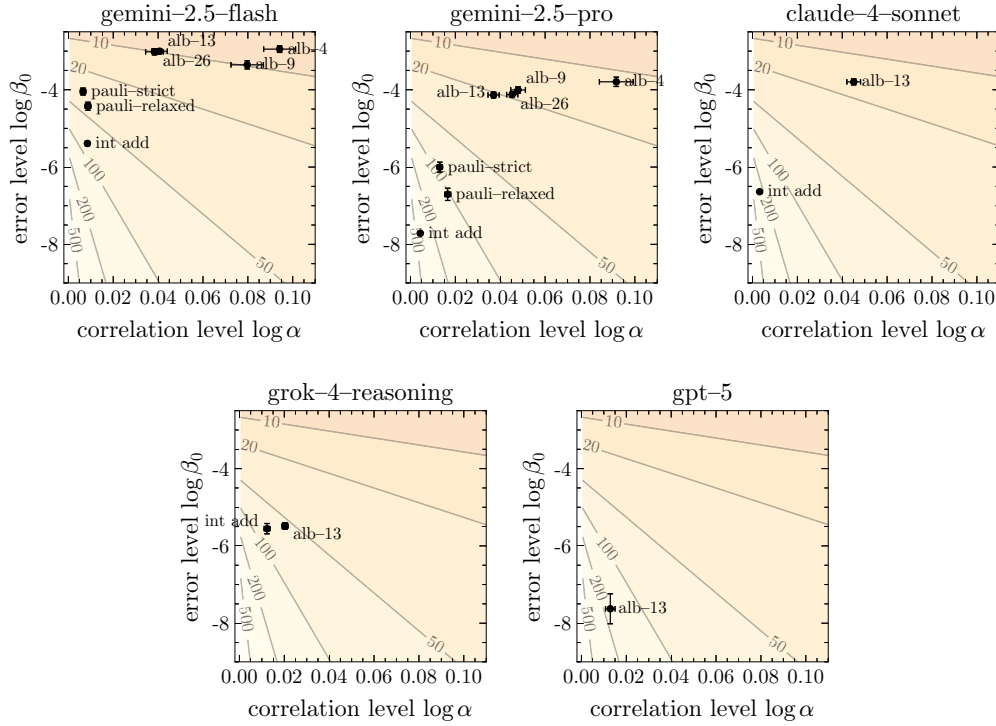


FIG. 7. Correlation-error map grouped by model. Each point is the maximum-likelihood $(\log \alpha, \log \beta_0)$ of Eq. (14) for one task, with 1σ error bars (quasiliquelihood); the horizontal and vertical axes represent the correlation level and error level, respectively. The background color and the gray contours give the crossover scale N_* implied by Eq. (14) at each point (numbers along the contours).

meaning that for all $N \geq N_{DC}$, one has $\text{SAR}_{DC}(N, k) > \text{SAR}(N)$. See Appendix D for a proof.

This result provides an explicit criterion for when the divide-and-conquer approach is beneficial. Under this strategy, as evident from Eq. (22), the intrinsic error rate β_0 remains unchanged, while the correlation amplification factor α is effectively rescaled to $\alpha_{\text{eff}} \simeq \alpha^{1/k}$ in the large- N regime. As a consequence, the characteristic accuracy-cliff scale is extended to

$$N_* \simeq 1 + \frac{\log(1/\beta_0)}{\log \alpha_{\text{eff}}} = 1 + k \frac{\log(1/\beta_0)}{\log \alpha}, \quad (25)$$

growing approximately linearly with k , according to Eq. (15).

Measurement. We test this implication on the Pauli string multiplication task, whose subproblems are exactly site local and whose recombination is unambiguous (Fig. 9); extending the test to the other task families is left to future work. Because every stage of the pipeline is recorded, the overhead is not a free parameter: From the same records, we measure the fraction of problems that were decomposed faithfully, the fraction whose k chunk products were all correct, and the probability that the final recombination is correct given correct chunk products, which is the overhead factor θ of Eq. (22). For gemini-2.5-pro, decomposition is faithful in 99% of the problems and $\theta = 0.998$ ($k = 2$) and 0.995 ($k = 3$) under the strict criterion; i.e., the overhead is negligible; for gemini-2.5-flash, decomposition is faithful in 98% ($k = 2$) but only 76% ($k = 3$) of the problems and $\theta = 0.82$ ($k = 2$) and 0.76 ($k = 3$), so the weaker model loses a substantial fraction of its correct chunk products in the recombination step. Under the phase-relaxed criterion, $\theta = 1.00$ for both models. The chunk stage itself is only semiquantitatively described by $[\text{SAR}(N/k)]^k$ with the single-shot fit. Under the

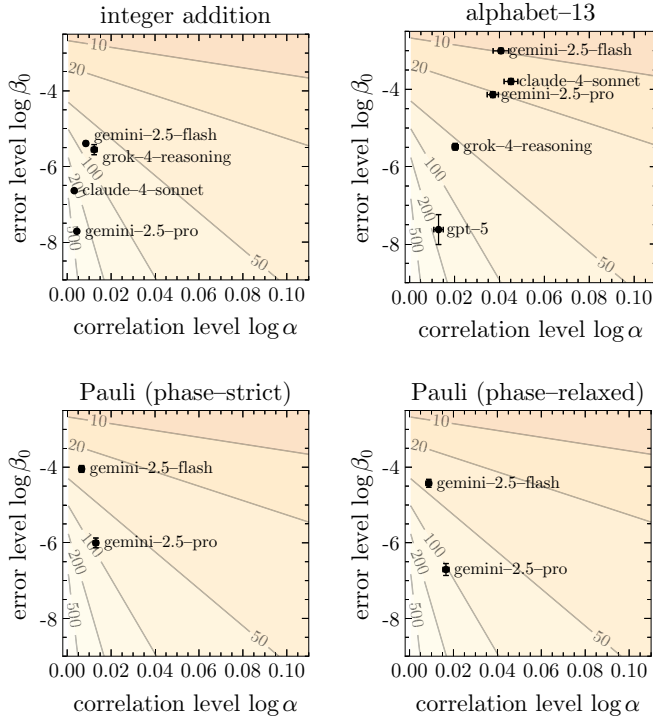


FIG. 8. Correlation-error map grouped by task; conventions as in Fig. 7.

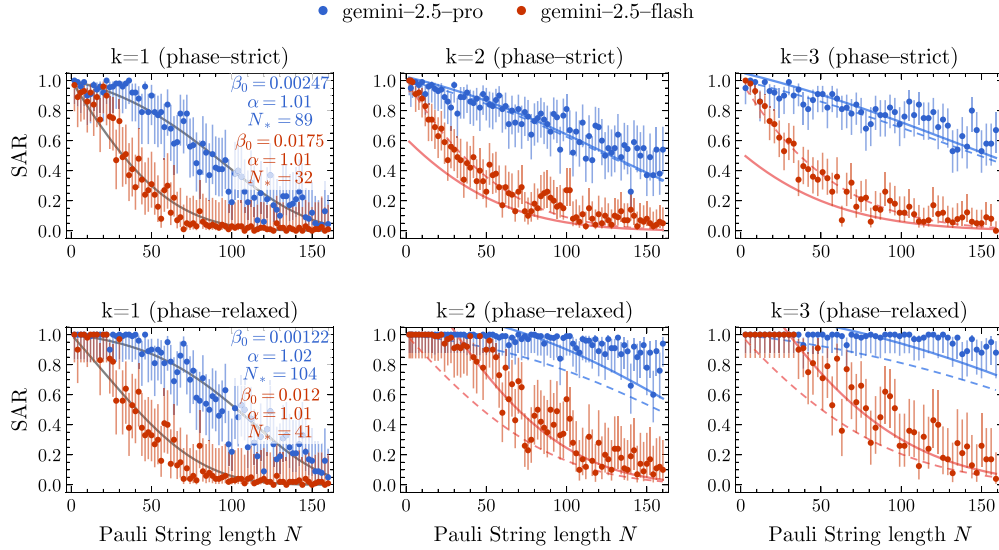


FIG. 9. Divide-and-conquer evaluation on the Pauli string multiplication task under the phase-strict (top) and phase-relaxed (bottom) criteria, using `gemini-2.5-pro` (blue) and `gemini-2.5-flash` (red). Points: fraction of correct instances with 95% intervals. The gray $k = 1$ curves are the maximum-likelihood fits of Eq. (14); the extracted (α, β_0) are then used to draw $\text{SAR}(N/k)^k$ for $k = 2$ and $k = 3$ without overhead (dashed) and multiplied by the overhead measured from the same records (solid; the product of the chunk-stage and recombination factors quoted in the text). For `pro`, the measured overhead is negligible and the data follow the prediction semiquantitatively; for `flash`, the recombination overhead ($\theta = 0.82, 0.76$) and a chunk-stage shortfall suppress the plateau under the strict criterion.

strict criterion, the measured probability that all chunks are correct exceeds this prediction by 3%–6% for `pro` and falls short of it by 25%–31% for `flash`, so Eq. (22) with a constant θ is a guide rather than a precise prediction. The observation that survives these caveats is the extension of the crossover scale. Reading N_* directly from the data (Fig. 9), `pro` moves from $N_* \approx 80$ ($k = 1$) to 136 ($k = 2$) and 155 ($k = 3$) under the strict criterion, and from 94 to beyond the measured range ($N > 160$) under the relaxed criterion; `flash` moves from 39 to 71 and 77 under the relaxed criterion, whereas under the strict criterion it shows no plateau for $k \geq 2$, consistent with its recombination overhead.

In summary, the empirical scaling law does more than describe the accuracy cliff; it also indicates how to mitigate it. Segmentation reduces the compounding of correlated errors and extends the *reliable length scale* of a sufficiently capable model by a factor comparable to k , at the price of an overhead that is model dependent and can be measured.

V. RELATED WORKS

Single-pass reasoning and capacity limits. Recent research has shown that even when models can locally apply correct reasoning steps, global accuracy collapses as reasoning chains lengthen. This phenomenon parallels information-theoretic capacity limits of single forward passes. The Fano-style upper bound framework [13] formalizes this collapse for multihop question answering, proving that once task complexity exceeds the model’s per-pass information capacity, accuracy must drop sharply—mirroring our empirical SAR crossover. Their work interprets the reasoning chain as a noisy communication channel, bounded by mutual information, while our SAR analysis characterizes the same breakdown empirically

and offers an effective statistical-physics description in which correlated error propagation along the output plays the central role. Together, they highlight the inevitability of reliability loss in single-pass reasoning under finite model capacity.

Prompting paradigms and decomposition. A wide range of prompting methods—Direct prompting, Chain-of-Thought (CoT) [8,16], ReAct [17], Plan-and-Solve [18], and Self-Ask [19]—operate in the *single-pass* regime, assuming the full reasoning chain fits within one generation. In contrast, *multicall* frameworks such as Self-Refine [20], InfoQA [13], and agentic systems for programming or reasoning [21–23] mitigate this bottleneck by decomposing reasoning into smaller, verifiable subcalls. Our SAR-based analysis provides a quantitative explanation of why such multicall methods succeed: By resetting the error-accumulation process, they effectively bound per-step complexity below the model’s capacity threshold, preventing the exponential or superexponential degradation seen in single-pass reasoning.

Self-checking, verification, and robust reasoning. Other orthogonal strategies—such as self-consistency sampling, debate-style verification, and tool-augmented reasoning—attempt to mitigate compounding errors by introducing redundancy or external validation [17,24–26]. SAR offers a unified diagnostic lens for such methods: Improved reliability corresponds to reduced intrinsic error rate β_0 , while better verification or decomposition corresponds to smaller accumulation factor α . Thus, SAR quantitatively links algorithmic interventions to empirical stability regimes.

VI. CONCLUSION AND FUTURE WORK

We introduced the SAR as a principled metric for quantifying reliability in deterministic reasoning tasks performed by large language models. Through three controlled

benchmarks—cyclic letter replacement, integer addition, and Pauli string multiplication—we demonstrated that SAR provides a quantitative view of how local inaccuracies compound with sequence length. Empirical results reveal a reliability crossover that is sharper than a simple exponential decay; the independent-error law is rejected on every model-task pair, which indicates correlated rather than independent-error accumulation. The scale N_* of the crossover is robustly determined, independently of the functional form used to fit it, while its shape varies moderately between model-task pairs.

To interpret this behavior, we introduced a minimal effective model in which task-position correctness variables interact through dense random couplings of Sherrington-Kirkpatrick type and compete with an external field set by the prompt. Evaluated by direct enumeration, the model reproduces superlinear error accumulation and an accuracy cliff qualitatively, and it provides a two-parameter description of each model-task pair in terms of an intrinsic error rate β_0 and an error-accumulation factor α . We regard it as a phenomenological picture of correlated error accumulation.

Future directions include extending SAR to stochastic reasoning pipelines, integrating it with self-verification or multical decomposition schemes, and applying the statistical-physics approach to study model scaling laws and architectural modifications. Two concrete next steps are (1) extending the divide-and-conquer test beyond Pauli string multiplication to the cyclic-letter-replacement and integer-addition tasks, (2) making the connection to network internals microscopic—for instance, by attempting to extract the effective couplings J_{ij} from attention patterns or testing whether controlled perturbations of attention sparsity shift the accumulation factor α in the predicted direction; and (3) studying a causal variant of the effective model with couplings only from earlier to later positions. More broadly, SAR provides a foundation for treating reasoning reliability as a measurable and scalable property, enabling systematic analysis and engineering of models that sustain compositional accuracy over increasing complexity.

ACKNOWLEDGMENTS

We would like to thank Chen Nie for the engineering support. We would like to acknowledge helpful discussion with Victor Albert and Hsin-Yuan Huang on [27]. W.H. additionally thanks Tianxiao Hu for helpful discussions.

DATA AVAILABILITY

The data that support the findings of this article are openly available at [28].

APPENDIX A: SUCCESS RATE ANALYSIS OF PAULI STRING TEST

Suppose an intellectual agent attempts to compute the product of two Pauli strings of length N . At each site, two types of operations are required:

(1) A Pauli multiplication step, where $P_i \times Q_i$ is reduced to a Pauli operator R_i and a local phase factor ϕ_i . Let the probability of correctly performing this step be p_σ .

(2) A phase multiplication step, where the accumulated phase is updated as $\Phi \mapsto \Phi \times \phi_i$. Let the probability of correctly performing a single phase update be p_ϕ .

We first model the accumulated phase as a Markov chain on the cyclic group $G = \{1, i, -1, -i\}$ with the state order $(1, i, -1, -i)$. At each step, the algorithm multiplies the current phase by one of the group elements, but due to noise it returns an incorrect element with probability p_ϕ , and the correct (true) element with probability $1 - p_\phi$. Hence, each of the three incorrect elements occurs with probability $\frac{p_\phi}{3}$. Let $\mathbb{1}$ denote the 4×4 identity matrix and J the all-ones matrix. When the intended multiplier is 1, the one-step transition kernel is the circulant matrix

$$K = (1 - p_\phi) \mathbb{1} + \frac{p_\phi}{3} (J - \mathbb{1}). \quad (\text{A1})$$

Define the four permutation (left-shift) matrices $\{S_\phi : \phi \in G\}$ that implement left multiplication by ϕ on the state order; explicitly,

$$S_1 = \mathbb{1}, \quad S_i = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$S_{-1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad S_{-i} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Thus, if the multiplier at a step is $\phi \in G$, the corresponding one-step transition matrix is

$$T(\phi) = S_\phi K. \quad (\text{A2})$$

Because every permutation matrix commutes with K , i.e., $S_\phi K = K S_\phi, \forall \phi \in G$, the product of stepwise transitions reduces to a similarity transform. For any sequence of true multipliers $\phi_1, \dots, \phi_N$, we have

$$T(\phi_N) \cdots T(\phi_1) = (S_{\phi_N} \cdots S_{\phi_1}) K^N \\ = S_\Phi K^N, \quad \Phi := \phi_N \cdots \phi_1. \quad (\text{A3})$$

Therefore, the success rate of phase computing task is given by

$$P_\phi(N) = e_\Phi^\top S_\Phi K^N e_1, \quad (\text{A4})$$

where e_Φ is the basis vector of Φ in the state order $(1, i, -1, -i)$ and $K^N = \frac{1}{4}J + \left(\frac{3-4p_\phi}{3}\right)^N (1 - \frac{1}{4}J)$. Explicitly, this gives

$$P_\phi(N) = \frac{1}{4} + \frac{3}{4} \left(\frac{3-4p_\phi}{3}\right)^N. \quad (\text{A5})$$

Combining this with the individual Pauli multiplication success rate, the overall sequence accuracy rate is

$$\text{SAR}(N) = (1 - p_\sigma)^N \left[\frac{1}{4} + \frac{3}{4} \left(\frac{3-4p_\phi}{3}\right)^N \right]. \quad (\text{A6})$$

The requirement to accumulate local phase factors across all sites is especially instructive, as it directly reflects the need for *memory handling* in agent design: Just as an LLM must correctly maintain and update a state variable across multiple

reasoning steps, the phase register here must be updated consistently at every site.

APPENDIX B: AN EFFECTIVE RANDOM-COUPLING MODEL FOR THE SAR

In this Appendix, we define and analyze the minimal energy-based model of Sec. III B: its perturbative expansion to leading order in the coupling (Appendix B 4), its direct evaluation by exhaustive enumeration (Appendix B 5), and the relation between the empirical estimator of SAR and the disorder averages used in the model (Appendix B 1).

Let $\mathcal{D} = \{(x_{(k)}, y_{(k)})\}_{k=1}^K$ be a dataset consisting of input-output sequence pairs, where each input sequence $x_{(k)}$ is deterministically mapped to a unique correct output sequence $y_{(k)}$. An LLM trained on this dataset defines a conditional probability distribution $p_\theta(y | x)$ over output sequences.

1. Definitions and independent-token baseline

The SAR is defined as the geometric mean of the model's probability on correct sequences:

$$\text{SAR}(p_\theta; \mathcal{D}) = \exp\left(\frac{1}{K} \sum_{k=1}^K \log p_\theta(y_{(k)} | x_{(k)})\right) \in (0, 1]. \quad (\text{B1})$$

Introduce Ising variables for correctness:

$$s_i = \begin{cases} +1, & y_i \text{ is correct,} \\ -1, & \text{otherwise,} \end{cases} \quad s = (s_1, \dots, s_N) \in \{\pm 1\}^N,$$

so all-correct output sequence corresponds to $s_i = +1$ for all $i = 1, \dots, N$, which we denote jointly as $s = \mathbf{1}$. If each token were generated independently with a token-level accuracy rate p_{tok} , the expected sequence accuracy would scale as

$$\text{SAR} = p_{\text{tok}}^N, \quad (\text{B2})$$

which decays exponentially with sequence length N . The corresponding energy model for the Ising variables is simply

$$E[s] = -h \sum_{i=1}^N s_i, \quad (\text{B3})$$

such that

$$\text{SAR} = \frac{e^{-E[s=\mathbf{1}]}}{\sum_s e^{-E[s]}} = \left(\frac{1}{1 + e^{-2h}}\right)^N, \quad (\text{B4})$$

implying that the external field h parametrizes the token-level accuracy as

$$p_{\text{tok}} = \frac{1}{1 + e^{-2h}}. \quad (\text{B5})$$

However, this independent-position prediction fails to capture the crossover behavior observed in experiments.

Empirical estimator versus model averages. The quantity measured in Sec. II is the fraction of problem instances whose complete answer is correct. Its expectation is the *arithmetic* instance average of the per-instance success probability, $\mathbb{E}_x p_\theta(y_x | x)$, where y_x is the unique correct answer to instance x . In the effective model, problem instances are represented by disorder realizations of the couplings, so the model counterpart of the empirical SAR is the arithmetic disorder

average $\text{SAR}_{\text{ari}} = \mathbb{E}_J p_J[s = \mathbf{1}]$. The replica calculation below is most naturally carried out for the geometric (“quenched”) average $\text{SAR}_{\text{geo}} = \exp(\mathbb{E}_J \log p_J[s = \mathbf{1}])$ of Eq. (20). The two averages coincide to leading order in J_0 (the disorder variance of $\log p_J$ enters only at $\mathcal{O}(J_0^2 e^{-4h})$) and whenever SAR is close to one; through the crossover, the arithmetic average is larger, because it is dominated by favorable realizations, and the two are compared numerically in Appendix B 5. Neither average is the quantity defined at the beginning of this subsection (a geometric mean over instances of the model's probability of the correct answer), which is a property of the language model's output distribution rather than of the binary correctness outcomes; we retain that definition only as motivation.

2. Energy model (dense random couplings of Sherrington-Kirkpatrick type)

In our experiments, the SAR shows a pronounced dependence on output length N . While one might expect a simple exponential decay, $\text{SAR} \sim e^{-\alpha N}$, our results reveal a clear *accuracy cliff*: SAR remains high for short sequences and then sharply declines around a characteristic length. This behavior indicates that sequence-level reliability in LLMs arises not from independent errors but from correlated effects governing the shift from reliable to unreliable generation in deterministic tasks. We hypothesize that the deviation arises from dense correlations among the output positions during sequence generation. Because attention couples tokens across the sequence, dependencies beyond the independent-position picture can arise; we capture these in an effective way, without claiming a specific microscopic origin. We represent them by effective random coupling strengths J_{ij} between the Ising variables s_i (the position-correctness labels), whose random signs encode constructive and destructive interference. Such couplings modify the collective statistics of sequence correctness, producing a nonexponential decay and the observed crossover in SAR.

We therefore propose the following energy model:

$$E_J[s] = - \sum_{i < j} J_{ij} s_i s_j - h \sum_{i=1}^N s_i, \quad (\text{B6})$$

where $s_i = \pm 1$ for $i = 1, 2, \dots, N$ are Ising variables representing the correctness of the i th position y_i of the output generated by the LLM.

The coupling strengths J_{ij} are drawn independently from a Gaussian distribution with zero mean and finite variance:

$$\mathbb{E}_J[J_{ij}] = 0, \quad \mathbb{E}_J[J_{ij}^2] = J_0^2. \quad (\text{B7})$$

They model the noisy all-to-all correlations between output positions, and J_0 characterizes the noise strength. The variance J_0^2 is taken independent of N (see the discussion in Sec. III B); with the conventional SK normalization J_0^2/N , the disorder energy would scale like the field energy and, as shown in Appendix B 5, no accuracy cliff would arise. The

external field h parametrizes the position-level accuracy rate via $p_{\text{tok}} = 1/(1 + e^{-2h})$: A strong field h corresponds to a high position-level accuracy, which is typically assumed. The energy function is that of the SK model of spin-glass physics, up to the normalization of the couplings.

For a particular random realization of J_{ij} , the probability of observing the Ising sequence s is

$$p_J[s] = \frac{1}{Z_J} e^{-E_J[s]}, \quad (\text{B8})$$

where Z_J is the partition function,

$$Z_J = \sum_s e^{-E_J[s]}. \quad (\text{B9})$$

In the perturbative analysis, we use the geometric disorder average of the probability of the all-correct configuration (Appendix B 1):

$$\text{SAR}_{\text{geo}} := \exp(\mathbb{E}_J \log p_J[s = \mathbf{1}]). \quad (\text{B10})$$

3. Replica trick and disorder average

From Eq. (B10), it is straightforward to show that

$$\begin{aligned} \log \text{SAR} &= \mathbb{E}_J \log p_J[s = \mathbf{1}] = \mathbb{E}_J \log \left(\frac{e^{-E_J[s=\mathbf{1}]}}{Z_J} \right) \\ &= \mathbb{E}_J(-E_J[s = \mathbf{1}] - \log Z_J) \\ &= \mathbb{E}_J \left(\sum_{i < j} J_{ij} + hN - \log Z_J \right). \end{aligned} \quad (\text{B11})$$

Using the disorder mean $\mathbb{E}_J[J_{ij}] = 0$, we have

$$\log \text{SAR} = hN - \mathbb{E}_J \log Z_J. \quad (\text{B12})$$

a. Replica trick

The expectation $\mathbb{E}_J \log Z_J$ in Eq. (B12) is intractable directly. Physicists therefore use the *replica trick*:

$$\mathbb{E}_J \log Z_J = \lim_{n \rightarrow 0} \partial_n \mathbb{E}_J[Z_J^n]. \quad (\text{B13})$$

By definition,

$$\begin{aligned} \mathbb{E}_J[Z_J^n] &= \mathbb{E}_J \left[\left(\sum_{[s]} \exp \left(\sum_{i < j} J_{ij} s_i s_j + h \sum_i s_i \right) \right)^n \right] \\ &= \sum_{\{s^a\}} \exp \left(\sum_{a=1}^n \sum_{i < j} J_{ij} s_i^a s_j^a + h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right), \end{aligned} \quad (\text{B14})$$

where the spin variables have been replicated to s_i^a with $i = 1, \dots, N$ and $a = 1, \dots, n$. Here, N is the number of sites (sequence length) and n the number of replicas.

b. Disorder average

Averaging Z_J^n over i.i.d. Gaussian couplings J_{ij} with $\mathbb{E}_J[J_{ij}] = 0$, $\mathbb{E}_J[J_{ij}^2] = J_0^2$, we obtain

$$\begin{aligned} \mathbb{E}_J[Z_J^n] &= \sum_{\{s^a\}} \prod_{i < j} \mathbb{E}_{J_{ij}} \exp \left(\sum_{a=1}^n J_{ij} s_i^a s_j^a \right) \exp \left(h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right) \\ &= \sum_{\{s^a\}} \prod_{i < j} \exp \left(\frac{J_0^2}{2} C_{ij}^2 \right) \exp \left(h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right), \end{aligned} \quad (\text{B15})$$

where $C_{ij} \equiv \sum_{a=1}^n s_i^a s_j^a$ and we used the Gaussian moment-generating function $\mathbb{E}_{J_{ij}} e^{J_{ij} C_{ij}} = \exp(\frac{J_0^2}{2} C_{ij}^2)$. Thus,

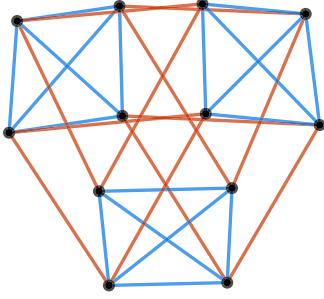
$$\begin{aligned} \mathbb{E}_J[Z_J^n] &= \sum_{\{s^a\}} \exp \left(\frac{J_0^2}{2} \sum_{i < j} C_{ij}^2 + h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right) \\ &= \sum_{\{s^a\}} \exp \left(\frac{J_0^2}{2} \sum_{i < j} \left(\sum_{a=1}^n s_i^a s_j^a \right)^2 + h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right). \end{aligned} \quad (\text{B16})$$

Expanding the quadratic replica term gives

$$\begin{aligned} \sum_{i < j} \left(\sum_{a=1}^n s_i^a s_j^a \right)^2 &= \sum_{i < j} \left(\sum_{a=1}^n (s_i^a s_j^a)^2 + 2 \sum_{a < b} s_i^a s_j^a s_i^b s_j^b \right) \\ &= \sum_{i < j} \left(n + 2 \sum_{a < b} s_i^a s_j^a s_i^b s_j^b \right) \\ &= \frac{nN(N-1)}{2} + 2 \sum_{i < j} \sum_{a < b} s_i^a s_j^a s_i^b s_j^b. \end{aligned} \quad (\text{B17})$$

Introducing the constant $\Omega \equiv \exp(J_0^2 N(N-1)/4)$ to absorb the $\mathcal{O}(n)$ term, we obtain the compact form

$$\mathbb{E}_J[Z_J^n] = \Omega^n \sum_{\{s^a\}} \exp \left(J_0^2 \sum_{i < j} \sum_{a < b} s_i^a s_j^a s_i^b s_j^b + h \sum_{a=1}^n \sum_{i=1}^N s_i^a \right). \quad (\text{B18})$$

FIG. 10. Example: $K_4 \times K_3$ graph.

c. Geometric interpretation

Equation (B18) defines a plaquette four-spin interaction on the graph product $K_N \times K_n$ (complete graphs on N and n vertices, respectively). The construction is self-dual under the exchange $N \leftrightarrow n$.

4. Small- J_0 expansion

Assuming that the noise strength J_0 is weak, we expand the replicated partition function in powers of J_0 :

$$\mathbb{E}_J Z_J^n = \Omega^n \sum_{\{s\}} e^{h \sum_{i,a} s_i^a} \left[1 + J_0^2 \sum_{i < j, a < b} s_i^a s_j^a s_i^b s_j^b + \mathcal{O}(J_0^4) \right]. \quad (\text{B19})$$

Because the Boltzmann factor factorizes over spins,

$$e^{h \sum_{i,a} s_i^a} = \prod_{i,a} e^{h s_i^a},$$

the spin summations can be carried out independently. For a single spin, we define

$$z := \sum_{s=\pm 1} e^{hs} = 2 \cosh h, \quad m := \langle s \rangle = \frac{1}{z} \sum_{s=\pm 1} s e^{hs} = \tanh h. \quad (\text{B20})$$

The term $\langle \sum_{i < j, a < b} s_i^a s_j^a s_i^b s_j^b \rangle = \frac{1}{4} N(N-1)n(n-1)m^4$ counts the number of plaquettes on the product graph $K_N \times K_n$, corresponding to choosing one edge from K_N and one from K_n independently (Fig. 10); each summand represents a single four-spin plaquette interaction. Hence,

$$\mathbb{E}_J Z_J^n = \Omega^n (2 \cosh h)^{nN} \left[1 + \frac{J_0^2}{4} N \times (N-1)n(n-1)m^4 + \mathcal{O}(J_0^4) \right]. \quad (\text{B21})$$

a. Replica limit

Taking the replica limit as in Eq. (B13),

$$\mathbb{E}_J \log Z_J = \log \Omega + N \log(2 \cosh h) - \frac{J_0^2}{4} N(N-1) \tanh^4 h + \mathcal{O}(J_0^4). \quad (\text{B22})$$

b. Result for SAR

Substituting Eq. (B22) into Eq. (B12) yields

$$\log \text{SAR}_{\text{geo}} = N(h - \log(2 \cosh h)) - \frac{J_0^2}{4} N(N-1)(1 - \tanh^4 h) + \mathcal{O}(J_0^4), \quad (\text{B23})$$

which is Eq. (21) [note $h - \log 2 \cosh h = -\log(1 + e^{-2h})$]. We have checked both coefficients against exact enumeration (Appendix B5): With a common set of Gaussian random numbers, $\mathbb{E}_J \log p_J[\mathbf{1}]$ is a smooth function of J_0 whose J_0^2 coefficient agrees with Eq. (B23) to within the statistical precision of the extraction (10%–15% for $N = 6$ –10, $h = 1$ –3). The $\mathcal{O}(J_0^4)$ coefficient, which involves five topological classes of plaquette pairs on $K_N \times K_n$, was quoted in an earlier version of this work; the numerical extraction shows that the expansion is asymptotic only for $J_0^2 N \ll 1$, well before the crossover, and we therefore do not rely on it and evaluate the model exactly instead.

For $h \gg 1$, using $\tanh h \approx 1 - 2e^{-2h}$, the leading contribution in e^{-2h} reads

$$\log \text{SAR}_{\text{geo}} \simeq -e^{-2h} N \left[1 + 2J_0^2(N-1) + \mathcal{O}(J_0^4) \right]. \quad (\text{B24})$$

The $J_0^2 N(N-1)$ contribution is the superlinear (correlation) term; its sign is fixed and positive in $-\log \text{SAR}$, so that disorder always accelerates the decay. To this order, the model predicts the quadratic law $\exp(-aN - bN^2)$ of Sec. III A. Exponentiating the relative correction,

$$\begin{aligned} \log \text{SAR} &\simeq -e^{-2h} N \left[1 + 2J_0^2(N-1) \right] \\ &\rightarrow -e^{-2h} N e^{2J_0^2(N-1)}, \end{aligned} \quad (\text{B25})$$

reproduces the form of the empirical law [Eq. (14)] with $\beta_0 = e^{-2h}$ and $\alpha = e^{2J_0^2}$. We stress that Eq. (B25) is a phenomenological closed form rather than a resummation of the perturbative series; the exact curves of Appendix B5 show that it places the crossover too early and makes it too sharp. Its usefulness is limited to motivating the two-parameter form of Eq. (14) and to interpreting β_0 and α qualitatively.

5. Direct finite- N evaluation

To assess the model beyond perturbation theory, we evaluate $p_J[s = \mathbf{1}] = e^{-E_J[\mathbf{1}]} / Z_J$ exactly, by exhaustive enumeration of all 2^N configurations for $N \leq 19$, and average over $M = 200$ realizations of the couplings. To obtain smooth curves, the same Gaussian random numbers are used for every N (the couplings for size N are the leading $N \times N$ block of one 19×19 realization), so that neighboring lengths share the disorder. Both the geometric average SAR_{geo} of Eq. (B10) and the arithmetic average $\text{SAR}_{\text{ari}} = \mathbb{E}_J p_J[\mathbf{1}]$ are recorded. Figure 11(a) shows the results for $(J_0, h) = (0.3, 2.5)$, $(J_0, h) = (0.5, 2.5)$, $(J_0, h) = (0.5, 3.5)$, $(J_0, h) = (0.8, 3.5)$, chosen so that the crossover falls inside the enumerable range; the shaded bands are the interquartile ranges of $p_J[\mathbf{1}]$ over realizations.

The exact curves display the qualitative phenomenology of the data: a plateau near $\text{SAR} = 1$ followed by a crossover whose scale is set by J_0 and h and is an order of magnitude

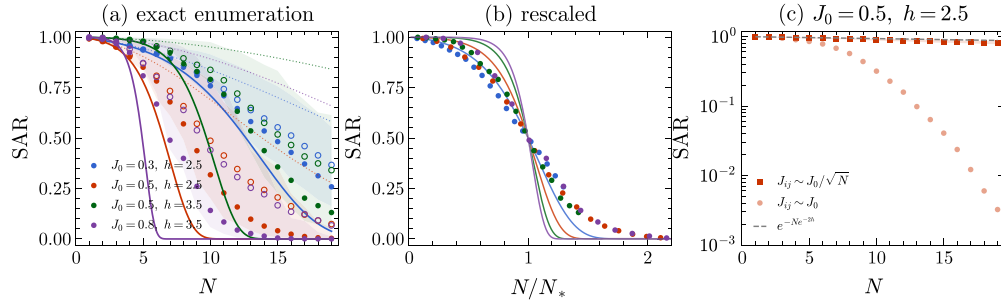


FIG. 11. Direct finite- N evaluation of the random-coupling model. (a) $\text{SAR}(N)$ from exhaustive enumeration ($N \leq 19, 200$ disorder realizations) for four parameter sets; filled symbols: geometric average $\exp \mathbb{E}_J \log p_J[\mathbf{1}]$, open symbols: arithmetic average $\mathbb{E}_J p_J[\mathbf{1}]$, and bands: interquartile range over realizations. Solid lines: the exponentiated closed form [Eq. (B25)] with the same (J_0, h) and no free parameter; dotted: the leading-order quadratic law [Eq. (B24)]. (b) The exact geometric curves against N/N_* , together with the rescaled closed form for the corresponding $s = 2J_0^2 N_*$. (c) Contrast for $(J_0, h) = (0.5, 2.5)$ between the N -independent coupling variance used in this work (circles) and the conventional SK normalization $J_{ij} \sim J_0/\sqrt{N}$ (squares, log scale): With the latter, $-\log \text{SAR}$ is linear in N and no cliff develops.

shorter than the independent-error scale $N_*^{(\alpha=1)} = \ln 2 e^{2h}$ (Table I). The crossover is well described by the two-parameter form [Eq. (14)] with effective parameters, but the effective accumulation factor α_{eff} is markedly smaller than $e^{2J_0^2}$ suggested by exponentiating the leading-order correction: The closed form [Eq. (B25)] places the crossover 15%–30% too early and makes it far too sharp, while the unexponentiated quadratic law [Eq. (B24)] is accurate only for N well below N_* and places the crossover too late. The rescaled exact curves [Fig. 11(b)] have widths $w = 0.55$ – 0.66 in units of N_* (for the three sets whose crossover is completed within $N \leq 19$); i.e., the model produces a crossover at the sharp end of the range $w = 0.6$ – 1.6 found for the LLM data; the broader crossovers of most model-task pairs are not reproduced by the model with a single (J_0, h) . The arithmetic average exceeds the geometric one through the crossover (by 0.07–0.2 at $N \simeq N_*$, growing with J_0) and its crossover scale is 10%–35% larger, but the two show the same qualitative behavior. Finally, Fig. 11(c) shows that with the conventional SK normalization $J_{ij} \sim J_0/\sqrt{N}$ the disorder merely renormalizes the exponential rate [$-\log \text{SAR}/N \simeq 0.0109$, compared with $\log(1 + e^{-2h}) = 0.0067$ without couplings] and no cliff develops; the superlinear accumulation is a consequence of the N -independent coupling variance assumed in Sec. III B, as stated there.

TABLE I. Crossover scales of the exactly enumerated model. N_*^{geo} , N_*^{ari} : from the geometric and arithmetic disorder averages; “closed”: the exponentiated form [Eq. (B25)]; “quad”: the leading-order quadratic law; $\alpha = 1$: independent errors, $\ln 2 e^{2h}$. α_{eff} is the accumulation factor of a fit of Eq. (14) to the exact geometric curve, to be compared with $e^{2J_0^2}$; w is the crossover width of the exact curve in units of N_* .

(J_0, h)	N_*^{geo}	N_*^{ari}	Closed	Quad	$\alpha = 1$	$e^{2J_0^2}$	α_{eff}	w
(0.3, 2.5)	14.7	16.2	12.6	21.7	103	1.20	1.108	–
(0.5, 2.5)	8.5	9.8	6.5	13.9	103	1.65	1.113	0.66
(0.5, 3.5)	13.4	15.3	9.7	38.5	760	1.65	1.155	0.55
(0.8, 3.5)	6.9	9.4	4.9	24.5	760	3.60	1.145	0.64

APPENDIX C: COMPARISON OF CROSSOVER LAWS

Table II lists, for every model-task curve, the quasi-AIC differences between the double-exponential law [Eq. (14)] and (1) its $\alpha = 1$ special case, (2) the quadratic law $\exp(-aN - bN^2)$, (3) the stretched exponential $\exp[-(N/N_0)^\gamma]$, and (4) the logistic crossover $[1 + e^{(N-N_0)/w}]^{-1}$. All laws are fitted by maximizing the binomial log-likelihood $\sum_N [k_N \log p(N) + (n_N - k_N) \log(1 - p(N))]$ of the instance counts; the quasi-AIC is $\text{QAIC} = 2P - 2 \log L/\hat{c}$, with P the number of fitted parameters and $\hat{c}$ the design effect of the curve (Sec. II), and a positive ΔQAIC means that the alternative fits worse than Eq. (14). Quasi-BIC gives the same ranking; the leave-one-length-out predictive likelihood agrees on the rejection of $\alpha = 1$ on every curve and selects the same best two-parameter law on 13 of the 19 curves, the stretched and quadratic laws exchanging places on the remaining six. The last column gives the crossover scale N_* of Eq. (14) together with the range spanned by the four two-parameter laws. The $\alpha = 1$ law is rejected on every curve ($\Delta \text{QAIC} \geq 35$); among the two-parameter laws, the stretched exponential has the lowest QAIC on 14 curves, Eq. (14) on 3 and the logistic on 2; the quadratic law is never the best but has a lower QAIC than Eq. (14) on 14 of the 19 curves. The parameters of Eq. (14), their standard errors, and the design effects for every curve are provided with the analysis code.

APPENDIX D: PROOF OF THEOREM 1

Proof. Start with the condition

$$N \geq N_{\text{DC}} = 1 + \frac{1}{\log \alpha} \left(\log \left(1 - \frac{2 \log \theta_{N,k}}{\beta_0} \right) + \frac{\log 2}{1 - 1/k} \right), \quad (\text{D1})$$

with $\alpha > 1$, $\beta_0 > 0$, $0 < \theta_{N,k} \leq 1$, $k \geq 2$. Each term in N_{DC} is non-negative,

$$\log \alpha > 0, \quad \log \left(1 - \frac{2 \log \theta_{N,k}}{\beta_0} \right) \geq 0, \quad \frac{\log 2}{1 - 1/k} > 0. \quad (\text{D2})$$

TABLE II. Model comparison for every model-task curve. n : number of instances; $\hat{c}$: batch design effect; ΔQAIC of the $\alpha = 1$ exponential, the quadratic, stretched-exponential, and logistic laws relative to Eq. (14) (positive: worse); N_* of Eq. (14) and its range over the four two-parameter laws.

Task	Model	n	$\hat{c}$	$\Delta\text{QAIC}_{\alpha=1}$	$\Delta\text{QAIC}_{\text{quad}}$	$\Delta\text{QAIC}_{\text{str}}$	$\Delta\text{QAIC}_{\text{log}}$	N_* (range)
Letter-4	gemini-2.5-pro	4900	1.9	+260	-20	-33	-21	12 (11-12)
Letter-4	gemini-2.5-flash	4580	1.0	+252	-88	-94	-110	7 (7-8)
Letter-9	gemini-2.5-pro	4900	1.7	+292	-33	-34	-9	17 (17-17)
Letter-9	gemini-2.5-flash	4890	1.6	+207	-35	-36	-15	10 (9-10)
Letter-13	gemini-2.5-pro	4900	1.5	+257	-36	-37	+7	21 (20-21)
Letter-13	gemini-2.5-flash	4880	1.3	+171	-83	-88	-72	10 (10-11)
Letter-26	gemini-2.5-pro	4900	1.2	+425	-44	-46	-19	19 (18-19)
Letter-26	gemini-2.5-flash	4900	1.8	+121	-52	-60	-44	10 (10-11)
Letter-13	gpt-5	3010	4.2	+64	+1	-1	+4	167 (148-167)
Letter-13	grok-4	8960	2.5	+270	+9	+11	+13	55 (53-55)
Letter-13	claude-4-sonnet	4900	1.5	+293	-79	-80	-49	16 (15-16)
Addition	gemini-2.5-pro	24830	1.2	+1599	+37	+4	+105	344 (332-344)
Addition	gemini-2.5-flash	8970	1.5	+391	-26	-30	+36	79 (75-80)
Addition	grok-4	2860	2.1	+137	+22	+30	+9	73 (68-73)
Addition	claude-4-sonnet	22400	1.0	+982	-48	-52	+163	244 (232-247)
Pauli (strict)	gemini-2.5-pro	7970	4.8	+147	-3	-4	-8	89 (84-89)
Pauli (relaxed)	gemini-2.5-pro	8000	5.2	+182	+2	-21	-4	104 (97-104)
Pauli (strict)	gemini-2.5-flash	7950	4.9	+35	-10	-17	-13	32 (32-37)
Pauli (relaxed)	gemini-2.5-flash	8000	5.9	+56	-22	-28	-16	41 (41-45)

Therefore, Eq. (D1) implies the following weaker conditions:

$$N > 1, \quad (\text{D3})$$

$$N > 1 + \frac{1}{\log \alpha} \log \left(1 - \frac{2 \log \theta_{N,k}}{\beta_0} \right) > 1 + \frac{1}{\log \alpha} \log \left(-\frac{2 \log \theta_{N,k}}{\beta_0} \right), \quad (\text{D4})$$

$$N > \frac{1}{\log \alpha} \frac{\log 2}{1 - 1/k}, \quad (\text{D5})$$

which are nevertheless more directly useful.

Equation (D5) implies

$$\begin{aligned} N(1 - 1/k) \log \alpha &> \log 2 \Rightarrow \alpha^{N(1-1/k)} > 2 \\ &\Rightarrow \alpha^{-N(1-1/k)} < 1/2 \\ &\Rightarrow (1 - \alpha^{N/k-N}) > 1/2. \end{aligned} \quad (\text{D6})$$

Equation (D4) implies

$$\begin{aligned} (N-1) \log \alpha &> \log \left(-\frac{2 \log \theta_{N,k}}{\beta_0} \right) \Rightarrow \frac{1}{2} \beta_0 \alpha^{N-1} \\ &> -\log \theta_{N,k}. \end{aligned} \quad (\text{D7})$$

Putting Eqs. (D3), (D6), and (D7) together, we have

$$\begin{aligned} \beta_0 N (\alpha^{N-1} - \alpha^{N/k-1}) &= \beta_0 \alpha^{N-1} N (1 - \alpha^{N/k-N}) \\ &> \beta_0 \alpha^{N-1} \times 1 \times \frac{1}{2} \\ &> -\log \theta_{N,k}. \end{aligned} \quad (\text{D8})$$

Therefore, the logarithmic gain $\Delta(N, k) = \log \theta_{N,k} + \beta_0 N (\alpha^{N-1} - \alpha^{N/k-1}) > 0$ is positive. ■

- [1] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, *et al.*, Holistic evaluation of language models, *Trans. Mach. Learn. Res.* (2023), <https://openreview.net/forum?id=iO4LZibEqW>.
- [2] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in *Text Summarization Branches Out* (Association for Computational Linguistics, Barcelona, Spain, 2004), pp. 74–81.
- [3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, edited by P. Isabelle, E. Charniak, and D. Lin (Association for Computational Linguistics, Philadelphia, PA, 2002), pp. 311–318.
- [4] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, Training verifiers to solve math word problems, [arXiv:2110.14168](https://arxiv.org/abs/2110.14168).
- [5] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, Measuring mathematical problem solving with the math dataset, in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, edited by J. Vanschoren and S. Yeung (Neural Information Processing Systems Foundation, 2021), Vol. 1.
- [6] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, *et al.*, Solving quantitative reasoning problems

- with language models, *Adv. Neural Inf. Process. Syst.* **35**, 3843 (2022).
- [7] A. Havrilla, S. C. Raparthy, C. Nalmpantis, J. Dwivedi-Yu, M. Zhuravinskiy, E. Hambro, and R. Raileanu, GloRe: When, where, and how to improve LLM reasoning via global and local refinements, in *Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 2024, Proceedings of Machine Learning Research* (PMLR, 2024), Vol. 235, pp. 17719–17733.
- [8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, *Adv. Neural Inf. Process. Syst.* **35**, 24824 (2022).
- [9] P. Astolfi, M. Careil, M. Hall, O. Mañas, M. Muckley, J. Verbeek, A. R. Soriano, and M. Drozdal, Consistency-diversity-realism Pareto fronts of conditional image generative models, [arXiv:2406.10429](https://arxiv.org/abs/2406.10429).
- [10] B. Lake and M. Baroni, Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks, in *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 2018, Proceedings of Machine Learning Research* (PMLR, 2018), Vol. 80, pp. 2873–2882.
- [11] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, Lost in the middle: How language models use long contexts, *Trans. Assoc. Comput. Linguist.* **12**, 157 (2024).
- [12] M. Turpin, J. Michael, E. Perez, and S. Bowman, Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting, *Adv. Neural Inf. Process. Syst.* **36**, 74952 (2023).
- [13] K. Wan, L. Gao, H. Mu, P. Nakov, Y. Wang, and X. Chen, A Fano-style accuracy upper bound for LLM single-pass reasoning in multi-hop QA, [arXiv:2509.21199](https://arxiv.org/abs/2509.21199).
- [14] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, Self-consistency improves chain of thought reasoning in language models, in *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda (OpenReview.net, 2023).
- [15] W. Hou, L. Zhou, H.-Y. Hu, Y. Chen, Y.-Z. You, and X.-L. Qi, PyCliffordMCP Prompt templates, run records, and analysis pipeline for the sequence accuracy rate benchmarks, GitHub, 2026, https://github.com/EdenCodeInc/PyCliffordMCP/tree/main/dev/pauli_string_multiplication.
- [16] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, Large language models are zero-shot reasoners, *Adv. Neural Inf. Process. Syst.* **35**, 22199 (2022).
- [17] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, React: Synergizing reasoning and acting in language models, in *The Eleventh International Conference on Learning Representations (ICLR 2023)*, Kigali, Rwanda (OpenReview.net, 2023).
- [18] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim, Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Toronto, 2023), Vol. 1, Long Papers, pp. 2609–2634.
- [19] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis, Measuring and narrowing the compositionality gap in language models, in *Findings of the Association for Computational Linguistics: EMNLP 2023* (Association for Computational Linguistics, Singapore, 2023), pp. 5687–5711.
- [20] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang, *et al.*, Self-refine: Iterative refinement with self-feedback, *Adv. Neural Inf. Process. Syst.* **36**, 46534 (2023).
- [21] S. Kim, S. Moon, R. Tabrizi, N. Lee, M. W. Mahoney, K. Keutzer, and A. Gholami, An LLM compiler for parallel function calling, in *Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 2024, Proceedings of Machine Learning Research* (PMLR, 2024), Vol. 235, pp. 24370–24391.
- [22] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges, *Vicinearth* **1**, 9 (2024).
- [23] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong *et al.*, Chatdev: Communicative agents for software development, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Bangkok, 2024), Vol. 1, Long Papers, pp. 15174–15186.
- [24] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths, Embers of autoregression: Understanding large language models through the problem they are trained to solve, *Proc. Natl. Acad. Sci. USA* **121**, e2322420121 (2024).
- [25] K. Wan, H. Mu, R. Hao, H. Luo, T. Gu, and X. Chen, A cognitive writing perspective for constrained long-form text generation, in *Findings of the Association for Computational Linguistics ACL 2025* (Association for Computational Linguistics, Vienna, Austria, 2025), pp. 9832–9844.
- [26] Z. Xie, R. Xing, Y. Wang, J. Geng, H. Iqbal, D. Sahnan, I. Gurevych, and P. Nakov, Fire: Fact-checking with iterative retrieval and verification, in *Findings of the Association for Computational Linguistics: NAACL 2025* (Association for Computational Linguistics, Albuquerque, New Mexico, 2025), pp. 2901–2914.
- [27] bench.science Benchmarking the future of science (accessed 3 October 2026), <https://bench.science/>.
- [28] W. Hou, L. Zhou, H.-Y. Hu, Y. Chen, Y.-Z. You, and X.-L. Qi, PyCliffordMCP, GitHub repository, 2026, <https://github.com/EdenCodeInc/PyCliffordMCP>.