

---

# Beyond Exponential Decay: Rethinking Error Accumulation in Large Language Models

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 A common pessimistic argument holds that autoregressive language models suffer  
2 exponential decay in correctness over long outputs: if each token has independent  
3 error probability  $e$ , then  $(1 - e)^n \rightarrow 0$  as  $n$  grows. The argument is clean to state  
4 and widely cited. It is also brittle, and three lines of recent empirical work make the  
5 cracks visible. The first is that only a small subset of tokens—roughly 5% to 10%  
6 in the studies that have actually measured it—genuinely depends on long-range  
7 context; the rest get more predictable, not less, as context accumulates. The second  
8 is geometric: LLM embeddings organize into stratified low-dimensional manifolds,  
9 so once a model is working inside one semantic region it tends to stay there  
10 even when individual tokens slip. The third concerns what happens when models  
11 do err on the consequential tokens—errors turn out to be idiosyncratic across  
12 samples rather than systematic, which is why majority-vote ensembles recover  
13 so much accuracy. Pulling these together gives a two-rate model,  $P(\text{correct}) \approx$   
14  $(1 - e_{\text{key}})^k \cdot (1 - e_{\text{non}})^{n-k}$ , in which  $k$  scales sublinearly with  $n$  and  $e_{\text{non}}$  approaches  
15 zero with sufficient context. The predicted decay is, at worst, stretched-exponential;  
16 often power-law; and when  $k$  saturates at some task-specific  $k_{\text{max}}$ , constant in  $n$ .  
17 A number of recent capabilities—anchor compression at 99% context reduction,  
18 128K-token retrieval on consumer GPUs, self-consistency gains on reasoning  
19 benchmarks—then read as natural consequences of one structural fact rather than  
20 independent engineering wins: long-context reliability hinges on a handful of  
21 decision points, not on uniform per-token accuracy.

## 22 1 Introduction

23 Autoregressive language models, the argument goes, are doomed for long outputs. If each token has  
24 even a 1% error rate then a 100-token chain is correct only  $(0.99)^{100} \approx 37\%$  of the time, and longer  
25 chains decay exponentially toward zero [LeCun, 2023, Dziri et al., 2023]. The argument is clean, and  
26 it has been used to predict a hard ceiling on what autoregressive systems can do.

27 What modern LLMs actually do is harder to square with this picture. Multi-page outputs hold  
28 together. Mid-generation self-correction is routine—later tokens revise earlier interpretations, a  
29 behavior Gwern [2023] called out as incompatible with monotonically increasing error. The attention  
30 patterns are lopsided: 96% of cumulative attention weight in Llama-3 concentrates on  $\sim 1\%$  of the  
31 context [Liu et al., 2024], and only about 9% of tokens in natural text show meaningful dependence  
32 on distant context [Fang et al., 2024]. With the right test-time strategy, a 1B-parameter model can  
33 match or beat a 405B model on specific reasoning tasks [Venture Research, 2025]. None of this is  
34 consistent with uniform per-token error.

Our reading is that the independent-error model fails because it averages over a heterogeneous population of tokens. A small fraction—key tokens—carry the load of long-range dependency and global coherence. The rest are constrained by local syntax and the accumulating context, and their error rate goes to zero rather than to a constant. Three threads of recent work make this concrete: Fang et al. [2024] on which tokens actually need long context, Li and Sarwate [2025] on the stratified-manifold organization of embeddings, and Wang et al. [2023] on the convergence of correct reasoning paths under sampling. Combined, they produce a refined model:

$$P(\text{correct}) \approx (1 - e_{\text{key}})^k \cdot (1 - e_{\text{non}})^{n-k}, \quad (1)$$

where  $k$  is the number of key tokens,  $e_{\text{key}}$  their error rate, and  $e_{\text{non}}$  the much lower error rate for the remaining  $n - k$ . Because  $k$  scales sublinearly with  $n$ —logarithmically, as a fractional power, or saturating at some  $k_{\text{max}}$ —and because  $e_{\text{non}} \rightarrow 0$  with context, the predicted decay is far gentler than  $(1 - e)^n$ .

**Contributions.** (1) A two-rate error model that distinguishes key from non-key tokens and makes the dependence on  $k(n)$  explicit (§3). (2) A synthesis of evidence from three recent research streams—attention sparsity, embedding geometry, and ensemble convergence—that supports each component of the model (§4). (3) An accounting of how this framework reframes existing systems-level results (anchor compression, retrieval-augmented attention, self-consistency, tool integration) as predictable consequences of a single structural fact rather than independent engineering wins (§5).

## 2 Related Work

**Error compounding.** The exposure-bias problem is old. Bengio et al. [2015] formalized it: models trained with teacher forcing must, at inference time, condition on their own potentially noisy output, so small errors can compound through state updates. LeCun [2023] extended the worry to autoregressive LLMs in the now-familiar form—if each token has error probability  $e$ , sequence-level correctness decays as  $(1 - e)^n$ . The bound is tight when independence really holds, and loose otherwise; transformers attending backward over their own output is exactly the case where it goes loose, and Gwern’s inner-monologue evidence [Gwern, 2023] of error rates that *decrease* at certain points in a sequence is hard to reconcile with strict monotonic accumulation. We keep the per-token decomposition but partition tokens by their long-range dependency.

**Long-context utilization.** Liu et al. [2023] documented the lost-in-the-middle effect—models disproportionately attend to material near the beginning and end of the prompt. The same skew was given a sharper edge by Fang et al. [2024], whose long-short difference (LSD) metric finds only  $\sim 9\%$  of tokens in natural text scoring  $\text{LSD} > 2$ . Liu et al. [2024] showed 96% of cumulative attention weight concentrating on  $\sim 1,000$  tokens out of 100,000 in Llama-3, and used the observation to enable 128K-token contexts on consumer GPUs. Anchor compression [Pang et al., 2024] reaches 99% context reduction with  $< 1.5\%$  accuracy loss. Each of these papers treats sparsity as an empirical fact about a particular system; we read them as direct evidence for  $k \ll n$ .

**Embedding geometry.** Li and Sarwate [2025] probed LLM embedding spaces with sparse mixture-of-experts and found a stratified-manifold structure: a union of low-dimensional submanifolds aligned with semantic domain. The story gets more concrete with Robinson et al. [2025], who distinguished signal from noise dimensions in token embeddings—perturbing signal dimensions reroutes outputs, perturbing noise dimensions does not. Viswanathan et al. [2024] link intrinsic dimensionality to prediction loss; high-confidence contexts collapse to lower-dimensional representations. And Gao et al. [2023], probing intermediate layers, recover the correct answer from the model’s internals more than 80% of the time even when the surface output was wrong. Compartmentalized representations, signal/noise separation, dimensional collapse under confidence: this is what lets the model absorb the small token-level slips without breaking coherence at the semantic level.

**Ensemble reasoning.** Self-consistency, in Wang et al. [2023]’s formulation, is almost embarrassingly simple: sample multiple reasoning paths, take the majority answer. The gain was +17.9 points on GSM8K, with no retraining. Tree-of-Thoughts [Yao et al., 2023] extends the idea with structured tree search at high-uncertainty branch points and pulls GPT-4 on Game-of-24 from 4% to 74%. Li et al. [2023] sort the effects by error type, separating systematic errors (knowledge gaps—similar

across samples) from idiosyncratic ones (variable across samples), and that distinction explains why ensembles work for reasoning but not for retrieval. The size of these effects is the part the independent-error model cannot absorb. If errors compounded uniformly, multiple samples would yield multiple failures, not a more accurate mode.

### 3 Theoretical Framework

We develop the two-rate model in three stages: the partition of tokens into key and non-key, the manifold dynamics that structure error correlation, and the redundancy gain that ensemble methods extract.

#### 3.1 Two-Rate Error Model

The independent-error argument assumes every token carries equal weight in the failure probability. It does not. Key tokens are the ones whose correctness genuinely depends on long-range context or global knowledge—factual claims, logical operators, points of co-reference, transitions between topics. Non-key tokens are governed by local regularity: syntax, frequent collocations, content already established by the surrounding text. Their error rate is small and decreases as more context accumulates.

In a sequence of length  $n$ , let  $k$  count the key tokens. The empirical estimate of  $k/n$  from Fang et al. [2024] is  $\sim 9\%$ ; adversarial-perturbation studies converge on a similar range [Morris et al., 2022]. We hypothesize that  $k$  grows sublinearly with  $n$ , possibly saturating at a task-specific  $k_{\max}$ : even a book-length argument operates within a finite knowledge frame requiring a bounded number of critical decisions. With per-token error rates  $e_{\text{key}}$  (large) and  $e_{\text{non}}$  (small, decreasing in context), Eq. 1 gives three regimes:

1. **Logarithmic key-token growth** ( $k \sim \log n$ ): polynomial decay  $n^{-c}$ , much slower than exponential.
2. **Fractional-power growth** ( $k \sim \sqrt{n}$ ): stretched-exponential decay.
3. **Saturating** ( $k \rightarrow k_{\max}$ ): reliability becomes constant in  $n$  once the key facts are committed.

The third regime is the strange one. Pang et al. [2024] achieved 99% context reduction with  $< 1.5\%$  accuracy loss—a result that requires  $k$  to be effectively bounded for the relevant tasks. There is no way to express that under exponential decay; the functional form forbids it.

#### 3.2 Stratified Manifold and Error Correlation

The  $(1 - e)^n$  form also assumes errors are independent across positions. They are not. The embedding space’s stratified structure [Li and Sarwate, 2025] introduces correlation, and we can read the dynamics off it directly. Treat the model’s hidden state as moving along a manifold  $M_C$  determined by the prevailing context. Most errors are minor—a synonym substitution, a grammatical slip, a momentary wobble that stays on  $M_C$  and has no downstream effect. Disruptive errors are different: a key-token mistake jumps the trajectory onto a different manifold  $M_{C'}$ , after which subsequent tokens cohere with the wrong commitment. The result is a fluent-but-wrong continuation. Errors cluster rather than scatter.

This correlation structure cuts the union-bound failure probability. For small  $k$ ,

$$P(\text{any disruptive error}) \leq k \cdot e_{\text{key}}, \quad (2)$$

which, when  $k \ll n$ , is substantially below  $1 - (1 - e)^n$ . The manifold structure also explains the compartmentalization Gao et al. [2023] document: intermediate layers preserve correct internal representations even when the surface output strays, because the hidden-state trajectory remains on the correct manifold even when token sampling does not.

#### 3.3 Self-Consistency and Redundancy Gain

When key-token errors are idiosyncratic across samples—different paths fail at different junctions—majority-vote ensembles can recover the correct answer. For  $m$  samples with error correlation

130  $\rho$ ,

$$e_{\text{key}}^{(\text{eff})} = f(\rho, m) \cdot e_{\text{key}}, \quad (3)$$

131 where  $f \rightarrow 1$  at  $\rho = 1$  (perfect correlation, no gain) and  $f \rightarrow e_{\text{key}}^{m-1}$  at  $\rho = 0$  (independent errors,  
132 exponential gain). Wang et al.’s self-consistency results [Wang et al., 2023] sit closer to the  $\rho = 0$   
133 regime than to  $\rho = 1$ , which is why a method that adds no parameters and no training data can lift  
134 GSM8K accuracy by 17.9 points: the gain is structural, not a property of the sampling temperature.

135 Costello et al. [2025] push this further: as a model’s Pass@1 accuracy improves through iterative  
136 self-training, the marginal gain from majority voting falls. Correct reasoning paths cluster on a narrow  
137 manifold; incorrect ones fan out. Ensemble methods only help when the model’s spread of guesses  
138 straddles the right region.

## 139 4 Empirical Evidence

140 Three predictions, three lines of evidence. The framework says  $k/n$  is small and stable; that token  
141 errors correlate via manifold geometry; and that correct reasoning paths converge while incorrect  
142 ones diverge. The supporting work for each comes from a different methodology, which is what  
143 makes the convergence interesting.

### 144 4.1 Key-Token Sparsity

145 The most direct measurement is Fang et al. [2024]’s LSD metric: 9% of tokens in natural text  
146 qualify as key ( $\text{LSD} > 2$ ), and perplexity restricted to those tokens correlates with downstream task  
147 performance at Pearson  $\approx -0.96$ . Perplexity on the other 91% tracks task success at essentially  
148 zero correlation. Adversarial robustness work, working from the opposite direction, lands in the  
149 same band—Morris et al. [2022] flip model decisions by perturbing 5%–10% of strategically chosen  
150 tokens, while random perturbation of much larger fractions does nothing. The 5%–10% figure recurs  
151 across methodologies built to measure different things; that is what one would expect if it were a  
152 structural property of natural language rather than a dataset artifact.

153 The systems-level corollary is sharp. Liu et al. [2024] reach 128K-token effective context on consumer  
154 GPUs by computing attention only over the top  $\sim 1,000$  tokens by attention mass, recovering  $> 90\%$   
155 of full-attention scores in the process. Anchor-LLM [Pang et al., 2024] compresses sequence  
156 information into a single token at 99% reduction with  $< 1.5\%$  accuracy loss. Neither method would  
157 work if token importance were uniform; both work because it is not.

### 158 4.2 Stratified Manifold Structure

159 Li and Sarwate [2025]’s sparse-MoE probe of frozen LLM embeddings finds a union of low-  
160 dimensional manifolds aligned with semantic domain (scientific text, dialogue, code), with mea-  
161 surably different intrinsic dimension across regions. Robinson et al. [2025]’s fiber-bundle analysis  
162 distinguishes regular-neighborhood tokens (manifold interior; perturbation has minimal effect) from  
163 irregular-neighborhood tokens (manifold junctions; perturbation reroutes generation). The mapping  
164 to our key/non-key partition is direct: irregular-neighborhood tokens are exactly the points where a  
165 small input change can move the trajectory between manifolds.

166 Viswanathan et al. [2024] link this to confidence: prompts that elicit low-intrinsic-dimensional  
167 representations correlate with lower prediction loss. The model’s state contracts as confidence rises—  
168 which is why the per-token error rate  $e_{\text{non}}$  should fall, not stay constant, as context accumulates.

169 Gao et al. [2023] probe intermediate layers and recover correct answers from  $> 80\%$  of cases where  
170 the final output was wrong. The information was preserved internally; only the surface form failed.  
171 This is the manifold-stays-correct behavior our framework predicts.

### 172 4.3 Convergent Reasoning Paths

173 Wang et al. [2023] showed that self-consistency lifts GSM8K by 17.9 points and SVAMP by 11.0,  
174 with no model change. The mechanism, on our reading, is that correct paths concentrate near a  
175 low-dimensional attractor while incorrect ones disperse—so majority voting picks out the attractor’s

mode. Yao et al. [2023]’s Tree-of-Thoughts moves GPT-4 on Game-of-24 from 4% to 74% by branching at high-uncertainty points specifically. Li et al. [2023] explain why these methods work for reasoning and not for knowledge retrieval: knowledge gaps produce systematic errors (every sample fails the same way); reasoning slips produce idiosyncratic ones (samples fail differently). The framework’s two error rates accommodate both:  $e_{\text{key}}$  rises uniformly when the model lacks the relevant fact, and ensembling cannot help.

A summary of how the systems-level results map onto the framework’s three pillars appears in Appendix B.

## 5 Practical Implications

The framework is not just a description of why  $(1 - e)^n$  is wrong. It also predicts where computation should go, how context should be managed, and what to evaluate.

**Sparse attention and context compression.** If  $k \ll n$ , attention does not need to be dense. Pang et al. [2024]’s anchor compression and Liu et al. [2024]’s top- $k$  attention are the existing demonstrations; Wu et al. [2024]’s dynamic KV-cache selection is another point on the same curve. These methods stop being clever tricks and become predictable: dense attention pays a quadratic cost to recover a sparse signal.

**Targeted compute at decision points.** Errors concentrate at manifold transitions, so compute should too. Moshkov et al. [2025]’s tool-integrated reasoning fires Python execution at high-entropy spans rather than uniformly. Adaptive computation time [Xin et al., 2023] lets confident tokens skip layers altogether, saving 40%–60% on SST-2 and TriviaQA. The adaptive-temperature decoding of Zhu et al. [2024] raises exploration at uncertain tokens and damps it elsewhere. Three different prescriptions, one underlying instruction: spend cycles where the manifold is about to fork.

**Strategic ensembles.** Wang et al. [2023]’s self-consistency, Yao et al. [2023]’s tree search, and Moshkov et al. [2025]’s GenSelect all exploit the convergence-of-correct-paths structure. The framework explains why they work—and predicts when they will not, namely when errors are systematic rather than idiosyncratic, as in pure knowledge-retrieval tasks [Li et al., 2023].

**Evaluation aligned with key tokens.** Plain perplexity averages over a population the model is trying to handle separately, which is why it underperforms as a predictor. Fang et al. [2024]’s LongPPL restricts perplexity to key tokens and lifts the correlation with downstream performance to  $r = -0.96$  from a near-zero baseline. Costello et al. [2025]’s success-plateau curves show what reliability looks like up close: sharp drops after extended plateaus—the staircase one would expect under our model, not the smooth exponential decay of the alternative.

A more speculative implication—modular reasoning architectures that route by manifold region, building on the alignment-not-scale results of Costello et al. [2025]—we defer to Appendix A.

## 6 Limitations

The framework is a synthesis, not a derivation. Three cautions belong on record. The two-rate model is descriptive:  $k$ ,  $e_{\text{key}}$ , and  $e_{\text{non}}$  are observable in principle but not yet jointly measured on a single benchmark, so the predicted decay regimes are arguments from supporting evidence rather than fitted curves. The manifold-structure account leans on Li and Sarwate [2025] and Robinson et al. [2025] for direct geometric evidence; both are recent, and replication on larger model scales would tighten the claim. The convergent-paths claim holds where Li et al. [2023] call errors idiosyncratic; we have no quantitative criterion for distinguishing the idiosyncratic from the systematic regime in advance, only the post-hoc observation that ensemble methods help in one and not the other.

## 7 Conclusion

The independent-error argument was always going to fail in one of two places. Either some tokens would matter more than others, breaking the per-token uniformity; or errors would correlate, breaking

the independence. As it turns out, both fail at once. The empirical literature has been steadily accumulating the evidence: 5%–10% of tokens carry the long-range dependency, embeddings stratify into manifolds that compartmentalize errors, and ensemble methods extract gains that any uniform-error model rules out.

The two-rate model in Eq. 1 is a small change to the algebra and a large change to the predicted behavior. Reliability depends on  $k$  key decisions, not on  $n$  tokens; and  $k$  scales sublinearly, often saturating. Modern LLMs hold coherence across thousands of tokens not because the underlying  $(1 - e)^n$  is being beaten by clever engineering, but because  $(1 - e)^n$  was the wrong functional form to begin with.

The systems-level consequences are already visible. Anchor compression, retrieval-augmented attention, tool-integrated reasoning, and self-consistency all gain explanatory unity once they are read as instances of a single principle: identify where the manifold forks, and put the compute there. Future work should make  $k(n)$  measurable on a fixed benchmark, derive tighter bounds from attention-pattern statistics rather than from heuristic union bounds, and connect token-level decay to the reasoning-step decay that bears more directly on agentic systems.

## References

- S Bengio, O Vinyals, N Jaitly, and N Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- C Costello, C Wells, E Grefenstette, and A Glaese. Think, prune, train, improve: Scaling reasoning without scaling models. *arXiv preprint arXiv:2504.18116*, 2025.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. Faith and fate: Limits of transformers on compositionality (2023). *arXiv preprint arXiv:2305.18654*, 2023.
- L Fang, Y Wang, Z Liu, C Zhang, S Jegelka, J Gao, B Ding, and Y Wang. What is wrong with perplexity for long-context language modeling? In *International Conference on Learning Representations (ICLR) 2025*, 2024.
- M Gao, S Zhou, and K Chang. When transformers know but don’t tell: Analyzing the know-but-don’t-tell phenomenon in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.
- B Gwern. Inner monologue demonstrates that error probability is not monotonic in sequence length. *Online discussion, LessWrong*, 2023.
- Yann LeCun. Why autoregressive language models are exponentially diverging and doomed. *LinkedIn post*, May 2023.
- X Li and A D Sarwate. Unraveling the localized latents: Learning stratified manifold structures in llm embedding space with sparse mixture-of-experts. *arXiv preprint arXiv:2502.13577*, 2025.
- Y Li, B Deng, and S Bengio. On the reliability of linguistic features for error prediction in llms. In *Proceedings of ACL 2023*, 2023.
- D Liu, Z Fang, S Li, and P Rai. Retrievalattention: Accelerating long-context llm inference via vector retrieval. *arXiv preprint arXiv:2409.10516*, 2024.
- 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. *arXiv preprint arXiv:2307.03172*, 2023.
- J Morris, E Lifland, Y Jin, and J Quinn. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of ACL 2022*, 2022.
- I Moshkov, D Hanley, I Sorokin, S Toshniwal, C Henkel, B Schifferer, W Du, and I Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

- 269 NVIDIA AI Research. Scaling laws drive smarter, more powerful ai. Technical report, NVIDIA  
270 Technical Blog, 2023.
- 271 J Pang, B Liu, Y Huang, Y Sheng, Q Wang, J Fu, and L Huang. Anchor-based large language models.  
272 In *Findings of ACL 2024*, 2024.
- 273 M Robinson, S Dey, and T Chiang. Token embeddings violate the manifold hypothesis. *arXiv*  
274 *preprint arXiv:2504.01002*, 2025.
- 275 Venture Research. How test-time scaling unlocks hidden reasoning abilities in small language models.  
276 Technical report, Technology Research Report, 2025.
- 277 K Viswanathan, Z Wang, L Yang, and A Anandkumar. The geometry of tokens in internal representa-  
278 tions of large language models. *Submitted to ICLR 2025*, 2024.
- 279 X Wang, J Wei, D Schuurmans, Q Le, E Chi, and D Zhou. Self-consistency improves chain of thought  
280 reasoning in language models. In *International Conference on Learning Representations (ICLR)*,  
281 2023.
- 282 Wei Wu, Zhuoshi Pan, Chao Wang, Liyi Chen, Yunchu Bai, Kun Fu, Hui Zhang, Yu Liu, and Hui  
283 Xiong. Tokenselect: Efficient long-context inference and length extrapolation for LLMs via  
284 dynamic token-level KV cache selection. *arXiv preprint arXiv:2411.02886*, 2024.
- 285 J Xin, Y Song, L Cao, and D Yu. Adaptive computation time for transformers via early-exit  
286 mechanisms. In *Proceedings of ACL 2023*, 2023.
- 287 S Yao, J Zhao, D Yu, N Du, I Shafran, K Narasimhan, and Y Cao. Tree of thoughts: Deliberate  
288 problem solving with large language models. In *Advances in Neural Information Processing*  
289 *Systems 36*, 2023.
- 290 Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Jia Li, Zhi Jin, and Hong Mei. Hot or cold? adaptive  
291 temperature sampling for code generation with large language models. In *Proceedings of the AAAI*  
292 *Conference on Artificial Intelligence*, 2024. arXiv:2309.02772.

## 293 A Architectural Implications: Modular Reasoning

294 The stratified-manifold view suggests an architectural prescription the main body only gestures at:  
295 instead of scaling monolithic models, route reasoning subtasks to specialized models aligned with  
296 manifold regions. The clearest existing evidence comes from alignment-not-scale work. Costello et al.  
297 [2025]’s Trace-Prune-Train pipeline shows that smaller models (2–9B parameters) iteratively fine-  
298 tuned on their own pruned reasoning traces can match models 30× larger on GSM8K—improving  
299 Gemma2-2B from 41.9% to 57.6% Pass@1, Gemma2-9B to 82% (matching LLaMA-3.1-70B), and  
300 LLaMA-3.1-70B to 91% (above GPT-4o’s 82%). Whether explicit routing on top of these aligned  
301 smaller models compounds the gains, or runs into a ceiling once the routed-to manifold is itself  
302 stratified, is the natural next question; we are not aware of a definitive empirical answer.

303 The pattern across these systems is consistent. Fitting the manifold beats expanding the parameter  
304 count, when the task population is narrow enough that the manifold is well-defined. The open  
305 question is how broad a domain a single specialized model can cover before its internal stratification  
306 reasserts itself and the routing problem reappears one level down. We do not attempt to settle that  
307 here; it is a question for empirical work that systematically varies routing granularity.

## 308 B Extended Case Studies in Advanced Reasoning Systems

309 The main body summarizes the systems-level evidence at a high level. Three case studies bear closer  
310 reading.

311 **AIMO-2 (Moshkov et al. [2025]).** The winning solution to the AI Mathematical Olympiad com-  
 312 bines all three pillars of the framework. Tool integration fires at high-entropy spans, addressing  
 313 key-token uncertainty directly; the dataset-creation pipeline exploits the structured nature of correct  
 314 reasoning paths to generate training data; GenSelect leverages the convergent-paths property to  
 315 pick the best candidate solution from many. The system reaches state-of-the-art on AIME and  
 316 Harvard-MIT Mathematics Tournament problems. Most striking, the GenSelect mechanism works  
 317 on *compressed summaries* of reasoning traces, which would be impossible if errors were spread  
 318 uniformly across the full output: the summaries would lack the discriminating signal.

319 **Test-time compute scaling (NVIDIA AI Research [2023], Venture Research [2025]).** Optimal  
 320 test-time strategy depends on problem difficulty and model size. Beam search dominates best-of-  
 321  $N$  on hard problems for sub-7B models; the order reverses on easier problems for larger models.  
 322 The framework reads this as a direct consequence of  $k$  varying with task: harder tasks have more  
 323 decision points to traverse correctly, rewarding search-based exploration; easier ones have few,  
 324 rewarding sampling-based ensembling. The two-orders-of-magnitude result—an optimized 1B model  
 325 outperforming a 405B model on specific reasoning tasks [Venture Research, 2025]—is exactly what  
 326 an independent-error model rules out and what the two-rate model predicts under saturation.

327 **Attention concentration in production models (Liu et al. [2024]).** Llama-3 places 96% of  
 328 cumulative attention weight on  $\sim 1,000$  tokens out of 100,000. RetrievalAttention exploits this  
 329 directly: at generation time, only the top-mass tokens enter attention, with the rest treated as a  
 330 nearest-neighbor lookup. The result is 128K effective context on a single RTX 4090. The structural  
 331 claim is the same one  $k \ll n$  encodes; the engineering is just paying attention to it.

332 A consolidated summary of how each result maps onto the framework’s three pillars appears in  
 333 Table 1.

Table 1: How systems-level results map onto the framework’s three pillars.

| Challenge             | Independent-error view                   | Two-rate view                          | Exemplar systems                                                            |
|-----------------------|------------------------------------------|----------------------------------------|-----------------------------------------------------------------------------|
| Long-context handling | Uniform attention over all tokens        | Sparse retrieval focused on key tokens | Anchor LLMs [Pang et al., 2024]; RetrievalAttention [Liu et al., 2024]      |
| Compute allocation    | Equal resources for all tokens           | Targeted at manifold transitions       | Tool integration [Moshkov et al., 2025]; ACT [Xin et al., 2023]             |
| Error reduction       | Independent samples, multiplicative gain | Branching at uncertain junctions       | Self-consistency [Wang et al., 2023]; GenSelect [Moshkov et al., 2025]      |
| Evaluation            | Uniform perplexity                       | Key-token-restricted metrics           | LongPPL [Fang et al., 2024]; success-plateau curves [Costello et al., 2025] |
| Architecture          | Scale monolithic models                  | Alignment over scale                   | TPT [Costello et al., 2025]                                                 |

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and §1 state three contributions: (i) the two-rate error model of Eq. 1 that distinguishes key from non-key tokens; (ii) a synthesis of empirical evidence from attention-sparsity, embedding-geometry, and ensemble-convergence work that supports each component of the model; (iii) an accounting of how existing systems-level results follow from the model. (i) is developed in §3; (ii) in §4; (iii) in §5 and Appendix B. The paper makes no empirical claims of its own and does not promise new state-of-the-art numbers; it is a synthesis-and-framework contribution and is scoped as such in §6.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: §6 states three specific limitations: (i)  $k$ ,  $e_{\text{key}}$ , and  $e_{\text{non}}$  are observable in principle but not yet jointly measured on a single benchmark, so the predicted decay regimes are arguments from supporting evidence rather than fitted curves; (ii) the manifold-structure account leans on Li and Sarwate [2025] and Robinson et al. [2025] for direct geometric evidence, both recent and not yet replicated at larger scales; (iii) the convergent-paths claim has no a-priori criterion for distinguishing the idiosyncratic from the systematic regime—only the post-hoc observation that ensembles help in one and not the other.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper develops a descriptive two-rate error model, not formal theorems. The assumptions are stated explicitly: (i)  $k \ll n$  and  $k$  scales sublinearly with  $n$  (§3, supported by the  $\sim 5\%$ – $10\%$  key-token estimates of Fang et al. [2024], Morris et al. [2022]); (ii)  $e_{\text{key}} \gg e_{\text{non}}$  and  $e_{\text{non}} \rightarrow 0$  with context (supported by Viswanathan et al. [2024]); (iii) error correlation via stratified-manifold structure (supported by Li and Sarwate [2025], Robinson et al. [2025]). Eq. 1 follows from the partition of tokens; the union-bound disruptive-error inequality follows from  $k$ -fold subadditivity. Both are derived inline in §3. There are no formal theorems requiring detached proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper is a synthesis of published empirical results combined with a descriptive theoretical framework; it runs no new experiments and trains no new models. Reproducibility of the cited works is the responsibility of those works' authors; we provide pinned citations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper releases no new datasets, models, or code. It is a synthesis paper; all underlying empirical results are owned by their respective authors and accessed through the cited references.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification: No model training, evaluation runs, or hyperparameter selection is performed by this paper. The experimental settings of cited works are documented in those works.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: The paper reports no original experimental results, hence no error bars. Where we cite numerical results from the literature (e.g., the +17.9-point GSM8K gain, the 99% context-reduction, the  $-0.96$  Pearson correlation), we cite the originating work.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: The paper performs no original computational experiments, so no compute disclosure is required for it. Compute requirements of the cited systems are documented in those works (e.g., Liu et al. [2024] reports an RTX 4090 sufficing for 128K-token effective context).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics. The work is theoretical and synthesis; it involves no human subjects, no scraped or sensitive data, no model training, and no deployed system. There are no fairness, privacy, or security risks specific to the paper.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: §5 discusses the framework’s positive implications: more efficient long-context handling, targeted compute allocation, and modular reasoning architectures. The negative-impact direction is structurally limited because the paper releases no model or dataset; the framework, like any reliability argument, could be misread as a license to deploy LLMs in higher-stakes settings than current evidence justifies, which we flag implicitly via the limitations section (§6).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No new pretrained models, generative artifacts, or scraped datasets are released. The contribution is a theoretical framework and a synthesis of cited work.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper uses no datasets, code, or models—it cites published work but does not use any artifact in a way that requires license disclosure. Citations follow NeurIPS conventions.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new assets are released by this paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects research is conducted, so no IRB approval is required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing assistance (drafting, polishing, and copy-editing) and not as part of the research methodology. The framework, the synthesis, and the choice of cited evidence are the authors'; LLMs were not involved in selecting evidence, deriving the model, or shaping the argument. Per the question's stated guidance, no declaration is required for editing-only usage; we answer [N/A] accordingly.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.