

More Test-Time Compute Can Hurt: Overestimation Bias in LLM Beam Search

Gal Dalal^{1*}, Assaf Hallak^{1*}, Gal Chechik^{1,2}, Yftah Ziser^{1,3}
¹NVIDIA Research, ²Bar-Ilan University, ³University of Groningen
*Equal contribution

Abstract

Wider beam search should improve LLM reasoning, but when should you stop widening? Prior work on beam width selection has focused on inference efficiency (Qin et al., 2025; Freitag and Al-Onaizan, 2017), without analyzing whether wider search can *hurt* output quality. We study each search step as scoring n candidates and keeping the top b of them. Best-of- n keeps a single winner ($b = 1$); beam search keeps several. Because the scorer is noisy, a worse candidate can outrank a better one, and that risk grows as more candidates are compared. How far you can usefully widen therefore depends on the scorer: how large the true quality gap is relative to the noise. Keeping several hypotheses is easier than naming one winner, so beam search is more forgiving than best-of- n , but that per-step analysis does not by itself tell you the best width for the full task. We compare perplexity scoring and a process reward model (PRM) across three 7–8B models and all ten MR-BEN subjects (5,737 questions). Perplexity yields no accuracy gain at any tested width; the PRM gains up to 8.9 points and continues improving through beam width 4 ($n = 16$ candidates per step). Across 444K search steps, the spacing of ranked scores agrees with a simple Gaussian ranking model. The same model and algorithm therefore support very different useful widths, depending on how reliable the scorer is.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable progress on complex reasoning tasks, in large part due to techniques that elicit and leverage step-by-step solutions (Wei et al., 2022). A major recent thread is *process supervision*: instead of judging only the final answer, a model (or learned verifier) provides feedback on intermediate steps, enabling search and credit assignment over reasoning traces. In structured domains such as

mathematics, Process Reward Models (PRMs) that score each step have proven particularly effective (Lightman et al., 2023; Cui et al., 2025).

At the same time, there is growing interest in *test-time compute scaling*: allocating additional inference-time computation to improve accuracy via sampling and search (Snell et al., 2024). Common instantiations include best-of- n sampling, reranking, and beam search over partial chains of thought. A pervasive implicit assumption is that *any* reasonable scoring signal can drive such search: generate more candidates, score them, and pick the best. This assumption has not been rigorously examined.

A natural candidate for a cheap scoring signal is *perplexity*, a model’s measure of predictive confidence. Several intrinsic signals have been proposed for self-evaluation, from entropy-based self-certainty (Kang et al., 2025) to Markovian reward adjustments (Ma et al., 2025). We focus on perplexity as the most widely used and simplest such signal. It requires no step-level training and is available for any autoregressive model, making it a representative test of whether training-free intrinsic signals can drive reasoning-time search. If perplexity sufficed, one could avoid the expense of training specialized PRMs. Yet it is unclear whether perplexity is informative enough at the granularity of intermediate reasoning steps, or whether its noise causes search to select fluent but unfaithful steps.

A related overestimation phenomenon has been observed in model-based reinforcement learning, where tree search with learned value functions can degrade policy performance (Dalal et al., 2021). Our analysis addresses the distinct setting of LLM beam search, where the scorer is a perplexity model or PRM and the search operates over natural language reasoning traces. We frame model selection as retaining the top b of n noisy candidates. This exposes a single order-statistic mechanism underlying both best-of- n selection ($b = 1$) and beam search

Perplexity vs PRM Scoring: Effect of beam width

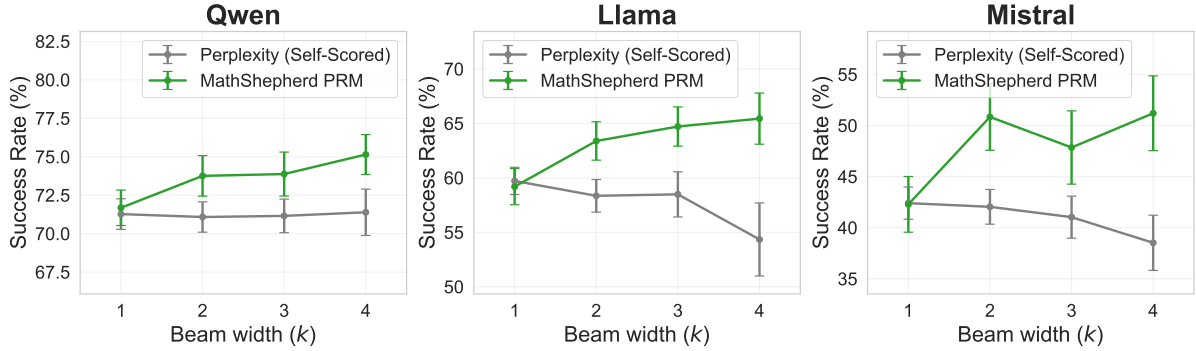


Figure 1: Per-model comparison of perplexity (gray) and PRM (green) scoring across beam widths. At width k the search scores $n = k^2$ candidates per step, so $k = 4$ is a pool of 16. Perplexity-guided performance is flat or declines, whereas all three PRM-guided models achieve their best tested mean at $k = 4$. The same models and search procedure therefore exhibit substantially different useful-width behavior under the two scorers.

($b = k$): incorrect candidates receive a selection-induced score bonus whose magnitude grows with search width and scorer noise. From this we derive a top-1 *selection-reliability* threshold and a top- b *retention-gap* threshold. They share the same machinery. Neither is an end-task useful-width law: both condition on a designated target already being in the pool.

Our study addresses three questions, each answered by a corresponding contribution:

1. **What limits search width?** We develop a unified order-statistic analysis of noisy selection. Retaining b of n candidates yields a conservative per-step bound determined by the expected b -th largest incorrect score. When $b = 1$, this score is the *maximum* over the incorrect candidates, recovering the Extreme-Value-Theory analysis of best-of- n selection and its selection-reliability threshold. Beam search instead retains multiple hypotheses: with $b = k$ and $n = k^2$, a correct path is pruned only if at least k incorrect candidates outrank it, so the relevant quantity is the k -th largest incorrect score rather than the maximum. These thresholds characterize per-step selection and retention reliability, not the end-to-end optimal beam width.
2. **Does scorer quality change the answer?** We validate the theory by comparing perplexity and PRM scoring across three model families and ten reasoning domains (Figure 1). Perplexity scoring yields no benefit at any beam width, while PRM scoring yields meaningful

gains.

3. **How should the beam width be chosen in practice?** We propose a principled approach to beam width selection based on diagnostic indicators such as score margins and pilot-run comparisons that signal whether a scorer can support wider search. While prior work adjusts beam width for inference speed (Qin et al., 2025), our criterion targets output quality. Our results emphasize improving scorer quality rather than widening the beam indiscriminately.

2 Background and Related Work

2.1 Beam Width Selection

Prior work on beam width selection has focused on inference efficiency. Freitag and Al-Onaizan (2017) introduced adaptive pruning strategies for neural machine translation, achieving up to 43% speedup without degrading quality. Qin et al. (2025) proposed Dynamic-Width Speculative Beam Decoding, dynamically adjusting beam count to balance latency and quality in LLMs. These approaches treat beam width as an *efficiency* parameter. We address a complementary question: when does wider search *degrade* output quality? The answer depends on the scorer’s signal-to-noise ratio. Top-1 selection has an explicit overoptimization threshold; top- k retention is more forgiving, but end-task useful width still has to be read off empirically.

2.2 Process Supervision for LLM Reasoning

Chain-of-Thought prompting showed that step-by-step reasoning improves LLM performance (Wei et al., 2022; Chen et al., 2022). Process Reward Models (PRMs) extend this by scoring intermediate steps (Lightman et al., 2023; Zheng et al., 2025), with Math-Shepherd (Wang et al., 2024) reducing annotation costs via Monte Carlo estimation. Cobbe et al. (2021) showed that best-of- N selection with a trained verifier matches a $30\times$ parameter increase.

PRMs face reward hacking and distributional shift (Zheng et al., 2025; Cui et al., 2025). Agrawal et al. (2025) show that PRM false positives impose an asymptotic ceiling on best-of- N performance, and Luo et al. (2025) establish that implicit Q-functions learned during fine-tuning produce over-optimistic estimates that degrade beam search. Both findings are consistent with the overestimation mechanism we formalize in Section 3.

2.3 Inference-Time Search

“Test-time compute scaling” allocates additional inference computation to improve performance (Snell et al., 2024). Wu et al. (2025) show that smaller models with advanced search achieve Pareto-optimal cost-performance trade-offs, and Liu et al. (2025) demonstrate that optimal strategies depend on the policy model, PRM, and problem difficulty. The success of models like DeepSeek-R1 (DeepSeek-AI, 2025) further motivates understanding these limits. These studies establish that more search *can* help. Our analysis characterizes two per-step reliability thresholds, selection and retention, rather than an end-task useful-width law. Tree of Thoughts (Yao et al., 2023) frames reasoning as search over a tree of partial solutions. Beam search and best-of- N sampling are common strategies in this paradigm (Xie et al., 2023; Kang et al., 2025), relying on a scoring function such as a PRM or a proxy like perplexity (Jelinek et al., 1977).

A parallel line of work studies *how* to spend a fixed inference budget rather than when extra candidates become harmful. Wang et al. (2025) allocate rollouts across reasoning directions instead of individual candidates. Light et al. (2025a) dynamically decompose traces so that harder steps receive finer sampling, and Light et al. (2025b) treat code generation as black-box search over solution space. These methods improve the organization of search; they do not identify when a

noisy scorer makes additional candidates harmful. Closest in spirit, Stroebel et al. (2026) show that imperfect verifiers impose a ceiling on resampling, with optimal sample counts often small. Our setting is complementary: we study ranking under noisy process scores rather than accept/reject against a verifier, and we separate noisy argmax (selection reliability) from noisy top- k pruning (retention). Komiyama et al. (2026) analyze the majority-vote limit of best-of- N ; our $\hat{n}_{\text{top-1}}$ is a selection-reliability threshold, not a prediction of where best-of- N stops helping once generation is modeled.

2.4 Overestimation Bias and Reward Overoptimization

Systematic overestimation from maximization over noisy estimates was first identified by Thrun and Schwartz (1993) in Q-learning, where function approximation combined with the max operator produces upward bias. Dalal et al. (2021) extended this to tree search with learned value functions, showing that deeper search can degrade policy performance in model-based RL. In the LLM setting, Gao et al. (2023) demonstrated reward overoptimization: optimizing against an imperfect reward model via best-of- N sampling degrades true performance past a critical N , a manifestation of Goodhart’s law. Their experiments correspond to the $b = 1$ special case of our framework. The critical N is analogous to $\hat{n}_{\text{top-1}}$ below, and the degradation is consistent with the same overestimation mechanism.

3 Theoretical Framework: Noisy Top- b Retention

We analyze noisy selection as a single primitive: score n candidates and retain the top b . Best-of- n (argmax) selection is $b = 1$. Beam search retains multiple hypotheses; our experiments use a width- k beam with $b = k$. Treating retention b and pool size n separately makes exploration part of the theory rather than a later add-on to a top-1 model.

The analysis is per selection step. It yields a conservative bound on dropping a target correct path, not a guarantee of end-to-end task accuracy.

3.1 Setup and Notation

We analyze beam search at the *reasoning-step* level: each node is a complete chain-of-thought step, as in Tree of Thoughts (Yao et al., 2023) and PRM-

guided search (Lightman et al., 2023), rather than a token.

Consider a single beam selection step. The current beam of k paths each generates k candidate continuations, producing $n = k^2$ total candidates. This symmetric construction is a design choice, not a theoretical requirement: tying the expansion factor to the beam width keeps k as the single tuning parameter and matches standard beam search implementations. The theory itself operates on the candidate pool size n and the retention count b , and applies regardless of how n decomposes into beams and expansions.

A scorer assigns each candidate i a score

$$R_i = \mu_i + \epsilon_i, \quad (1)$$

where $\mu_i \in \mathbb{R}$ is true quality and ϵ_i is scorer error. We distinguish *single-sample decoding* ($k = 1$), where no comparison occurs, from selection among multiple noisy candidates. Experiments use temperature 0.7, so “single-sample” is more precise than “greedy.”

3.2 Assumptions

Assumption 1 (Within-Step Gaussian Scorer Noise). *Within a selection step, scorer errors are i.i.d. and homoscedastic: $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \dots, n$.*

This is a tractable model of scorer uncertainty, not a consequence of using separate forward passes. The scale σ is the noise in the scorer’s own units; we later show selection reliability is governed by the scale-invariant ratio Δ/σ . We assume within-step independence for tractability, even though candidates in the same step may share prefixes; prefix-induced correlations and cross-depth dependence are not modeled.

Assumption 2 (Dominated Incorrect Quality). *Among the n candidates, fix one target correct candidate c with quality μ_c . Every incorrect candidate j satisfies $\mu_j \leq \mu_w < \mu_c$. Define the minimum gap $\Delta := \mu_c - \mu_w > 0$.*

Unlike an equal-quality two-class model, incorrect candidates may have a spectrum of qualities below μ_w . Raising every μ_j to the envelope μ_w can only make incorrect candidates harder to beat, so the homogeneous comparison is conservative. Analyzing one target path is likewise pessimistic when several acceptable paths are present. This is a worst-case envelope, not a random-quality model

in which μ_j is itself drawn from a distribution (we do not derive a convolution/tail form for that case).

3.3 Retention as an Order Statistic

Let $R_{(b)}$ be the b -th largest score among the $n - 1$ incorrect candidates. The target correct path is retained iff fewer than b incorrect candidates outscore it:

$$\text{correct retained} \iff R_c > R_{(b)}. \quad (2)$$

Couple each incorrect score to a homogeneous worst-case $\tilde{R}_j := \mu_w + \epsilon_j$ using the same noise, and let $\tilde{R}_{(b)}$ be the corresponding b -th largest. Assumption 2 gives $R_{(b)} \leq \tilde{R}_{(b)}$ pointwise, hence a larger drop probability in the homogeneous problem.

Define the *top- b selection bias* by the exact Gaussian order-statistic expectation

$$B_b(\sigma, m) := \sigma \mathbb{E}[Z_{m-b+1:m}], \quad m := n - 1, \quad (3)$$

where $Z_{r:m}$ is the r -th ascending order statistic of m i.i.d. standard normals. Approximations enter only when we want a closed form.

Lemma 3.1 (Order-Statistic Retention Bias). *Under Assumptions 1 and 2,*

$$\mathbb{E}[R_{(b)}] \leq \mu_w + B_b(\sigma, n - 1). \quad (4)$$

For intermediate ranks, the usual quantile approximation is

$$B_b(\sigma, m) \approx \sigma \Phi^{-1} \left(1 - \frac{b}{m+1} \right). \quad (5)$$

Proof outline. Domination $R_{(b)} \leq \tilde{R}_{(b)}$ follows by raising incorrect means to μ_w . Standardizing the homogeneous scores gives the exact expectation (3); (5) is the standard intermediate-order-statistic approximation (David and Nagaraja, 2003). Details in Appendix A.

The quantity that matters for retention is the cutoff, not the maximum. Define

$$\Delta_{\text{eff}}^{(b)} := \Delta - B_b(\sigma, n - 1). \quad (6)$$

Theorem 3.2 (Top- b Retention Bound). *Under Assumptions 1 and 2, if $\Delta_{\text{eff}}^{(b)} > 0$,*

$$\Pr(R_{(b)} > R_c) \leq \left(1 + \frac{(\Delta_{\text{eff}}^{(b)})^2}{2\sigma^2} \right)^{-1}. \quad (7)$$

Proof outline. Domination reduces the problem to $\Pr(\tilde{R}_{(b)} > R_c)$. A Gaussian order statistic has variance at most σ^2 (David and Nagaraja, 2003, §5.4), so $\text{Var}[\tilde{R}_{(b)} - R_c] \leq 2\sigma^2$. Cantelli’s inequality on the exact mean gap $\Delta_{\text{eff}}^{(b)}$ yields (7). See Appendix A.

The theorem is a conservative Cantelli bound, not an exact drop probability. We use $\Delta_{\text{eff}}^{(b)} > 0$ as a *predicted retention criterion*: the correct candidate remains above the pruning cutoff in expectation. That is a per-step survival diagnostic, not a criterion for whether beam search improves end-task accuracy relative to $k = 1$.

3.4 Top-1 Selection Recovers Extreme Value Theory

Setting $b = 1$ makes $R_{(1)} = \max_j R_j$ an extreme value. The Gumbel approximation of B_1 (Appendix B) gives the large-pool expansion $B_1(\sigma, n-1) \approx \sigma\sqrt{2\log(n-1)}$. Positive expected separation then requires $\Delta \gtrsim \sigma\sqrt{2\log(n-1)}$, or

$$\hat{n}_{\text{top-1}} \approx 1 + \exp\left(\frac{\Delta^2}{2\sigma^2}\right). \quad (8)$$

Under expansion $n = k^2$ this corresponds to $\hat{k}_{\text{top-1}} \approx \exp(\Delta^2/4\sigma^2)$. As $\Delta/\sigma \rightarrow 0^+$, $\hat{n}_{\text{top-1}} \rightarrow 2$; the corresponding beam width under $n = k^2$ tends toward 1. The formula is a top-1 *selection-reliability* threshold, consistent with reward overoptimization (Gao et al., 2023). It has the same conditioning as the retention bound: a designated correct candidate is already in the pool. It does not by itself determine where best-of- n stops helping once generation quality is modeled. For beam search the correct path need not win the argmax; it only needs to stay among the retained b hypotheses.

3.5 Beam Search: Top- k Retention

For our width- k beam, each of the k retained paths generates k continuations, giving $n = k^2$ candidates, of which the top $b = k$ are retained. A target correct path is therefore removed only if at least k incorrect candidates outscore it. Using the exact Gaussian order statistic, we define the corresponding *mean-retention threshold*

$$\hat{k}_{\text{retain}} := \max\{k \geq 2 : B_k(\sigma, k^2 - 1) < \Delta\}. \quad (9)$$

For $k \leq \hat{k}_{\text{retain}}$, the target’s expected score remains above the expected pruning cutoff. This is a per-step retention criterion: it conditions on the target

already being present in the candidate pool and does not by itself predict the end-to-end optimal beam width.

Using the quantile approximation from (5),

$$B_k(\sigma, k^2 - 1) \approx \sigma, \Phi^{-1}\left(1 - \frac{1}{k}\right), \quad (10)$$

which yields

$$\hat{k}_{\text{retain}} \approx \frac{1}{1 - \Phi(\Delta/\sigma)}. \quad (11)$$

This highlights the benefit of retaining multiple hypotheses. For top-1 selection, the relevant competitor is the *maximum* incorrect score, yielding $\hat{k}_{\text{top-1}} \approx \exp(\Delta^2/4\sigma^2)$ under $n = k^2$. For top- k retention, the relevant competitor is instead the k -th largest incorrect score. At high SNR, (11) grows as $\exp(\Delta^2/2\sigma^2)$ up to subexponential factors. Thus, at leading exponential order, the retention threshold grows approximately as the square of the top-1 beam-width threshold.

At the beam widths used in our experiments, we rely on the exact Gaussian order statistics rather than their large- k approximation. For example, at $k = 4$ the exact cutoff is $B_4 = 0.71\sigma$, close to the quantile approximation $\sigma\Phi^{-1}(3/4) \approx 0.67\sigma$.

4 Experiments

4.1 Setup

We evaluate on MR-BEN (Zeng et al., 2024): 5,737 questions across 10 subjects spanning biology, chemistry, medicine, physics, mathematics, and logic. We use three policy models: Qwen2.5-7B-Instruct (Alibaba Cloud, 2024), Llama-3.1-8B-Instruct (Meta AI, 2024), and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024). For scoring, we compare: (1) **Perplexity**: each model self-scores using negative log-perplexity of the full reasoning trace, so that higher scores indicate greater confidence; (2) **MathShepherd PRM**: math-shepherd-mistral-7b-prm (Wang et al., 2024), a process reward model outputting step-correctness probabilities. At each depth, beam selection uses only the most recent step’s score, not a cumulative aggregate. For perplexity, this is the negative log-perplexity of the entire sequence up to that step. For PRM, this is the correctness probability of the latest step.

We run beam search with beam width $k \in \{1, 2, 3, 4\}$. At each depth, each of the k beams generates k candidate continuations, yielding $n =$

k^2 total candidates from which the top k are selected. At $k = 1$, the policy produces a single sample per step and no selection is performed; we refer to this as *single-sample decoding*. The maximum depth is 24 and the temperature is 0.7 for all configurations. We run 3 seeds per configuration and report means and standard errors. Across 3 model families, 2 scoring paradigms, 10 subjects, 4 beam widths, and 3 seeds, this constitutes 720 full beam-search sweeps and roughly 444,000 logged beam-selection steps, a large, multi-domain base for the analysis. At $k=4$, a single subject–seed combination requires approximately one week of GPU time on a single A100, making beam widths beyond 4 computationally prohibitive with our current setup.

4.2 Main Results

Tables 1 and 2 present the central comparison. Perplexity-scored beam search provides no benefit: Qwen is flat (± 0.1 points), while Llama and Mistral *degrade* at wider beams (-5.4 and -3.9 points at $k=4$), consistent with noisy overestimation under a weak scorer. The model ranking Qwen $>$ Llama $>$ Mistral is stable across all beam widths, confirming that policy capability dominates when the reward signal is unreliable. Standard errors also grow with beam width under perplexity scoring (e.g., Llama: 1.3 at $k=1$ vs. 3.4 at $k=4$), consistent with greater variability under wider noisy selection.

Replacing perplexity with PRM scoring, while keeping the model and algorithm identical, transforms beam search scaling from flat or negative to strongly positive (Table 2, Figure 1). All three models benefit at $k=4$: Mistral gains $+8.9$ points (42.3% to 51.2%), Llama gains $+6.2$ points (59.2% to 65.4%), and Qwen gains $+3.4$ points (71.7% to 75.1%). The largest relative gain occurs for the weakest model (Mistral), suggesting PRM-guided search is especially valuable when the policy alone struggles.

4.3 Interpreting Results: Selection versus Retention

The theory separates noisy selection (8) from per-step retention (9). End-task beam width is a third quantity, which neither formula claims to predict. Our results are consistent with a scorer-dependent split:

- **Perplexity:** no tested beam width improves accuracy, and Llama and Mistral degrade. That is consistent with a weak-scorer selection-reliability

regime.

- **PRM:** all three models peak at the widest tested beam, $k=4$; Qwen and Llama scale monotonically, while Mistral dips at $k=3$ before recovering at $k=4$. A more reliable scorer makes both selection and retention more forgiving, and extra generation can help.

The two scorers therefore occupy different points on the same Δ/σ continuum. Across the 30 (model, subject) configurations, 10/30 perplexity runs have best $k = 1$, versus 0/30 for PRM. We do not estimate Δ/σ from raw score dispersion: perplexity and PRM live on different scales, and observed spread mixes quality variation with scorer error. The scale-free slack-ratio test below is the appropriate check of the order-statistic mechanism. Figure 2 visualizes the accuracy contrast. Per-subject breakdowns appear in Appendix C.

4.4 Consistency Across Domains

The scorer-dependent split is not an artifact of aggregation: it holds within each of MR-BEN’s ten domains. Table 3 reports the per-domain change from $k=1$ to $k=4$, averaged over models and seeds. PRM-guided search improves in *all ten* domains ($+3.6$ to $+15.6$), spanning biology, chemistry, medicine, physics, mathematics, and logic, while perplexity is flat or negative in most (-5.2 to $+6.2$). Full per-subject curves are in Appendix C.

4.5 Validating the Exploration Model

The theory makes an order-statistic prediction we can test directly: per-step sorted scores are empirical upper order statistics $R_{(1)} \geq R_{(2)} \geq \dots$. Aggregating over roughly 444,000 selection steps at $k=4$ ($n = 16$), the scale-free slack ratio $\rho_k := \mathbb{E}[R_{(1)} - R_{(k)}] / \mathbb{E}[R_{(1)} - R_{(2)}]$ tracks the homogeneous Gaussian benchmark (Table 4): observed ρ_4 ranges 1.9–2.5 versus 2.07 predicted. The empirical scores mix heterogeneous quality with scorer error and may be correlated, so this is consistency with the order-statistic mechanism rather than a direct test of Lemma 3.1. Using exact Gaussian order-statistic expectations at $n = 16$ (not the large- k tail expansion), the homogeneous cutoff falls from $B_1 = 1.74 \sigma$ for top-1 selection to $B_4 = 0.71 \sigma$ for top-4 retention, a 59% reduction. That is the finite- k mechanism behind (9): retaining k hypotheses moves the bar well below the maximum.

Model	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Gain ($k=4-k=1$)
Qwen	71.3 ± 1.0	71.1 ± 1.0	71.2 ± 1.1	71.4 ± 1.5	+0.1
Llama	59.7 ± 1.3	58.4 ± 1.5	58.5 ± 2.1	54.3 ± 3.4	-5.4
Mistral	42.4 ± 1.6	42.0 ± 1.7	41.0 ± 2.1	38.5 ± 2.7	-3.9

Table 1: Perplexity-scored beam search: Macro success rate (%) by model and beam width. Values are mean $\pm$ standard error across subjects and seeds. The “Gain” column reports the difference in accuracy between $k=4$ and $k=1$ (not to be confused with the latent quality gap Δ in the theory).

Model	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Gain (best- $k=1$)
Qwen	71.7 ± 1.2	73.8 ± 1.3	73.9 ± 1.4	75.1 ± 1.3	+3.4
Llama	59.2 ± 1.7	63.4 ± 1.8	64.7 ± 1.8	65.4 ± 2.4	+6.2
Mistral	42.3 ± 2.7	50.8 ± 3.3	47.9 ± 3.6	51.2 ± 3.7	+8.9

Table 2: PRM-scored beam search: Macro success rate (%) by model and beam width. Values are mean $\pm$ standard error across subjects and seeds. The “Gain” column reports the accuracy difference between the best beam width and $k=1$.

Domain	Perplexity gain	PRM gain
College Biology	+0.3	+8.1
College Chemistry	+3.1	+7.8
College Math	-5.2	+9.7
College Medicine	-1.3	+7.5
College Physics	-3.0	+10.8
High-School Biology	-2.5	+3.6
High-School Chemistry	+1.7	+8.2
High-School Math	+0.7	+15.6
High-School Physics	+6.2	+9.7
Logic	+2.7	+7.7

Table 3: Per-domain change in macro success rate from $k=1$ to $k=4$ (percentage points, averaged over three models and three seeds). PRM improves in every domain; perplexity does not.

Model	Scorer	ρ_3	ρ_4	ρ_5
<i>Gaussian theory</i> ($n=16$)		1.60	2.07	2.46
Llama	perplexity	1.52	1.94	2.44
Mistral	perplexity	1.66	2.21	2.90
Qwen	perplexity	1.68	2.28	3.09
Llama	PRM	1.75	2.44	3.20
Mistral	PRM	1.72	2.39	3.12
Qwen	PRM	1.76	2.47	3.32

Table 4: Rank-spacing ratios $\rho_k = \mathbb{E}[R_{(1)} - R_{(k)}] / \mathbb{E}[R_{(1)} - R_{(2)}]$ from $\approx 444K$ beam-selection steps ($k=4$, $n=16$), versus the homogeneous Gaussian benchmark. Perplexity ratios are close to the benchmark; PRM ratios are somewhat larger.

4.6 First-Step Selection at Larger Pool Sizes

Full multi-step beam search beyond $k = 4$ is prohibitive at our scale (about one A100-week per subject-seed at $k = 4$). To test the top-1 primitive (8) past the experimental pool $n = k^2 = 16$, we isolate the first selection step: sample N candidates at depth 0, keep the highest-scoring one, and continue greedily. This is not identical to generating N complete rollouts, but it is the same per-step argmax the theory bounds, run at $N \in \{2, 4, 8, 16, 32, 64\}$ on MR-BEN college chemistry (50 examples per cell). Table 5 reports the result.

PRM shows no systematic dip out to $N = 64$ on any of the three models. Perplexity peaks at $N \leq 16$ on all three; Llama perplexity declines 11.4 points from $N = 8$ to $N = 64$. The split is consistent with earlier saturation under perplexity

than PRM.

5 Case Study: Diagnosing Reward Inversion

The aggregate results above show *that* scorer quality determines beam search effectiveness. This section examines *how* the failure unfolds at the level of individual beam selections. We trace beam search behavior for Mistral-7B-Instruct-v0.3 on high_school_biology with perplexity scoring, where $k=4$ scored 59.7% versus 63.0% for single-sample decoding, despite substantially more search compute (see also Figures 10 and 11 in Appendix E).

5.1 Reward Inversion at the Trace Level

Problem cf68a215 illustrates noisy ranking at a single step. At depth 2, the beam-selection log shows:

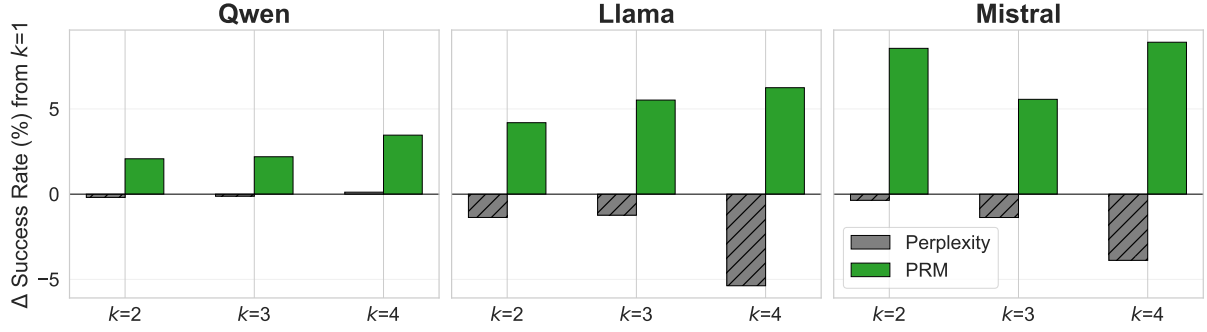


Figure 2: Per-model change from $k=1$ by scorer. Hatched bars indicate a negative change from $k=1$. The PRM unlocks search benefits where perplexity cannot.

Model	Scorer	$N=2$	4	8	16	32	64
Llama	PRM	40.0	34.0	36.0	44.0	46.0	44.0
Mistral	PRM	34.0	26.0	30.0	26.0	32.0	40.0
Qwen	PRM	48.0	54.0	50.0	42.0	40.0	56.0
Llama	PLX	38.0	36.0	50.7	47.3	38.7	39.3
Mistral	PLX	30.7	28.0	27.3	34.0	27.3	30.7
Qwen	PLX	45.3	48.7	53.3	50.0	48.0	50.0

Table 5: First-step selection accuracy (%) vs. candidate-pool size N on college chemistry (50 examples). Perplexity (PLX) rows are 3-seed means; PRM rows are one seed. Per-cell binomial SE is about 7 points, so neighboring N differences should be read with care. Bold marks the best N in each row.

```
[BEAM_SELECT] depth=2 | candidates=12
top_rewards=[-3.04, -3.05, -3.18, ...]
#1: reward=-3.0405 [ANS:C]
#2: reward=-3.0496 [ANS:A]
```

Incorrect C outranks correct A by a margin of 0.009 among 12 candidates.¹ At width $k=4$ both hypotheses remain in the beam: rank 2 is not pruned. The snippet is therefore a ranking diagnostic, not a top- k drop. The small gap illustrates how weakly the scorer separates these two paths.

The inversion appears in what the search *returns*. In 25 disagreement cases where $k=1$ was correct but $k=4$ was wrong, the completed $k=4$ trace consistently had a higher reward (lower perplexity) than the $k=1$ trace, as shown in Table 6. This is consistent with noisy top-1 selection at the final output stage, the $b=1$ special case, not intermediate pruning of the correct rank-2 candidate.

5.2 Score Margins as a Diagnostic

The trace-level analysis above reveals why perplexity scoring fails. Score margins provide a heuristic

¹The candidate count can fall below the nominal $n=k^2=16$ when fewer unique continuations are sampled in a step (action deduplication or generation collapses). The logger prints the top three scores before the width- k retention step.

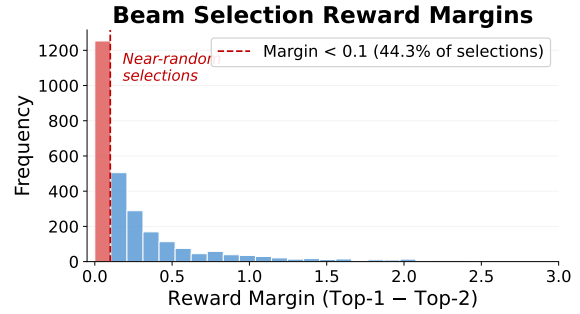


Figure 3: Reward margin distribution for perplexity-scored beam selections. The peak near zero shows that 44% of selections have a top-two margin below 0.1, a heuristic signal of ambiguous rankings.

Problem	$k=1$	Correct	$k=4$	Correct
cf68a215	-6.71	✓	-2.84	✗
c1eacc33	-4.65	✓	-2.19	✗
5c9182fd	-3.70	✓	-2.53	✗
681201fd	-4.49	✓	-2.82	✗

Table 6: Reward inversion at the returned path: $k=4$ finishes with a higher-scoring but incorrect trace.

diagnostic that does not require ground-truth labels. Across 2,770 beam selections in this domain, 44.3% had a reward margin < 0.1 between the top two candidates (Figure 3), indicating ambiguous rankings. An absolute margin of 0.1 is not a scale-free test of near-random selection, and a small top-two gap can also arise from two similar-quality candidates. We therefore treat consistently small margins as a qualitative warning, not as a direct estimate of $\Delta_{\text{eff}}^{(b)}$.

This observation connects to the practical guidance in Section 6: inspecting score margins on a small pilot run is a cheap check of whether rankings are sharp enough to support wider search. Com-

bined with a weak top-1 signal, end-task useful width is then likely near 1, as we observe for perplexity.

6 Discussion

6.1 Retention Capacity versus Empirically Useful Width

Our central theoretical results are the top-1 selection-reliability threshold (8) and the top- b retention-gap threshold (9). Beam width is a scorer-dependent design parameter, not a “wider is better” heuristic. Neither formula replaces the empirically optimal k . Both are governed by Δ/σ . At vanishing SNR, the top-1 pool threshold approaches two candidates, while the mean-retention threshold approaches beam width two; neither implies an end-task optimum above $k = 1$. End-task accuracy can still peak at $k = 1$, as we observe for perplexity.

6.2 Practical Guidance

Neither Δ nor σ is directly observable in ordinary deployment, and we do not claim that the asymptotic thresholds can be computed exactly from raw search traces. The theory instead provides three forms of guidance:

1. **Diagnostic indicators.** Top-two score margins provide a qualitative signal of ambiguity in selecting the highest-scoring candidate. If margins are consistently small, as we observe for 44% of perplexity selections, rankings are often ambiguous. Score dispersion mixes genuine quality variation with scorer noise, so it is a heuristic rather than a direct estimator of σ or $\Delta_{\text{eff}}^{(b)}$.
2. **Pilot runs.** A practical approach is to compare $k = 2$ against $k = 1$ on a small validation set and inspect the full ranked score profile, not only the winner. If widening produces no improvement while cutoff margins shrink, the scorer likely has insufficient SNR to support broader search. With ground-truth labels, one can estimate the separation between correct and incorrect score distributions more directly.
3. **Invest in the scorer, not only the beam width.** Retention capacity and the top-1 pool limit are exponentially sensitive to Δ/σ at high SNR. Improving scorer reliability can expand the range over which additional search remains reliable, whereas widening against a noisy scorer can be counterproductive.

6.3 Toward Bias-Aware Adaptive Beam Search

The unified formulation suggests a more principled adaptation rule than uniformly widening the beam. Given an estimate of scorer noise, one can compare the observed or estimated quality separation with the rank-dependent bias $B_b(\hat{\sigma}, n - 1)$ and stop increasing the expansion or retention width when the mean retention gap approaches zero. A stronger correction could explicitly shrink rank-selected scores by their expected order-statistic bias before propagating them across depths. Because this bias is induced by selection, a uniform subtraction from every candidate would not change the ranking; any useful correction must depend on rank or on the search decision itself. We leave such adaptive and bias-corrected algorithms to future work.

7 Conclusion

This study asks: *how wide should you search?* We model noisy selection as retaining b of n scored candidates. Best-of- n is $b = 1$ and recovers a top-1 selection-reliability threshold (8). Width- k beam search uses $b = k$ and $n = k^2$; the correct path need only beat the k -th best incorrect score. That makes a target easier to retain than to select as the unique winner, especially at high SNR. Neither threshold is a prediction of the empirically best k .

Empirically, scorer quality decides whether extra search helps. Perplexity scoring provides no benefit at any tested width across three models and ten domains; PRM scoring yields gains of up to 8.9 points and continues improving through $k = 4$. Rank-spacing ratios over 444K selection steps are close to the homogeneous Gaussian benchmark. Beam width should be matched to scorer reliability, not maximized. These claims concern reasoning-time beam search with noisy scorers; other test-time-compute paradigms need separate analysis.

Reproducibility. Code is available at <https://github.com/NVlabs/beam-overestimation>, as a fork of LLM Reasoners (Hao et al., 2024).

Limitations

Theory assumptions. The analysis considers a single target path and uses a worst-case envelope for heterogeneous incorrect qualities. It assumes i.i.d. homoscedastic Gaussian scorer errors within each selection step and does not model correlations

across shared prefixes or across search depth. Extending the analysis to dependent, heteroscedastic, and multi-step errors, or to a distribution over latent candidate qualities, is left for future work.

Experimental scope. We evaluate multi-step beam search on three 7–8B models, widths up to $k = 4$, and a single temperature $T = 0.7$; we therefore do not empirically locate a turning point for PRM-guided search. Perplexity is the representative training-free scorer, so other intrinsic scoring signals may behave differently.

Ethical Considerations

This work analyzes inference-time search strategies for LLM reasoning. We do not collect or use personal data, and all experiments use publicly available models and benchmarks. The primary ethical consideration is computational cost: beam search at width k requires $O(k^2)$ scorer evaluations per step. Our work mitigates this concern by identifying when wider search is wasteful, helping avoid unnecessary computation.

References

- Aakriti Agrawal, Souradip Chakraborty, Huazheng Qiu, Sharath Chandra Raparthy, Aditya Raghav, Tomek Kaliszewski, Mengdi Wang, Shuai Bai, Jing Shang, Furong Huang, Dinesh Manocha, and Amrit Bedi. 2025. Cut the overcredit: Precision first process rewards for reasoning LLMs. In *International Conference on Learning Representations*.
- Alibaba Cloud. 2024. Qwen2.5-7B-Instruct model card. Hugging Face.
- J. M. Blair, C. A. Edwards, and J. H. Johnson. 1976. Rational Chebyshev approximations for the inverse of the error function. *Mathematics of Computation*, 30(136):827–830.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Stuart Coles. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer.
- Ganqu Cui, Lifan Zhang, Bingxiang Wang, Zefan Zhang, Hanbin Liu, Yangyi Pan, Yuqing Li, Qiufeng Wang, Yang Yue, Zhiyuan Liu, and Maosong Sun. 2025. PRIME: Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.
- Gal Dalal, Assaf Hallak, Steven Dalton, Iuri Frosio, Shie Mannor, and Gal Chechik. 2021. Improve agents without retraining: Parallel tree search with off-policy correction. *Advances in Neural Information Processing Systems*, 34.
- Herbert A. David and Haikady N. Nagaraja. 2003. *Order Statistics*, 3rd edition. John Wiley & Sons.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- R. A. Fisher and L. H. C. Tippett. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.
- Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*.

- Boris Gnedenko. 1943. Sur la distribution limite du terme maximum d’une série aléatoire. *Annals of Mathematics*, 44(3):423–453.
- Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. LLM Reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.
- Frederick Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-N selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*.
- Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. 2026. Best-of-infinity: Asymptotic performance of test-time LLM ensembling. In *International Conference on Learning Representations*.
- Jonathan Light, Wei Cheng, Benjamin Riviere, Yue Wu, Masafumi Oyamada, Mengdi Wang, Yisong Yue, Santiago Paternain, and Haifeng Chen. 2025a. DISC: Dynamic decomposition improves LLM inference scaling. In *Advances in Neural Information Processing Systems*.
- Jonathan Light, Yue Wu, Yiyu Sun, Wenchao Yu, Yanchi Liu, Xujiang Zhao, Ziniu Hu, Haifeng Chen, and Wei Cheng. 2025b. SFS: Smarter code space search improves LLM inference scaling. In *International Conference on Learning Representations*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*.
- Runze Liu, Junhao Li, Jian Jiao, Xun Ye, Jie Wang, and Tong Zhang. 2025. Can 1b LLM surpass 405b LLM? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*.
- Haotian Luo, Lingkai Li, Guang Yang, Zhuohua Zhang, Yue Yang, and Dawn Song. 2025. Supervised optimism correction: Be confident when LLMs are sure. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Ziyang Ma, Qingyue Yuan, Zhenglin Wang, and Deyu Zhou. 2025. Large language models have intrinsic meta-cognition, but need a good lens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Meta AI. 2024. Llama-3.1-8B-Instruct model card. Hugging Face.
- Mistral AI. 2024. Mistral-7B-Instruct-v0.3 model card. Hugging Face.
- Zongyue Qin, Zifan He, Neha Prakriya, Jason Cong, and Yizhou Sun. 2025. Dynamic-width speculative beam decoding for LLM inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.
- Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. 2026. The limits of inference scaling through resampling. In *International Conference on Learning Representations*.
- Sebastian Thrun and Anton Schwartz. 1993. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*.
- Peiyi Wang, Lei Li, Zhihong Shao, R.X. Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*.
- Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. 2025. Every rollout counts: Optimal resource allocation for efficient test-time scaling. In *Advances in Neural Information Processing Systems*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. In *International Conference on Learning Representations*.
- Yuxi Xie, Anirudh Goyal, Wenye Zheng, and Min-Yen Kan. 2023. Self-evaluation guided beam search for reasoning. In *Advances in Neural Information Processing Systems*, volume 36.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. 2024. MR-Ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs. In *Advances in Neural Information Processing Systems*, volume 37.

Congmin Zheng, Jihao Huang, Boyuan Chen, Zhihong Shao, Deli Chen, and Zhifang Sui. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.

A Proof of Lemma 3.1 and Theorem 3.2

Proof of Lemma 3.1. For every incorrect candidate j , Assumption 2 gives $R_j = \mu_j + \epsilon_j \leq \mu_w + \epsilon_j = \tilde{R}_j$ under the coupling that uses the same noise realization. Order statistics are coordinatewise monotone, hence $R_{(b)} \leq \tilde{R}_{(b)}$ pointwise and

$$\mathbb{E}[R_{(b)}] \leq \mathbb{E}[\tilde{R}_{(b)}].$$

Write $m = n - 1$. Because the $\tilde{R}_j$ are i.i.d. $\mathcal{N}(\mu_w, \sigma^2)$, standardization gives

$$\tilde{R}_{(b)} \stackrel{d}{=} \mu_w + \sigma Z_{m-b+1:m},$$

where $Z_{m-b+1:m}$ is the $(m - b + 1)$ -th ascending order statistic of m standard-normal draws. Therefore

$$\mathbb{E}[\tilde{R}_{(b)}] - \mu_w = \sigma \mathbb{E}[Z_{m-b+1:m}] = B_b(\sigma, m),$$

which proves (4). For intermediate ranks, the standard quantile approximation centers the order statistic near the population quantile with plotting position $(m - b + 1)/(m + 1) = 1 - b/(m + 1)$ (David and Nagaraja, 2003), giving (5). Under $b = k$ and $n = k^2$, this position is exactly $1 - 1/k$, yielding (10). The further Gaussian tail expansion $\Phi^{-1}(1 - 1/k) \approx \sqrt{2 \log k}$ (Blair et al., 1976) is not used at experimental k . $\square$

Proof of Theorem 3.2. The coupling above gives

$$\Pr(R_{(b)} > R_c) \leq \Pr(\tilde{R}_{(b)} > R_c).$$

Define $Z := \tilde{R}_{(b)} - R_c$. By definition of B_b ,

$$\mathbb{E}[Z] = \mu_w + B_b(\sigma, n - 1) - \mu_c = -\Delta_{\text{eff}}^{(b)}.$$

The incorrect scores are independent of the correct candidate's noise. Moreover, the variance of an order statistic of i.i.d. Gaussian variables is at most the parent variance σ^2 (David and Nagaraja, 2003, §5.4). Hence

$$\text{Var}[Z] = \text{Var}[\tilde{R}_{(b)}] + \text{Var}[R_c] \leq 2\sigma^2.$$

When $\Delta_{\text{eff}}^{(b)} > 0$, Cantelli's inequality gives

$$\begin{aligned} \Pr(Z > 0) &= \Pr(Z - \mathbb{E}[Z] > \Delta_{\text{eff}}^{(b)}) \\ &\leq \frac{\text{Var}[Z]}{\text{Var}[Z] + (\Delta_{\text{eff}}^{(b)})^2} \leq \left(1 + \frac{(\Delta_{\text{eff}}^{(b)})^2}{2\sigma^2}\right)^{-1}. \end{aligned}$$

Combining this with the stochastic-domination inequality proves (7). Setting $b = 1$ recovers the top-1 Cantelli bound. $\square$

B Top-1 Gumbel Approximation

Proof. For $b = 1$, the homogeneous cutoff is the maximum of $m = n - 1$ i.i.d. $\mathcal{N}(\mu_w, \sigma^2)$ scores. By the Fisher–Tippett–Gnedenko theorem (Fisher and Tippett, 1928; Gnedenko, 1943; Coles, 2001), the normalized maximum converges to a Gumbel distribution. Using location

$$a_m = \mu_w + \sigma \Phi^{-1}\left(1 - \frac{1}{m}\right)$$

and scale

$$c_m = \sigma \left[\Phi^{-1}\left(1 - \frac{1}{em}\right) - \Phi^{-1}\left(1 - \frac{1}{m}\right) \right],$$

the Gumbel expectation $a_m + \gamma_{\text{EM}} c_m$ approximates $B_1(\sigma, m)$. The large- m expansion is $B_1(\sigma, m) \approx \sigma \sqrt{2 \log m}$ (Blair et al., 1976), which yields (8). This is the $b = 1$ extreme-value specialization of (3), not a separate model of beam search. $\square$

C Per-Subject Curves by Model

The per-domain summary (Table 3) and order-statistic slack ratios (Table 4) appear in the main text. Below are the full per-subject success curves.

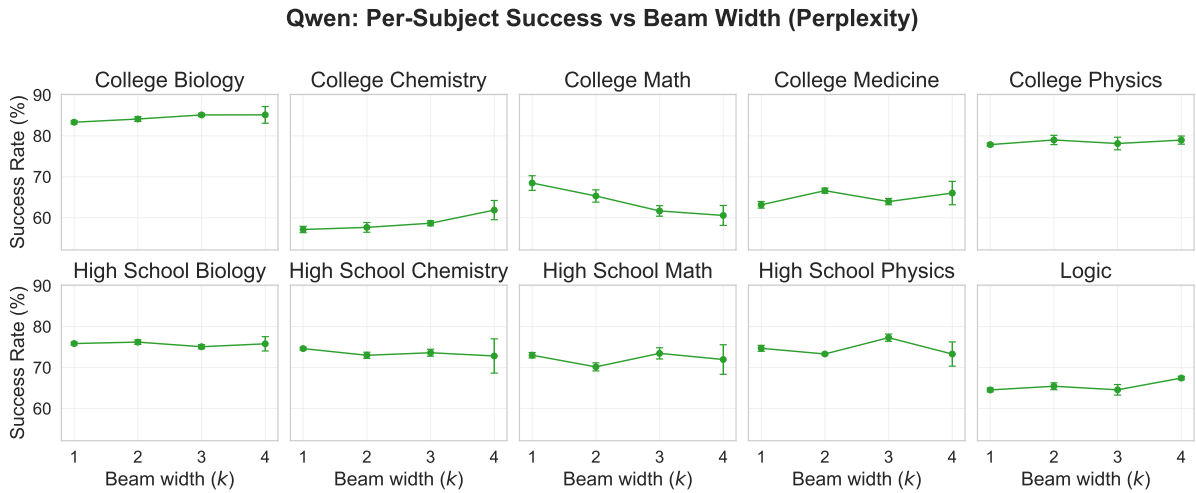


Figure 4: Per-subject success rate vs. beam width for **Qwen2.5-7B-Instruct** with perplexity scoring.

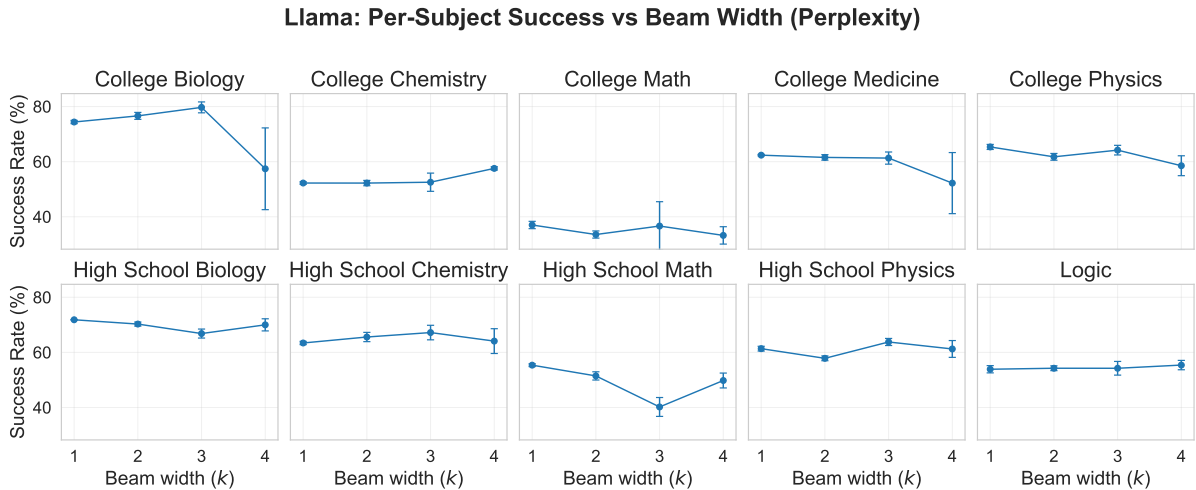


Figure 5: Per-subject success rate vs. beam width for **Llama-3.1-8B-Instruct** with perplexity scoring.

Mistral: Per-Subject Success vs Beam Width (Perplexity)

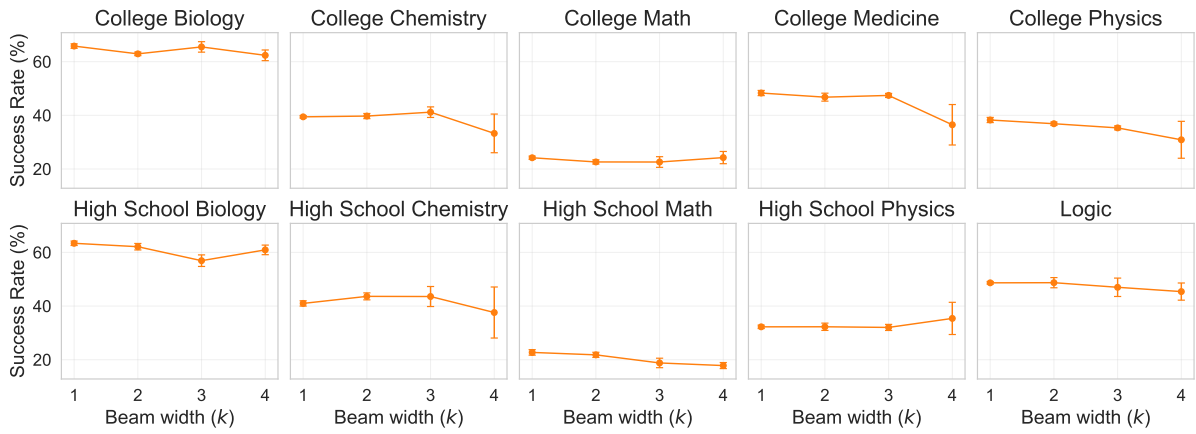


Figure 6: Per-subject success rate vs. beam width for **Mistral-7B-Instruct-v0.3** with perplexity scoring.

C.1 PRM Scoring

Qwen: Per-Subject Success vs Beam Width (PRM)

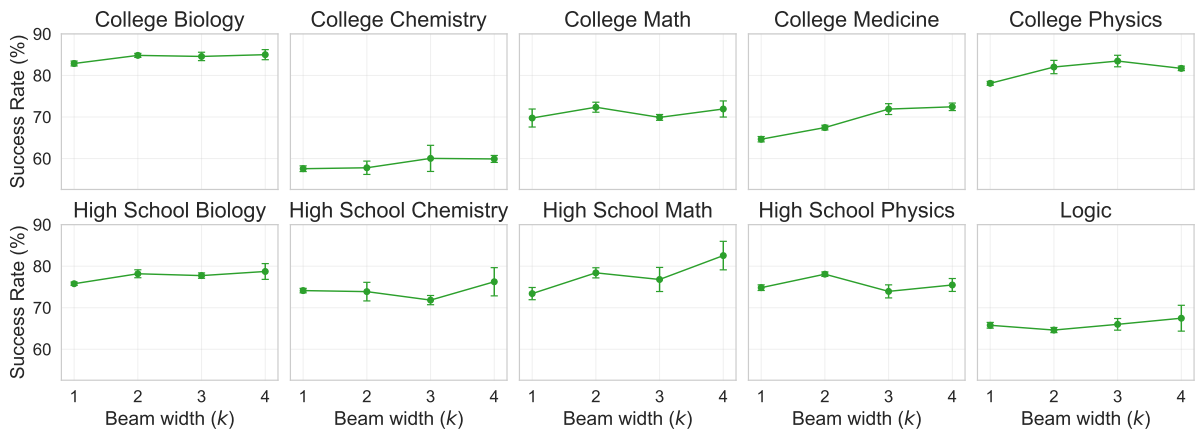


Figure 7: Per-subject success rate vs. beam width for **Qwen2.5-7B-Instruct** with PRM scoring.

Llama: Per-Subject Success vs Beam Width (PRM)

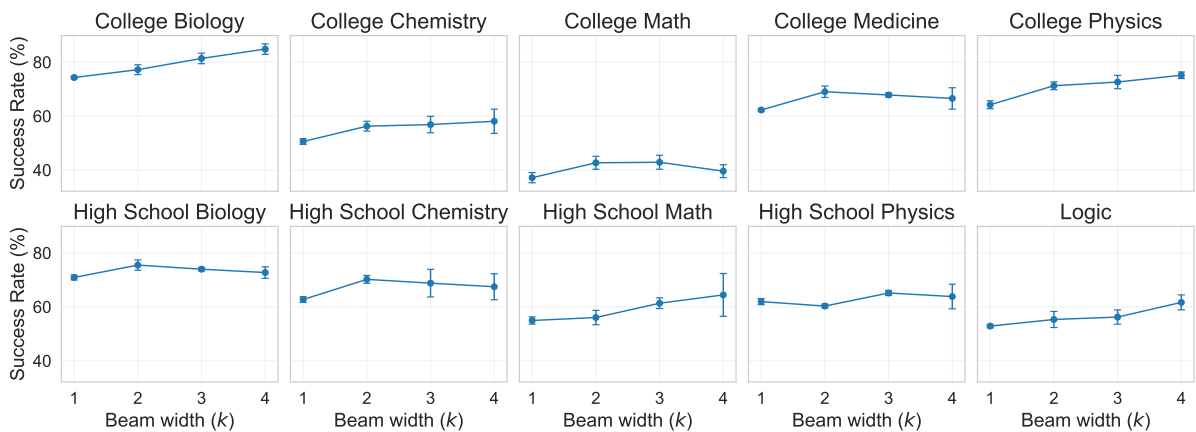


Figure 8: Per-subject success rate vs. beam width for **Llama-3.1-8B-Instruct** with PRM scoring.

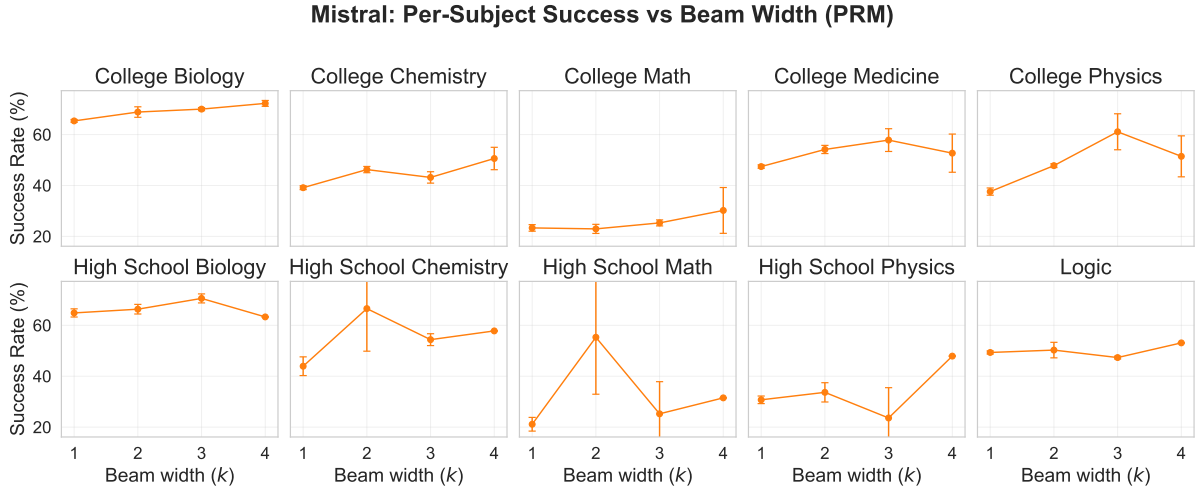


Figure 9: Per-subject success rate vs. beam width for **Mistral-7B-Instruct-v0.3** with PRM scoring.

D Implementation Details

- **Beam width:** $k \in \{1, 2, 3, 4\}$.
- **Expansion factor (number of actions per beam):** equal to the beam width k , giving a candidate pool of $n = k^2$ per selection step.
- **Maximum depth:** 24 steps.
- **Temperature:** 0.7 for all configurations (including $k=1$).
- **Reward aggregation:** last: only the most recent step’s score is used for beam selection (no cumulative aggregation across depths).
- **Perplexity scorer:** negative log-perplexity over the full sequence (prompt + all completed steps) at each depth.
- **PRM scorer:** math-shepherd-mistral-7b-prm, using the correctness probability emitted at the last step tag.

All models are loaded from the Hugging Face Hub. Experiments are managed via Weights & Biases.

E Additional Case Study Figures

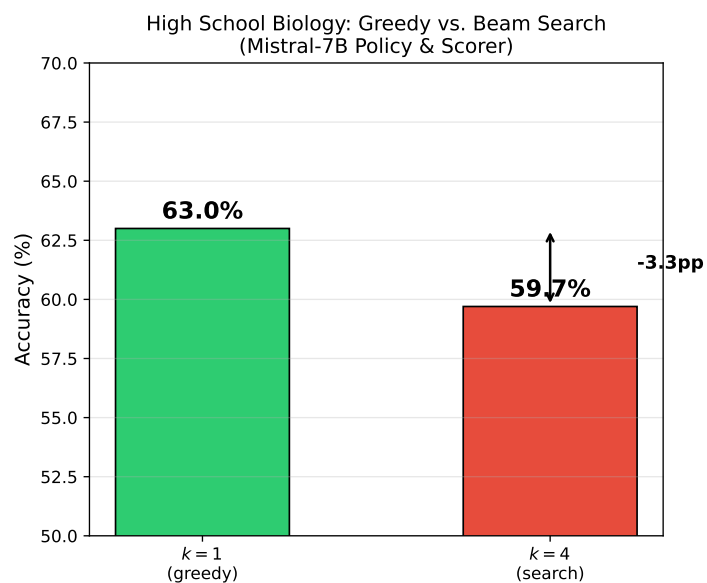


Figure 10: Accuracy: single-sample ($k=1$) vs. beam ($k=4$) on high_school_biology. Beam search underperforms despite evaluating $k^2 = 16$ candidates per step.

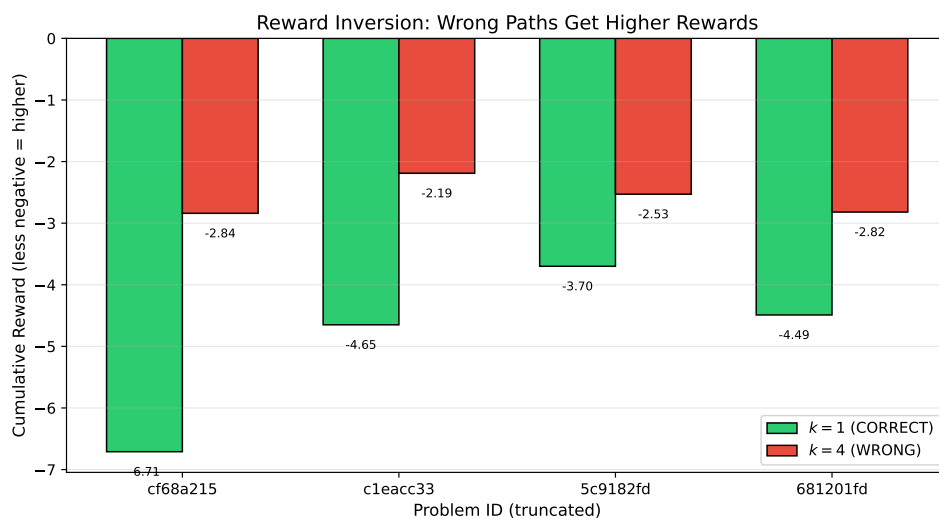


Figure 11: Reward inversion: incorrect paths returned by $k=4$ have higher rewards than correct paths from $k=1$.