One Temperature to Rule Them All?

Aviv Orly* **Ori Shem-Ur*** **Yaron Oz**
School of Physics and Astronomy
Tel Aviv University, Israel
{avivorly, orishemor}@mail.tau.ac.il yaronoz@tauex.tau.ac.il

Abstract

Inference-time sampling is a central mechanism for improving large language models, with temperature controlling the tradeoff between reliability and diversity. While recent work suggests that varying temperature across samples can improve multi-sample inference, simple fixed-temperature schemes remain the standard in practice. We investigate this discrepancy in the canonical pass@K setting, asking when temperature diversity is useful, and when a single tuned temperature suffices. To this end, we analyze per-question success probabilities across multiple models and datasets, tracking how they vary with sampling temperature. We identify several consistent properties of these profiles, including that harder questions are less likely to share efficient sampling temperatures, although this trend is only logarithmic in hardness. We then derive an efficient method for evaluating the theoretically optimal multi-temperature allocation scheme, and show that although such a scheme improves performance, its gains over single-temperature baselines are typically modest when temperature is tuned as a function of the sampling budget K. We explain this phenomenon through mathematical mechanisms that we call dulling effects, and show how they ensure the robustness of temperature choices across different question distributions. Our results clarify when multi-temperature sampling is useful, and how temperatures should ideally be chosen.

1 Introduction

Modern machine learning has made remarkable progress over the past decade. In particular, large language models (LLMs) have emerged as general-purpose foundations for a wide range of tasks, including question answering, summarization, program synthesis, and mathematical reasoning [1–12]. Most contemporary LLMs are trained as next-token predictors, optimizing the likelihood of human-generated text, often followed by reinforcement learning-based alignment objectives [13–18]. Despite the apparent simplicity of this paradigm, these models exhibit increasingly capable forms of inference, including multi-step reasoning and problem solving at human-level and beyond [19–24].

One major driver of these recent improvements in LLM performance is the shift from relying solely on training-time scaling toward investing more computation at inference time [25–27]. Modern systems increasingly allocate additional resources to inference through techniques such as chain-of-thought reasoning and related structured prompting methods [20–23, 28]. These inference-time strategies can substantially improve accuracy without retraining the underlying model [25, 29–31]. As a result, there is growing interest in *test-time* (or *inference-time*) compute, which trades extra sampling, reasoning, and verification for better performance [25–27].

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

A central family of inference-time methods is *multiple sampling*, where the model is queried several times for the same problem, and the resulting candidates are compared, filtered, or aggregated [7, 9, 10, 26, 29]. Multiple sampling can be implemented in various ways, depending on the task structure and the available supervision signal.

For a multiple-sampling strategy to provide meaningful benefits, rather than repeatedly producing the same result, we require a mechanism that introduces variation. A standard approach for achieving such diversity is the use of *temperature*. Instead of deterministically selecting the most likely next token, sampling draws from a temperature-scaled softmax distribution, i.e., a Boltzmann-like distribution [32–34]. In this approach, higher temperatures increase randomness and encourage the exploration of less likely continuations, which can increase the probability of discovering novel or unconventional solutions [34–38]. This however, comes at the cost of reliability, as the temperature increases, samples are more likely to contain errors, inconsistencies, or nonsensical outputs [34–39]. The effectiveness of multiple sampling thus depends on balancing exploration and exploitation.

Recent work has highlighted that this trade-off is not uniform across inputs. Different questions can exhibit substantially different *optimal* sampling behavior. Some are best solved with low-temperature, highly deterministic inference, while others benefit from increased exploration [27, 37, 39–41]. This observation motivates a natural extension of classical multiple sampling: using multiple temperatures within the same sampling process.

Several works investigating this approach, have relied on dynamically adjusting the temperature during inference based on the properties of the input [39, 42, 43], some even at a token-by-token level [44–46]. While these approaches are effective, they can also introduce a substantial computational cost. In contrast [27] demonstrated that this additional cost can be avoided, as even pre-committed allocation schemes, such as sampling temperatures uniformly across the relevant temperature range, can offer substantial benefits over single-temperature strategies.

While these results suggest that multi-temperature sampling should yield substantial gains, single-temperature sampling remains the dominant approach in practical applications. The goal of this paper is to explain this apparent disparity, and to expand our understanding of temperature selection in LLM inference.

Contributions:

We show that the probability of solving LLM questions as a function of temperature exhibits a simple structure, typically characterized by a single local maximum in temperature, and that questions can be classified according to whether this optimum occurs at zero or at a positive temperature (Section 3.1).

We then study question behavior as a function of hardness, defined as the number of samples required for solution (Equation 3), showing that the average optimal temperature varies significantly with hardness (Section 3.2.1). We further show that both the range of efficient temperatures, and their overlap across questions shrink with hardness, albeit only logarithmically (Section 3.2.2).

We then derive and prove a method for efficiently approximating the optimal multi-sampling strategy for the pass@K setting (Section 4.1), and compare it with existing single, and multi-temperature approaches. While single-temperature strategies typically underperform, much of the gap can be recovered by adapting the temperature to the number of samples (Section 4.2).

Finally, we explain these effects, and identify practical advantages of these mechanisms. First, the optimal sampling parameters depend more strongly on the computational budget than on individual question hardness, reducing the need to estimate the latter. This robustness also allows the sampling parameters to be selected without closely matching the difficulty distributions of the tuning and test datasets (Section 4.3).

2 Preliminaries

2.1 Notation

We begin by introducing the main notation used throughout the paper. We denote by $\tau \geq 0$ the temperature used during the inference process, and by $\mathcal{Q}$ the distribution over the questions, for which we study how this choice affects performance. The individual questions drawn from this distribution are then denoted by $q \sim \mathcal{Q}$.

We assign $P_q : [0, \infty) \to [0, 1]$ for the probability that a question q is solved in a single attempt, and write P when the dependence on q is clear from context. Accordingly, $P(\tau)$ denotes the probability of solving the question in a single attempt for τ. We refer to the function $P(\tau)$ as the *temperature profile* of the question. Since in many of our computations the failure probability is more convenient than the success probability, we designate for it the notation $\bar{P} = 1 - P$.

2.2 Pass@K

We conduct our analysis of multi-sampling approaches within the canonical *pass@*K metric. In this setting, we generate $K \in \mathbb{N}$ independent solutions, and consider the task successful if at least one of them is correct [7, 26, 47, 48]. Accordingly, for K separate attempts, we can express the probability of failure over the corresponding K temperatures $\tau_1, \ldots, \tau_K$, as:

$$\bar{\mathcal{P}}(\tau_1, \ldots, \tau_K) = \bar{P}(\tau_1) \cdots \bar{P}(\tau_K) \ . \tag{1}$$

In the common special case where all attempts use the same temperature τ, this reduces to:

$$\bar{\mathcal{P}} = \bar{P}(\tau)^K \ . \tag{2}$$

While the pass@K approach is especially effective when the correctness can be verified automatically, for example by running unit tests for code generation, or checking a formal solution [7–9, 12, 49], it can also be extended using learned verifiers, including auxiliary LLMs that evaluate candidate solutions [10, 25, 30, 31].

Furthermore, many high-performing model variants rely heavily on this general pattern of sampling multiple candidates followed by selection [9, 10, 25, 26, 29, 31]. In fact, many of the properties of the pass@K framework extend to related multi-sampling approaches, such as chain-of-thought reasoning [20, 21, 29]. As such, this framework is both useful in its own right, and serves as a representative setting for understanding the general properties of more advanced multi-sampling methods.

2.3 Empirical Setup

We evaluate four models, Qwen3-0.6B, Qwen3-8B, Mistral 7B Instruct, and GPT-OSS 20B, on three benchmarks: LiveCodeBench (V1-V6) [49] for code generation, MATH-500 [11, 31] for mathematics, and Hi-ToM [50] for multiple-choice higher-order theory of mind. For each model and question, we draw typically 1,500 samples at each of 16 to 38 temperatures between 0.075 and 4.2, and grade them automatically by unit tests, final answers, and chosen options, respectively (Appendix C).

2.4 Other Sampling Hyperparameters

While we focus our analysis on sampling temperature, as it is the principal hyperparameter governing inference-time sampling [27, 39], our results can be generalized to other sampling hyperparameters, including min_p and top_h [51, 52], as discussed for min_p in Appendix I.

3 Properties of Question Profiles

We begin by presenting our findings on the characteristics of question temperature profiles, defined in Section 2.1, along with an observation on the typical overlap of the temperature ranges in which different questions can be solved efficiently. To formalize this, we introduce several characteristics that capture the fundamental properties of a question. The consistency of these patterns across all datasets and models we consider implies that they may reflect a more universal property of large language model inference.

3.1 Unimodality and Typical Profile Shapes

Our first observation is that question profile shapes are qualitatively simple, and can be well approximated by one of two canonical forms. Typically, profiles either begin at 1 for $\tau = 0$ and then monotonically decay toward zero, or they begin with $P(0) = 0$, increase monotonically to some optimal temperature, and then decay.

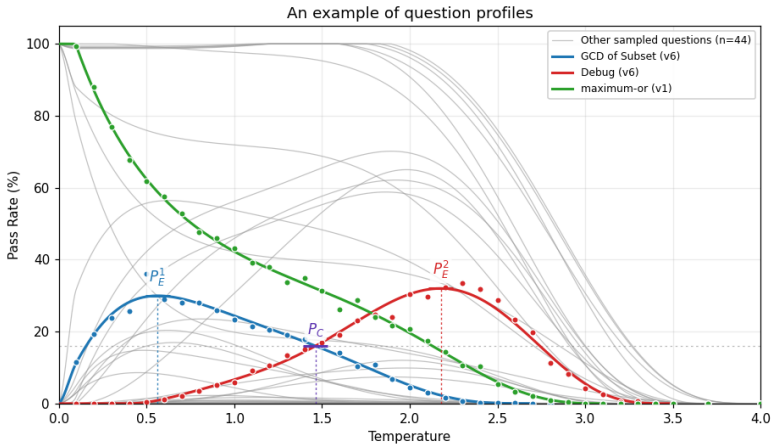


Figure 1: Temperature profiles for 51 Qwen3-8B questions. The green curve, which attains its maximum at $\tau = 0$, illustrates the first canonical form, whereas the red and blue curves, which peak at positive temperatures with corresponding probabilities P_E^1 and P_E^2, illustrate the second.

We illustrate these profiles in Figure 1, which shows the probability of solving different live-code questions as a function of their temperature for the Qwen3-8B model.

In our experiments, presented in full in Appendix D, only a negligible fraction of the questions considered across all models deviated from this pattern. These deviations were generally minor, with most not clearly distinguishable from statistical uncertainty.

While the two profile types appear very different, they arise from a shared underlying property, namely, unimodality, or the existence of a single local maximum. At $\tau = 0$, once external sources of randomness are removed, the model behaves deterministically, implying that $P(0) \in \{0, 1\}$. At the other extreme, as $\tau \to \infty$, performance degrades toward random behavior, and $P(\tau)$ approaches near zero [34–39]. These boundary conditions, together with unimodality, imply that the two profile types we have found are the only possible qualitative forms, which explains our observation.

An important common sub-profile, consists of questions with $P(0) = 0$, but with a maximum very close to $\tau = 0$. These questions are typically easy, in the sense that they have relatively high single-sample success probability, and appear to require only a small amount of stochasticity to nudge away from the incorrect deterministic solution.

Estimating the distribution for each question required careful statistical analysis, especially for very difficult questions with only a few positive solutions, where poor estimation can dramatically distort the resulting temperature profiles. We describe this analysis in detail in Appendix D.

3.2 Question Behavior by Hardness

A natural property to consider for a question in the context of multiple-sampling, is how many samples are required for it to be solved reliably. In the pass@K setting, this can be naturally defined as the number of attempts required to achieve a success probability of at least $1 - \frac{1}{e}$. Using Equation 2, and relaxing the integrality constraint on K, this yields according to the notation in Section 2.1:

$$H_q(\tau) = -\frac{1}{\ln\left(1 - P_q(\tau)\right)} = -\frac{1}{\ln\left(\bar{P}_q(\tau)\right)}, \tag{3}$$

where $H_q(\tau)$ denotes the hardness of question q at temperature τ. We further define the overall hardness of a question (with a slight abuse of notation) as $H_q = \inf_{\tau>0} H_q(\tau)$. To classify questions by hardness, we denote the restriction of the distribution $\mathcal{Q}$ to questions whose hardness satisfies $H_- \leq H_q < H_+$ as: $\mathcal{Q}_{H_-,H_+}$.

Throughout this work, we will see the usefulness of this notion in the context of the pass@K framework. We begin with the following two subsections, which present important trends in question behavior as a function of their hardness.

3.2.1 Average Ideal Temperature

Our first observation in this section is that the average optimal temperature for solving different questions varies as a function of hardness. We illustrate this effect in the left panel of Figure 2 and in Figures 5, 6, and 7, where we plot the optimal temperature for each question as a function of its hardness, along with a sliding average over all questions. To isolate the effect of the easy questions whose optimal temperature does not lie very close to $\tau = 0$, we also plot a separate sliding average for this subset.

As shown in the figure, the optimal temperature required to solve questions varies not only across individual questions, but also in its average behavior across different hardness levels. This distinction is important, as these two sources of variation have different implications for the design of multi-temperature sampling strategies, as we discuss in Section 4.3.2.

3.2.2 Efficient Temperature Range

Our second observation in this section is that as hardness increases, both the variability of efficient temperatures across questions, and the fraction of questions that share a common efficient temperature decrease, albeit only logarithmically.

To formalize these observations, we define the efficient temperature range of a question, as the set of temperatures whose success probability is at least half of its optimal success probability:

$$E_q = \left\{ \tau \geq 0 \mid P_q(\tau) \geq \frac{1}{2} \sup_{\tau' \geq 0} P_q(\tau') \right\} . \tag{4}$$

As we discuss in Section 4.3.1, when considering two questions in the pass@K setting, it is typically more efficient to use two different temperatures, rather than a single shared temperature, only when their efficient ranges do not overlap. As such, this definition provides a useful characterization of the separation between question profiles within a distribution.

In the middle panel of Figure 2, and in Figures 8, 9, and 10, we examine how the sizes of these ranges vary with hardness, where we plot for each question its hardness against its efficient range size, and fit a linear trend as a function of log-hardness. The resulting relationship indicates a logarithmic decay in efficient range size as the hardness increases.

This observation suggests that a similar trend may hold for the overlap between questions' efficient ranges. To quantify this, we consider, for each hardness level H, the maximal fraction of questions with hardness H, that can be efficiently solved using the same temperature:

$$R_{\mathcal{Q}}(H) = \lim_{\Delta H \to 0} \sup_{0 \leq \tau} \Pr_{q \sim \mathcal{Q}_{H, H+\Delta H}} (\tau \in E_q) . \tag{5}$$

In the right panel of Figure 2, and in Figure 11, we plot an empirical approximation of this quantity using a finite hardness window ΔH on a logarithmic hardness scale over our dataset.

A logarithmic trend is still observed, though it is less pronounced than the decay in efficient range sizes. This effect is driven by the variation in the average optimal temperature across hardness levels. In particular, the cluster of easy questions whose optimal temperature lies very close to $\tau = 0$ obscures the overall trend at small K for some of the models. These results support and refine the observation of [27] that as questions become harder, they share fewer efficient temperatures.

4 Multiple Temperature Sampling

Having established the typical structure of question temperature profiles, we now analyze the relative effectiveness of pre-committed multi-temperature strategies in the pass@K setting. This includes the ideal sampling strategy, which we show can be efficiently approximated.

We show that although these approaches provide a measurable advantage over fixed single-temperature strategies, this advantage is relatively limited, and a large portion of it can be recovered by using a single temperature adapted to the available computational budget. We then explain the underlying reasons for this behavior, and how it can be leveraged for practical applications.

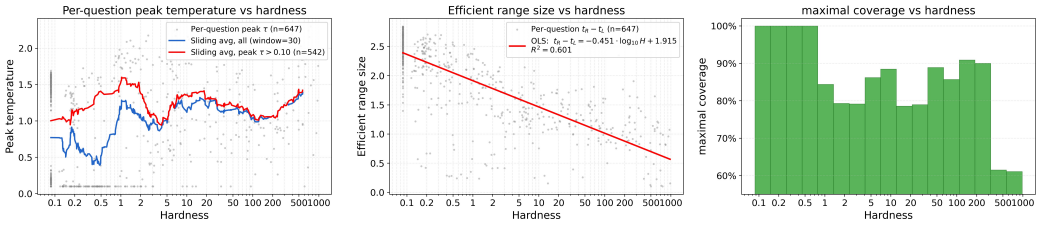


Figure 2: Three hardness-based views on Qwen3-8B on LiveCodeBench: peak temperature (left), efficient range size (middle), and maximal overlap ratio of efficient ranges (right) vs. hardness ($\log_{10}$ scale). Left: blue/red curves show sliding averages over all instances, with/without small ideal-temperature instances. Middle: linear fit over instances. Right: maximal coverage in log-uniform hardness bins. Temperatures vary with hardness, while range size and overlap shrink logarithmically.

4.1 The Ideal Multiple-Temperature Allocation Scheme

We begin this section by showing that, although the space of possible multi-temperature strategies is potentially infinite, in practice, it can be approximated effectively. In the pass@K setting, this strategy is fully defined by the selection of temperatures that maximize the success probability, or equivalently, minimize the failure probability, which according to Equation 1, can be written as:

$$\bar{\mathcal{P}}\left(\tau_1, \ldots, \tau_K\right) = \mathbb{E}_{q\sim\mathcal{Q}} \bar{\mathcal{P}}_q\left(\tau_1, \ldots, \tau_K\right) \ . \tag{6}$$

While directly optimizing over a continuous space of temperatures is challenging, we can discretize the set of candidate temperatures to a small set of $l \in \mathbb{N}$ fixed values $0 \leq \gamma_1, \ldots, \gamma_l$, while still preserving near-optimal performance, due to the linear dependence on the $\bar{P}(\tau)$s in Equation 1, as shown in Theorem B.1.

This allows us to shift from optimizing over the choice of temperatures, to optimizing the allocation of samples across the fixed temperature set, denoted by $K_1, \ldots, K_l \in \mathbb{N}_0$, subject to:

$$K_1 + \cdots + K_l = K \ . \tag{7}$$

Under this formulation, the average failure probability becomes:

$$\bar{\mathcal{P}}\left(K_1, \ldots, K_l\right) = \mathbb{E}_{q\sim\mathcal{Q}} \prod_{i=1}^{l} \bar{P}_q\left(\gamma_i\right)^{K_i} = \mathbb{E}_{q\sim\mathcal{Q}} \exp\left(\sum_{i=1}^{l} K_i \ln\left(\bar{P}_q\left(\gamma_i\right)\right)\right) \ . \tag{8}$$

As $\mathcal{Q}$ may be infinite, this optimization problem remains challenging. However, by standard concentration arguments, and the symmetry in the ordering of temperature choices, we show in Theorem B.2 that it is sufficient to approximate the expectation using a relatively small, randomly sampled subset of questions.

We can therefore simplify the problem to a finite optimization task by denoting the probability of failure in a single attempt for each question $i = 1, \ldots, n$ and temperature index $j = 1, \ldots, l$ by:

$$\bar{P}_{ij} = \bar{P}_{q_i}\left(\gamma_j\right) \ , \tag{9}$$

and find that the optimization problem reduces to finding integers $K_1, \ldots, K_l$ satisfying Equation 7, that minimize the average failure probability:

$$\bar{\mathcal{P}}\left(K_1, \ldots, K_l\right) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(\sum_{j=1}^{l} K_j \ln\left(\bar{P}_{ij}\right)\right) \ . \tag{10}$$

If we relax the integrality constraint on $K_1, \ldots, K_l$, the problem becomes convex, and can be solved efficiently, even for large numbers of questions and candidate temperatures. Then, finding the actual integer solution is straightforward in this resource-allocation setting, since proximity and rounding results imply that good integer solutions can be recovered near the relaxed optimum [53–55].

Because this algorithm and its associated bounds apply to any parameters satisfying the mild conditions of Theorem B.1, their applicability extends beyond sampling temperature to other sampling parameters such as min_p, or top_h, and to models that do not exhibit the properties identified in Section 3, as we discuss further for min_p in Appendix I.

4.2 Temperature Sampling Scheme Comparison

Having derived the optimal temperature allocation scheme, we now compare its average pass rate against [27]'s uniform sampling scheme, and against single-temperature strategies, in Figures 3, 14, 15, and 16. Even with our extensive datasets, the errors remain substantial, due to the differences in question difficulty and temperature sensitivity. However, after normalizing the performance relative to a baseline scheme, in our case, the uniform sampling scheme, these differences become significantly clearer, and the underlying trends more pronounced. We provide the details of this analysis in Appendix F.1.

As shown in Figure 3, the optimal allocation scheme has about a 10% advantage over each fixed single-temperature strategy for at least some values of K. However, if we allow the temperature to change as a function of K, this advantage becomes much more limited. It is important to note that although these advantages are modest, in real-world deployment settings, even a small improvement in the pass rate can translate into substantial practical and economic value.

Another consistent finding is that while the uniform scheme underperforms relative to finely tuned single-temperature strategies, it does not deviate substantially, as it always allocates some samples to the appropriate temperature region. This observation supports the usefulness of the scheme proposed by [27], in settings where fine-tuning temperature choices is impractical.

4.3 The Dulling Effects

While the separation observed in Section 3.2.2 is relatively small, on the order of $10\% - 20\%$, it is still substantially larger than the modest advantage that multi-temperature sampling methods exhibit over finely tuned single-temperature strategies. In this section, we explain this disparity by identifying two effects, which we term *dulling effects*. These two effects are mechanisms that diminish the potential gains from multi-temperature sampling relative to what the apparent separation in the data would suggest.

We establish these effects under the question-profile properties identified in Section 3, which we have shown to capture a relatively universal phenomenon. However, the conditions underlying these effects can be generalized beyond these specific properties, particularly in the case of the second effect discussed in Section 4.3.2. Furthermore, we show that these effects not only help explain the observed phenomenon, but also have important practical implications.

4.3.1 The Short-Sampling Effect

The first effect we identify, which we term the *Short-Sampling Effect*, is best illustrated in a simplified setting of two questions sharing similar Gaussian-like profile shapes, as shown by the red and blue curves in Figure 1. We assume these profiles have a clear maximum and rapidly decaying tails, as these assumptions are consistent with the typical behavior observed in Section 3.1.

In this setting, it is straightforward to show that the number of distinct optimal temperatures in Equation 10 cannot exceed the number of questions, as formalized in Lemma B.3, which in this case is two. Consequently, the optimal strategy reduces to either allocating samples across two "expert" temperatures, each close to the optimum of one question (represented by P_E^1, P_E^2 in the figure), or allocating all samples to a single compromise temperature (represented by $P_C \leq P_E^1, P_E^2$, which we assume is the same for both questions).

If we also assume for simplicity that the two profiles are identical, meaning $P_E^1 = P_E^2 = P_E$ we can capture the simplest manifestation of the short sampling effect via the following proposition:

Proposition 4.1 (The Short-Sampling Effect)**.** Given the setup above, where we assume for simplicity that K is even, and define:

$$\Delta P = \frac{P_E - 2P_C}{2} \,, \tag{11}$$

then, the difference in success probability between using the two expert temperatures, with $\frac{K}{2}$ samples assigned to each, and using a single compromise temperature with all K samples assigned to it, which we denote by $\Delta\mathcal{P}$, satisfies to first order:

$$\Delta\mathcal{P} = e^{-\frac{P_E K}{2}} \left(e^{K\Delta P} - 1 + O\left(K P_E^2\right) \right) \,. \tag{12}$$

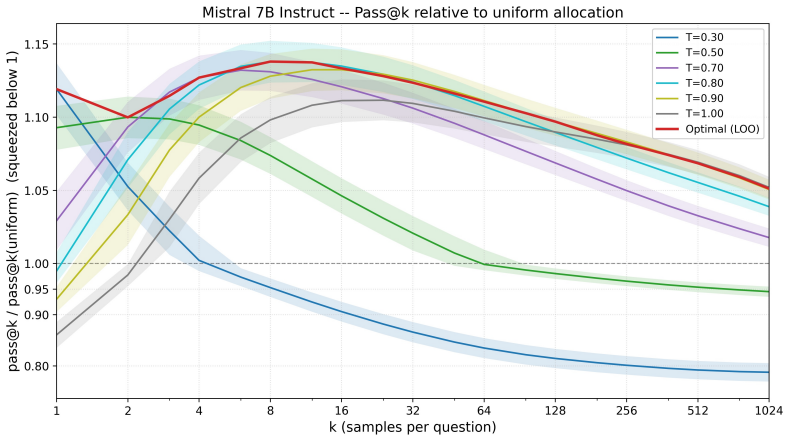


Figure 3: Pass@K on $\log_{10}$ scale for Mistral 7B on LiveCodeBench under fixed-temperature strategies (blueish) vs. the optimal allocation (red), normalized by the uniform-temperature baseline (dotted). While the optimal allocation consistently outperforms fixed strategies for some K, for each K at least one achieves comparable performance.

Consequently, to the first order, the two-expert strategy outperforms the compromise strategy if and only if:

$$P_E > 2P_C \ . \tag{13}$$

Although this proposition, which we prove in Appendix B.2.2 considers only a highly simplified setting, we also show that its core components extend to more realistic scenarios. In particular, we analyze settings with multiple questions (Proposition B.1), asymmetries (Proposition B.2), and non-vanishing tails (Proposition B.3), and show that the same underlying intuition persists, highlighting the universality of the effect.

From these results, two consistent phenomena emerge. First, the threshold at which the compromise temperature begins to outperform the expert temperatures occurs when P_C is on the order of $\frac{P_E}{2}$, which motivates Equation 4 as a natural characterization of question separation.

Second, we observe that over-utilizing multi-temperature strategies incurs a much larger penalty than under-utilizing them. As seen in Equation 12, splitting resources across suboptimal temperatures leads to exponential losses, whereas the loss from using a single compromise temperature grows only linearly, since $e^{K\Delta P} - 1 \simeq K\Delta P$. This asymmetry, which is analogous to short selling in financial markets both in its consequences and underlying mathematical structure, partially explains the robustness of single-temperature strategies.

4.3.2 The Goldilocks Sampling Effect

Our second dulling effect is the *Goldilocks sampling effect*, which states that for any number of samples $K \in \mathbb{N}$, the questions most sensitive to the choice of sampling strategy are those whose hardness is on the same scale as K. This observation motivates our decision to characterize questions by their hardness throughout the paper, as questions of substantially different hardness can be analyzed largely independently.

We formalize this effect by considering a question with single-sample success probability $0 < P < 1$, and analyzing how its overall success probability changes as a function of the number of samples. Relaxing $K \in \mathbb{R}$ for analytical convenience, and using Equation 2, we obtain:

$$\frac{d\mathcal{P}}{dK} = -\bar{P}^K \ln\left(\bar{P}\right) \ , \tag{14}$$

which captures the *marginal gain* in success probability from adding additional samples. The Goldilocks effect is then described as follows:

Proposition 4.2 (The Goldilocks Sampling Effect)**.** Given the setup described above, where the number of samples K is fixed, then the marginal gain in success probability is maximized as a function of the question hardness when:

$$K = -\frac{1}{\ln\left(\bar{P}\right)} = H \ , \tag{15}$$

where H is the question hardness, as defined in Equation 3.

Furthermore, if the sampling scale deviates from the hardness scale by a factor $\alpha > 0$, i.e. $K = \alpha H$, then the normalized marginal gain is:

$$\left(\frac{d\mathcal{P}_{\text{opt}}}{dK}\right)^{-1} \frac{d\mathcal{P}_{\alpha}}{dK} = \alpha e^{1-\alpha} \ . \tag{16}$$

While the proof, provided in Appendix B.2.3 is a straightforward calculus exercise, it has an elegant interpretation. Questions that are significantly harder than the sampling budget K are almost never solved, regardless of the temperature allocation scheme. Conversely, questions that are much easier are also almost always solved, again regardless of the temperature allocation scheme. As a result, only questions whose hardness is on the same scale as K meaningfully contribute to the differences between sampling strategies.

This transition is also quite sharp. By analyzing the range of α for which the expression in Equation 16 remains at least $\frac{1}{e}$, we find that $0.16 \lesssim \alpha \lesssim 3.15$ is the range in which the marginal gain remains stable. Taking the ratio between these endpoints yields an efficient range spanning roughly a factor of 20 in hardness. Given the large variability in question difficulty observed in Section 3, this defines a relatively well-localized regime of relevant questions. As this effect is independent of the structure of the question profiles, it applies very broadly, extending to virtually any inference-time sampling hyperparameter.

With this effect we complete our explanation of why well-tuned single-temperature strategies are often sufficient. From Section 3.2.2, we know that questions of similar hardness have substantially overlapping temperature profiles, meaning that by the short-sampling effect, multi-temperature sampling provides only limited additional benefit within a fixed hardness regime. Although the average optimal temperature varies with hardness, the Goldilocks sampling effect now shows that we can largely ignore hardness regimes far from the relevant scale of K. This explains why varying the temperature as a function of K can capture most of the benefits of the ideal sampling strategy.

We demonstrate this effect in Figures 4, 29, 30, and 31 by comparing the optimal temperature with one whose success probability is 10% lower, and determining the factor β by which questions outside the hardness range for some $0 < D$:

$$\log K - D \leq \log H \leq \log K + D \tag{17}$$

must be reweighted for the suboptimal temperature to outperform the optimal one, as explained in Appendix H.1. As shown in Figures 4, 29, 30, and 31, the required reweighting grows approximately exponentially with D, demonstrating that the relative performance of different temperatures is determined primarily by questions whose hardness is close to K. This effect persists even for very large K, as shown in Appendix H.2, extending its relevance to the regimes considered in [9, 26].

This effect has an important practical implication. Since performance at budget K is primarily driven by questions whose hardness is on the scale of K, temperature-selection strategies should be relatively insensitive to mismatches between the hardness distributions of the training and deployment data. This suggests that even substantially skewed training distributions can yield temperature choices that transfer effectively, reducing the need for exact alignment between training and deployment data.

We test this prediction in Appendix G, where we randomly reweight questions according to hardness, optimize temperatures under the reweighted distribution, and then evaluate the resulting strategy on the original distribution. There we see that strategy is indeed robust to substantial distortion. Even when the KL divergence exceeds 3, on average roughly 90% of reweightings lead to less than a 10% performance loss. For KL divergences between 1 and 2, this fraction rises to 99%.

Furthermore, this effect suggests that selecting temperatures for a given question does not require knowing the question's hardness, but only the available computational budget. Although this observation is less relevant for pre-committed strategies, we believe that our results could be combined with adaptive strategies, as discussed in the Conclusion, where this property may prove valuable.

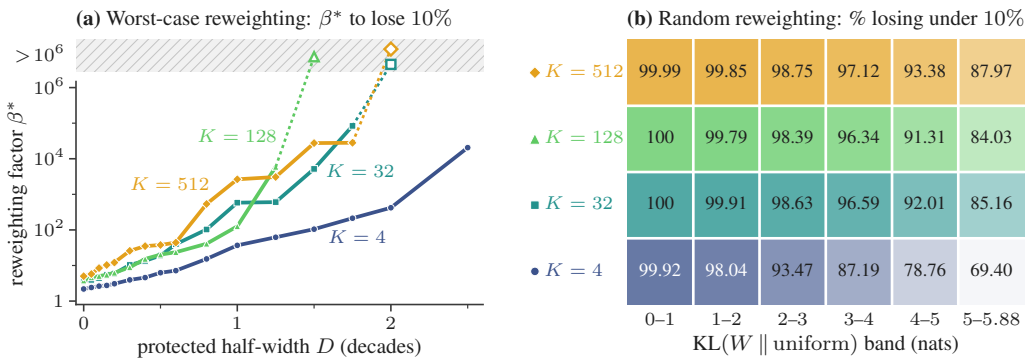


Figure 4: Sensitivity of the test pass@K score to the question hardness distribution used for temperature selection (Mistral 7B Instruct, LiveCodeBench). Left: Goldilocks protection. Questions whose hardness is within a factor 10^D of K keep their weight, a 10% loss requires reweighting the others by β^*, which grows exponentially with D. Right: KL divergence. Even far from the test distribution in KL divergence, most random training distributions W cost under 10%.

5 Conclusion

In this work, we studied pre-committed multi-temperature sampling strategies for LLM inference in the canonical pass@K setting. We began by analyzing how per-question success probability varies with sampling temperature, which we refer to as a question's temperature profile. Across models and datasets, these profiles exhibit a remarkably simple structure, typically with a single local maximum occurring either at $\tau = 0$ or at a positive temperature, followed by decay.

We then analyzed these profiles through the lens of hardness, defined as the effective number of samples required for a reliable solution. This perspective revealed several consistent trends. In particular, we saw that the average optimal temperature varies significantly with hardness, while efficient temperature ranges shrink as questions become harder, although only logarithmically. As a result, we found that substantial overlap between efficient temperature ranges remains even for difficult questions.

Building on these observations, we derived an efficient approximation to the theoretically optimal multi-temperature allocation scheme. We showed that although this strategy outperforms single-temperature approaches, its gains are typically small once the temperature is tuned as a function of the sampling budget K. Thus, much of the benefit of temperature diversity can be recovered by selecting a single well-tuned temperature.

We explained this behavior through two dulling effects. The first shows that overusing multiple temperatures typically incurs a larger penalty than underusing them. The second shows that performance at budget K is primarily governed by questions whose hardness is on the same scale as K. Together, these effects explain why single-temperature strategies remain strong in practice. They also suggest an important practical implication: multi-temperature strategies can be selected effectively without accurately knowing the hardness of the question, or even the full hardness distribution of the deployment data. This reduces the resources required to construct the training dataset, or to analyze specific questions during inference.

Overall, our results show that multi-temperature sampling is most useful when questions have well-separated efficient temperature ranges, but that such separation is limited in typical LLM settings. We discuss the limitations of our work, along with potential generalizations and broader impacts, in Appendix A.

Looking forward, extending this analysis beyond the pass@K setting, and combining it with adaptive or higher-dimensional inference strategies are important directions for future work. In such settings, greater separation may emerge as dimensionality increases, potentially enhancing the practical value of our approximation to the optimal sampling scheme.

Acknowledgments and Disclosure of Funding

This work is supported in part by the Israel Science Foundation Excellence Center, the US-Israel Binational Science Foundation, and the Israel Ministry of Science. The authors declare no competing interests.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[5] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[8] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[9] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022.

[10] 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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[12] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857, 2022.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019.

[14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, 2022.

[16] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[18] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[19] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213, 2022.

[22] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023.

[23] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[24] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551, 2023.

[25] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

[26] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[27] Yuheng Wu, Azalia Mirhoseini, and Thierry Tambe. On the role of temperature sampling in test-time scaling. *arXiv preprint arXiv:2510.02611*, 2025.

[28] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, 2024.

[29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[30] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[31] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, pages 39578–39601, 2024.

[32] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.

[33] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[34] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[35] John Hewitt, Christopher D. Manning, and Percy Liang. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, 2022.

[36] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

[37] Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, 2024.

[38] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898, 2018.

[39] Weihua Du, Yiming Yang, and Sean Welleck. Optimizing temperature for language models with multi-sample inference. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 14648–14668. PMLR, 13–19 Jul 2025.

[40] Nicola Cecere, Andrea Bacciu, Ignacio Fernández-Tobías, and Amin Mantrach. Monte Carlo temperature: A robust sampling strategy for LLM's uncertainty quantification methods. In *Proceedings of the 5th Workshop on Trustworthy NLP*, pages 305–320, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

[41] Noam Itzhak Levi. A simple model of inference scaling laws. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33984–33998. PMLR, 13–19 Jul 2025.

[42] Pei-Hsin Wang, Sheng-Iou Hsieh, Shih-Chieh Chang, Yu-Ting Chen, Jia-Yu Pan, Wei Wei, and Da-Chang Juan. Contextual temperature for language modeling. *arXiv preprint arXiv:2012.13575*, 2020.

[43] Chengkun Cai, Xu Zhao, Yucheng Du, Haoliang Liu, and Lei Li. T^2 of thoughts: Temperature tree elicits reasoning in large language models. *arXiv preprint arXiv:2405.14075*, 2024.

[44] Yuqi Zhu, Jia Li, Ge Li, YunFei Zhao, Jia Li, Zhi Jin, and Hong Mei. Hot or cold? Adaptive temperature sampling for code generation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 437–445, 2024.

[45] Shimao Zhang, Yu Bao, and Shujian Huang. EDT: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*, 2024.

[46] Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[47] Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. SPoC: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[48] Joshua Kazdan, Rylan Schaeffer, Youssef Allouah, Colin Sullivan, Kyssen Yu, Noam Levi, and Sanmi Koyejo. Efficient prediction of pass@k scaling in large language models. *arXiv preprint arXiv:2510.05197*, 2025.

[49] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, pages 58791–58831, 2025.

[50] Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore, 2023. Association for Computational Linguistics.

[51] Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *International Conference on Learning Representations*, pages 70333–70366, 2025.

[52] Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-H decoding: Adapting the creativity and coherence with bounded entropy in text generation. In *Advances in Neural Information Processing Systems*, volume 38, pages 28482–28513, 2025.

[53] Dorit S. Hochbaum and J. George Shanthikumar. Convex separable optimization is not much harder than linear optimization. *Journal of the ACM*, 37(4):843–862, 1990.

[54] Satoko Moriguchi, Akiyoshi Shioura, and Nobuyuki Tsuchimura. M-convex function minimization by continuous relaxation approach: Proximity theorem and algorithm. *SIAM Journal on Optimization*, 21(3):633–668, 2011.

[55] Rama Cont and Massoud Heidari. Optimal rounding under integer constraints. *arXiv preprint arXiv:1501.00014*, 2015.

[56] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[57] Constantin Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Mathematische Annalen*, 64:95–115, 1907.

[58] Alexander Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[59] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[60] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[61] OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[62] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

A Assumptions and Limitations

We begin the appendix by discussing the main assumptions and limitations of our work, along with their potential implications and directions for future investigation.

The first limitation concerns empirical coverage. While our numerical analysis is extensive and spans multiple models and datasets, the space of LLM architectures, training procedures, and real-world tasks is extremely large. Although the consistency of our results suggests that the observed phenomena are likely to be general, we cannot fully establish their universality without significantly broader evaluation. In particular, extending this analysis to additional datasets, more recent large-scale models, and qualitatively different generative paradigms (such as diffusion-based or hybrid models) is an important direction for future work. Further empirical studies, especially on state-of-the-art systems, would help determine how broadly these effects apply.

A second limitation lies in the relationship between our theoretical framework and real-world behavior. While all of our theoretical results are formally derived and rely on explicitly stated assumptions, these assumptions are not derived rigorously from first principles. For example, properties such as unimodality of temperature profiles and the structure of success probabilities are consistently observed across models and datasets, and we provide theoretical justification for why such behavior is expected. However, they are not proven to hold universally. As a result, although our theory provides strong intuition and closely matches observed behavior, formally characterizing the extent to which it applies to all real-world LLMs remains an open problem. Bridging this remaining gap between empirically grounded theory and full theoretical guarantees is an important direction for future work.

In the case of the short-sampling dulling effect in Section 4.3.1 and Appendix B.2.2, there is an additional limitation. While we were able to prove this effect under a variety of assumptions, and demonstrate it across diverse representative cases, we did not obtain a single unified theorem that fully characterizes the phenomenon in all settings. Instead, our analysis consists of several complementary results, each capturing a different aspect of the effect. Although these results provide strong evidence and a consistent explanation of the behavior, a more complete and general characterization remains open. Nevertheless, the assumptions are not arbitrary, but are supported by consistent empirical evidence and underlying structural arguments about LLM behavior. As such, we believe they capture the dominant factors governing the observed phenomena, and that the resulting conclusions remain informative even beyond the exact regimes in which they are formally justified.

More broadly, while our empirical results strongly support the theoretical analysis, additional computational validation would further strengthen the conclusions. In particular, exploring more extreme regimes, larger sample budgets, and more diverse distributions of questions could help refine the observed trends and identify potential edge cases. Nevertheless, the consistency of these results across multiple formulations and settings suggests that the effect captures a fundamental mechanism rather than an artifact of specific assumptions. In particular, the qualitative behavior it predicts aligns closely with both empirical observations and the structure of the optimization problem, giving us confidence that the conclusions remain valid in practice.

Overall, we believe that our work provides strong evidence for the existence and practical relevance of the phenomena we describe. However, fully understanding their generality, as well as their precise theoretical foundations, will require further empirical and theoretical investigation.

B Theory

B.1 Ideal Temperatures Sampling Approximation Derivation

In this section, we complete the analysis of Section 4.1, where we introduced an approach for approximating the ideal temperature sampling distribution that maximizes the total success probability. Here, we show how the approximation error scales with properties of the dataset, and under temperature discretization. We demonstrate that, in our setting, this approach provides a reliable approximation.

The optimal sampling scheme is defined for any $K \in \mathbb{N}$, by the optimal (not necessarily distinct) temperatures $0 \leq \tau_1, \ldots, \tau_K$ that maximize the success probability, or equivalently, minimize its

complement, as defined in Equation 1:

$$\bar{\mathcal{P}}\left(\tau_1,\ldots,\tau_K\right)=\mathbb{E}_{q\sim\mathcal{Q}}\prod_{k=1}^{K}\bar{P}_q\left(\tau_k\right)=\mathbb{E}_{q\sim\mathcal{Q}}\exp\left(\sum_{k=1}^{K}\ln\left(\bar{P}_q\left(\tau_k\right)\right)\right)\ . \tag{18}$$

As discussed in the main text, the primary difficulty with this formulation, is that it involves optimization over a continuous and unbounded space of temperatures. However, under relatively mild assumptions, this problem can be discretized significantly, as formalized in the following theorem:

Theorem B.1 (Temperatures Discretization)**.** Assume the setup presented in Section 4.1, a probability $0<\rho<1$, a truncation temperature $T>0$, and a discretization scale $\Delta>0$, this defines a grid:

$$\mathcal{G}=\{0,\Delta,2\Delta,\ldots,M\Delta\}\ , \tag{19}$$

where $M=\left\lceil\frac{T}{\Delta}\right\rceil$.

Suppose that within this probability $1-\rho$, over the choice of questions $q\sim\mathcal{Q}$, we know that the failure probability satisfies, for every $0\leq\tau,\tau'\leq T$ such that $|\tau-\tau'|\leq\Delta$, that it is proportionally Lipschitz bounded:

$$\bar{P}_q(\tau')\leq\exp\left(L|\tau-\tau'|\right)\bar{P}_q(\tau)\ , \tag{20}$$

and that beyond the truncation temperature i.e. $T<\tau$, for some $\alpha>0$ the endpoint of the truncated grid satisfies:

$$\bar{P}_q(M\Delta)\leq\exp(\alpha)\bar{P}_q(\tau)\ . \tag{21}$$

Denote the optimal failure probabilities over the continuous temperature space and over the discrete grid by:

$$\bar{\mathcal{P}}^{\star}=\inf_{\tau_1,\ldots,\tau_K\geq0}\bar{\mathcal{P}}\left(\tau_1,\ldots,\tau_K\right),\quad\bar{\mathcal{P}}^{\star}_{\mathcal{G}}=\min_{\gamma_1,\ldots,\gamma_K\in\mathcal{G}}\bar{\mathcal{P}}\left(\gamma_1,\ldots,\gamma_K\right)\ . \tag{22}$$

Then we find that for every $K\in\mathbb{N}$, the error of the best assignment over the discrete temperatures is bounded by:

$$0\leq\bar{\mathcal{P}}^{\star}_{\mathcal{G}}-\bar{\mathcal{P}}^{\star}\leq\left(e^{K(L\Delta+\alpha)}-1\right)\bar{\mathcal{P}}^{\star}+\rho\ . \tag{23}$$

The proof idea, presented in full in Appendix J.1, is to show that any continuous temperature selection can be replaced by nearby grid temperatures while incurring only a controlled relative increase in failure probability. This relative error compounds across the K temperatures, yielding a factor of $e^{K(L\Delta+\alpha)}$. Questions for which this control fails contribute at most ρ, since failure probabilities are bounded by 1. Consequently, the optimal grid-based solution is close to the optimal continuous solution, up to the multiplicative discretization error and the mass of such "bad" questions.

The strength of this theorem lies in the fact that our final bound in Equation 23 is proportional to the failure probability. Thus, although the multiplicative term grows with K, the overall deviation decreases as long as L, Δ, and α are sufficiently small.

In our setting, as shown in Section 3.1, the profiles tend to decay rapidly for relatively small τ, so Equation 21 is trivially satisfied for small α. Equation 20 is well satisfied for small L, particularly for harder questions, since we consider the proportional Lipschitz continuity of $1-P$. This means that the approximation in Equation 23 is robust even for large K and very small probabilities.

Having established that the problem can be effectively reduced to a finite set of temperatures, we can now investigate the effect of operating with a limited number of questions.

Theorem B.2 (Finite Question Distribution Reduction)**.** Assume the setup of Theorem B.1, and let $n\in\mathbb{N}$ questions $\mathcal{Q}'=\{q_1,\ldots,q_n\}$ be drawn independently from $\mathcal{Q}$. For any choice of grid temperatures $\gamma_1,\ldots,\gamma_K\in\mathcal{G}$, define the empirical failure probability, as in Equation 8, by:

$$\bar{\mathcal{P}}_{\mathcal{Q}'}\left(\gamma_1,\ldots,\gamma_K\right)=\frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{K}\bar{P}_{q_i}\left(\gamma_k\right)\ . \tag{24}$$

Let an empirical minimizer of this objective be denoted by:

$$(\hat{\gamma}_1,\ldots,\hat{\gamma}_K)\in\arg\min_{\gamma_1,\ldots,\gamma_K\in\mathcal{G}}\bar{\mathcal{P}}_{\mathcal{Q}'}\left(\gamma_1,\ldots,\gamma_K\right)\ . \tag{25}$$

We denote the result of plugging these temperatures back into the population objective as:

$$\bar{\mathcal{P}}'_{\mathcal{Q}}=\bar{\mathcal{P}}_{\mathcal{Q}}\left(\hat{\gamma}_1,\ldots,\hat{\gamma}_K\right)\ . \tag{26}$$

Also, for every probability $0 < \rho' < 1$, define:

$$\Lambda = \log \binom{K+M}{M} + \log \left(\frac{1}{\rho'} \right) . \tag{27}$$

Then, with probability at least $1 - \rho'$ over the draw of $\mathcal{Q}'$, the difference between this failure probability and the optimal failure probability over the grid, defined in Equation 22, is bounded by

$$0 \leq \bar{\mathcal{P}}'_{\mathcal{Q}} - \bar{\mathcal{P}}^{\star}_{\mathcal{G}} \leq 2\sqrt{\frac{\Lambda \bar{\mathcal{P}}^{\star}_{\mathcal{G}}}{n}} + \frac{10\Lambda}{3n} . \tag{28}$$

The proof idea, also presented in Appendix J.1, is similar to standard bounds on the error incurred when approximating a population distribution using a finite sample over a discrete set of temperatures. However, instead of bounding the error for a fixed choice of temperatures, we need to control the error of the temperatures selected by minimizing the empirical objective. To this end, we compare every possible grid assignment to the best grid assignment under the full population objective. If an assignment has much larger population failure probability than the optimum, then it is unlikely to look better than the optimum on the sampled questions.

A Bernstein-type concentration bound makes this comparison sharper by using the fact that the relevant random variables are bounded failure probabilities, whose variance is controlled by their mean. Finally, we take a union bound over the unordered grid assignments, whose number is $\binom{K+M}{M}$, since the order of the K temperatures does not affect the objective. This shows that, with high probability, the empirical minimizer cannot be much worse under the true distribution than the best grid assignment.

The theorem shows that the gap between the true and empirical distributions grows only slowly as a function of K, but decreases rapidly with n. It also highlights which questions are most important to estimate accurately in order to minimize the error, as discussed in Section 4.3.2.

Combined with Theorem B.1, this result implies that our evaluation of the ideal distribution, based on a relatively small dataset and a discretized set of temperatures, is very reliable.

B.2 Detailed Analysis of the Dulling Effects

In this part of the appendix, we complete the analysis from Section 4.3, by providing the full proofs of the main results, and elaborating on the associated discussions.

B.2.1 The Temperature Saturation Property

Before analyzing the dulling effects themselves, we first establish an important structural property of the optimization problem defined in Section 4.1. Specifically, we show that the number of distinct temperatures in the optimal solution is bounded by the number of questions in the dataset. This is formalized in the following lemma:

Lemma B.3 (The Temperature Saturation Property)**.** Given the setup described in Section 4.1, for a finite dataset of questions with $n \in \mathbb{N}$, the number of distinct temperatures in an optimal solution is at most n.

The intuition behind this result is that after sampling n questions, each temperature is seen only through its n log-failure probabilities on those questions. Thus, the relaxed optimization problem is a convex optimization problem of only n effective dimensions, meaning we can only have up to n independent temperatures. A complete and rigorous proof is provided in Appendix J.2.

B.2.2 Detailed Analysis of the Short-Sampling Effect

We can now take a closer look at the Short-Sampling Dulling Effect, introduced in Section 4.3.1. While the high-level intuition was discussed there, a more careful analysis is required to make the argument precise.

Having justified the reduction to the three-temperature case in Lemma B.3, we now proceed to prove Proposition 4.1.

Proof - Proposition 4.1. The proof of this proposition follows from a direct algebraic computation. By Equation 10, the probability of solving a question using the compromise temperature is

$$\mathcal{P}_{\text{comp}} = 1 - (1 - P_C)^K \ . \tag{29}$$

For the two-expert strategy, symmetry implies that the optimal allocation is uniform, so each expert temperature receives $\frac{K}{2}$ samples. Applying Equation 10 again yields

$$\mathcal{P}_{\text{expert}} = 1 - (1 - P_E)^{\frac{K}{2}} \ . \tag{30}$$

Taking the difference between the two strategies, we obtain:

$$\Delta\mathcal{P} = \mathcal{P}_{\text{expert}} - \mathcal{P}_{\text{comp}} = (1 - P_C)^K - (1 - P_E)^{\frac{K}{2}} \ . \tag{31}$$

Using the definition of ΔP from Equation 11, we can rewrite this as

$$\Delta\mathcal{P} = \left(1 - \frac{P_E}{2} + \Delta P\right)^K - (1 - P_E)^{\frac{K}{2}} \ . \tag{32}$$

Finally, applying the first-order approximation

$$\log(1 - x) = -x + O\left(x^2\right), \tag{33}$$

and noting that $\Delta P = O(P_E)$ since $P_C \leq P_E$, we recover Equation 12, completing the proof. □

As seen from the proof, the threshold around $P_C = \frac{P_E}{2}$ arises naturally from the low-probability approximation. In this regime, the success probability appears in the exponent, scaled by the number of samples K. Consequently, increasing the number of samples allocated to a temperature by some factor is, to first order, equivalent to increasing the single-sample success probability by the same factor.

Proposition B.1 (The Weighted Short-Sampling Effect)**.** Given the setup described above, when the expert temperatures is optimized in the continuous relaxation, then up to first-order, the result is equivalent to Proposition 4.1, with the substitution:

$$P_E \mapsto P_E - \frac{1}{K} \log\left(4r(1 - r)\right) \ . \tag{34}$$

Or equivalently:

$$\bar{P}_E \mapsto \sqrt[K]{4r(1 - r)}\bar{P}_E \ . \tag{35}$$

The idea of the proof (given in full in Appendix J.3) is as follows. The compromise temperature remains unaffected, since all questions share the same effective behavior under it. In contrast, the expert allocation becomes asymmetric due to the weighting. However, since the problem still reduces to a two-variable optimization, it can be solved explicitly, yielding the $\sqrt{4r(1 - r)}$ factor. Since the contribution of the number of questions appears multiplicatively in the exponent, this factor enters accordingly.

While this proposition shows that asymmetry between question groups can slightly favor expert allocations, the effect is quite limited. To see this, consider the regime $K \simeq H \simeq \frac{1}{P}$, which, as discussed in Section 4.3.2, is the relevant scaling.

In this regime, the contribution of the $\log\left(4r(1 - r)\right)$ term is small. For example, to double the effective success probability, which in the Gaussian case, corresponds roughly to increasing the efficient range width by a factor of about 1.4. This means we already require the relatively imbalanced setting of $r \simeq 10.2\%$. To double the effective efficient range, we would get that we need an extreme imbalance of $r \simeq 2.81 \cdot 10^{-6}\%$. Moreover, this gain only benefits one side. Increasing the effective efficient range of one group necessarily reduces that of the other. This further highlights the robustness of the short-sampling effect, as even significant asymmetries translate into only marginal overall gains.

This weighted setting therefore represents an extreme case in favor of multi-temperature strategies. In more realistic scenarios, introducing additional questions with intermediate profiles would only strengthen the performance of the compromise temperature, as it benefits from broader coverage. This

further supports the robustness of the efficient range-based notion of separation and the generality of the short-sampling effect.

The second source of complexity is the fact that the questions' profiles may not be identical. To study this case, we analyze a setting with two questions having different expert probabilities P_E^1, P_E^2 and different compromise probabilities P_C^1, P_C^2. We restrict our attention to the two questions case, as the previous analysis shows that extending to multiple questions does not qualitatively change the conclusions.

Proposition B.2 (The Non-Symmetric Short-Sampling Effect)**.** Given the setup described above, the result is equivalent to Proposition 4.1, with

$$\Delta P = \frac{\frac{1}{2}P_E^1 + \frac{1}{2}P_E^2 - P_C^1 - P_C^2}{2} \ . \tag{36}$$

We do not present the proof explicitly, as the analyses follows the same lines as that for the proof of Proposition B.1.

This result shows that even when question profiles differ, as long as their efficient ranges overlap, the compromise temperature tends to dominate. In this sense, our definition of separation via efficient range overlap provides only a lower bound. In practice, the preference for compromise temperatures can be even stronger.

Finally, we examine the validity of the simplifying assumption that the success probability outside a question's comprise region can be treated as negligible.

To this end, we consider a simplified model in which the temperature profiles are Gaussians, with identical amplitude A and variance σ^2, but with shifted means:

$$P_1(\tau) = A \exp\left(-\frac{(\tau-\tau_1)^2}{2\sigma^2}\right), \qquad P_2(\tau) = A \exp\left(-\frac{(\tau-\tau_2)^2}{2\sigma^2}\right) \ . \tag{37}$$

We denote by τ_C the intersection point of the two profiles. Since the two Gaussians have the same amplitude and variance, this point is given by the midpoint.

Proposition B.3 (Gaussian Tail Correction)**.** In the Gaussian setting described above, the two-experts strategy outperforms the compromise temperature, in the first-order Gaussian-tail approximation, if and only if:

$$P_E > \sqrt{e} P_C \simeq 1.65 P_C \ . \tag{38}$$

This result shows that $\frac{P_E}{2}$ is a reasonable approximation for the critical value of P_C in realistic cases, where the two efficient ranges have comparable shape. Once the Gaussian tail is taken into account, the threshold shifts from $0.5P_E$ to approximately $\frac{1}{\sqrt{e}}P_E \simeq 0.61P_E$, which changes the effective efficient range size by a factor of only about 1.18.

Although real temperature profiles are not exactly Gaussians, the empirical profiles we observed in Section 3.1 exhibit similar behavior. Once they drop to roughly half of their maximum, they continue to decay rapidly. This supports the approximation used in the simplified short-sampling argument. The Gaussian case is solved analytically in Appendix J.3.

It is difficult to combine all of these arguments into a single, fully general theorem, as pathological counterexamples can always be constructed. Nevertheless, the results presented here demonstrate the robustness of the dulling effects. They also justify the use of efficient range overlap as a meaningful measure of separation, and explain why the cost of introducing too many temperatures typically outweighs the cost of using too few.

B.2.3 Proof of The Goldilocks Sampling Effect

We conclude this section by proving our main theoretical result of the paper, Proposition 4.2.

The proof of this proposition is just a straightforward calculus exercise. We begin by inverting the definition of hardness to express the probability as a function of hardness, rather than the other way around as in Equation 3:

$$\bar{P} = e^{-1/H} \ . \tag{39}$$

Substituting this into the marginal gain defined in Equation 14 yields:

$$\frac{d\mathcal{P}}{dK} = \frac{1}{H} e^{-K/H} \,. \tag{40}$$

This expression already reveals the main behavior of the problem. If $H \ll K$, the question is easy relative to the sampling budget, and most of the success probability has already been captured. If $H \gg K$, the question is too difficult, and additional samples provide little benefit. The largest marginal gain occurs between these two regimes.

To locate this point, we fix K and maximize the expression with respect to H. Taking the derivative and setting it to zero yields:

$$H = K \,. \tag{41}$$

This is the unique critical point, and it corresponds to a maximum, as we can verify by the second derivative, establishing Equation 15.

To quantify the decay away from this optimal scale, we parameterize

$$K = \alpha H \quad \Longleftrightarrow \quad H = \frac{K}{\alpha} \,. \tag{42}$$

Substituting this into Equation 40, we obtain:

$$\frac{d\mathcal{P}_\alpha}{dK} = \frac{\alpha}{K} e^{-\alpha} \,. \tag{43}$$

At the optimal scale, where $H = K$ (i.e., $\alpha = 1$), the marginal gain is:

$$\frac{d\mathcal{P}_{\text{opt}}}{dK} = \frac{1}{K} e^{-1} \,. \tag{44}$$

Normalizing by this optimal value gives

$$\left(\frac{d\mathcal{P}_{\text{opt}}}{dK}\right)^{-1} \frac{d\mathcal{P}_\alpha}{dK} = \alpha e^{1-\alpha} \,, \tag{45}$$

which proves Equation 16 and completes the proof.

C Computational Resources

We used the `vLLM` inference engine for batched sampling from each of the four base models, on NVIDIA L40S, A40, GeForce RTX 5090, and GeForce RTX 4090 GPUs (24 to 48 GB of memory each), for a total of about 14,000 GPU-hours.

Code and reproducibility. Our code is available at `https://github.com/avivorly/one-temperature`. It contains a vLLM sampling script over a temperature grid, the MATH-500 grader, the unbiased pass@K estimator, and the temperature-allocation comparison (best single temperature, uniform mixing, and the optimal allocation) with leave-one-out and 10-fold held-out evaluation. Its README gives the exact commands and lists the parts it does not include. All benchmarks and models are public; their sources, versions and licenses are listed in the paper checklist.

C.1 GPT OSS 20B

Dataset	Included questions	Samples per (q, τ)	Temperature grid
LiveCodeBench	766	$\sim$1.5k; harder questions $\sim$10k, sometimes more	16: 0.1, 0.2, . . ., 1.2, 1.4, 1.5, 1.7, 2.0
MATH-500	500	1500	the same 16
Hi-ToM	678	1500	31: 0.075, 0.15, . . ., 1.95, 2.0, 2.2, 2.5, 3.0, 4.0

C.2 Mistral 7B Instruct

Dataset	Included questions	Samples per (q, τ)	Temperature grid
LiveCodeBench	359	$\sim$1.5k; harder questions $\sim$10k, sometimes more	16: 0.1, 0.2, . . ., 1.2, 1.4, 1.5, 1.7, 2.0
MATH-500	455	1500	the same 16
Hi-ToM	615	1500	31: 0.075, 0.15, . . ., 1.95, 2.0, 2.2, 2.5, 3.0, 4.0

C.3 Qwen3-8B

Dataset	Included questions	Samples per (q, τ)	Temperature grid
LiveCodeBench	646	$\sim$1.5k; harder questions $\sim$10k, sometimes more	38: 0.1, 0.2, . . ., 3.5, 3.7, 4.0, 4.2
MATH-500	496	750	20: 0.1, 0.2, . . ., 2.0
Hi-ToM	466	1500	31: 0.075, 0.15, . . ., 1.95, 2.0, 2.2, 2.5, 3.0, 4.0

C.4 Qwen3-0.6B

Dataset	Included questions	Samples per (q, τ)	Temperature grid
LiveCodeBench	85	approximately doubled on some particularly hard questions	20: 0.1, 0.2, . . ., 2.0
MATH-500	492	1500	the same 20
Hi-ToM	743	1500	31: 0.075, 0.15, . . ., 1.95, 2.0, 2.2, 2.5, 3.0, 4.0

C.5 Datasets

From LiveCodeBench [49], we use the six datasets V1 to V6, with 975 coding questions, and Qwen3-0.6B uses 110 of them. MATH-500 is the 500-question test split of the MATH benchmark [11] defined for PRM800K [31]. Hi-ToM [50] is a multiple-choice benchmark of higher-order theory-of-mind reasoning with 1,200 questions. On Hi-ToM, we analyze the questions whose best single-temperature pass rate was below 0.30 in a first stage of 500 samples per question and temperature (1000 for Mistral 7B Instruct).

The token limit is 4096 new tokens on MATH-500 (2048 for Qwen3-8B) and 1000 on Hi-ToM (500 for Mistral 7B Instruct). On MATH-500, every model received the instruction "Solve the following math problem efficiently and clearly. Reason step by step, and put your final answer within `\boxed{}`." in its chat template, and Qwen3-8B received it followed by a sentence that states its limit. On Hi-ToM, GPT OSS 20B and the two Qwen3 models received each question's prompt exactly as the dataset gives it, and Mistral 7B Instruct received it followed by a fixed instruction that states its limit. On MATH-500 and Hi-ToM, GPT OSS 20B used the Harmony chat format with reasoning effort `medium` and the response primed to open in the `final` channel, and ran in bfloat16 with the checkpoint's native MXFP4 quantization. On these two datasets, Mistral 7B Instruct used the Mistral chat template and ran in float16, and the two Qwen3 models used the Qwen3 chat template with thinking mode disabled and ran in bfloat16 (Qwen3-8B) and float16 (Qwen3-0.6B).

On MATH-500, the answer is the content of the last `\boxed{}` span of the completion, and a sample with no such span fails. A sample passes if its answer equals the reference answer after normalization, or if an equivalence layer of generic rules and a time-bounded symbolic check with SymPy accepts it. On Hi-ToM, where the prompt lists all 15 options, the answer is the first valid option in the completion, that is, the first mention of one of the question's own options, by its name or by its letter given as an answer label. A completion that names no valid option fails.

D Question Profile Analyses

We begin this appendix with a detailed exposition of our analysis of the most fundamental objects we consider, which underpin the remainder of the text.

Goal and choice of basis

Our starting point is a set of discrete pass-rate measurements on a temperature grid, one curve per sampled question q_i, and our goal is to recover the smooth profile $P_{q_i}(\tau)$ that best denoises those measurements. Concretely, for question q_i we observe s_{ij} successes out of ν trials at each sampled temperature τ_j, and we model their source as an underlying smooth function $P_{q_i}(\tau)$ that we wish to estimate faithfully.

We do not want to assume a specific functional form as a sigmoid, an exponential decay, a Gaussian peak, because any such ansatz would bias the recovered peak position, peak height, and tail behavior toward what the family can express, as conclusions drawn afterwards would reflect the chosen form as much as the data.

We therefore fit in a **general basis**: a polynomial space rich enough to express any sufficiently smooth P_{q_i} up to a chosen degree, with **shape constraints** that encode minimal physical priors but no functional form. The data, rather than an ansatz, determines the curve.

Confidence intervals for each measurement

Each measurement is binomial, $s_{ij} \sim \mathrm{Binomial}(\nu, P_{q_i}(\tau_j))$. For the per-point error bars we use **exact Clopper-Pearson 1σ intervals**: at the two-sided tail probability $\alpha_{\mathrm{CP}} = 1 - \mathrm{erf}(1/\sqrt{2}) \approx 0.317$ corresponding to 1σ, the interval $[\mathrm{lo}, \mathrm{hi}]$ for a binomial proportion is given by quantiles of the Beta distribution,

$$\mathrm{lo} \;=\; \mathrm{Beta}^{-1}_{s,\,\nu-s+1}\!\left(\tfrac{\alpha_{\mathrm{CP}}}{2}\right), \qquad \mathrm{hi} \;=\; \mathrm{Beta}^{-1}_{s+1,\,\nu-s}\!\left(1-\tfrac{\alpha_{\mathrm{CP}}}{2}\right), \tag{46}$$

with $\mathrm{lo} = 0$ when $s = 0$ and $\mathrm{hi} = 1$ when $s = \nu$. We prefer this exact construction over Wald or Wilson approximations because Clopper-Pearson has guaranteed coverage at the boundary counts (s near 0 or ν) where Gaussian-style intervals systematically under-cover.

Clopper-Pearson intervals are asymmetric around $\hat{p} = s/\nu$, especially near the boundary. We keep the asymmetry rather than collapse it: each datum carries both half-widths $\sigma_- = \hat{p} - \mathrm{lo}$ and $\sigma_+ = \mathrm{hi} - \hat{p}$. Symmetrizing to a single $\sigma_{\mathrm{eff}} = (\mathrm{hi} - \mathrm{lo})/2$ would discard information that the Clopper-Pearson computation already produced and would shift the implied 1σ band off-center.

Loss: weighted least squares with side-dependent variance

We fit by **weighted least squares (WLS)**, with each residual normalized by the actual CP half-width on its side. Without side-dependent weights, the asymmetric Clopper-Pearson information preserved in the previous subsection would simply be discarded. Concretely:

$$\mathcal{L}(P_{q_i}) \;=\; \sum_j \left(\frac{P_{q_i}(\tau_j) - \hat{p}_{ij}}{\sigma_\pm(P_{q_i}, \hat{p}_{ij})}\right)^2, \qquad \sigma_\pm \;=\; \begin{cases} \sigma_+ & \text{if } P_{q_i}(\tau_j) > \hat{p}_{ij}, \\ \sigma_- & \text{otherwise}. \end{cases} \tag{47}$$

This loss is convex in the polynomial coefficients, C^1 at $P_{q_i}(\tau_j) = \hat{p}_{ij}$ (the gradient from both sides is zero regardless of which σ applies, so only the second derivative jumps), and faithful to the binomial geometry near $s = 0$ and $s = \nu$ where the CP asymmetry is most pronounced.

The full binomial negative-log-likelihood would be statistically more rigorous, but its curvature diverges as $p \to 0$ or 1, breaking the closed-form step size used below. The two-piece Gaussian above is a tractable proxy.

Shape constraints

We use the **Bernstein basis of degree** 10 on a per-question domain bracket $[\tau_{\mathrm{lo}}, \tau_{\mathrm{hi}}]$, taken as the support of the non-zero observations padded by one temperature step on each side; $P_{q_i} \equiv 0$ outside. Bernstein is chosen over the monomial or Legendre bases because shape constraints transfer cleanly from the function to the coefficients $\{c_j\}$: non-negative coefficients yield a non-negative function, and unimodal coefficient sequences yield a unimodal function, which makes constrained optimization tractable. Degree 10 is large enough to express the observed shapes and small enough to avoid the conditioning blow-up that plagues higher-degree polynomial bases on a fine grid. Moreover, degree

Table 1: Shape constraints translated from the function level to the Bernstein-coefficient level. The $\{c_j\}$ are the Bernstein-basis weights of the fitted profile.

Constraint	Function-level meaning	Bernstein-coefficient meaning
Non-negativity	$P_{q_i}(\tau) \geq 0$ for all τ	All coefficients $c_j \geq 0$
Unimodality	$P_{q_i}(\tau)$ has a single peak	Sequence $\{c_j\}$ is unimodal

10 is feasible because the shape regularization itself (the non-negativity and unimodality projections below) reduces the effective number of free polynomial parameters; without those constraints a lower degree would be safer, but with them 10 is empirically optimal.

We impose two minimal physical priors on P_{q_i}:

- **Non-negativity.** P_{q_i} is a probability.
- **Unimodality.** Sampling temperature trades exploration against exploitation, so on physical grounds each question has a single optimal regime; multimodal P_{q_i} would correspond to two distinct mechanisms favoring different temperatures, which we do not see in the data.

Both constraints are enforced as **projections during optimization**, not as additive penalties: the current Bernstein coefficient vector is periodically projected onto the cone of non-negative unimodal sequences via isotonic regression around each candidate peak position, retaining the projection closest in Euclidean distance. We deliberately do not add an explicit smoothness penalty: the bounded polynomial degree combined with the two priors already produces visibly smooth curves, and any explicit smoothness term would compete with the data fit at the peak and bias the recovered peak height downward.

Optimization

The WLS loss is minimized by projected gradient descent on the Bernstein coefficients, with a Lipschitz-safe step size derived from the largest eigenvalue of the WLS Hessian (estimated by power iteration). Initialization is at zero, which biases the iterates toward small-norm solutions before the projections engage. In the early iterations the projection is soft-blended into the iterate so the data fit can settle before the shape priors take effect; later iterations apply it as a hard projection. The optimization runs for a fixed budget chosen generously enough that further iteration changes the coefficients negligibly.

The output, per question, is a smooth, non-negative, unimodal $P_{q_i}(\tau)$: a degree-10 polynomial inside its bracket and zero outside.

Question filtering and sample-budget conventions

Minimum peak success probability (applies to every dataset and every plot). A single global filter is applied to every dataset and every figure in Appendices E to H: questions whose fitted peak success probability $\sup_\tau P_q(\tau)$ is below 10^{-3} are excluded. This filter applies once, in advance, before any plot is drawn or any table is built.

We have observed questions whose hardness H_q (Equation 3) extends another order of magnitude beyond this cutoff, but including them would be misleading for two reasons:

1. *Fit instability.* Below 10^{-3} peak success probability, the available data per question is so sparse that the polynomial fit becomes chaotic and is no longer representative of the underlying response curve.
2. *Goldilocks irrelevance.* By the Goldilocks sampling effect (Proposition 4.2), questions whose hardness H (Equation 3) is much larger than $K_{\max}$ are negligible at the K values reported in Appendices E to H.

Unsolvable questions. Questions that no temperature ever solves (peak success probability exactly zero) are excluded everywhere, from every plot, every table, and every denominator. They contribute nothing to any number reported below.

Maximum sample budget. All pass@K traces, leave-one-out (LOO) points, and ratio-to-uniform curves in Appendices E to H are evaluated up to $K_{\max} = 1024$. Larger values of K are uninformative for the deployed setting and were clipped at the report-build stage to keep the optimal-allocation LOO solve tractable.

E Section 3 Expanded Results

In this section, we present the analyses and results from Section 3 after deriving the question profiles, for the four models on the three datasets, LiveCodeBench, MATH-500, and Hi-ToM. We begin by describing the analysis procedure and then present the results.

E.1 Statistical Analysis Approach

Sliding window. Throughout this paper, "sliding window of size N" (where N is even) refers to the N questions closest in hardness to a given anchor point: the $N/2$ questions with the next-higher hardness and the $N/2$ questions with the next-lower hardness. At the edges of the data the window is clipped symmetrically (so the window may shrink, but never becomes asymmetric). Quantities reported "per window" average over those $\leq N$ neighbors.

How to read the per-question peak temperature plot. Each grey dot is one included question, plotted at its hardness H_q (Equation 3, x-axis on a log scale) versus its peak temperature $\tau_q^\star = \arg\max_\tau P_q(\tau)$ (y-axis, linear).

Sliding averages. The blue line is a symmetric sliding mean of $\tau_q^\star$ taken over a window of 30 consecutive questions in the H-sorted list (the same definition as the global "sliding window of size N" in Section E.1, with $N = 30$). At the edges the window is clipped symmetrically so it shrinks but never becomes asymmetric. The red line is the same running mean, but restricted to the subset of questions with $\tau_q^\star > 0.1$; this isolates the trend in $\tau_q^\star$ once the spike of near-zero-peak questions has been removed. Each line leaves out the 15 questions at each end of its H-sorted list, where its window is not full. The figures compute H_q with the peak pass rate capped at 99.999%, so every question at or above that rate sits at $H = 0.0869$. On LiveCodeBench and MATH-500, each line then starts at the first remaining question whose next question is at least 5% harder. This removes questions at $H = 0.0869$ and at most two other questions per line. The Hi-ToM lines use every full window, because no Hi-ToM question sits at $H = 0.0869$.

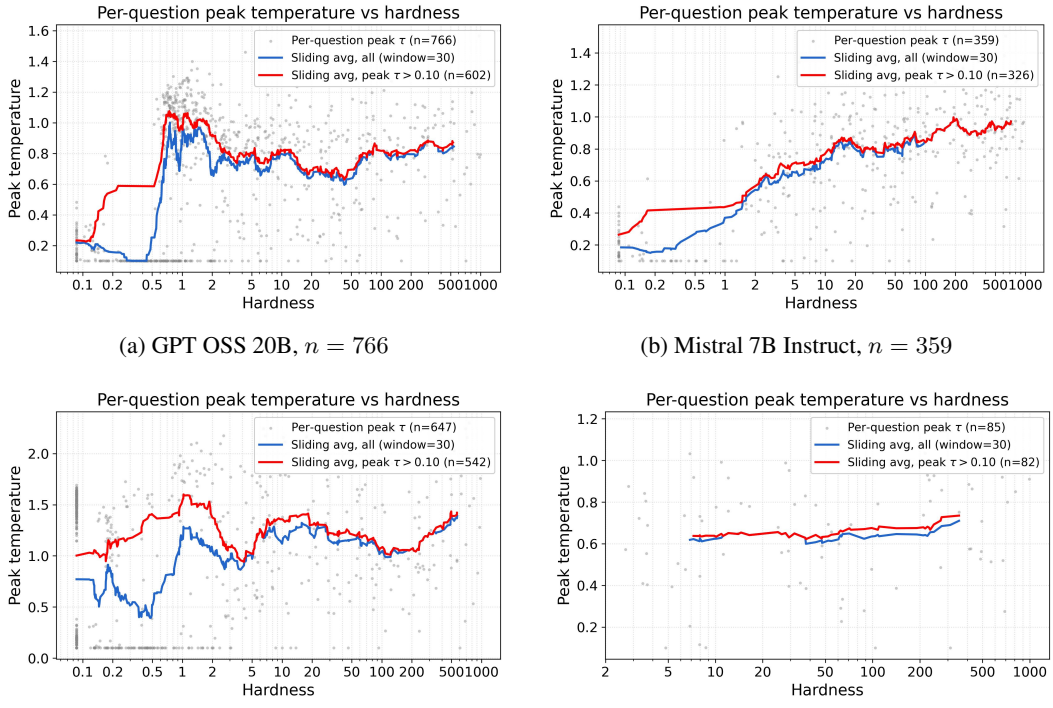
What to look for. A trend in either curve as H increases indicates that the hardness of a question carries information about its preferred sampling temperature. The red curve is the more informative of the two when the dataset contains a large near-zero cluster, because the all-questions blue line is otherwise dominated by it.

How to read the efficient-range-size plot. Each grey dot is one included question, plotted at its hardness H_q (Equation 3, x-axis on a log scale) versus the size $|E_q|$ of its efficient temperature range (Equation 4, y-axis on a linear scale). For a unimodal profile the efficient range is approximately an interval $[\,t_L,\, t_R\,]$, so we report $|E_q| = t_R - t_L$ directly.

OLS fit. The red line is the ordinary least-squares regression of $|E_q|$ on $\log_{10} H_q$; the slope, intercept, and R^2 are printed in the legend. The slope is the rate at which the efficient range shrinks per decade of hardness.

What to look for. A clearly negative slope confirms the shrinking-range trend stated in the main text (Section 3.2.2): harder questions have narrower efficient temperature windows. The R^2 summarizes how tight that log-hardness trend is across the dataset.

How to read the maximum-coverage plot. We split the hardness range $[\min_q H_q,\ \max_q H_q]$ into log-uniform bins (by default $\lfloor n/20 \rfloor$ bins, where n is the number of included questions). For each bin b, let M_b be the number of in-bin questions, and let $\mathcal{C}(\tau)$ be the number of those questions whose efficient temperature range E (Equation 4) contains temperature τ. We compute $\max_\tau \mathcal{C}(\tau)$ via a sweep-line over the range endpoints and plot $\max_\tau \mathcal{C}(\tau)/M_b$ as a bar spanning the bin's width.



(a) GPT OSS 20B, $n = 766$

(b) Mistral 7B Instruct, $n = 359$

(c) Qwen3-8B, $n = 647$

(d) Qwen3-0.6B, $n = 85$

Figure 5: Per-question peak temperature vs. hardness on LiveCodeBench, with sliding averages over all questions (blue) and over the large ideal temperature subset (red), across all four models. n is the number of included questions.

Bar position and height. Each bar is centered on its bin center and spans the bin's width on the H axis. Height $= 1$ means *some* single temperature lies inside the half-max range of every in-bin question simultaneously; height $= 1/M_b$ means at most one in-bin question is covered by any given temperature.

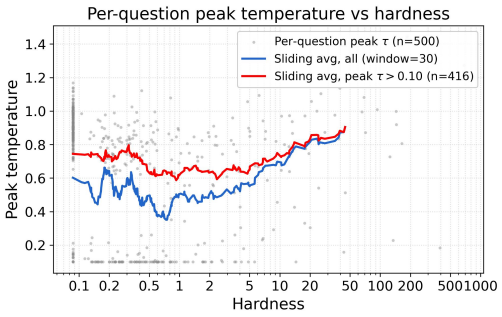
What to look for. Bars should stay near 1 at low H (broad, mutually-overlapping ranges) and fall toward 0 at high H (narrow ranges, no shared sweet spot), the same shrinking-range trend the previous two plots quantify question-by-question.

E.2 Extended Results

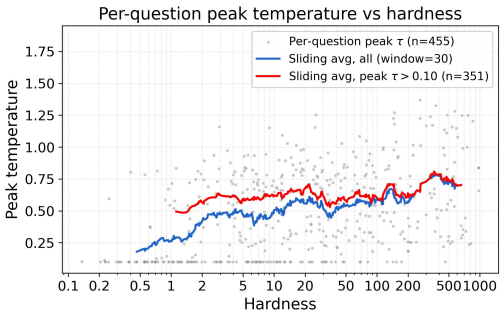
Here, we present all results from the analysis of the questions in Section 3, as explained in Appendix E.1.

On LiveCodeBench, across all four models, the peak temperature trends rise gently with hardness once near-zero peaks are excluded (Figure 5); the efficient temperature range shrinks roughly log-linearly with H with a clearly negative OLS slope for every model (Figure 8); and the maximum-coverage bars correspondingly fall from near 1 at low H toward 0 at high H (Figure 11), so harder questions become both narrower and harder to cover with a single shared temperature. The intermediate-H deviations in Figure 11 are consistent with the near-zero-$\tau_q^\star$ easy cluster discussed in Section 3.2.2.

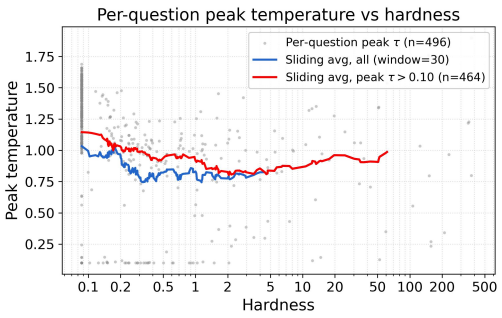
On MATH-500, the red sliding average of the peak temperature ends higher than it starts for GPT OSS 20B (0.74 to 0.90), Mistral 7B Instruct (0.49 to 0.70), and Qwen3-0.6B (0.61 to 0.85), and lower for Qwen3-8B (1.14 to 0.99) (Figure 6); the OLS slope of the efficient range size is negative for every model, from -0.393 ($R^2 = 0.768$) for Qwen3-8B to -0.048 ($R^2 = 0.023$) for Mistral 7B Instruct (Figure 9); and the maximum-coverage bars start at 1 and reach their lowest value in one of the two highest-H bins, 0.500, 0.667, 0.333, and 0.625 in the order of the panels (Figure 12).



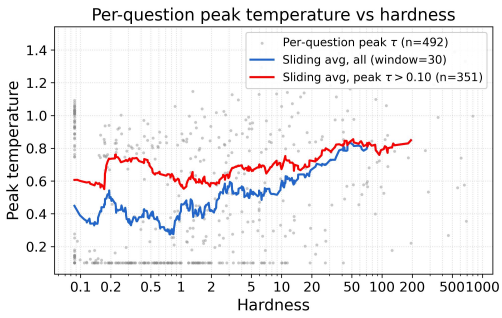
(a) GPT OSS 20B, $n = 500$



(b) Mistral 7B Instruct, $n = 455$

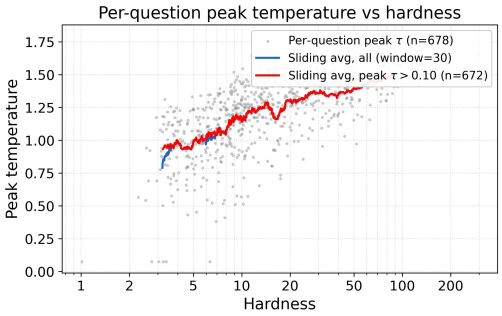


(c) Qwen3-8B, $n = 496$

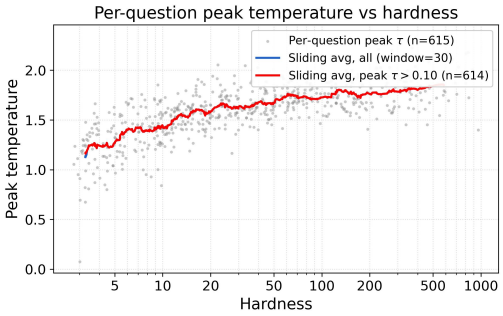


(d) Qwen3-0.6B, $n = 492$

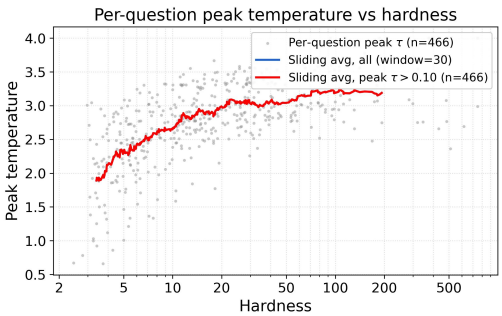
Figure 6: Per-question peak temperature vs. hardness on MATH-500, drawn as in Figure 5. n is the number of included questions.



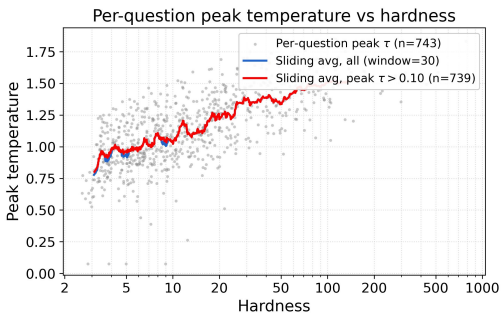
(a) GPT OSS 20B, $n = 678$



(b) Mistral 7B Instruct, $n = 615$

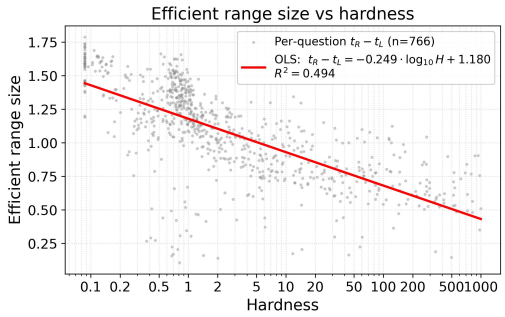


(c) Qwen3-8B, $n = 466$

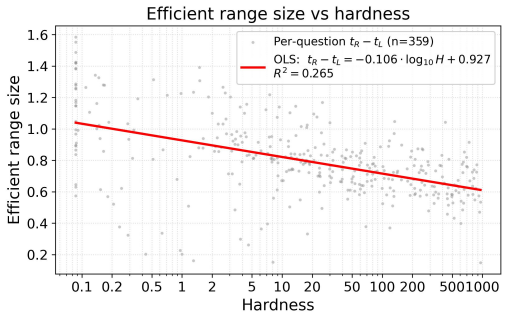


(d) Qwen3-0.6B, $n = 743$

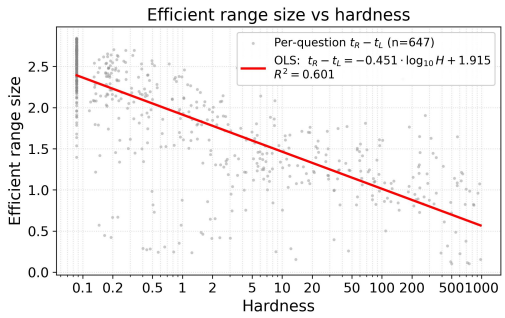
Figure 7: Per-question peak temperature vs. hardness on Hi-ToM, drawn as in Figure 5, with the sliding averages over every full window (Appendix E.1). n is the number of included questions.



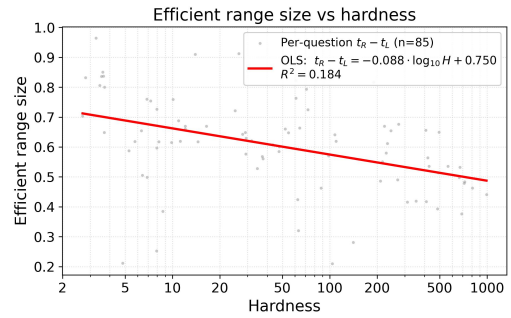
(a) GPT OSS 20B, $n = 766$



(b) Mistral 7B Instruct, $n = 359$

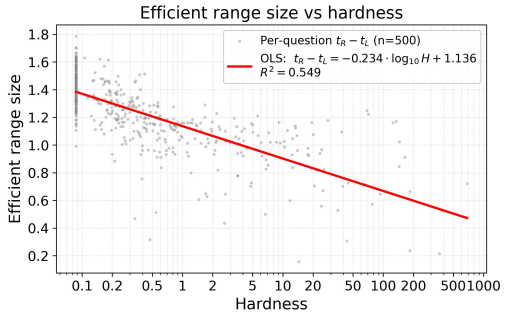


(c) Qwen3-8B, $n = 647$

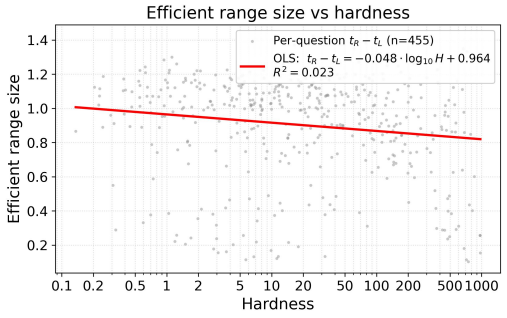


(d) Qwen3-0.6B, $n = 85$

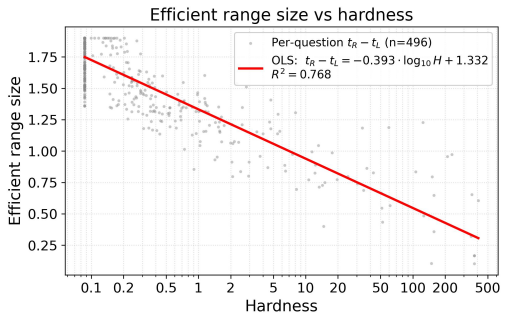
Figure 8: Per-question efficient range size vs. hardness on LiveCodeBench, with the OLS fit on $\log_{10} H$ (slope and R^2 in each legend), across all four models. n is the number of included questions.



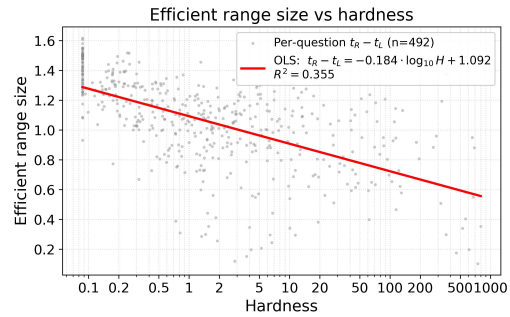
(a) GPT OSS 20B, $n = 500$



(b) Mistral 7B Instruct, $n = 455$

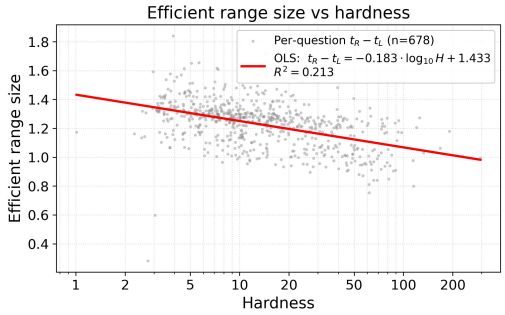


(c) Qwen3-8B, $n = 496$

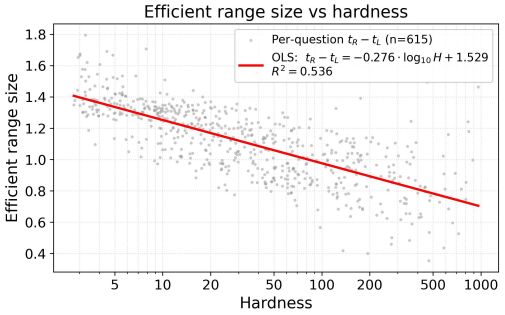


(d) Qwen3-0.6B, $n = 492$

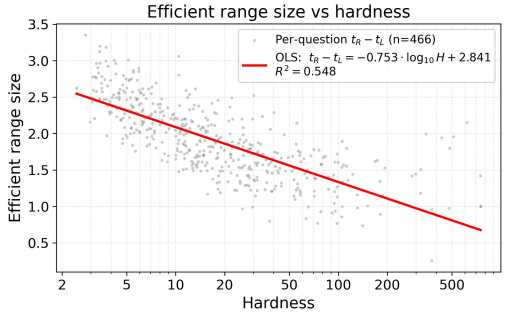
Figure 9: Per-question efficient range size vs. hardness on MATH-500, drawn as in Figure 8. n is the number of included questions.



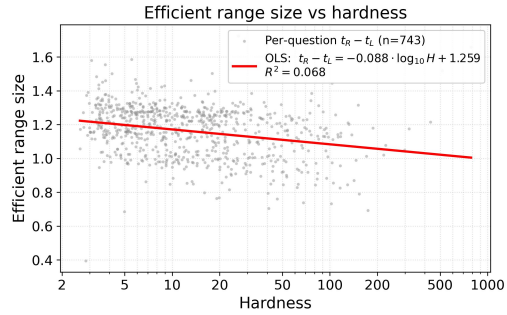
(a) GPT OSS 20B, $n = 678$



(b) Mistral 7B Instruct, $n = 615$

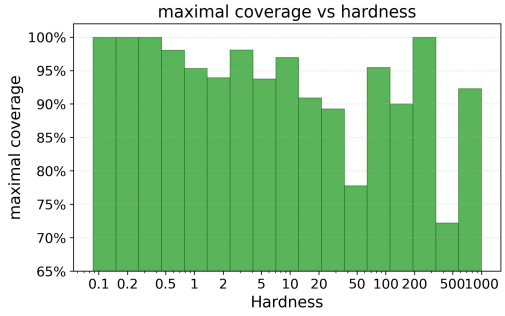


(c) Qwen3-8B, $n = 466$

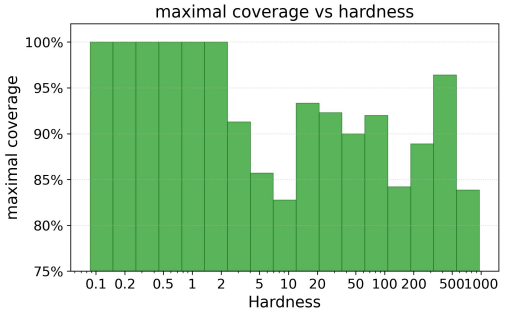


(d) Qwen3-0.6B, $n = 743$

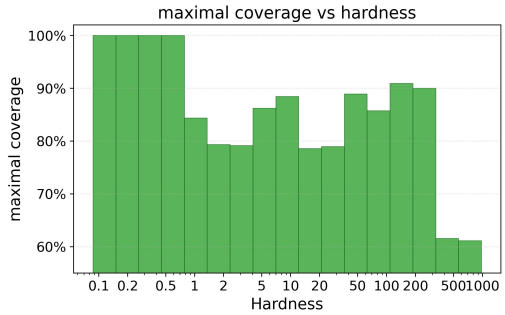
Figure 10: Per-question efficient range size vs. hardness on Hi-ToM, drawn as in Figure 8. n is the number of included questions.



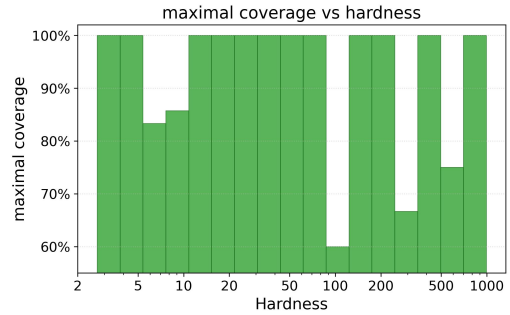
(a) GPT OSS 20B, $n = 766$



(b) Mistral 7B Instruct, $n = 359$

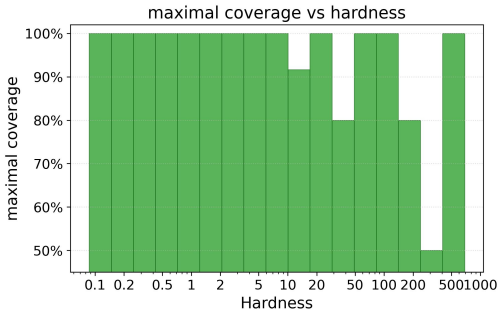


(c) Qwen3-8B, $n = 647$

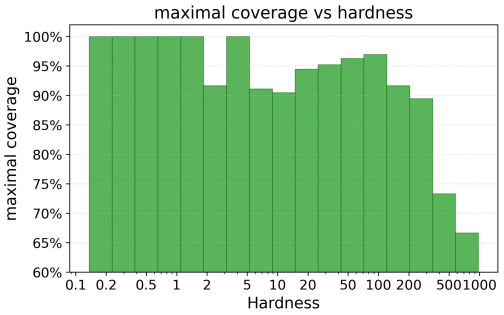


(d) Qwen3-0.6B, $n = 85$

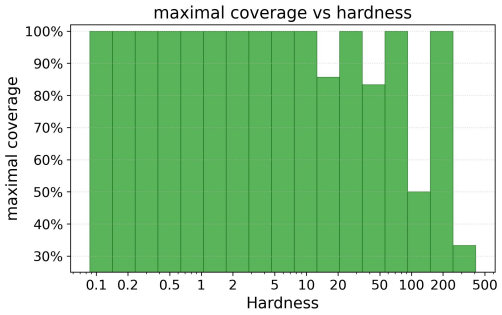
Figure 11: Empirical $R_{\mathcal{Q}}(H)$ from Eq. 5 on LiveCodeBench for all four models. Questions are placed in 17 log-uniform hardness bins. n is the number of included questions.



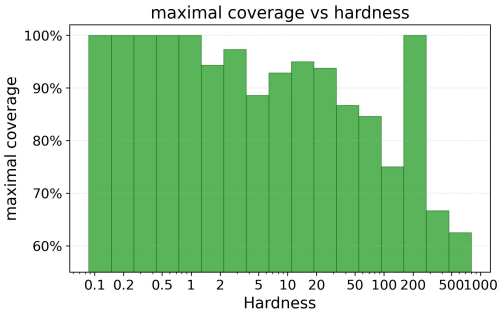
(a) GPT OSS 20B, $n = 500$



(b) Mistral 7B Instruct, $n = 455$

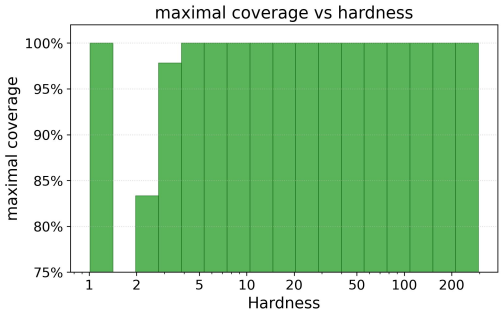


(c) Qwen3-8B, $n = 496$

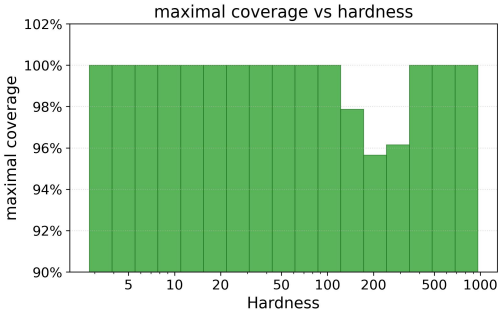


(d) Qwen3-0.6B, $n = 492$

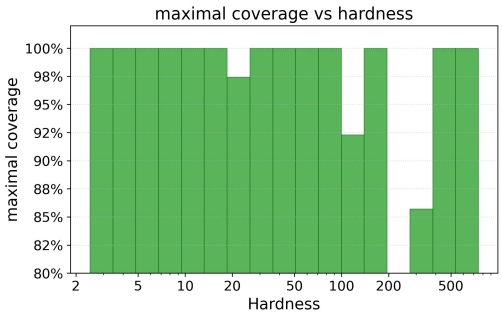
Figure 12: Empirical $R_{\mathcal{Q}}(H)$ on MATH-500, drawn as in Figure 11. n is the number of included questions.



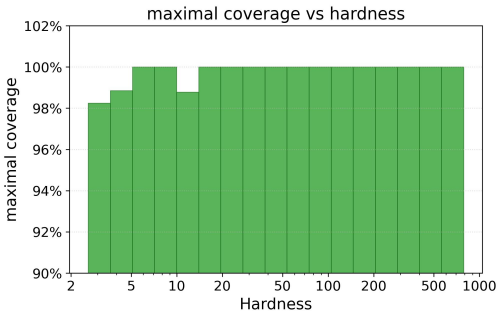
(a) GPT OSS 20B, $n = 678$



(b) Mistral 7B Instruct, $n = 615$



(c) Qwen3-8B, $n = 466$



(d) Qwen3-0.6B, $n = 743$

Figure 13: Empirical $R_{\mathcal{Q}}(H)$ on Hi-ToM, drawn as in Figure 11. n is the number of included questions. A bin with no question has no bar, so the GPT OSS 20B and Qwen3-8B panels have 16 bars.

On Hi-ToM, the red sliding average of the peak temperature rises with hardness for every model, from 0.93 to 1.50 for GPT OSS 20B, 1.17 to 1.85 for Mistral 7B Instruct, 1.89 to 3.19 for Qwen3-8B, and 0.80 to 1.52 for Qwen3-0.6B (Figure 7); the OLS slope of the efficient range size is negative for every model, from -0.753 ($R^2 = 0.548$) for Qwen3-8B to -0.088 ($R^2 = 0.068$) for Qwen3-0.6B (Figure 10); and every maximum-coverage bar lies between 0.83 and 1 (Figure 13).

F Temperature Sampling Schemes Comparisons

In this section, we explain how we compared the different temperature-sampling schemes. We begin by describing the analysis procedure and then present the results.

F.1 Statistical Analysis Approach

Per-temperature pass@K score

Once each profile $P_{q_i}(\tau)$ has been fitted, we evaluate sampling strategies via the pass@K metric: the expected fraction of questions solved when each question receives K independent attempts under a given allocation across temperatures. For a single temperature τ_j used for all K attempts, the probability that question q_i is solved at least once is $1 - \bar{P}_{q_i}(\tau_j)^K$, where $\bar{P}_{q_i}(\tau) = 1 - P_{q_i}(\tau)$ is the single-attempt failure probability. The dataset-level score at temperature τ_j is

$$\mathcal{P}_K(\tau_j) \;=\; 1 - \bar{\mathcal{P}}_K(\tau_j) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[1 - \bar{P}_{q_i}(\tau_j)^K\right] \;, \tag{48}$$

Bootstrap error bars

We use bootstrap because the scores reported here are averages over a sampled question set, and we want a confidence interval that reflects how those averages would change if a different set of questions had been drawn. Resampling questions with replacement gives a consistent estimate of that variability without any parametric assumption on the per-question score distribution.

We attach a confidence band to each per-temperature score using the **paired nonparametric bootstrap**. We draw R_{boot} resamples of the question set with replacement; in each resample b, we recompute $\mathcal{P}_K(\tau_j)$ for every τ_j from the same resampled question set, and the standard error is the empirical standard deviation of $\mathcal{P}_K^{(b)}(\tau_j)$ across replicates. Pairing across temperatures, that is, using one common resample for all τ_j, removes question-level variation that would otherwise inflate every interval; the resulting ratios between temperatures are especially stable. For this reason, we display bootstrap bands only on the ratio-to-uniform plots, where the band is attached to the superiority proposition itself. An absolute per-temperature band answers a different question: how uncertain that one temperature's pass rate is in isolation. Without pairing, such bands would be wide and largely overlap, hiding genuine strategy differences. The comparison of interest is whether one allocation strategy outperforms another at the same K. Paired resampling keeps each question's difficulty common to both sides of the ratio, so the interval reflects uncertainty in the relative advantage rather than in the absolute score.

Computationally, we exploit a count-then-matmul identity that reduces the bootstrap inner loop to a single matrix multiplication: for any per-question quantity f_i, the bootstrap sum over the n slots of resample b equals $\sum_{i=1}^{n} f_i \cdot c_{b,i}$, where $c_{b,i}$ is the number of times question q_i was drawn in resample b. This collapses each bootstrap pass into a single BLAS-optimized matrix multiplication, orders of magnitude faster than direct loop-based indexing.

Optimal temperature allocation

For a fixed total budget of K attempts, we ask how those attempts should be distributed across the l candidate temperatures $\tau_1, \ldots, \tau_l$ to maximize the dataset-level pass@K score according to Equation 10.

Convexity and uniqueness. Each summand $\exp\left(\sum_j K_j \ln \bar{P}_{ij}\right)$ is the composition of a convex nondecreasing outer function with an affine map of the allocation, hence convex. The objective is a finite sum of convex terms over a polyhedron, so the relaxation of Equation 10 is convex. The

Hessian has entries

$$\frac{\partial^2\big(n\,\bar{\mathcal{P}}\big)}{\partial K_j\,\partial K_{j'}} \;=\; \sum_{i=1}^{n} \ln \bar{P}_{ij} \cdot \ln \bar{P}_{ij'} \cdot \exp\!\left(\sum_{k=1}^{l} K_k \ln \bar{P}_{ik}\right) , \tag{49}$$

which is positive definite whenever the matrix $(\ln \bar{P}_{ij})_{i,j}$ has full column rank. On our data this matrix has rank l (smallest singular value $\approx 3.3 \times 10^{-2}$), so the objective is strictly convex and the minimizer is unique. Slater's condition holds at the uniform sampling scheme $K_j = K/l$, which makes the KKT conditions necessary and sufficient for global optimality.

Numerical solver. We solve this continuous relaxation with SLSQP (sequential least-squares programming), supplying the analytic gradient

$$\frac{\partial\big(n\,\bar{\mathcal{P}}\big)}{\partial K_j} \;=\; \sum_{i=1}^{n} \ln \bar{P}_{ij} \cdot \exp\!\left(\sum_{k=1}^{l} K_k \ln \bar{P}_{ik}\right) . \tag{50}$$

Iterates are warm-started across K by carrying the previous continuous optimum forward. Because each K_j is in practice a non-negative integer, we follow the SLSQP step with a combinatorial integer search inside a radius-2 neighborhood of the rounded continuous optimum.

Certificate of global optimality. At every reported K, the continuous optimum is verified by three independent diagnostics:

- **KKT residuals.** Equality residual $\left|\sum_j K_j - K\right|$, bound violations, stationarity gap on the active set, and dual slack on the inactive set, evaluated against the gradient above.
- **First-order gap bound.** The a posteriori convexity bound on the objective gap obtained by linearizing $\bar{\mathcal{P}}$ at the candidate allocation and minimizing the linearization over the simplex.
- **Independent solver cross-check.** The same problem is re-solved with the interior-point conic solver CLARABEL (via CVXPY), and the gap between the two reported objective values is recorded.

On the dataset used for the headline results, all three quantities remain below 10^{-7} across the entire K grid, certifying that the reported allocation achieves the unique global optimum to floating-point tolerance.

How to read the ratio-to-uniform plot. Every y value is a pass@K score divided by the same dataset's uniform-scheme pass@K at the same K. The uniform scheme spreads K equally over the temperatures of the model's grid on that dataset (Appendix C). The dashed grey line at $y = 1$ is the uniform baseline: values above 1 outperform uniform at that K and values below 1 underperform. The x-axis is the total budget K on a log scale.

Colored curves. One curve per single-temperature allocation: all K attempts placed at one fixed τ. To keep the plot readable, only temperatures that are the dominant choice in the optimal allocation at some K are drawn; the full per-temperature pass-rates for every τ in the dataset's grid are tabulated in Section F.2 for LiveCodeBench. The shaded band around each curve is the paired-bootstrap $\pm 1\sigma$ interval on the ratio (see the *Bootstrap error bars* paragraph above); pairing keeps each question's difficulty common to numerator and denominator, so the band measures uncertainty in the relative advantage rather than in the absolute score. The ratios of Figures 15 and 16 come from the paper's rendering code. Figure 16 is drawn with dashed reference curves, colors by temperature, exact legend values, and tick labels over the full range.

Red trace. The leave-one-out out-of-sample optimal allocation. At each K the convex relaxation of Equation 10 is fit on the held-in questions and scored on the held-out one; the score is then divided by the same in-sample uniform-scheme total used for the colored curves. Figures 14, 15, and 16 draw the red trace without a band.

What to look for. Compare the red trace against the highest colored curve at each K. Where they coincide, no allocation beats that single temperature; where the red trace sits clearly above, mixing temperatures pays off out-of-sample.

How to read the summary tables. The summary table of Appendix F.2 gives, for each benchmark, two quantities per model at $K \in \{1, 16, 256, 1024\}$, both from the same ratio-to-uniform computation as the figures. These are the Optimal (LOO) / uniform ratio of the red trace, and the best single temperature, followed in parentheses by its ratio to uniform. The best single temperature at K is the temperature, among all temperatures of the model's grid, whose single-τ allocation has the highest ratio to uniform at that K on the model's full included question set. It is therefore selected and scored in-sample. The LOO optimum is scored on held-out questions. This temperature is not necessarily drawn in the figures, whose colored curves (the solid curves in Figure 16) show only the dominant temperatures of the optimal allocation.

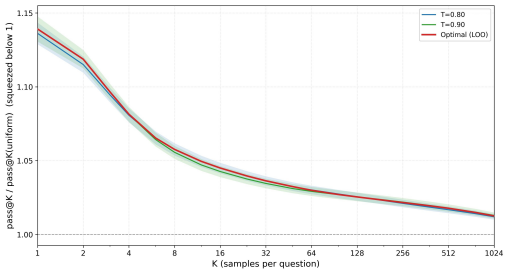
How to read the pass@K tables. Tables 2 to 6 report, for LiveCodeBench, per-K pass-rates for each allocation strategy: the per-temperature single-τ allocations, and three row-aggregate strategies, namely the best single τ, the uniform sampling scheme, and the optimal allocation. **Bold** marks the row-best aggregate strategy among the three; per-temperature columns are read column-wise to locate each K's peak temperature and are never bolded. The Best τ and Optimal columns each show two numbers per row. *In-sample* (top): the strategy (best single τ or convex allocation, Eq. 10) chosen on all included questions and scored on the same set, optimistic since the strategy saw its own test data. *Out-of-sample* (bottom, smaller): the leave-one-out mean, for each question we re-fit the strategy on the remaining $n_{\text{inc}} - 1$ questions and score the held-out one, then average. Close numbers indicate a robust choice; a large gap indicates the in-sample value is fitting noise. Each per-model table places the same K rows in two stacked sub-tables to fit a portrait page: low temperatures above, high temperatures together with the three row-aggregate columns below.

F.2 Extended Results

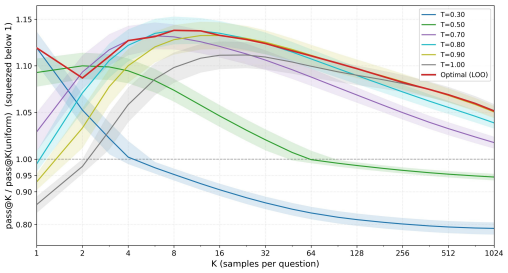
We show the ratio-to-uniform curves of the four models in Figure 14 for LiveCodeBench, in Figure 15 for MATH-500, and in Figure 16 for Hi-ToM. The table below gives, for each benchmark and model, the two quantities defined in Appendix F.1 at $K \in \{1, 16, 256, 1024\}$, and Tables 2 to 6 give the complete pass@K values on LiveCodeBench.

On all three benchmarks, the red trace tracks the colored upper envelope, which shows the limited gain of multi-temperature allocation over a K-tuned single temperature, the central observation of Section 4.2. We evaluate all curves from the fitted profiles up to $K_{\max} = 1024$ (Appendix D), which exceeds the per-question, per-temperature sample count only for Qwen3-8B on MATH-500, where $\nu = 750$, so its points at $K = 768$ and $K = 1024$ are fit-based.

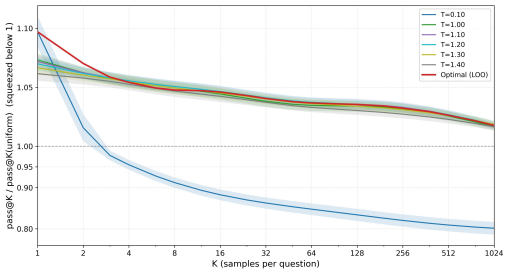
	Model		$K = 1$	$K = 16$	$K = 256$	$K = 1024$
LiveCodeBench	GPT OSS 20B	Optimal (LOO) / uniform	1.139	1.045	1.022	1.013
		best single τ (in-sample)	0.9 (1.139)	0.8 (1.045)	0.9 (1.022)	0.9 (1.013)
	Mistral 7B Instruct	Optimal (LOO) / uniform	1.119	1.133	1.082	1.052
		best single τ (in-sample)	0.3 (1.119)	0.8 (1.135)	0.9 (1.083)	1.0 (1.052)
	Qwen3-8B	Optimal (LOO) / uniform	1.097	1.046	1.032	1.018
		best single τ (in-sample)	0.1 (1.097)	1.2 (1.046)	1.1 (1.033)	1.3 (1.019)
	Qwen3-0.6B	Optimal (LOO) / uniform	2.099	1.500	1.137	1.082
		best single τ (in-sample)	0.7 (2.174)	0.7 (1.516)	0.7 (1.147)	0.7 (1.089)
MATH-500	GPT OSS 20B	Optimal (LOO) / uniform	1.058	1.015	1.003	1.002
		best single τ (in-sample)	0.2 (1.058)	0.8 (1.015)	1.0 (1.004)	1.2 (1.003)
	Mistral 7B Instruct	Optimal (LOO) / uniform	1.341	1.111	1.041	1.018
		best single τ (in-sample)	0.2 (1.341)	0.5 (1.113)	0.7 (1.041)	0.7 (1.018)
	Qwen3-8B	Optimal (LOO) / uniform	1.045	1.012	1.000	0.998
		best single τ (in-sample)	0.9 (1.045)	1.0 (1.012)	1.0 (1.002)	1.0 (0.998)
	Qwen3-0.6B	Optimal (LOO) / uniform	1.233	1.049	1.011	1.003
		best single τ (in-sample)	0.2 (1.233)	0.6 (1.051)	0.9 (1.011)	1.0 (1.003)
Hi-ToM	GPT OSS 20B	Optimal (LOO) / uniform	1.679	1.320	1.025	1.001
		best single τ (in-sample)	1.125 (1.679)	1.275 (1.320)	1.5 (1.025)	1.575 (1.001)
	Mistral 7B Instruct	Optimal (LOO) / uniform	1.727	1.447	1.153	1.061
		best single τ (in-sample)	1.425 (1.758)	1.575 (1.447)	1.8 (1.153)	1.875 (1.062)
	Qwen3-8B	Optimal (LOO) / uniform	2.084	1.740	1.217	1.075
		best single τ (in-sample)	2.5 (2.084)	3.0 (1.740)	3.0 (1.217)	3.0 (1.075)
	Qwen3-0.6B	Optimal (LOO) / uniform	1.768	1.315	1.033	1.004
		best single τ (in-sample)	1.05 (1.768)	1.2 (1.315)	1.5 (1.033)	1.575 (1.004)



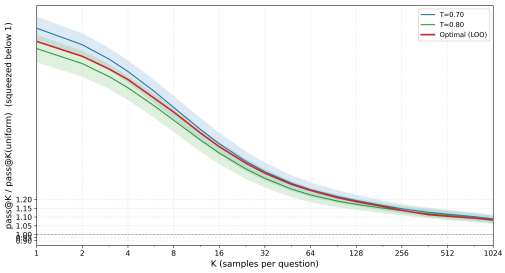
(a) GPT OSS 20B, $n = 766$



(b) Mistral 7B Instruct, $n = 359$

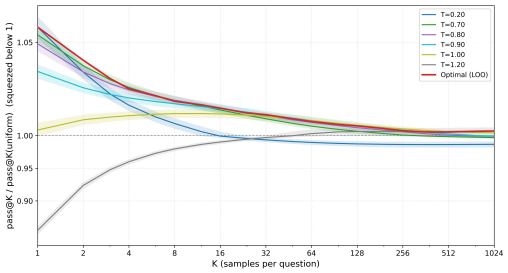


(c) Qwen3-8B, $n = 647$

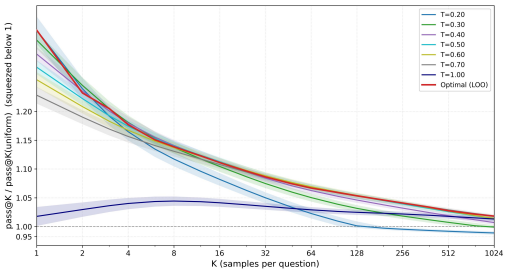


(d) Qwen3-0.6B, $n = 85$

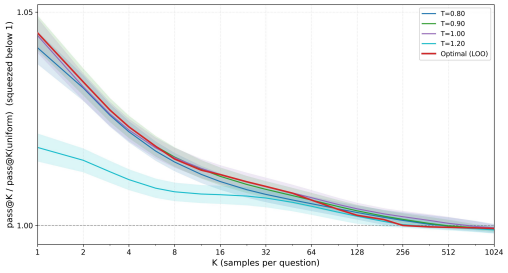
Figure 14: Ratio to uniform pass@K, $\mathrm{rtu}(K) = \mathrm{pass@}K(\mathrm{strategy})/\mathrm{pass@}K(\mathrm{uniform})$, on LiveCodeBench for all four models, read as described in Appendix F.1. The red trace is the Section 4.1 optimum, evaluated leave-one-out (LOO) at each K of the colored-curve grid. n is the number of included questions.



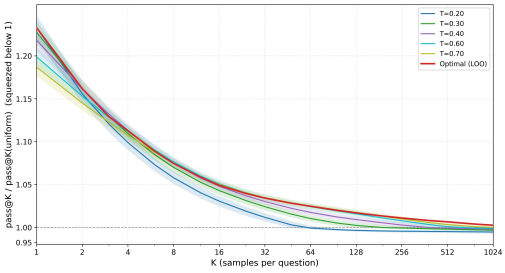
(a) GPT OSS 20B, $n = 500$



(b) Mistral 7B Instruct, $n = 455$



(c) Qwen3-8B, $n = 496$



(d) Qwen3-0.6B, $n = 492$

Figure 15: Ratio to uniform pass@K on MATH-500 for all four models, defined as in Figure 14; n is the number of included questions. For Qwen3-8B, the uniform scheme spreads K over 20 temperatures here and over 38 in Figure 14c, so the absolute ratios of the two panels are not comparable.

Table 2: Pass@K for GPT OSS 20B. Low temperatures $\tau \in [0.1, 0.8]$ above, high temperatures $\tau \in [0.9, 2.0]$ together with the row-aggregate columns below.

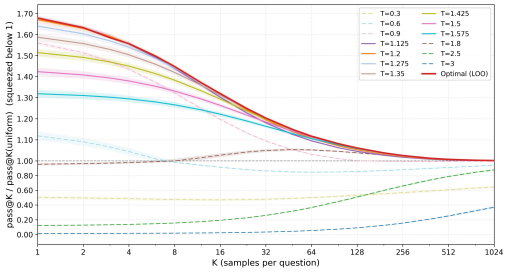
K	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.344	0.361	0.372	0.381	0.389	0.396	0.402	0.407
2	0.424	0.466	0.491	0.508	0.521	0.532	0.542	0.549
3	0.467	0.522	0.555	0.575	0.590	0.602	0.612	0.619
4	0.495	0.560	0.595	0.617	0.633	0.645	0.654	0.660
6	0.531	0.607	0.646	0.669	0.685	0.696	0.704	0.709
8	0.556	0.638	0.678	0.701	0.716	0.727	0.734	0.738
12	0.589	0.678	0.718	0.740	0.754	0.764	0.771	0.774
16	0.611	0.704	0.743	0.765	0.778	0.788	0.794	0.796
24	0.641	0.737	0.775	0.795	0.807	0.816	0.822	0.824
32	0.662	0.759	0.795	0.814	0.826	0.835	0.840	0.843
48	0.691	0.787	0.820	0.838	0.850	0.858	0.864	0.866
64	0.709	0.804	0.836	0.854	0.866	0.874	0.879	0.882
96	0.734	0.826	0.856	0.874	0.886	0.894	0.899	0.902
128	0.750	0.840	0.869	0.886	0.899	0.907	0.913	0.916
192	0.771	0.857	0.885	0.903	0.915	0.924	0.930	0.933
256	0.784	0.868	0.896	0.914	0.926	0.935	0.941	0.945
384	0.802	0.882	0.910	0.928	0.940	0.949	0.955	0.959
512	0.814	0.891	0.919	0.937	0.949	0.958	0.964	0.968

K	**0.9**	**1.0**	**1.1**	**1.2**	**1.4**	**1.5**	**1.7**	**2.0**	**Best τ**	**Uniform**	**Optimal (in / out)**
1	0.408	0.402	0.387	0.360	0.270	0.214	0.101	0.007	0.408 0.408	0.358	**0.408** **0.408**
2	0.551	0.545	0.528	0.498	0.394	0.323	0.167	0.013	0.551 0.551	0.492	**0.551** **0.551**
3	0.620	0.613	0.597	0.568	0.464	0.390	0.215	0.019	0.620 0.620	0.565	**0.620** **0.620**
4	0.660	0.654	0.638	0.610	0.509	0.434	0.251	0.024	0.660 0.660	0.611	**0.660** **0.660**
6	0.708	0.701	0.686	0.660	0.565	0.491	0.303	0.033	0.709 **0.709**	0.665	**0.709** 0.709
8	0.736	0.729	0.714	0.690	0.599	0.527	0.339	0.040	0.738 0.738	0.698	**0.738** **0.738**
12	0.772	0.765	0.750	0.727	0.641	0.571	0.388	0.052	0.774 0.774	0.737	**0.774** **0.774**
16	0.794	0.787	0.774	0.751	0.667	0.599	0.420	0.061	0.796 0.796	0.762	**0.796** **0.796**
24	0.823	0.816	0.803	0.782	0.700	0.634	0.460	0.075	0.824 0.824	0.793	**0.824** **0.824**
32	0.841	0.835	0.822	0.802	0.722	0.656	0.485	0.084	0.843 0.843	0.813	**0.843** **0.843**
48	0.865	0.859	0.848	0.828	0.749	0.685	0.515	0.097	0.866 0.866	0.839	**0.866** **0.866**
64	0.881	0.876	0.864	0.845	0.767	0.703	0.533	0.107	0.882 **0.882**	0.856	**0.882** 0.881
96	0.902	0.897	0.887	0.868	0.791	0.727	0.554	0.119	0.902 **0.902**	0.878	**0.902** 0.902
128	0.916	0.911	0.901	0.884	0.807	0.742	0.566	0.128	0.916 0.913	0.893	**0.916** **0.916**
192	0.934	0.930	0.921	0.904	0.827	0.762	0.580	0.141	0.934 **0.934**	0.912	**0.934** 0.933
256	0.945	0.942	0.934	0.918	0.841	0.776	0.588	0.149	0.945 **0.945**	0.925	**0.945** 0.945
384	0.960	0.957	0.949	0.935	0.858	0.793	0.597	0.161	0.960 **0.960**	0.941	**0.960** 0.960
512	0.969	0.967	0.959	0.946	0.870	0.805	0.602	0.169	0.969 **0.969**	0.952	**0.969** 0.969

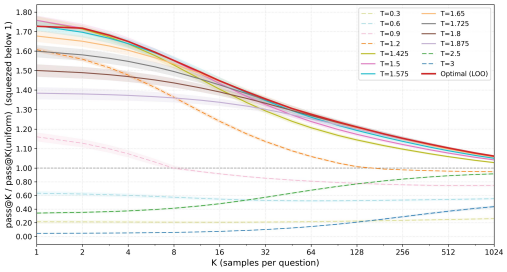
Table 3: Pass@K for Mistral 7B Instruct. Low temperatures $\tau \in [0.1,\ 0.8]$ above, high temperatures $\tau \in [0.9,\ 2.0]$ together with the row-aggregate columns below.

K	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**
1	0.178	0.184	0.186	0.185	0.181	0.177	0.171	0.164
2	0.198	0.216	0.227	0.234	0.238	0.238	0.236	0.231
3	0.210	0.236	0.254	0.265	0.273	0.277	0.277	0.274
4	0.219	0.252	0.274	0.289	0.299	0.305	0.308	0.306
6	0.233	0.275	0.303	0.323	0.338	0.347	0.352	0.353
8	0.243	0.292	0.325	0.349	0.366	0.378	0.386	0.388
12	0.256	0.316	0.357	0.386	0.408	0.424	0.434	0.439
16	0.266	0.334	0.380	0.413	0.438	0.457	0.470	0.475
24	0.278	0.360	0.413	0.452	0.482	0.505	0.521	0.529
32	0.287	0.378	0.436	0.480	0.514	0.540	0.557	0.566
48	0.298	0.405	0.470	0.520	0.560	0.589	0.609	0.619
64	0.307	0.424	0.495	0.549	0.591	0.623	0.644	0.656
96	0.318	0.451	0.529	0.588	0.635	0.669	0.692	0.706
128	0.327	0.470	0.553	0.616	0.665	0.701	0.725	0.739
192	0.339	0.497	0.586	0.653	0.705	0.743	0.768	0.784
256	0.349	0.516	0.608	0.679	0.733	0.771	0.797	0.814
384	0.363	0.542	0.639	0.714	0.770	0.808	0.835	0.853
512	0.373	0.561	0.660	0.738	0.795	0.833	0.860	0.879

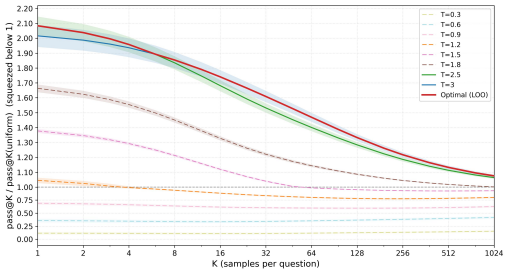
K	**0.9**	**1.0**	**1.1**	**1.2**	**1.4**	**1.5**	**1.7**	**2.0**	**Best τ**	**Uniform**	**Optimal (in / out)**
1	0.154	0.143	0.129	0.112	0.072	0.052	0.018	0.000	0.186 0.186	0.166	**0.186** **0.186**
2	0.223	0.211	0.195	0.173	0.119	0.088	0.034	0.001	0.238 **0.238**	0.216	**0.238** 0.235
3	0.267	0.256	0.238	0.215	0.153	0.116	0.047	0.001	0.277 **0.277**	0.248	**0.277** 0.275
4	0.300	0.289	0.271	0.247	0.180	0.139	0.058	0.002	0.308 0.308	0.273	**0.308** **0.308**
6	0.349	0.338	0.320	0.295	0.221	0.174	0.077	0.003	0.353 **0.353**	0.311	**0.353** 0.352
8	0.385	0.374	0.356	0.330	0.252	0.202	0.093	0.003	0.388 0.388	0.341	**0.388** **0.388**
12	0.437	0.427	0.409	0.382	0.297	0.242	0.118	0.005	0.439 0.439	0.386	**0.439** **0.439**
16	0.474	0.466	0.448	0.420	0.331	0.273	0.138	0.006	0.475 **0.475**	0.419	**0.475** 0.475
24	0.529	0.520	0.502	0.474	0.380	0.317	0.169	0.009	0.529 0.519	0.468	**0.529** **0.528**
32	0.567	0.559	0.541	0.513	0.416	0.350	0.192	0.011	0.567 **0.567**	0.504	**0.567** 0.566
48	0.621	0.613	0.596	0.567	0.467	0.397	0.227	0.014	0.621 **0.621**	0.556	**0.621** 0.620
64	0.658	0.651	0.634	0.605	0.503	0.430	0.253	0.017	0.658 **0.658**	0.592	**0.658** 0.658
96	0.709	0.703	0.687	0.658	0.552	0.477	0.290	0.021	0.709 0.709	0.643	**0.709** **0.709**
128	0.744	0.739	0.724	0.695	0.587	0.509	0.316	0.025	0.744 0.744	0.678	**0.744** **0.744**
192	0.791	0.788	0.774	0.746	0.636	0.554	0.354	0.030	0.791 **0.791**	0.726	**0.791** 0.790
256	0.822	0.821	0.808	0.781	0.671	0.585	0.381	0.033	0.822 **0.822**	0.759	**0.822** 0.821
384	0.863	0.863	0.852	0.828	0.719	0.629	0.419	0.039	0.863 0.854	0.803	**0.864** **0.863**
512	0.890	0.891	0.881	0.858	0.752	0.659	0.445	0.044	0.891 **0.891**	0.833	**0.891** 0.890



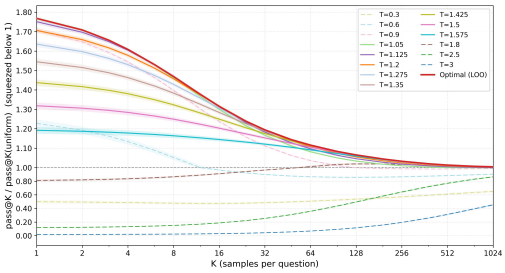
(a) GPT OSS 20B, $n = 678$



(b) Mistral 7B Instruct, $n = 615$



(c) Qwen3-8B, $n = 466$



(d) Qwen3-0.6B, $n = 743$

Figure 16: Ratio to uniform pass@K on Hi-ToM for all four models, defined as in Figure 14; n is the number of included questions. Solid curves show the temperatures that are the dominant choice in the optimal allocation at some K, namely every grid temperature from 1.125 to 1.575 for GPT OSS 20B, from 1.425 to 1.875 for Mistral 7B Instruct, 2.5 and 3.0 for Qwen3-8B, and from 1.05 to 1.575 for Qwen3-0.6B. Dashed curves show the other temperatures of a reference set drawn in every panel, $\tau \in \{0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.5, 3.0\}$. A temperature has the same color in every panel.

Table 4: Pass@K for Qwen3-8B, low-temperature half ($\tau \in [0.1, 1.9]$). $\tau \in [0.1, 1.0]$ above, $\tau \in [1.1, 1.9]$ below. Continued in Table 5; the row-aggregate columns are computed across the full 38-temperature sweep and appear there.

K	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.585	0.585	0.583	0.582	0.580	0.578	0.576	0.575	0.574	0.573
2	0.615	0.626	0.632	0.637	0.639	0.641	0.642	0.643	0.643	0.643
3	0.628	0.645	0.656	0.663	0.668	0.672	0.675	0.676	0.678	0.679
4	0.636	0.658	0.671	0.681	0.688	0.693	0.696	0.699	0.701	0.702
6	0.647	0.675	0.693	0.705	0.713	0.720	0.725	0.729	0.731	0.733
8	0.655	0.688	0.707	0.720	0.730	0.738	0.744	0.748	0.751	0.753
12	0.666	0.705	0.726	0.741	0.752	0.761	0.768	0.774	0.778	0.780
16	0.674	0.716	0.738	0.754	0.766	0.776	0.784	0.790	0.795	0.798
24	0.686	0.732	0.755	0.772	0.785	0.796	0.805	0.812	0.817	0.821
32	0.694	0.742	0.767	0.784	0.798	0.809	0.819	0.827	0.832	0.836
48	0.706	0.756	0.783	0.801	0.816	0.829	0.839	0.847	0.853	0.856
64	0.713	0.766	0.793	0.812	0.829	0.842	0.853	0.861	0.867	0.870
96	0.722	0.779	0.808	0.829	0.847	0.861	0.872	0.881	0.886	0.890
128	0.728	0.788	0.818	0.840	0.860	0.875	0.886	0.894	0.900	0.903
192	0.736	0.801	0.833	0.856	0.877	0.893	0.904	0.912	0.918	0.921
256	0.742	0.810	0.843	0.868	0.889	0.904	0.916	0.924	0.929	0.933
384	0.749	0.821	0.856	0.882	0.905	0.920	0.931	0.939	0.944	0.947
512	0.754	0.830	0.865	0.892	0.915	0.930	0.940	0.948	0.953	0.955

K	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**
1	0.572	0.571	0.569	0.567	0.563	0.557	0.549	0.538	0.522
2	0.643	0.643	0.642	0.641	0.638	0.635	0.630	0.623	0.613
3	0.679	0.679	0.679	0.677	0.675	0.672	0.667	0.661	0.654
4	0.703	0.703	0.702	0.701	0.699	0.695	0.691	0.685	0.678
6	0.734	0.733	0.733	0.731	0.729	0.725	0.721	0.715	0.708
8	0.754	0.754	0.753	0.751	0.749	0.745	0.740	0.734	0.727
12	0.781	0.781	0.780	0.778	0.775	0.771	0.766	0.760	0.753
16	0.799	0.799	0.798	0.796	0.793	0.789	0.784	0.777	0.770
24	0.823	0.823	0.822	0.820	0.816	0.812	0.807	0.801	0.793
32	0.838	0.838	0.837	0.835	0.832	0.828	0.822	0.816	0.808
48	0.858	0.859	0.858	0.855	0.852	0.848	0.842	0.836	0.828
64	0.872	0.873	0.872	0.869	0.865	0.861	0.855	0.849	0.842
96	0.892	0.892	0.890	0.887	0.883	0.879	0.873	0.867	0.859
128	0.905	0.905	0.903	0.900	0.896	0.891	0.884	0.878	0.871
192	0.922	0.922	0.921	0.917	0.912	0.907	0.901	0.894	0.887
256	0.934	0.933	0.932	0.928	0.924	0.919	0.912	0.905	0.897
384	0.948	0.948	0.946	0.943	0.939	0.934	0.927	0.920	0.912
512	0.956	0.956	0.955	0.953	0.948	0.944	0.937	0.931	0.922

Table 5: Pass@K for Qwen3-8B, high-temperature half ($\tau \in [2.0,\ 4.2]$); continuation of Table 4. $\tau \in [2.0,\ 2.9]$ above, $\tau \in [3.0,\ 4.2]$ together with the row-aggregate columns below. Aggregates are computed across the full 38-temperature sweep.

K	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9
1	0.501	0.475	0.444	0.409	0.369	0.326	0.281	0.236	0.191	0.150
2	0.601	0.584	0.563	0.536	0.502	0.462	0.415	0.363	0.307	0.250
3	0.643	0.630	0.613	0.592	0.564	0.530	0.488	0.438	0.381	0.319
4	0.669	0.657	0.642	0.624	0.600	0.570	0.532	0.486	0.431	0.369
6	0.699	0.689	0.676	0.660	0.640	0.615	0.583	0.543	0.493	0.434
8	0.719	0.709	0.696	0.682	0.663	0.640	0.612	0.575	0.530	0.475
12	0.744	0.734	0.722	0.708	0.691	0.671	0.645	0.613	0.572	0.523
16	0.761	0.751	0.739	0.725	0.709	0.690	0.665	0.635	0.597	0.550
24	0.784	0.773	0.762	0.748	0.732	0.713	0.691	0.662	0.626	0.583
32	0.799	0.789	0.777	0.763	0.747	0.729	0.707	0.679	0.645	0.603
48	0.819	0.809	0.797	0.782	0.766	0.749	0.728	0.701	0.668	0.628
64	0.833	0.822	0.810	0.796	0.779	0.762	0.741	0.716	0.683	0.644
96	0.850	0.840	0.828	0.814	0.797	0.779	0.759	0.734	0.703	0.665
128	0.862	0.853	0.841	0.826	0.809	0.790	0.771	0.747	0.715	0.678
192	0.878	0.869	0.857	0.843	0.826	0.806	0.786	0.763	0.731	0.695
256	0.889	0.880	0.868	0.855	0.837	0.817	0.797	0.775	0.742	0.706
384	0.903	0.894	0.883	0.870	0.852	0.832	0.811	0.790	0.758	0.721
512	0.913	0.904	0.893	0.880	0.862	0.842	0.821	0.800	0.768	0.731

K	3.0	3.1	3.2	3.3	3.4	3.5	3.7	4.0	4.2	Best τ	Uniform	Optimal (in / out)
1	0.113	0.081	0.055	0.035	0.021	0.012	0.003	0.000	0.000	0.585 0.585	0.534	**0.585** **0.585**
2	0.195	0.144	0.101	0.066	0.041	0.023	0.005	0.000	0.000	0.643 0.642	0.606	**0.648** **0.648**
3	0.256	0.194	0.139	0.093	0.058	0.033	0.008	0.000	0.000	0.679 0.677	0.642	**0.681** **0.680**
4	0.302	0.235	0.172	0.117	0.074	0.043	0.010	0.000	0.000	0.703 0.701	0.666	**0.704** **0.702**
6	0.367	0.295	0.223	0.157	0.102	0.060	0.015	0.001	0.000	0.734 0.730	0.697	**0.734** **0.731**
8	0.410	0.337	0.261	0.188	0.126	0.076	0.020	0.001	0.000	0.754 0.752	0.718	**0.754** **0.752**
12	0.463	0.392	0.314	0.236	0.163	0.102	0.029	0.001	0.000	0.781 0.779	0.745	**0.781** **0.780**
16	0.494	0.427	0.350	0.269	0.191	0.124	0.037	0.002	0.000	0.799 **0.799**	0.764	**0.799** 0.799
24	0.531	0.469	0.394	0.313	0.232	0.156	0.050	0.003	0.001	0.823 **0.823**	0.789	**0.823** 0.822
32	0.553	0.494	0.422	0.341	0.259	0.180	0.062	0.004	0.001	0.838 **0.838**	0.805	**0.838** 0.838
48	0.580	0.523	0.456	0.377	0.294	0.213	0.081	0.005	0.001	0.859 **0.859**	0.827	**0.859** 0.858
64	0.596	0.542	0.477	0.399	0.317	0.235	0.096	0.007	0.002	0.873 **0.873**	0.841	**0.873** 0.872
96	0.618	0.564	0.502	0.428	0.345	0.264	0.118	0.011	0.002	0.892 0.890	0.860	**0.892** **0.891**
128	0.633	0.579	0.518	0.446	0.363	0.284	0.134	0.014	0.003	0.905 0.905	0.874	**0.905** **0.905**
192	0.652	0.597	0.539	0.468	0.386	0.309	0.156	0.020	0.004	0.922 **0.922**	0.892	**0.923** 0.922
256	0.665	0.610	0.552	0.482	0.402	0.326	0.170	0.025	0.006	0.934 **0.934**	0.904	**0.934** 0.933
384	0.680	0.628	0.571	0.501	0.423	0.347	0.187	0.034	0.008	0.948 **0.948**	0.920	**0.948** 0.947
512	0.691	0.640	0.583	0.514	0.438	0.360	0.196	0.041	0.010	0.956 0.954	0.931	**0.957** **0.956**

Table 6: Pass@K for Qwen3-0.6B. Low temperatures $\tau \in [0.1, 1.0]$ above, high temperatures $\tau \in [1.1, 2.0]$ together with the row-aggregate columns below.

K	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
1	0.011	0.022	0.032	0.041	0.047	0.051	0.051	0.048	0.043	0.036
2	0.020	0.041	0.061	0.077	0.087	0.093	0.094	0.089	0.079	0.066
3	0.029	0.059	0.086	0.107	0.122	0.130	0.130	0.124	0.111	0.093
4	0.036	0.075	0.108	0.134	0.152	0.161	0.162	0.154	0.139	0.117
6	0.050	0.102	0.146	0.179	0.201	0.213	0.214	0.204	0.185	0.158
8	0.061	0.125	0.177	0.215	0.240	0.254	0.255	0.244	0.222	0.191
12	0.078	0.162	0.226	0.271	0.299	0.315	0.317	0.305	0.279	0.242
16	0.091	0.190	0.263	0.312	0.343	0.360	0.362	0.350	0.322	0.282
24	0.110	0.232	0.316	0.372	0.406	0.424	0.426	0.413	0.384	0.340
32	0.124	0.264	0.355	0.415	0.451	0.469	0.471	0.458	0.428	0.382
48	0.145	0.310	0.410	0.476	0.515	0.532	0.533	0.520	0.490	0.442
64	0.162	0.345	0.449	0.519	0.559	0.576	0.576	0.563	0.534	0.485
96	0.189	0.395	0.504	0.578	0.619	0.634	0.635	0.622	0.594	0.546
128	0.210	0.430	0.541	0.618	0.658	0.673	0.675	0.662	0.636	0.589
192	0.240	0.476	0.591	0.670	0.709	0.723	0.727	0.718	0.693	0.648
256	0.261	0.507	0.625	0.704	0.744	0.757	0.763	0.756	0.733	0.689
384	0.291	0.548	0.668	0.748	0.789	0.804	0.812	0.807	0.786	0.743
512	0.312	0.575	0.697	0.778	0.821	0.836	0.846	0.841	0.820	0.779

K	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**	**Best τ**	**Uniform**	**Optimal (in / out)**
1	0.027	0.019	0.012	0.007	0.003	0.001	0.000	0.000	0.000	0.000	0.051 0.049	0.023	**0.051** **0.049**
2	0.051	0.036	0.023	0.013	0.006	0.002	0.001	0.000	0.000	0.000	0.094 **0.091**	0.045	**0.094** 0.091
3	0.073	0.052	0.033	0.019	0.009	0.003	0.001	0.000	0.000	0.000	0.130 0.127	0.065	**0.130** **0.127**
4	0.092	0.066	0.043	0.024	0.012	0.005	0.001	0.000	0.000	0.000	0.162 **0.158**	0.084	**0.162** 0.158
6	0.125	0.092	0.060	0.035	0.017	0.007	0.002	0.000	0.000	0.000	0.214 **0.213**	0.118	**0.214** 0.209
8	0.153	0.113	0.076	0.044	0.022	0.009	0.003	0.001	0.000	0.000	0.255 **0.253**	0.148	**0.255** 0.251
12	0.197	0.149	0.102	0.061	0.031	0.013	0.004	0.001	0.000	0.000	0.317 **0.317**	0.198	**0.317** 0.313
16	0.232	0.176	0.123	0.076	0.039	0.017	0.005	0.001	0.000	0.000	0.362 **0.362**	0.239	**0.362** 0.358
24	0.283	0.219	0.155	0.099	0.054	0.023	0.008	0.001	0.000	0.000	0.426 **0.426**	0.301	**0.426** 0.423
32	0.321	0.251	0.180	0.117	0.065	0.029	0.010	0.002	0.000	0.000	0.471 **0.471**	0.347	**0.472** 0.468
48	0.376	0.298	0.217	0.144	0.083	0.039	0.014	0.003	0.000	0.000	0.533 0.527	0.413	**0.534** **0.531**
64	0.417	0.332	0.245	0.165	0.097	0.048	0.018	0.004	0.000	0.000	0.576 0.564	0.459	**0.578** **0.574**
96	0.475	0.384	0.285	0.195	0.118	0.060	0.024	0.006	0.001	0.000	0.635 0.626	0.522	**0.636** **0.632**
128	0.517	0.421	0.315	0.217	0.134	0.070	0.029	0.007	0.001	0.000	0.675 **0.675**	0.565	**0.676** 0.671
192	0.577	0.476	0.358	0.249	0.157	0.084	0.038	0.010	0.001	0.000	0.727 **0.727**	0.624	**0.728** 0.724
256	0.619	0.516	0.391	0.272	0.174	0.094	0.044	0.013	0.002	0.000	0.763 **0.763**	0.665	**0.764** 0.757
384	0.675	0.571	0.440	0.305	0.198	0.109	0.053	0.018	0.002	0.000	0.812 **0.812**	0.721	**0.814** 0.802
512	0.712	0.608	0.476	0.329	0.215	0.121	0.059	0.021	0.003	0.000	0.846 **0.846**	0.758	**0.849** 0.837

G Sensitivity of optimal temperature sampling detection to dataset hardness-distribution variance

G.1 Statistical Analysis Approach

The pass@K scores reported in this paper are population averages over a particular set of questions with a particular hardness profile. A natural concern is whether the optimal sampling temperature is an artifact of *this* hardness profile. Would a different temperature win if the question population were reweighted, say by shifting weight toward harder questions or by concentrating on a narrower difficulty band?

The Goldilocks sampling effect (Proposition 4.2) predicts that the answer is *no*, with exponential margin. The contribution of a question with hardness H to pass@K is suppressed by the factor $\alpha\, e^{1-\alpha}$ with $\alpha = K/H$ (Eq. 16), which falls off exponentially as α moves away from 1. Questions outside the band $\alpha \approx 1$ are therefore exponentially silent, and any reweighting that redistributes mass between out-of-scale questions should have an exponentially suppressed effect on the argmax over τ. The optimal temperature is therefore expected to be robust to all but the most surgical reweightings, and increasingly so as K grows. The purpose of this section is to quantify that prediction empirically. We perturb the per-question weights, measure how much the chosen temperature degrades, and repeat the experiment across $K \in \{4,\ 32,\ 128,\ 512\}$ to see how the empirical sensitivity scales with the pass@K regime.

Experiment. We work on the data of the four models on each of our three datasets, LiveCodeBench, MATH-500, and Hi-ToM, restricted to the included questions whose peak pass-rate is at least 0.1% on the temperature grid. On LiveCodeBench, these are 766 questions for GPT OSS 20B, 359 for Mistral 7B Instruct, 647 for Qwen3-8B, and 85 for Qwen3-0.6B. In the same order, they are 500, 455, 496, and 492 on MATH-500, and 678, 615, 466, and 743 on Hi-ToM. For Qwen3-8B, we use a later release of its LiveCodeBench data, with 119,228,250 samples, in which 647 questions pass the filter. The table of Appendix C.3 counts the earlier release, with 100,568,250 samples, in which 646 questions pass it. For each dataset, each model, and each $K \in \{4,\ 32,\ 128,\ 512\}$, we compare two ways of choosing the best single temperature:

- **Reference A.** The best single τ_A as defined in Section 4.2, picked under uniform per-question weights.
- **Reweighted B.** A random per-question weight vector W is drawn from a Dirichlet smooth-base sampler over the included questions of the model; τ_B is the same argmax as in Reference A, but with each per-question term reweighted by W.

Crucially, only the *choice* of τ_B uses W. Once τ_B is chosen, score_B is computed back under uniform per-question weights, so score_B and score_A live on the same scale and are directly comparable. We report the ratio $\mathrm{score}_B/\mathrm{score}_A \in (0,\ 1]$ as a function of $\mathrm{KL}(W \,\|\, \mathrm{uniform})$. For each dataset and model, we draw 10,000 random W's with the same Dirichlet smooth-base sampler and seed, in log space so that no W underflows to zero, elect τ_B in log space so that no pass@K rounds to 1, and reuse the W's across the four K panels of its figure, so any difference between panels is attributable to K alone. Figure 4(b) repeats the top row of Figure 17 with 24 million W's; this moves its cells by at most 1.0 points.

The Dirichlet smooth-base sampler sweeps $\mathrm{KL}(W \,\|\, \mathrm{uniform})$ from 0 nats up to the entropy ceiling $\ln n$, where n is the number of included questions, so every panel covers the full reachable KL range. On LiveCodeBench, the ceilings are $\ln(766) \approx 6.64$ nats for GPT OSS 20B, $\ln(359) \approx 5.88$ for Mistral 7B Instruct, $\ln(647) \approx 6.47$ for Qwen3-8B, and $\ln(85) \approx 4.44$ for Qwen3-0.6B. On MATH-500 and Hi-ToM, they run from $\ln(455) \approx 6.12$ to $\ln(743) \approx 6.61$ nats, and the last KL band of each figure ends at its own.

The figures plot the W's with $0 \le \mathrm{KL} < \ln n$, and leave out the near one-hot W's with KL at the ceiling. On LiveCodeBench, these are 97 W's for GPT OSS 20B and 167 for Mistral 7B Instruct, all at the ceiling. The Qwen3-8B figure shows all 10,000. Of the 10,000 W's of Qwen3-0.6B on LiveCodeBench, the figure leaves out 2,005, all near one-hot W's at the ceiling, and plots the other 7,995. On Hi-ToM, the figures leave out 392 W's for GPT OSS 20B and 91 for Qwen3-0.6B, all at the ceiling, and the six other figures of MATH-500 and Hi-ToM leave out none. These two figures thus plot 9,608 and 9,909 W's.



Figure 17: Distribution of $\text{score}_B/\text{score}_A$ as a function of $\text{KL}(W \,\|\, \text{uniform})$ on the Mistral 7B dataset (359 included questions, peak pass-rate $\geq 0.1\%$), with W drawn from a Dirichlet smooth-base sampler over the questions. Each panel corresponds to one value of $K \in \{4, 32, 128, 512\}$; the same $10{,}000$ random W's are used in every panel, so differences reflect only the change of K. Each panel is a 6×6 column-normalized heatmap (KL bands on x, ratio bands on y); the cell label reports the percent of W's in that band, and the fill is on a log-scaled viridis from $\leq 0.01\%$ to 100%. The cell label 100 is reserved for columns in which *every* W lands in that ratio band; otherwise the label is the two-decimal percentage. At $K = 4$ the argmax is brittle (top-row occupancy falls to 69.71% at the highest KL band); at $K \in \{32, 128\}$ the rightmost band still keeps $\sim 84\%$ of W's within 10% of score_A; at $K = 512$ the rightmost band keeps 87.87%, marginally the most robust regime.

Reading the plot. Each panel of Figures 17 to 28 is a column-normalized heatmap. The x-axis is $\text{KL}(W \,\|\, \text{uniform})$ in nats, bucketed into 0-1, 1-2, 2-3, 3-4, 4-5, and 5 up to the entropy ceiling, which is the right edge: 5-6.64 for GPT OSS 20B, 5-5.88 for Mistral 7B Instruct, and 5-6.47 for Qwen3-8B on LiveCodeBench, and 5 up to the ceiling for every model on MATH-500 and Hi-ToM. For Qwen3-0.6B, the ceiling $\ln(85) \approx 4.44$ is below 5, so its KL bands are 0-1, 1-2, 2-3, 3-4, and 4-4.44. The y-axis is $\text{score}_B/\text{score}_A$, bucketed into ratio bands. On LiveCodeBench, the ratio bands start at each model's lowest ratio over its four panels, rounded down to a multiple of 0.1. They are 0.4-0.5, ..., 0.9-1.0 for Mistral 7B Instruct, 0.7-0.8, 0.8-0.9, and 0.9-1.0 for GPT OSS 20B and Qwen3-8B, and 0.2-0.5, 0.5-0.6, ..., 0.9-1.0 for Qwen3-0.6B. On MATH-500 and Hi-ToM, every heatmap uses the same ten ratio bands of width 0.1, from 0.0-0.1 to 0.9-1.0, like the equal bands of Figures 42 to 44. Each cell reports the percentage of W's that, conditional on landing in the cell's KL band, ended up in the cell's ratio band; by construction every column sums to exactly 100. Cell labels are two-decimal percentages, with one exception: the cell shows 100 only when *every* W in that column landed in that ratio band. A column where $9{,}999$ out of $10{,}000$ W's landed in it reads 99.99, never 100. The fill is a log-scaled viridis from $\leq 0.01\%$ (cell floor) to 100%.

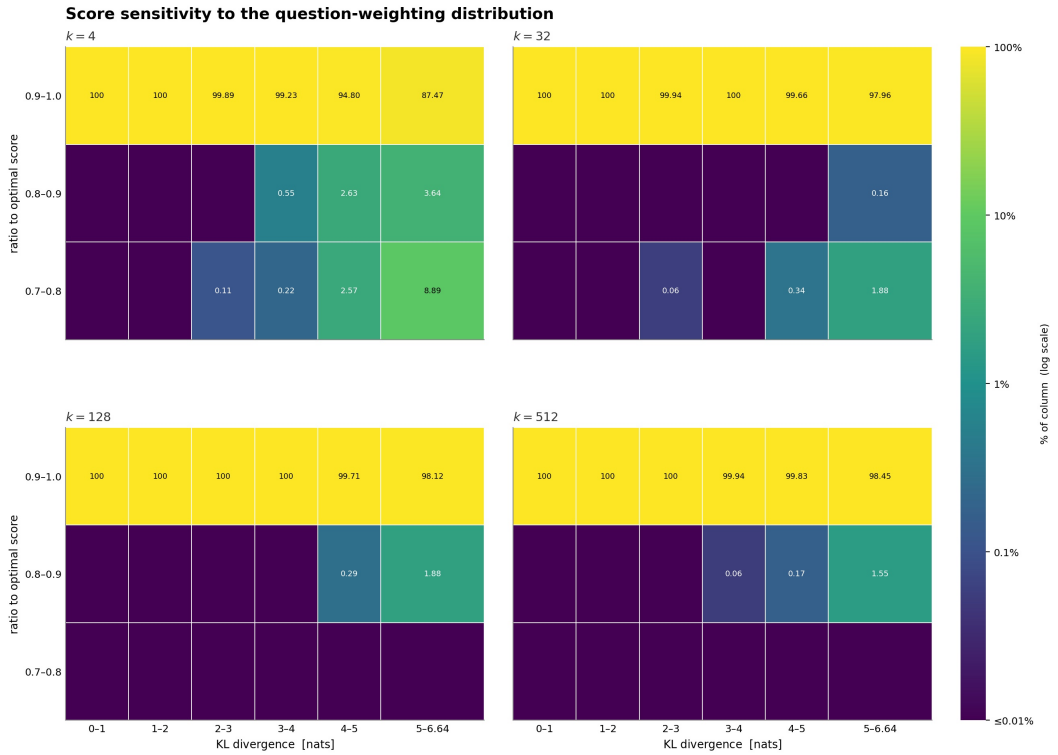


Figure 18: Distribution of $\text{score}_B/\text{score}_A$ as a function of $\text{KL}(W \,\|\, \text{uniform})$ for GPT OSS 20B (766 included questions, entropy ceiling $\ln(766) \approx 6.64$ nats), constructed as in Figure 17.

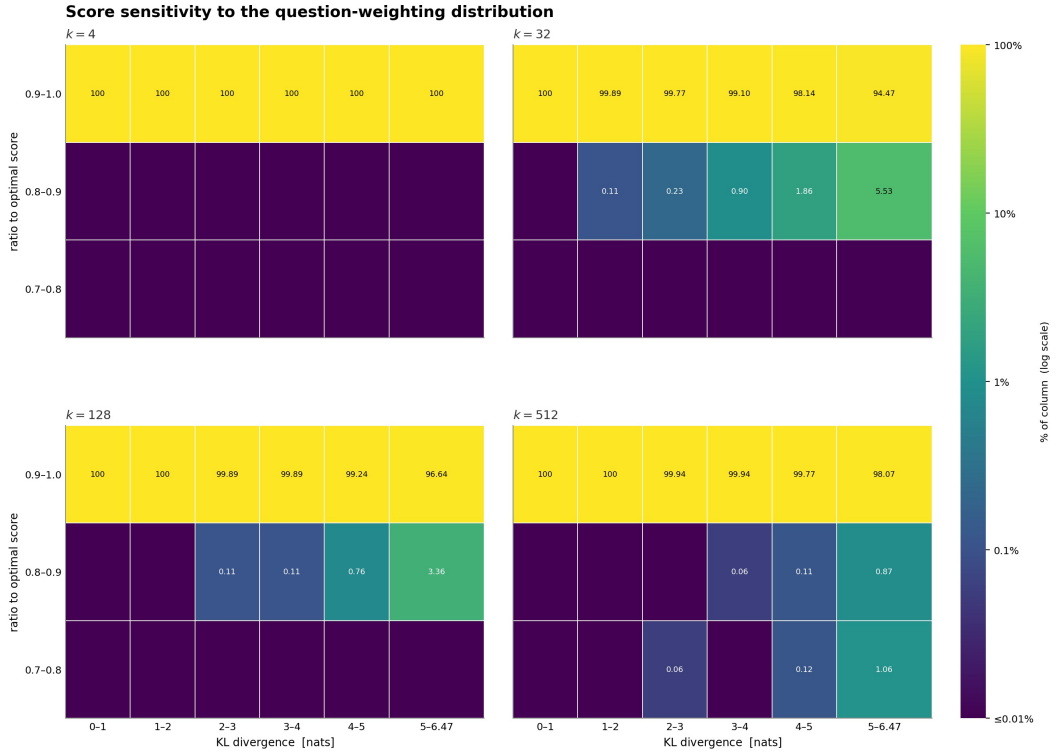


Figure 19: Distribution of $\text{score}_B/\text{score}_A$ as a function of $\text{KL}(W \,\|\, \text{uniform})$ for Qwen3-8B (647 included questions, entropy ceiling $\ln(647) \approx 6.47$ nats, later data release, see text), constructed as in Figure 17.

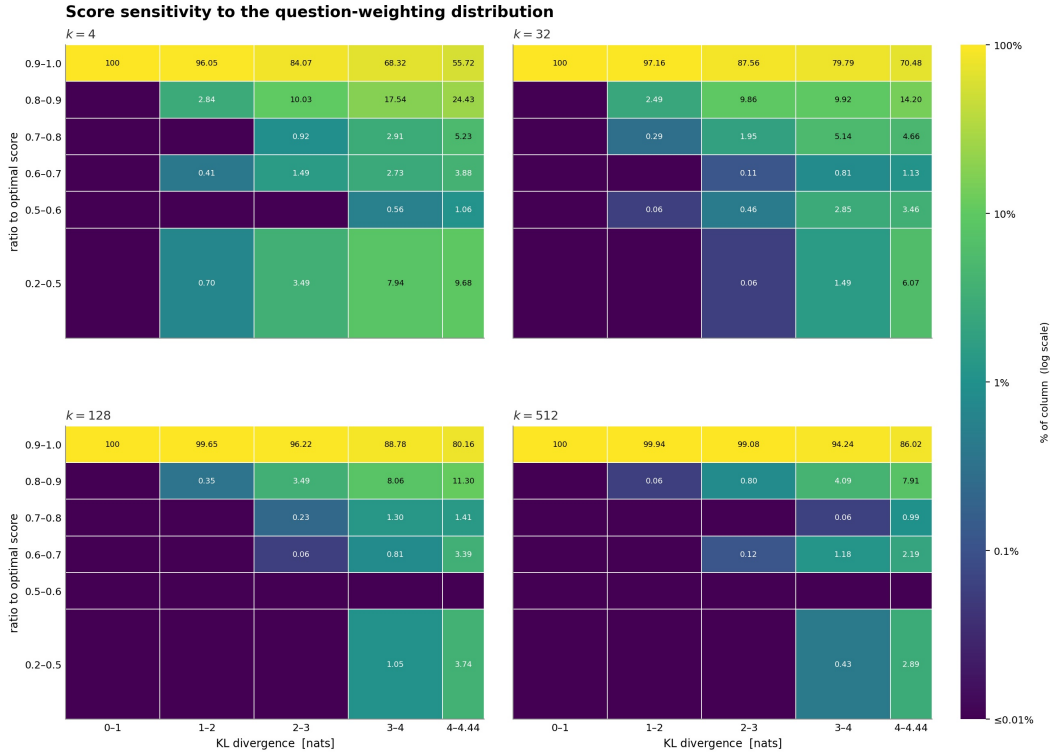


Figure 20: Distribution of $\text{score}_B/\text{score}_A$ as a function of $\text{KL}(W \,\|\, \text{uniform})$ for Qwen3-0.6B (85 included questions, entropy ceiling $\ln(85) \approx 4.44$ nats), constructed as in Figure 17.

G.2 Extended Results

Figures 17, 18, 19, and 20 show the heatmaps of Mistral 7B Instruct, GPT OSS 20B, Qwen3-8B, and Qwen3-0.6B on LiveCodeBench. Figures 21 to 24 show them on MATH-500, and Figures 25 to 28 on Hi-ToM, in the same order.

Findings on LiveCodeBench. Robustness of the optimal temperature is itself a function of the pass@K regime. At every K, low KL keeps $\text{score}_B/\text{score}_A$ near 1: in the lowest KL band, at least 99.8% of the W's of each model have a ratio of 0.9 or more. What changes with K is the rate at which high-KL columns slide downward. In every panel, the top row holds the W's with ratio ≥ 0.9, i.e., whose score_B is within 10% of score_A.

- At $K = 4$, the top-row occupancies in the 1-2 KL band and the highest band are 100% and 87.47% for GPT OSS 20B, 98.08% and 69.71% for Mistral 7B Instruct, 100% in both for Qwen3-8B, and 96.05% and 55.72% for Qwen3-0.6B. For Qwen3-8B at $K = 4$, all $10{,}000$ W's have ratio ≥ 0.9 (the lowest is 0.905), so its whole top row reads 100. For Mistral 7B Instruct, the choice at $K = 4$ is brittle. The top-row occupancy already drops to 99.86% in the lowest KL band, then slides monotonically to 69.71% at the entropy ceiling. The flat per-question pass curves at small K make the argmax easy to move with small reweightings.
- At $K = 32$ and $K = 128$, the table below gives the top-row occupancy in the highest KL band for each model. For Mistral 7B Instruct, the picture at these two values of K is similar: the top row stays above 98% up through $\text{KL} \leq 3$ nats and dips to roughly 84% and 83% respectively in the highest KL band.

The following table gives, for each model and K, the reference temperature τ_A, its score score_A, and the top-row occupancy (ratio 0.9-1.0, in $\%$) in the highest KL band, under "top". From left to right, this band is 5-6.64, 5-5.88, 5-6.47, and 4-4.44 nats.



Figure 21: Mistral 7B Instruct on MATH-500 (455 included questions), constructed as in Figure 17.

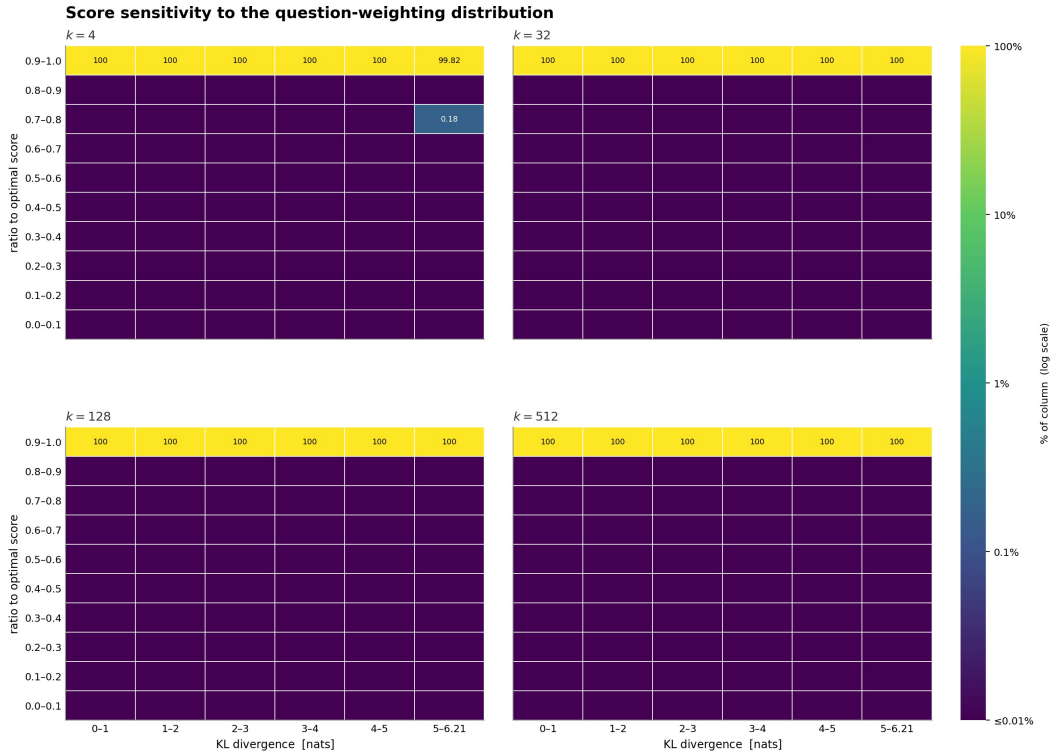


Figure 22: GPT OSS 20B on MATH-500 (500 included questions), constructed as in Figure 17.



Figure 23: Qwen3-8B on MATH-500 (496 included questions), constructed as in Figure 17.



Figure 24: Qwen3-0.6B on MATH-500 (492 included questions), constructed as in Figure 17.



Figure 25: Mistral 7B Instruct on Hi-ToM (615 included questions), constructed as in Figure 17.

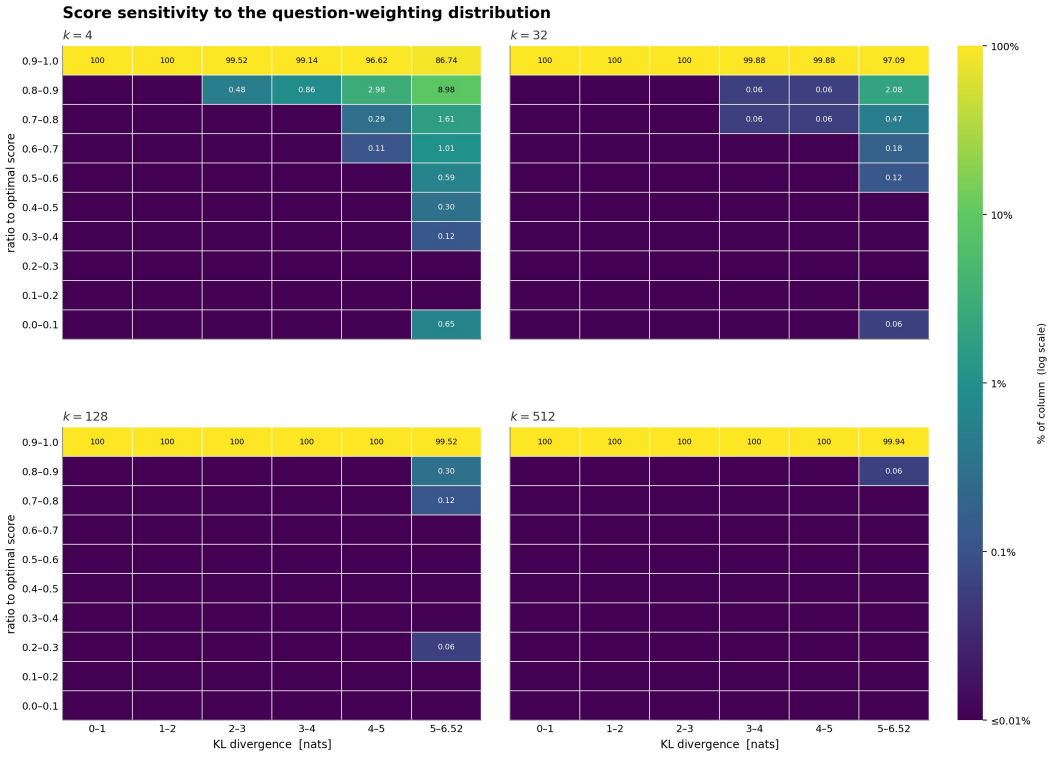


Figure 26: GPT OSS 20B on Hi-ToM (678 included questions), constructed as in Figure 17.



Figure 27: Qwen3-8B on Hi-ToM (466 included questions), constructed as in Figure 17.

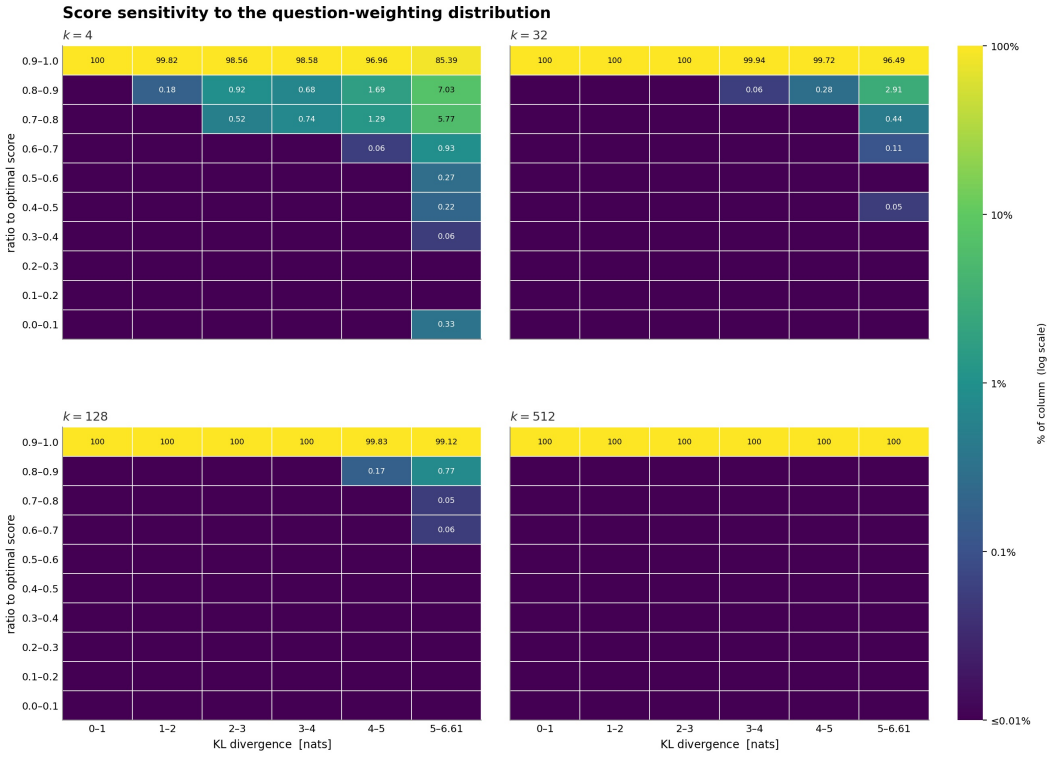


Figure 28: Qwen3-0.6B on Hi-ToM (743 included questions), constructed as in Figure 17.

	GPT OSS 20B			Mistral 7B Instruct			Qwen3-8B			Qwen3-0.6B		
K	τ_A	score_A	top	τ_A	score_A	top	τ_A	score_A	top	τ_A	score_A	top
4	0.90	505.86	87.47	0.70	110.44	69.71	1.20	454.64	100	0.70	13.77	55.72
32	0.80	645.46	97.96	0.90	203.60	84.18	1.20	542.34	94.47	0.70	40.06	70.48
128	0.80	701.40	98.12	0.90	267.11	83.31	1.10	585.45	96.64	0.70	57.35	80.16
512	0.90	742.11	98.45	1.00	319.77	87.87	1.10	618.73	98.07	0.70	71.92	86.02

- At $K = 512$, the same occupancies are 100% and 98.45% for GPT OSS 20B, 100% and 87.87% for Mistral 7B Instruct, 100% and 98.07% for Qwen3-8B, and 99.94% and 86.02% for Qwen3-0.6B. For Mistral 7B Instruct, the choice at $K = 512$ is marginally the most robust. The top-row occupancies essentially match $K = 32$ and $K = 128$ across the lower KL bands, and 87.87% of W's in the highest KL band still lose under 10%.

Findings on MATH-500 and Hi-ToM. On MATH-500, the top-row occupancy in the highest KL band is at least 92.31% for every model and K. It is lowest for Mistral 7B Instruct at every K, with 93.43%, 97.48%, 92.31%, and 95.36% at $K = 4$, 32, 128, and 512. On Hi-ToM, it is between 78.31% for Mistral 7B Instruct and 87.76% for Qwen3-8B at $K = 4$, and it rises at $K = 32$ and again at $K = 128$ for every model. At $K = 512$, it is between 99.94% for Mistral 7B Instruct and GPT OSS 20B and 100% for Qwen3-8B and Qwen3-0.6B.

H Sensitivity to Reweighting Outside a Protected Hardness Window

In Appendix G, we reweighted all questions at random. In this appendix, we test the Goldilocks sampling effect directly. For each K, we protect the questions whose hardness is within D decades of K by keeping their weights at $W_q = 1$, and let the weight of every other question range over $[1/\beta, \beta]$. We then find β^*, the smallest β at which a temperature that loses at least 10% under uniform weights becomes the best. We show β^* against D for each dataset, model, and K in Figures 29, 30, and 31. Figure 4(a) in Section 4.3.2 shows the curves of Mistral 7B Instruct on LiveCodeBench at a 10% loss.

H.1 Statistical Analysis Approach

Test. For a question q, let $s_q(\tau) = 1 - \bar{P}_q(\tau)^K$ be its pass@K at temperature τ, and for weights W let $S_W(\tau) = \sum_q W_q s_q(\tau)$. As in Appendix G, τ_A maximizes the score S_U under uniform weights, and $\text{score}_A = S_U(\tau_A)$. The temperatures that lose at least 10% form the set $C = \{\tau : S_U(\tau) \leq 0.9\,\text{score}_A\}$, and for a 1% loss we replace 0.9 by 0.99. The protected window holds the questions with $|\log_{10} H_q - \log_{10} K| \leq D$, which keep $W_q = 1$. The other questions have $W_q \in [1/\beta, \beta]$, and we do not normalize $\sum_q W_q$.

A temperature $\tau_c \in C$ wins under W if $S_W(\tau_c) - S_W(\tau') \geq 10^{-9}$ for every $\tau' \notin C$, so ties do not count. We define β^* as the smallest β on the grid 1.005^i, capped at 10^6, at which some $\tau_c \in C$ wins under some such W, and write $\beta^* > 10^6$ when no β up to 10^6 suffices. Since a larger D protects more questions, β^* is nondecreasing in D. We use $K \in \{4, 32, 128, 512\}$ and 23 values of D from 0 to 6, spaced more finely near 0, and stop each curve at the first D whose window holds every question.

Data. We run the test on all 12 combinations of our three datasets, LiveCodeBench, MATH-500, and Hi-ToM, with the four models. For every question, $P_q(\tau)$ is the profile fit of Appendix D to the question's own sample counts, and we keep the questions that pass the global 10^{-3} peak success-probability filter. On MATH-500 and Hi-ToM, these are the included questions of Appendix C. On LiveCodeBench, we use a corrected parsing of the sample counts, which gives 359, 766, 677, and 86 included questions for Mistral 7B Instruct, GPT OSS 20B, Qwen3-8B, and Qwen3-0.6B, respectively. Appendix G uses the earlier parsing, with 359, 766, 647, and 85. The correction changes the profiles of 4, 18, 197, and 3 questions of these models, respectively, and adds 30 included questions for Qwen3-8B and one for Qwen3-0.6B.

Exact certification. We verify every point with a question outside the window in exact rational arithmetic on the stored values of $s_q(\tau)$. For each finite β^*, an explicit W, with $W_q = 1$ inside the

window and every other weight in $[1/\beta^*, \beta^*]$, shows that some $\tau_c \in C$ wins, and a dual certificate shows that no $\tau_c \in C$ wins one grid step below. Such a certificate is a set of nonnegative weights $\lambda_{\tau'}$ on the temperatures $\tau' \notin C$ that sum to 1. With $g_q = \sum_{\tau' \notin C} \lambda_{\tau'} \left(s_q(\tau_c) - s_q(\tau')\right)$, no W allowed at β lets τ_c win if

$$\sum_{q \text{ protected}} g_q + \sum_{q \text{ outside}} \max\left(\beta g_q, \frac{g_q}{\beta}\right) < 10^{-9} \quad . \tag{51}$$

For each point with $\beta^* > 10^6$ and questions outside the window, such certificates show that no $\tau_c \in C$ wins at $\beta = 10^6$. Over the 12 combinations of dataset and model, this covers all 1,129 finite points and all 373 points with $\beta^* > 10^6$ and questions outside the window. An independent implementation re-checks every certificate in exact arithmetic, and every finite β^* and every verdict $\beta^* > 10^6$ is unchanged when the pass@K values are computed exactly from the fitted pass rates. These counts are those of the fitted pass rates. When we resample the sample counts of every question and temperature and refit them, the 90% interval of β^* spans, in the median, less than a factor of 2 on each side only for β^* below about 10^3, and the number of points above 10^6 has a 90% interval of 365 to 518.

The loss of the chosen temperature. The value of β^* shows when a temperature that loses at least 10% can become the best, but not how much a reweighting can cost τ_A. For the same windows and weights, we therefore also compute the worst-case loss

$$R(\beta, D) = \max_W \left(1 - \frac{S_W(\tau_A)}{\max_\tau S_W(\tau)}\right) \ , \tag{52}$$

over the W with $W_q = 1$ inside the window and $W_q \in [1/\beta, \beta]$ outside it. We compute R exactly on the double-precision values of $s_q(\tau)$, for $K \in \{1, 4, 16, 32, 64, 128, 256, 512, 1024\}$, $\beta \in \{1.5, 2, 3, 10, 30, 100, 1000\}$, and $D \in \{0, 0.25, 0.5, 1, 1.5, 2, 3\}$.

One side at a time. The Goldilocks weight $\alpha e^{1-\alpha}$ of Equation 16, with $\alpha = K/H_q$, falls exponentially in α only for questions much easier than K, where $\alpha \to \infty$. For questions harder than K, where $\alpha \leq 1$, it lies between α and $e\alpha$, and each further decade of hardness divides it by $10\, e^{-0.9\alpha}$, a factor between 4 and 10. We therefore also reweight one side of K at a time. For d from 0 to 3, only the questions whose hardness at their best temperature is more than d decades above K, or only those more than d decades below it, take weights in $[1/\beta, \beta]$, and every other question keeps $W_q = 1$. We find β^* as before, with the same exact certification, for $K \in \{4, 32, 128, 512\}$ at losses of 10% and 1%, and for every power of 2 from $K = 1$ to 1024 at losses of 1%, 5%, 10%, and 20%.

Tilting the mix. We also tilt every weight by hardness, with $W_q = 10^{\gamma(\log_{10} H_q - \log_{10} K)}$, which multiplies the weight by 10^γ per decade of hardness and favors harder questions for $\gamma > 0$. We use the 132 combinations of dataset, model, and K, for every power of 2 from $K = 1$ to 1024.

Natural sub-datasets. We then test shifts that occur naturally. We take the sub-datasets that the labels of each dataset define, namely the subjects and levels of MATH-500, the order, deception, story length, and prompting type of Hi-ToM, and the difficulty, platform, and version of LiveCodeBench, together with their pairwise intersections. We keep those with at least 20 questions, which gives 448 combinations of sub-dataset and model, and use $K \in \{1, 4, 16, 64, 256, 1024\}$. On each sub-dataset, we compare the temperature chosen on the full dataset with the best temperature for the sub-dataset alone.

The role of K. Since $s_q(\tau) = 1 - e^{-K/H_q(\tau)}$, multiplying K by a factor has exactly the same effect on this test as dividing every hardness by that factor. How β^* depends on K is therefore a matter of where K sits among the hardnesses of the questions.

Simulated hardness distributions. The hardness of our measured questions stays below 3×10^3, so to test larger budgets, we repeat the test on three simulated datasets of 1000 questions each. Their hardness follows one of three shapes on a linear hardness axis, up to a cutoff of 500,000 (Figure 32). The peaked shape has its maximum at $H = 25{,}000$, the reversed shape has its minimum there, both equal $1/e$ at $H = 50{,}000$, and the flat shape is constant. We place the i-th question at the

$(i - 0.5)/1000$ quantile of each shape, normalized as a density, which puts 966, 985, and 980 of the 1000 questions above $H = 10^4$.

In order of hardness, the simulated questions take their temperature profiles in turn from the 109 Mistral 7B Instruct LiveCodeBench questions with hardness between 10^2 and 10^3. We scale each template T in hardness by $\bar{P}_i(\tau) = \bar{P}_T(\tau)^{H_T/H_i}$, which gives $H_i(\tau) = (H_i/H_T)\, H_T(\tau)$ at every temperature. We use $K \in \{4, 32, 512, 10^4, 10^5, 10^6\}$ and D up to 7.5, and leave the test otherwise unchanged.

Reading the plots. In Figures 29, 30, and 31, open markers on the top row of each panel have $\beta^* > 10^6$, and the last one of each curve is the first D whose window holds every question. In Figure 32, the left panels show each shape on a linear hardness axis up to the cutoff of 500,000, with $H = 25{,}000$ and $H = 50{,}000$ marked, and the right panels show the hardness of the 1000 simulated questions of each shape, in bins of a quarter decade.

H.2 Extended Results

β^* against D. Over the 88 curves with at least five finite points, the median least-squares slope of $\log_{10} \beta^*$ against D for each dataset and loss lies between 1.36 and 2.39 decades per decade of D, and at $K = 4$ and a 10% loss the slopes lie between 0.9 and 2.5. These slopes are averages for windows centred at K, with hardness measured at each question's best temperature. In 42 of the 88 curves, a quadratic term in D has a nominal p-value below 0.01 and bends the fit by more than 0.25 decades from its chord, and in 35 of these the curve steepens as D grows.

Points with $\beta^* > 10^6$. Of the 373 points with $\beta^* > 10^6$ and questions outside the window, 248 have no finite β^* for the pass@K values as stored in double precision, which we prove with certificates for which Equation 51 holds at every β. When the pass@K values are instead computed exactly from the fitted pass rates, 117 of them still have no finite β^*, and the other 131 need $\beta > 10^{12}$. The remaining 125 points have a finite β^* above 10^6. For GPT OSS 20B at a 10% loss, with the fitted pass rates, no temperature in C can win under any weights on MATH-500 at $K = 32$, 128, and 512, and on Hi-ToM at $K = 512$ a win needs $\beta > 10^{12}$ at every D. When the sample counts are resampled, with pass@K stored in double precision, the curve of MATH-500 at $K = 32$ has no reachable loss in 9 of 20 resamples, and those at $K = 128$ and 512 in 18 of 20. In Figures 30 and 31, these four curves lie on the top row. In Figure 31, the curve of Qwen3-8B at $K = 512$ and a 10% loss has a reachable loss and $\beta^* > 10^6$ at every D, so it lies on the top row as well.

The loss of the chosen temperature. At $\beta = 10$ and $D = 1.5$, R is at most 0.94% on all 12 combinations of dataset and model at every K on the grid from 4 to 1024. With no protected window, R is 18% to 36% in the median combination at each of these K, and more than 56% on LiveCodeBench with Qwen3-0.6B. The value of 0.94% is not stable. When we resample the questions of each dataset, the 90% interval of this maximum is 0.96% to 4.10%, and when we resample the sample counts it is 0.84% to 1.14%. Between grid points, it reaches 1.34% at $K = 5$. A pool that we construct by adversarial search, with all ten of its summary statistics within the ranges of our datasets, loses 60.1% at $K = 512$ with $D = 1.5$. In each of 20 resamples of the questions and 20 resamples of the sample counts, the smallest D that keeps R at most 1% on all 12 combinations at every K on the grid from 4 to 1024 is at most 2.36, so $D = 2.5$ suffices in all of them. On the original data, with $D = 2.5$, R is exactly 0 at every K on the grid from 4 to 1024 and at most 0.18% at every integer K from 4 to 1024. This is a property of our datasets rather than a guarantee. At $K = 1$, $D = 1.5$ allows $R = 7.9\%$. Finally, R can be large even below β^*, through temperatures that lose less than 10% under uniform weights. At a 10% loss, R reaches 10% at 339 of the 3,219 grid points with $\beta < \beta^*$ that we compared.

One side at a time. For $K \in \{4, 32, 128, 512\}$, reweighting only the harder questions makes a temperature in C the best at some $\beta^* \le 10^6$ at 97.5% of the 315 points with $d \le 1$ and at least five questions outside. As d grows, $\log_{10} \beta^*$ rises by a median of 1.20 decades per decade of d, with an interquartile range of 0.87 to 1.44 over the 67 curves with at least three finite points, and by 1.12 over the 368 such curves of the larger grid. Reweighting only the easier questions still flips the choice at 22 of the 43 points with $d = 1$ at a 10% loss, counting the 4 curves on which no weighting reaches the loss when pass@K is stored in double precision. From $d = 2.5$, no $\beta \le 10^6$ yields a 10% loss at

any of the 34 points, and a 1% loss is reached at only 1 of 34. At these four values of K, each of the 28 flips with $d \geq 1$ at a 10% loss promotes a temperature colder than τ_A. In the median flip, all of the gain comes from questions that are easy at their own best temperature but within a decade of K at τ_A.

For deciding which of the easier questions can be reweighted without effect, hardness at a question's best temperature is therefore a poor yardstick. When we measure it at τ_A instead, reweighting only the questions more than 1.5 decades easier than K there, whose pass@K at τ_A exceeds $1 - 2 \cdot 10^{-14}$, by any factor up to 10^6, never makes a temperature that loses 1%, 5%, 10%, or 20% the best on our 12 combinations of dataset and model, at any K we test from 1 to 1024. This holds exactly, both on the double-precision values of $s_q(\tau)$ and with pass@K computed exactly from the fitted pass rates. At $K \in \{4, 32, 128, 512\}$ and at losses of 1% and 10%, it also holds in each of 20 resamples of the sample counts. It holds at all 4,177 such points of simulated pools built from our measured profiles, and up to $K = 1024$ it fails only in simulated pools whose questions each have their own steep best temperature. At $K = 4096$, one pool built from measured profiles flips at 1.5 decades. With 1 decade instead of 1.5, it still holds at a 10% loss at 101 of 102 points, but at a 1% loss 60 of the 102 points flip.

Part of this is a theorem. Write $M_A(\tau_c) = \sum_q (s_q(\tau_c) - s_q(\tau_A))$ over the questions that keep their weights, and $E_A = \sum_q e^{-K/H_q(\tau_A)}$ over the reweighted questions. A reweighted question can favor τ_c over τ_A by at most its failure probability $e^{-K/H_q(\tau_A)}$ at τ_A. So if $M_A(\tau_c) < 0$ for every $\tau_c \in C$, no $\tau_c \in C$ becomes the best for $\beta \leq \min_{\tau_c \in C} |M_A(\tau_c)|/E_A$. If each of the n_{out} reweighted questions is more than d decades easier than K at τ_A, then $E_A < n_{\text{out}}\, e^{-10^d}$. For $K \in \{4, 32, 128, 512\}$ and $d \geq 1$, the condition $M_A < 0$ holds at all 133 points at a 10% loss at which some question is reweighted, and this bound alone exceeds 10^6 at 95 of them. Without the condition the statement can fail. In a constructed pool of two groups of questions, a temperature that loses more than 10% becomes the best by down-weighting reweighted questions on which it is worse.

When the reweighted questions are easy at every temperature that competes, a bound follows from the loss alone. Let ℓ be the loss, so that $S_U(\tau_c) \leq (1 - \ell)\, S_U(\tau_A)$ for every $\tau_c \in C$, and suppose that we reweight only $n_{\text{out}} \geq 1$ questions, each more than d decades easier than K both at τ_A and at every $\tau \in C$, so that $s_q(\tau) > 1 - e^{-10^d}$ at all of them. Every reweighted question then has $|s_q(\tau_c) - s_q(\tau_A)| < e^{-10^d}$, and $|W_q - 1| \leq \beta - 1$, so every allowed W gives

$$\begin{aligned} S_W(\tau_c) - S_W(\tau_A) &\leq -\ell\, S_U(\tau_A) + (\beta - 1)\, n_{\text{out}}\, e^{-10^d} \\ &< n_{\text{out}} \left((\beta - 1)\, e^{-10^d} - \frac{\ell}{1 - \ell} \left(1 - e^{-10^d} \right) \right) , \end{aligned} \tag{53}$$

since $S_U(\tau_A) \geq S_U(\tau_c)/(1-\ell) > n_{\text{out}}(1-e^{-10^d})/(1-\ell)$. The right-hand side is at most 0 whenever $\beta \leq 1 + \frac{\ell}{1-\ell}\,(e^{10^d} - 1)$, so for these β no $\tau_c \in C$ beats τ_A, and therefore $\beta^* > 1 + \frac{\ell}{1-\ell}\,(e^{10^d} - 1)$. This bound is doubly exponential in d, and it exceeds 10^6 for every $d \geq 1.21$ at a 10% loss and every $d \geq 1.27$ at a 1% loss. On our data, however, such questions exist at only 24 of the 133 points above.

Tilting the mix. As measured, at $\gamma = 0.25$ the best temperature moves colder in none of the 132 combinations, and at $\gamma = -0.25$ it moves hotter in none. When we resample the questions or the sample counts, at least 129 of the 132 keep each direction in every resample, but all 132 do so in only 4 to 9 of 20. For γ from -1 to 1, the best temperature is nondecreasing in γ in 124 of the 132 combinations. In the other 8, all on MATH-500, it falls as γ grows under tilts of more than a factor of 1.8 per decade toward easier questions or at least 3.2 toward harder ones. In resamples, such falls also occur on LiveCodeBench and Hi-ToM. Toward harder questions, τ_A first loses 1% at a median tilt of a factor of 17 per decade, with 90% intervals of about 11 to 23 in resamples. This median is about 3 at $K = 1$ and about 2,100 at $K = 1024$, and higher at $K = 1024$ than at $K = 1$ in every resample, but the tilt needed is positively rank-correlated with K in only 9 of the 12 combinations of dataset and model. Toward easier questions, τ_A loses 1% at some tilt up to a factor of 10^4 per decade in only 23 of the 132 combinations, and in at most 30 in every resample. This direction holds approximately for moderate tilts by hardness within a dataset, but not across natural sub-datasets. There, in the 1,526 cases in which both the best temperature and the median hardness of the sub-dataset differ from those of the full dataset, the best temperature moves the way the hardness shift predicts in only 807, and on LiveCodeBench and MATH-500 it more often moves the other way.

Natural sub-datasets. The temperature chosen on the full dataset loses at least 10% in only 15 of the 2,688 cases, or 0.56%, all at $K \leq 4$, and at least 1% in 12.5% of them. Its median loss is below 0.1% on every dataset. The share of cases that lose at least 1% falls from 24.8% at $K = 1$ to 4.5% at $K = 64$, and is 5.8% and 5.1% at $K = 256$ and 1024. In each of 20 resamples of the questions of each dataset, fewer than 2% of the cases lose at least 10%, more than 80% of these are at $K \leq 4$, and the median loss stays below 0.25% on every dataset.

Reweighting whole groups is much weaker than reweighting single questions. On MATH-500 and Hi-ToM, no weighting of the levels of one factor, at any ratio, makes a temperature that loses 10% the best on the stored values, in all 144 cases. Weights on single questions do so in 122 of these cases, with no protected window and a median β^* of 7.5. On LiveCodeBench, some weighting of the levels of one factor makes a temperature that loses 10% the best on the stored values in 6 of the 72 cases, and each of these involves a group of fewer than 20 questions, or one or two questions without a difficulty label.

The role of K. Over K from 1 to 1024, β^* at $D = 0$ tends to grow with K, with a median Spearman correlation of 0.90 over the 48 combinations of dataset, model, and loss, but it never falls as K grows in only 7 of them. Pooling the 12 combinations of dataset and model and the 11 values of K at each loss, β^* at $D = 0$ follows the share of questions easier than K, with Spearman correlations of 0.76 to 0.82 over the four losses, more closely than $\log_{10} K$, with 0.51 to 0.63.

Simulated hardness distributions. For $K \leq 512$, which lies below the bulk of the simulated questions, the window holds few or no questions until D reaches that bulk. Until then, β^* stays within a factor of about 2 of its value at $D = 0$, which is between 2.0 and 2.5 at a 10% loss and between 1.1 and 1.5 at a 1% loss, and it then rises steeply (Figure 33). For $K = 10^5$ and 10^6, the least-squares slope of $\log_{10} \beta^*$ against D lies between 1.1 and 3.5 decades per decade, and 8 of these 12 curves rise slowly at first and steeply later. For these two values of K, β^* exceeds 10^6 by $D = 1.25$ to 2.5, with 5 to 172 questions still outside the window.

I min_p Sampling on Hi-ToM

The temperature sweeps of the main text and of Appendix C vary τ alone. In this appendix, we fix τ and vary `min_p`, a sampler parameter that removes unlikely tokens. A token is dropped if its probability is below `min_p` times that of the most likely token, and `min_p` $= 0$ is plain sampling. We ask whether one `min_p` value is enough at a fixed temperature, whether $\tau = 2.0$ does better than $\tau = 1.5$ with `min_p`, and whether the best `min_p` value does better than the best plain temperature. We study this on Hi-ToM, with Qwen3-0.6B at both temperatures and Mistral 7B Instruct at $\tau = 2.0$ alone.

Setup. We sampled with `vLLM` 0.19.0, `top_p` $= 1$, and `top_k` $= 0$, on the NVIDIA GPUs listed below, with all values of a question at $\tau = 2.0$ on one GPU. For Qwen3-0.6B, the chat template, token limit, and precision are those of Appendix C, and the sampler, sample counts, and GPUs are those given here. For Mistral 7B Instruct, the prompts, token limit, precision, and grader are those of that appendix.

Model	τ	kept questions	ν	GPUs (questions)
Qwen3-0.6B	1.5	648	500	RTX 5090, with one L40S in the screen
Qwen3-0.6B	2.0	648	400	RTX 5090 (351), L40S (270), RTX 4090 (27)
Mistral 7B Instruct	2.0	617	1500	RTX 5090 (330), L40S (287)

The grid at $\tau = 1.5$ has the 10 values `min_p` $\in \{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.2, 0.3, 0.5, 0.7\}$, and that at $\tau = 2.0$ has the 5 values `min_p` $\in \{0.0056, 0.0128, 0.0316, 0.0721, 0.178\}$. Since `vLLM` applies `min_p` after temperature, a token survives if and only if its untempered probability is at least `min_p`$^\tau$ times that of the most likely token. Within 1%, the $\tau = 2.0$ values have the untempered cutoffs of the $\tau = 1.5$ values 0.001, 0.003, 0.01, 0.03, and 0.1, which run from $3.2 \cdot 10^{-5}$ to 0.032. As a result, each such pair keeps nearly the same tokens, and $\tau = 2.0$ only samples more evenly among them. Because `min_p` can only remove diversity, we chose $\tau = 1.5$ at or above the large-K optimum of plain sampling for Qwen3-0.6B, 1.35 to 1.5 at $K = 256$ in the sweep of Appendix F.2, and $\tau = 2.0$ to test whether a higher temperature does better.

For Qwen3-0.6B, we screened all 1,200 questions with 100 samples at each $\tau = 1.5$ value, and kept a question if and only if the pass rate of its best value was below 0.30. In a second stage, every kept question received 400 samples at each value of both grids. For Mistral 7B Instruct, we used the 617 questions kept by its first stage in Appendix C.5, with no further screen. Pass rates are instance-weighted, and outside the ratio-to-uniform results we use the unbiased pass@K estimator.

Results at $\tau = 1.5$. One `min_p` value is enough for Qwen3-0.6B at $\tau = 1.5$. Figure 34 and the table below show pass@K as a ratio to the uniform scheme of Section 4.2, which spreads K equally over the 10 values, so a ratio above 1 is better. We compute the ratios from measured pass rates, because the degree-10 profile fit of Appendix D is underdetermined on 10 grid values. At every K from 1 to 384, the leave-one-out (LOO) optimum over all values is within 0.003 of the best single value in this ratio. The fixed value 0.003 is within 0.012 of it at these K, with the largest gap at $K = 1$, and exceeds the uniform scheme.

Plain sampling is below the uniform scheme until $K \approx 25$–30. The best single value decreases as K grows, from 0.01 for $K \leq 6$ to 0.003 for $K = 8$–16 and 0.001 for $K \geq 24$. Values of 0.2 and above lose to the uniform scheme at every K, with ratios of at most 0.99, so the table stops at 0.2. The best single value at each K is in bold, and the bootstrap standard deviation of every single-value entry is at most 0.030.

`min_p`	$K = 1$	$K = 16$	$K = 64$	$K = 384$
0	0.904	0.979	1.024	1.008
0.001	1.284	1.169	**1.081**	**1.014**
0.003	1.318	**1.174**	1.076	1.013
0.01	**1.329**	1.166	1.065	1.006
0.03	1.291	1.132	1.037	0.983
0.1	1.151	1.015	0.940	0.914
0.2	0.987	0.873	0.808	0.780
Optimal (LOO)	1.329	1.173	1.082	1.014
uniform pass@K	7.32%	58.38%	85.32%	97.94%

In absolute terms, with $\nu = 500$, the best single value raises pass@1 from 6.62% at `min_p` $= 0$ to 9.73%, a gain of 3.11 ± 0.14 percentage points (pp), or 47%, where $\pm$ is the paired standard error over questions. It gains 11.34 ± 0.51 pp in pass@16 and 4.63 ± 0.45 pp in pass@64, and the fixed value 0.003 gains 3.03, 11.34, and 4.21 pp at these K. In the second stage, every value from 0.001 up shortens the mean output by 35% to 46%, and the share of samples at the 1000-token limit falls from 6.4% to at most 0.5%.

$\tau = 2.0$ against $\tau = 1.5$. With `min_p`, $\tau = 2.0$ does better than $\tau = 1.5$ for Qwen3-0.6B at each $K \in \{1, 4, 16, 64\}$. We compare the best value at each temperature in the table below, picked per K on the 648 questions and their 400 second-stage samples per value. By a decision rule fixed in advance, $\tau = 2.0$ is better at K if and only if the gap exceeds 2 paired standard errors and is positive in at least 95% of 400 split-half checks. A check picks both best values on a random half of the questions and measures the gap on the other half. Worse is the mirror image, and anything else is no detectable difference.

	best at $\tau = 2.0$		best at $\tau = 1.5$			
K	`min_p`	pass@K	`min_p`	pass@K	gap (pp)	result for $\tau = 2.0$
1	0.0316	10.42%	0.01	9.71%	$+0.71 \pm 0.11$	better
4	0.0316	33.72%	0.01	31.51%	$+2.22 \pm 0.28$	better
16	0.0316	72.54%	0.003	68.97%	$+3.57 \pm 0.38$	better
64	0.0128	95.11%	0.001	92.92%	$+2.20 \pm 0.37$	better
256	0.0128	99.73%	0.001	99.25%	$+0.48 \pm 0.23$	borderline
400	0.0128	100.00%	0.001	99.69%	$+0.31 \pm 0.22$	no detectable difference

All 400 split-half gaps are positive at each $K \in \{1, 4, 16, 64\}$. At $K = 256$, the gap passes the 2-standard-error test, but the share of positive splits lies around the 95% bar and depends on the split seed, so the result is borderline. At $K = \nu$, pass@K is the share of questions with any correct sample, so it saturates.

A single setting carries most of the gain. Figure 35 shows each $\tau = 2.0$ value relative to $\tau = 1.5$ with $\texttt{min_p} = 0.003$, the fixed value above, so a curve above the dark line at 1 is ahead of this setting. Against it, the value 0.0316 gains between 0.77 ± 0.10 and 3.57 ± 0.38 pp at each $K \in \{1, 4, 16, 64\}$, and 0.31 ± 0.34 pp at $K = 256$, which is not significant.

The best $\texttt{min_p}$ value is higher at $\tau = 2.0$ than at $\tau = 1.5$ at every evaluated K. In untempered terms, the best cutoff is the same at both temperatures at $K = 1$ and 4, namely 10^{-3}, and higher at $\tau = 2.0$ at each evaluated K from 16 to 400. Between values with the same untempered cutoff, $\tau = 2.0$ gains 0.7 to 4.2 pp at each $K \in \{1, 4, 16, 64\}$ at the three cutoffs of 10^{-3} and above. At $1.6 \cdot 10^{-4}$, the result is mixed, from -0.38 ± 0.29 pp at $K = 4$ to $+2.56 \pm 0.37$ pp at $K = 64$. At the smallest cutoff, $3.2 \cdot 10^{-5}$, $\tau = 2.0$ loses up to 4.83 ± 0.36 pp at $K = 1$, 4, and 16. With $\texttt{min_p} = 0.0316$, the outputs at $\tau = 2.0$ stay close in length to those of the best values at $\tau = 1.5$, at 251.1 against at most 222.8 tokens per sample.

Comparison with the best plain temperature. For Qwen3-0.6B, the best $\texttt{min_p}$ value also does better than the best plain temperature of the sweep of Appendix F.2. This holds at every evaluated K at $\tau = 2.0$, by the decision rule above, with at least 99% of the 400 split-half gaps positive. At $\tau = 1.5$, it is ahead at each $K \in \{1, 4, 16, 64, 256\}$, and the gap at $K = 400$ is not significant. We give these gaps in the table below, on the 642 questions that both the screen and the sweep's first stage kept. To leave out the samples that selected the questions, we score this study on its 400 second-stage samples and the sweep on its 1000.

K	best plain τ	gap at $\tau = 2.0$ (pp)	gap at $\tau = 1.5$ (pp)
1	1.275	$+1.52 \pm 0.10$	$+0.79 \pm 0.07$
4	1.275	$+4.49 \pm 0.28$	$+2.22 \pm 0.19$
16	1.275	$+7.28 \pm 0.43$	$+3.68 \pm 0.29$
64	1.35	$+4.49 \pm 0.41$	$+2.27 \pm 0.34$
256	1.35	$+1.02 \pm 0.22$	$+0.54 \pm 0.22$
400	1.35	$+0.56 \pm 0.19$	$+0.25 \pm 0.16$

At $K = 1$, the best values reach 10.37% at $\tau = 2.0$ and 9.65% at $\tau = 1.5$, against 8.86% for the best plain temperature and 1.83% for plain $\tau = 2.0$ in the sweep. The split-half gap at $\tau = 1.5$ is positive in all 400 splits at each $K \in \{1, 4, 16, 64\}$, where the fixed value 0.001, not re-picked per K, exceeds the best plain temperature by at least 6.5 standard errors. On the 642 questions, the two setups agree where they overlap: plain sampling at $\tau = 1.5$ in this study differs from $\tau = 1.5$ in the sweep by -0.04 ± 0.06 pp in pass@1, and by less than one standard error at each evaluated K.

Mistral 7B Instruct. At $\tau = 2.0$, the best $\texttt{min_p}$ value does better for Mistral 7B Instruct than the best plain temperature at each $K \in \{1, 4, 16, 64, 256\}$, with relative gains of 19% at $K = 1$ and 14% at $K = 16$. At $K = 1024$ and 1500, there is no detectable difference, by the same decision rule, fixed in advance. We take as reference the best of the 31 plain temperatures of Appendix C, picked per K, and score both sides on all $\nu = 1500$ samples, which for the reference include the 1000 of the first stage. In Figure 36, a curve above the dark line at 1 is ahead of this reference.

	best value, 0.0056	best plain τ			
K	pass@K	τ	pass@K	gap (pp)	result for $\texttt{min_p}$
1	7.33%	1.575	6.17%	$+1.16 \pm 0.06$	better
4	23.72%	1.575	20.20%	$+3.52 \pm 0.17$	better
16	53.19%	1.575	46.52%	$+6.68 \pm 0.31$	better
64	78.56%	1.725	72.48%	$+6.08 \pm 0.43$	better
256	91.60%	1.8	89.81%	$+1.79 \pm 0.49$	better
1024	97.12%	2.0	98.00%	-0.88 ± 0.58	no detectable difference
1500	97.89%	2.0	99.19%	-1.30 ± 0.60	no detectable difference

All 400 split-half gaps are positive at each evaluated $K \leq 256$. However, at $K = 1024$ and 1500, there is no detectable difference from plain $\tau = 2.0$, as the gap at $K = 1024$ is within 2 standard errors, and at $K = 1500$, 87% of the splits are negative, below the 95% bar.

The best value is the smallest of the grid, 0.0056, at every K, and pass@K falls as $\texttt{min_p}$ rises. Since no value between 0 and 0.0056 was tested, the grid cannot locate the best cutoff. The value 0.0128

ties 0.0056 only at $K \leq 4$. For Qwen3-0.6B, on its 648 questions with at most 1000 new tokens, the best value at $\tau = 2.0$ is 0.0316 or 0.0128 at every evaluated K. Against plain sampling at $\tau = 2.0$, the value 0.0056 gains 3.41 ± 0.14 pp at $K = 1$, from 3.92%, and 15.57 and 11.42 pp at $K = 16$ and 64. Output length falls as `min_p` rises, from 449 tokens per sample at 0.0056, with 76% at the 500-token limit, to 238 at 0.178, with 3%.

Ratio to uniform at $\tau = 2.0$. Figure 37 shows $\mathrm{rtu}(K)$ at $\tau = 2.0$, defined as in Figure 34, where the uniform scheme spreads K equally over the 5 values. For Qwen3-0.6B, 0.0316 is the best single value at every $K \leq 32$, with a ratio of 1.107 at $K = 1$, and 0.0128 is the best from $K = 48$ on. The LOO optimum puts all samples on 0.0316 at every $K \leq 16$, and it is within 0.004 of the best single value at every K. The value 0.178 is below the uniform scheme at every K, with a largest ratio of 0.974. For Mistral 7B Instruct, 0.0056 is the best single value at every K, and the optimal allocation puts all samples on it at every $K \leq 384$, so its curve lies under the red trace. Its ratio rises from 1.057 at $K = 1$ to 1.109 at $K = 32$ and falls to 1.025 at $K = 1024$, while that of 0.178 falls from 0.884 to 0.558. The bootstrap standard deviation of every single-value entry is at most 0.013 for Qwen3-0.6B and 0.021 for Mistral 7B Instruct.

Question profiles and hardness. We repeat the analyses of Figures 1, 5, 8, and 11 with `min_p` in place of the temperature, on measured pass rates, since a profile fit is underdetermined on these grids. The hardness H_q is Equation 3 at the best value of the question on the grid, and the efficient range E_q is the set of contiguous grid values around the peak whose pass rate is at least half the peak, so $|E_q|$ counts grid values. Grid values are spaced equally on every `min_p` axis, and a sliding mean over `min_p` is a mean of grid position. All questions were kept as hard, so H_q starts near 2, and the easy side of the temperature figures has no counterpart here. For Mistral 7B Instruct, the hardness axes leave out the 19 questions below the inclusion floor, which leaves 598.

In Figure 38, the smallest value, 0 at $\tau = 1.5$ and 0.0056 at $\tau = 2.0$, samples at a fixed τ, so the profiles are not pinned at 0 or 1 at the left edge. A temperature profile is pinned at 0 or 1 at $\tau = 0$, where sampling is greedy. The peaks spread over the whole grid. At $\tau = 1.5$, 33 of the 648 questions peak at `min_p` $= 0$ and 24 at 0.7, although 576 have no correct sample at 0.7. At $\tau = 2.0$, 133 of the 648 Qwen3-0.6B questions and 119 of the 598 Mistral 7B Instruct questions peak at the largest value, 0.178. Each highlighted question is the easiest one peaking in its third, at 26.5% to 37.1%.

A hard question's peak rests on a few correct samples, and its H_q comes from the same maximum. We therefore describe a trend only if it passes two checks. First, in 400 random splits of the samples of every cell into halves, with the peak or range from one half and H_q from the other, it keeps its sign in at least 95% of the splits. Second, its slope lies outside the central 95% of the slopes of a null, on the side of its sign. In the null, every question has the same profile shape, the pass rate of each grid value pooled over all questions of the study, scaled so that its mean over the grid is the question's own mean pass rate and capped at 1, with binomial draws at the same ν, in 400 replicates, from which H_q, the peak, and E_q are recomputed. The first check alone is not enough, because noise grows with H_q, and on the null data, with 40 splits in each of 100 replicates, it passes in up to all of them.

In Figure 39, harder questions peak at smaller `min_p` in all three studies, that is, with fewer tokens removed. The OLS slope of the peak grid position on $\log_{10} H_q$ is -3.10 grid steps per decade at $\tau = 1.5$, and -1.98 and -1.51 at $\tau = 2.0$ for Qwen3-0.6B and Mistral 7B Instruct. It keeps its sign in all 400 splits, and no null replicate reaches it. The central 95% of the null slopes lies within $[-0.60, 0.54]$ in every study. Without the questions that peak at the smallest value, the slopes are -2.86, -1.35, and -1.48, and they pass both checks. The peak temperature trends of Figure 5 rise gently with hardness once near-zero peaks are excluded; on these hard questions, `min_p` moves the same way, toward more diverse sampling.

In Figure 40, the OLS slopes of the efficient range size on $\log_{10} H_q$ are -2.88, -1.69, and -1.96 grid values per decade. Noise alone also gives negative slopes, and the null gives median slopes of -3.25, -1.43, and -1.20. At $\tau = 2.0$, the range narrows with hardness beyond noise for both models, clearly for Mistral 7B Instruct, where no null replicate reaches its slope, and narrowly for Qwen3-0.6B, where 2 of 400 do. At $\tau = 1.5$, the measured slope is flatter than that of 398 of the 400 null replicates, but this null's common shape is wider than the profiles of the easiest questions, which overstates the narrowing from noise, so the test cannot tell a trend from noise there.

Figure 41 shows no trend beyond noise in any study. Among bins with at least 10 questions, the lowest bar is 0.73 at $\tau = 1.5$ and 0.70 for both models at $\tau = 2.0$. Any fall of the bars with H_q is within what noise alone produces, as the null slope of the bars on $\log_{10} H_q$, an unweighted OLS fit over all non-empty bins at their geometric centres, is at least as steep as the measured one in all 400 replicates at $\tau = 1.5$ and for Mistral 7B Instruct, and in 41 of 400 for Qwen3-0.6B at $\tau = 2.0$. At $\tau = 1.5$, the split-half check fails as well.

Reweighting the questions. We repeat the random reweighting of Appendix G with `min_p` in place of the temperature, using the same Dirichlet smooth-base sampler, seed, and bands, with 10,000 random W's per study reused across its four K panels. The grid has 10 or 5 values. A question's pass@K at each value is the unbiased estimator of its sample counts, and its peak pass rate, which sets the 0.1% filter and H, is its largest measured rate on the grid. In Appendix G, both come from the fitted profiles. The largest K is 512 for Mistral 7B Instruct, as in Appendix G, and 256 for Qwen3-0.6B, since $\nu = 500$ at $\tau = 1.5$ and $\nu = 400$ at $\tau = 2.0$. The 0.1% peak filter keeps all 648 questions of Qwen3-0.6B and 598 of the 617 of Mistral 7B Instruct, leaving out the 8 with no correct sample and the 11 whose best value has one correct sample in 1500, so the entropy ceilings are $\ln(648) \approx 6.47$ and $\ln(598) \approx 6.39$ nats. All of these questions were kept as hard, with H between 2.15 and 749.5 at their best value, so this test holds no easy questions.

Figures 42, 43, and 44 show that low KL costs little. In the 0-1 band, every W has a ratio of at least 0.9 at every K in all three studies. In the highest band, the top-row occupancy rises with K in each study, from 78.53%, 78.40%, and 82.44% at $K = 4$ to 99.88%, 100%, and 98.04% at the largest K, for Qwen3-0.6B at $\tau = 1.5$ and $\tau = 2.0$ and Mistral 7B Instruct, respectively. For Qwen3-0.6B at $\tau = 2.0$ and $K = 256$, every value is within 10% of the best, with a lowest ratio of 0.9075, so the whole top row reads 100. Here a label of 100 means that every W of the column has a ratio of at least 0.9.

Here τ_A and τ_B are `min_p` values of the grid, 10 at $\tau = 1.5$ and 5 at $\tau = 2.0$, so each panel's ratios take at most 10 or 5 distinct values. Every panel shows all 10,000 W's, since none has a computed KL below 0 or at the ceiling. Every panel has the same nine ratio bands, from 0.1-0.2 to 0.9-1.0, and its six KL bands are drawn at equal width, so all cells have the same size. The lowest band that holds a W is 0.1-0.2, 0.7-0.8, and 0.5-0.6 for Qwen3-0.6B at $\tau = 1.5$ and $\tau = 2.0$ and Mistral 7B Instruct, respectively, and 163, 178, and 162 of the 216 cells of each figure hold no W.

Reweighting outside a protected window. We also run the test of Appendix H with `min_p` in place of the temperature, with the same D grid, losses, β grid, and exact certification, on the questions and pass@K values of the reweighting above. The hardness H_q of a question is measured at its best value. Figure 45 shows β^* against D.

With no protected window, a weighting that changes no weight by more than a factor of 1.78 to 237 makes a value that loses 10% the best, and a factor of 1.11 to 10.27 suffices for a 1% loss. β^* exceeds 10^6 from $D = 0.8$ to 1.75 at a 10% loss and from $D = 1.25$ to 2.0 at a 1% loss. When the hardness and pass@K of each question come from separate random halves of its samples, with the inclusion floor applied to the hardness half and the sweep run without exact certification, this D is unchanged or one grid step away in 192 of 200 cases, 10 splits of each of the 20 curves with $K \leq \nu/2$. In the other 8 cases, all for Mistral 7B Instruct at a 10% loss, it moves by more than one grid step. The largest are at $K = 4$, where this D is 0.8 on all samples and splits give 2.0 and 2.5, and at $K = 512$, where a split gives 0.5 against 1.25. Since the hardness of these questions spans less than 2.6 decades, the first D whose window holds every question is 1.25 to 2.5, and 6 of the 23 curves reach the top row only at that D.

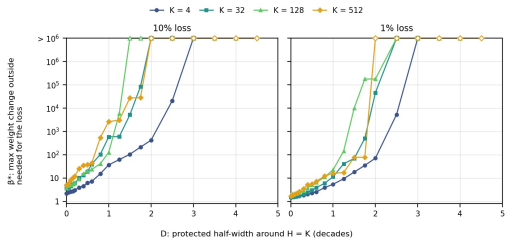
Limitations. The study covers two models, one dataset, and two temperatures, and the best combination of τ and `min_p` may lie at an untested temperature, such as one above 2.0 or between 1.5 and 2.0. The 648 and 617 questions are those kept as hard, so their absolute pass rates are not benchmark scores. Since every best value above is an in-sample maximum, its margin is optimistic, and the standard errors ignore the pick. The split-half checks guard against this in the comparisons between temperatures and with the best plain temperature, and the results of the decision rule depend on the split seed in two cases, both at $K = 256$: between the two temperatures, and between $\tau = 1.5$ and the best plain temperature. We compared one-dimensional slices, `min_p` at fixed τ and τ at `min_p` $= 0$, and did not run `min_p` at the best plain temperatures, 1.275 to 1.35 for Qwen3-0.6B and,

at $K \leq 256$, 1.575 to 1.8 for Mistral 7B Instruct. For Qwen3-0.6B, these come from measured pass rates on the 642 questions. The table of Appendix F.2 uses fitted profiles on all 743 questions kept by its first stage.

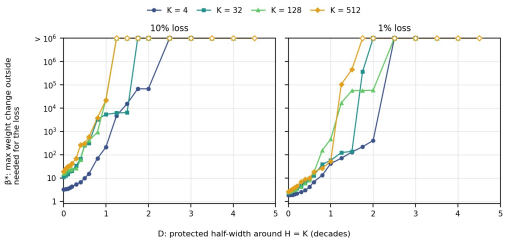
For Qwen3-0.6B, this study and the sweep also differ in hardware, seeds, and ν. On all 1,200 questions, plain sampling in the $\tau = 1.5$ screen passes 18.78% of its samples against 18.48% for $\tau = 1.5$ in the sweep. We have not resolved this gap of 0.30 pp (paired $z = 2.70$). It is $+0.35$ pp on the 1,090 questions that the screen ran on RTX 5090 GPUs and -0.19 pp on the 110 on the L40S GPU, too few to tell the two GPUs apart.

For Mistral 7B Instruct, the reference samples combine the 1000 that selected the questions, on RTX 5090 GPUs, with 500 more, drawn on L40S GPUs or from a continuation of the first stage's run. Selecting on low observed rates biases the reference's pass rates slightly down, while the `min_p` samples are fresh, so the selection mildly favors `min_p`. We have not measured the size of this bias. It can act only on questions near the 0.30 threshold, mostly at small K. Over all 1500 samples, 30 of the 617 questions have a best single-temperature pass rate of at least 0.25, and none of 0.30 or more.

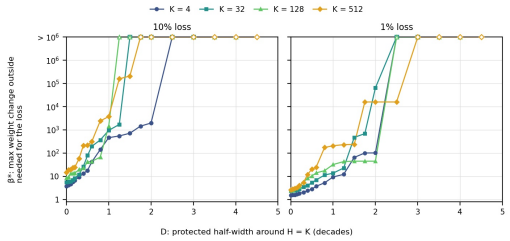
For $\tau = 2.0$ against $\tau = 1.5$ and for Mistral 7B Instruct, the results of the decision rule hold on the questions of each GPU type. For Qwen3-0.6B, the 27 RTX 4090 questions are enough to decide at $K = 16$ and too few at the other K, and the borderline result at $K = 256$ becomes no detectable difference. For Mistral 7B Instruct, the L40S gap at $K = 256$ just clears 2 standard errors.



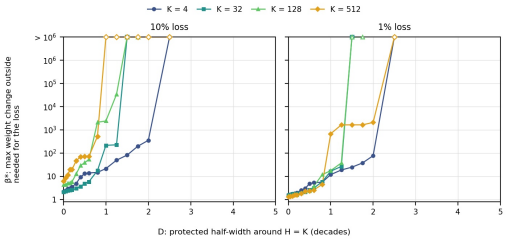
(a) Mistral 7B Instruct, $n = 359$



(b) GPT OSS 20B, $n = 766$

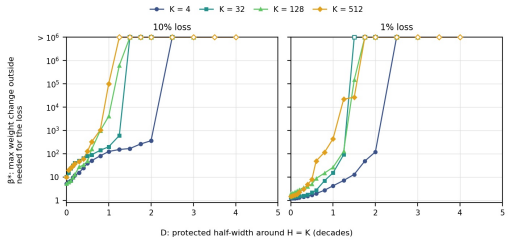


(c) Qwen3-8B, $n = 677$

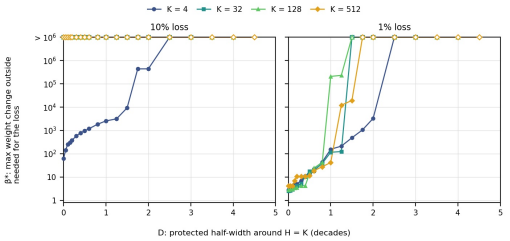


(d) Qwen3-0.6B, $n = 86$

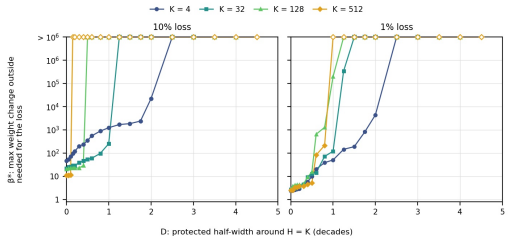
Figure 29: β^* against D on LiveCodeBench, at a 10% loss (left in each panel) and a 1% loss (right), with one curve per K. n is the number of included questions.



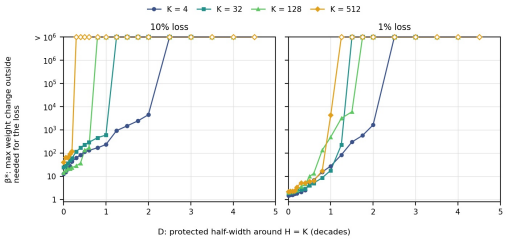
(a) Mistral 7B Instruct, $n = 455$



(b) GPT OSS 20B, $n = 500$

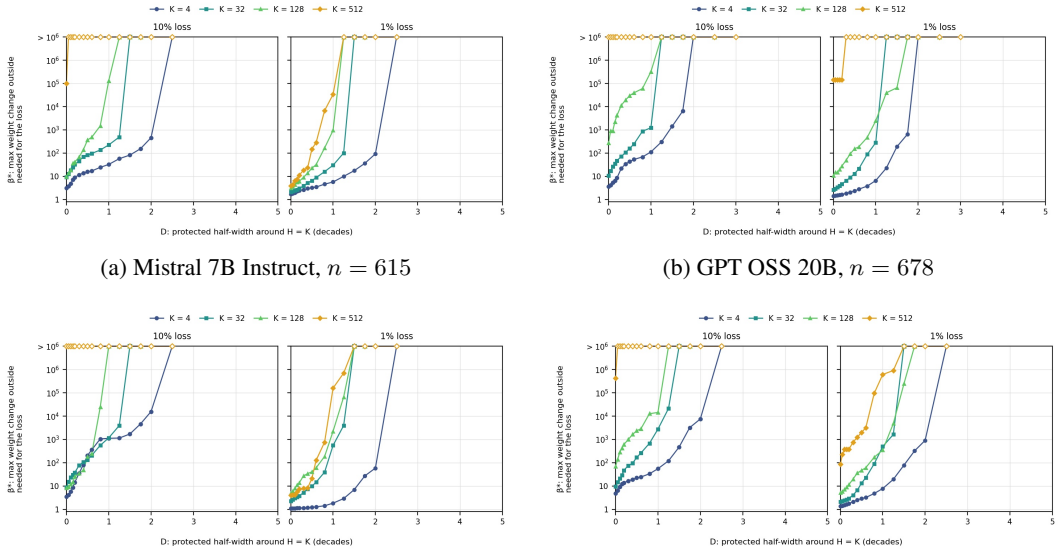


(c) Qwen3-8B, $n = 496$



(d) Qwen3-0.6B, $n = 492$

Figure 30: β^* against D on MATH-500, drawn as in Figure 29. n is the number of included questions. For GPT OSS 20B at a 10% loss, the curves of $K = 32$, 128, and 512 have no reachable loss with the fitted pass rates and lie on the top row.



(a) Mistral 7B Instruct, $n = 615$

(b) GPT OSS 20B, $n = 678$

(c) Qwen3-8B, $n = 466$

(d) Qwen3-0.6B, $n = 743$

Figure 31: β^* against D on Hi-ToM, drawn as in Figure 29. n is the number of included questions. For GPT OSS 20B at a 10% loss, with pass@K computed exactly from the fitted pass rates, the curve of $K = 512$ needs $\beta > 10^{12}$ at every D and lies on the top row.

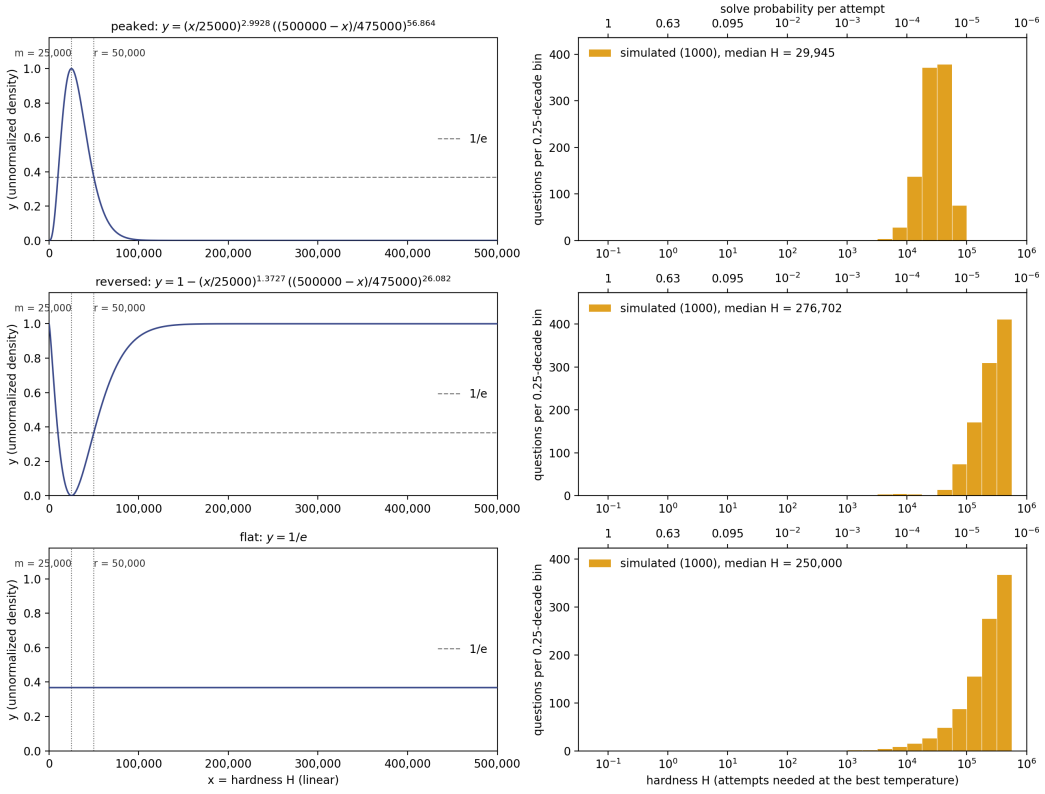
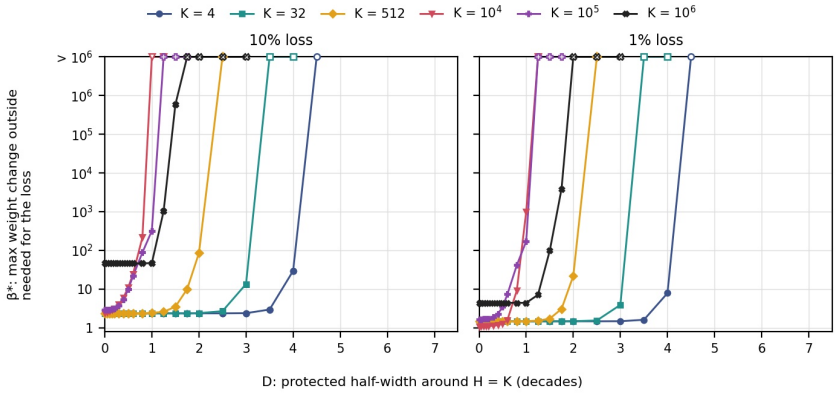
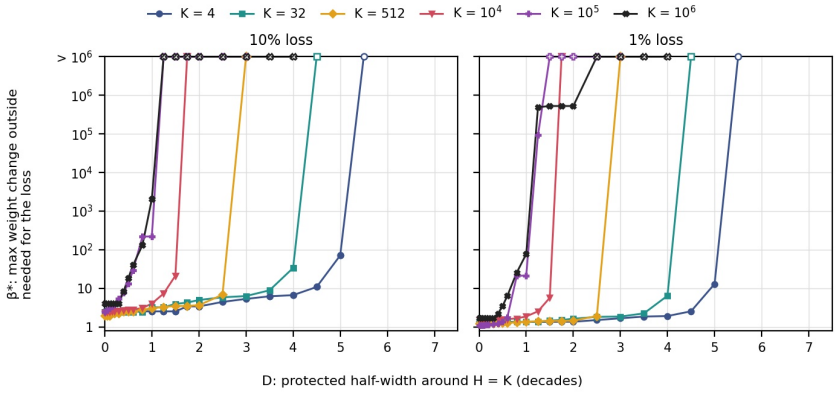


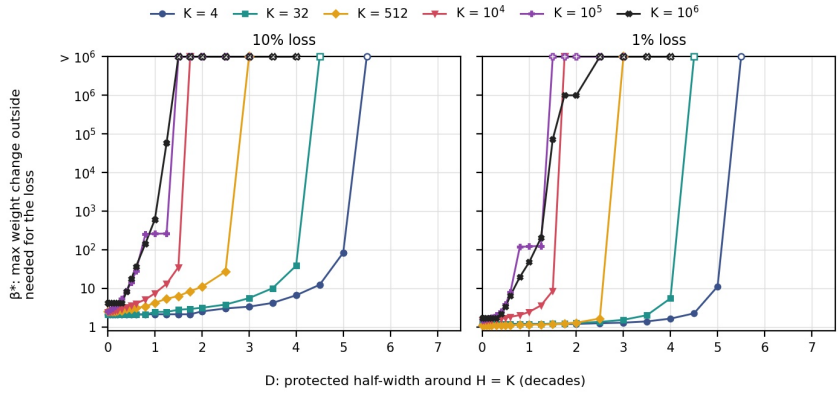
Figure 32: The three simulated hardness shapes, peaked (top), reversed (middle), and flat (bottom).



(a) Peaked



(b) Reversed



(c) Flat

Figure 33: β^* against D for the three simulated shapes, at a 10% loss (left in each panel) and a 1% loss (right), with one curve per K, drawn as in Figure 29. Each shape has 1000 simulated questions, with hardness up to a cutoff of $500{,}000$ and temperature profiles from Mistral 7B Instruct.

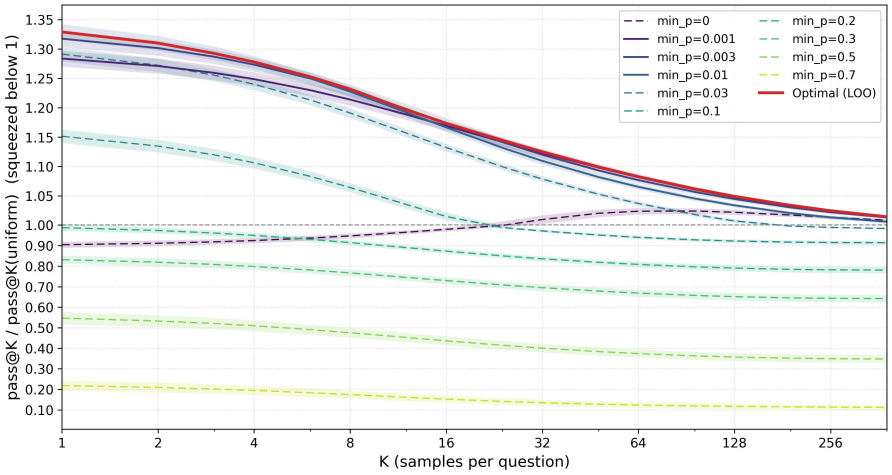


Figure 34: Ratio to uniform pass@K, $\mathrm{rtu}(K)$, of Qwen3-0.6B on Hi-ToM at $\tau = 1.5$, defined as in Figure 14 with `min_p` in place of the temperature. Colored curves are single-`min_p` allocations, each with its paired-bootstrap $\pm 1\sigma$ band. Solid curves show the values that are the dominant choice in the optimal allocation at some K, namely 0.001, 0.003, and 0.01, and dashed curves show the other values. The red trace is the LOO optimum. The population is the 648 kept questions, with $\nu = 500$, and the vertical axis is compressed below 1.

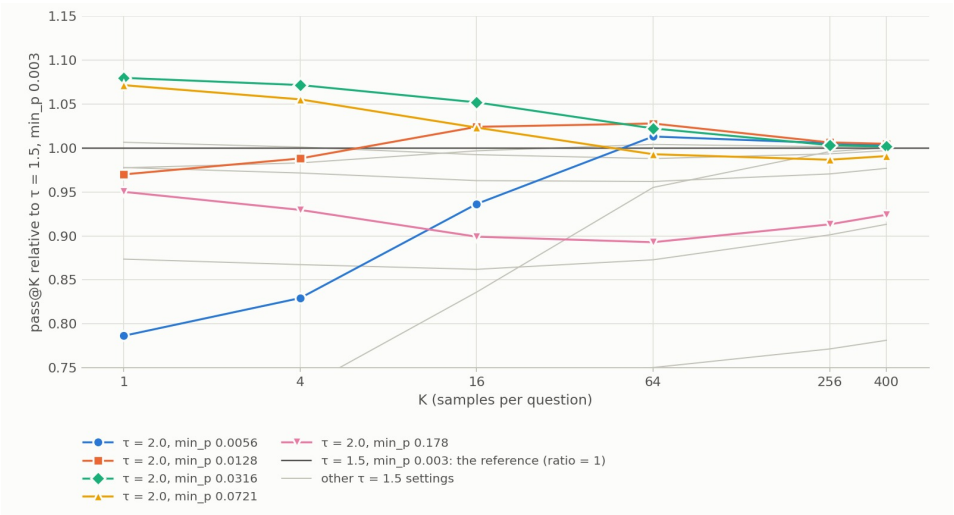


Figure 35: Ratio of pass@K to that of $\tau = 1.5$ with `min_p` $= 0.003$, for Qwen3-0.6B on Hi-ToM. Colored curves are the 5 values at $\tau = 2.0$, and the dark line at 1 is the reference. Faint curves are the other $\tau = 1.5$ values. Those of `min_p` $= 0$ and 0.2 are cut at the lower edge, and those of 0.3, 0.5, and 0.7 lie entirely below the plotted range, so 6 of the 9 are visible. Points are ratios of two means over the 648 questions, with 400 second-stage samples and no error bars.

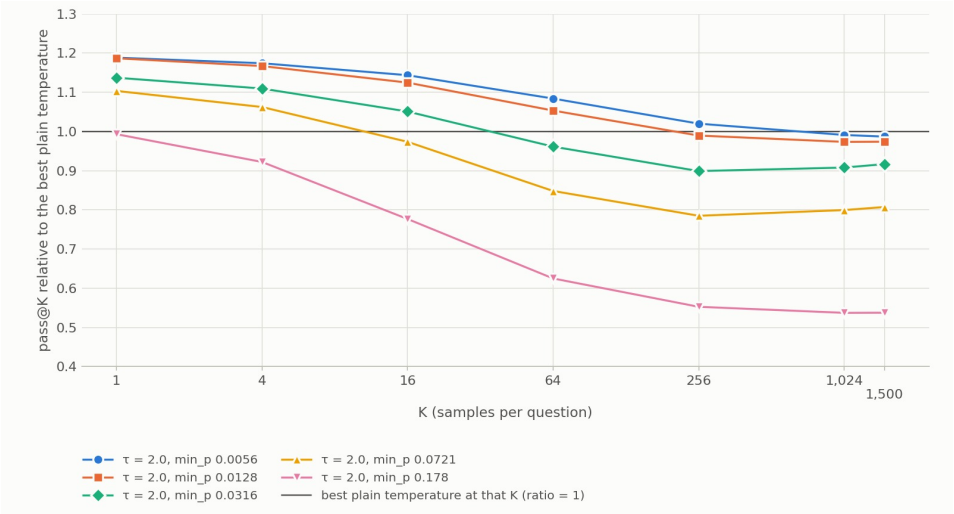
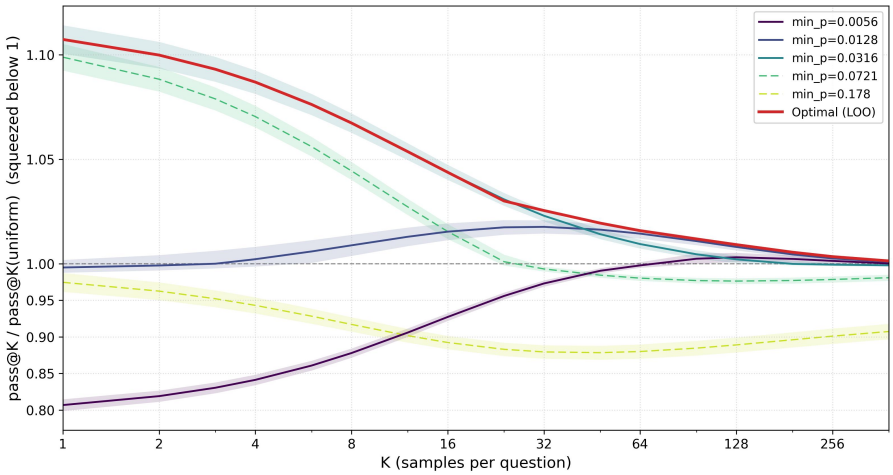
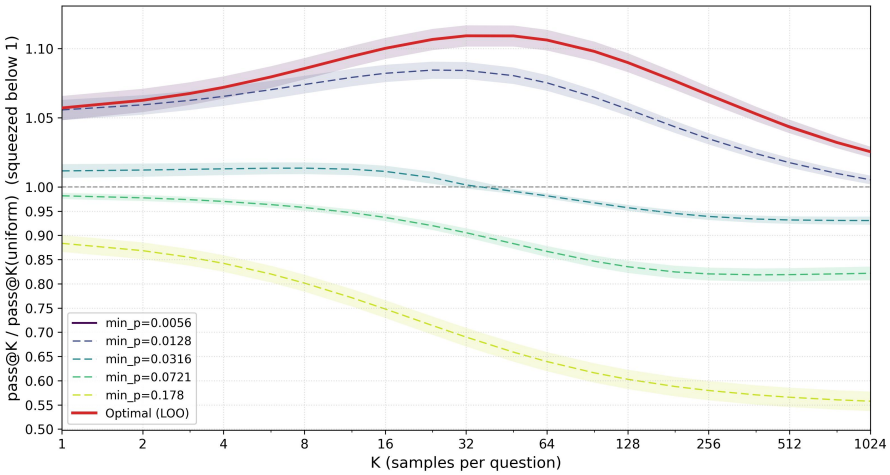


Figure 36: Ratio of pass@K with `min_p` at $\tau = 2.0$ to that of the best plain temperature of Appendix F.2, for Mistral 7B Instruct on Hi-ToM. Colored curves are the 5 values, and the dark line at 1 is the best plain temperature at each K, picked on the same questions. Points are ratios of two means over the 617 questions, with $\nu = 1500$ and no error bars.

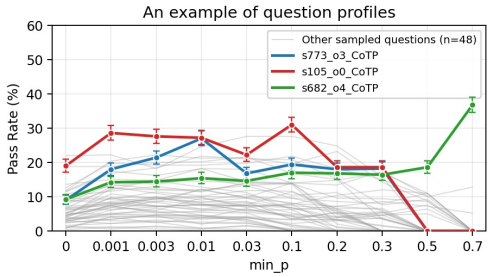


(a) Qwen3-0.6B, $n = 648$, $\nu = 400$

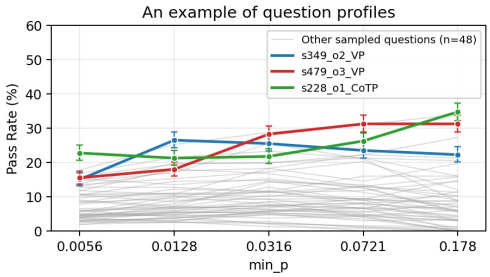


(b) Mistral 7B Instruct, $n = 598$, $\nu = 1500$

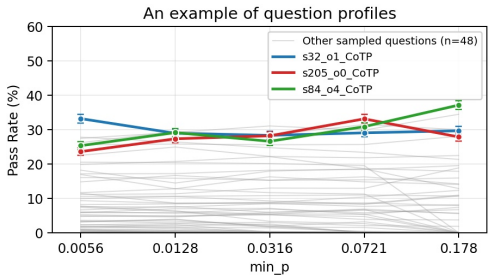
Figure 37: Ratio to uniform pass@K, $\mathrm{rtu}(K)$, on Hi-ToM at $\tau = 2.0$, defined as in Figure 34 and computed from measured pass rates. n is the number of included questions. Colored curves are single-`min_p` allocations, each with its paired-bootstrap $\pm 1\sigma$ band. Solid curves show the values that are the dominant choice in the optimal allocation at some K, dashed curves the others, and the red trace is the LOO optimum. K runs to 384 for Qwen3-0.6B and to 1024 for Mistral 7B Instruct. For Mistral 7B Instruct, 19 questions have a best pass rate below the 0.1% inclusion floor of the paper's analysis, 8 of them with no correct sample, and the curve of 0.0056 lies under the red trace. The vertical axis is compressed below 1.



(a) Qwen3-0.6B, $\tau = 1.5$

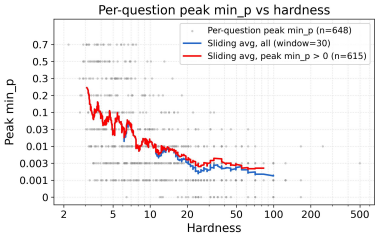


(b) Qwen3-0.6B, $\tau = 2.0$

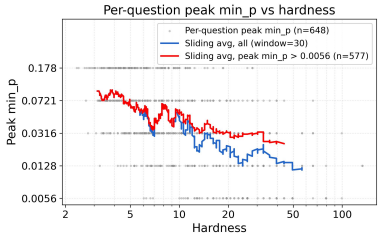


(c) Mistral 7B Instruct, $\tau = 2.0$

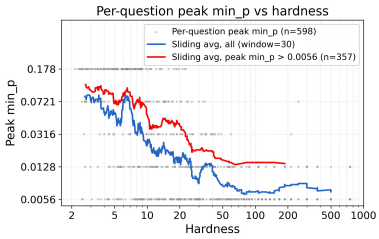
Figure 38: `min_p` profiles of 51 Hi-ToM questions in each study, the counterpart of Figure 1. Lines join measured pass rates, with no fit, and grid values are spaced equally on the horizontal axis. In each third of the grid, with 4, 3, and 3 values at $\tau = 1.5$ and 2, 2, and 1 at $\tau = 2.0$, we highlight, among the questions that peak there, the one with the most correct samples at its peak, with the peak at the smallest `min_p` among tied values and ties between questions broken by id. Bars are exact Clopper-Pearson 1σ intervals, and the vertical axis stops at 60%. The 48 grey questions are drawn at random, with a fixed seed, from the other questions of the study.



(a) Qwen3-0.6B, $\tau = 1.5$

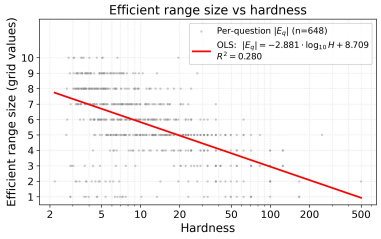


(b) Qwen3-0.6B, $\tau = 2.0$

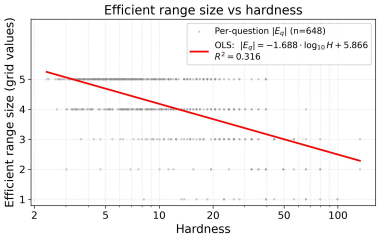


(c) Mistral 7B Instruct, $\tau = 2.0$

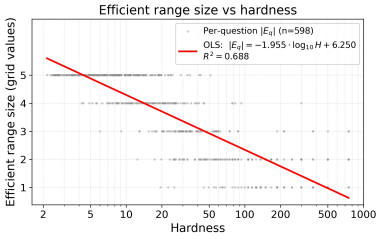
Figure 39: Per-question peak `min_p` against hardness, with sliding averages over all questions (blue) and over the questions whose peak lies above the smallest grid value (red), the counterpart of Figure 5. The peak is the `min_p` of the largest measured pass rate, the smallest among tied values. Grid values are spaced equally on the vertical axis, so the averages are means of grid position. The window is 30 questions, and the averages are drawn only where the window is full. No further low-hardness points are dropped, since H_q takes discrete values here.



(a) Qwen3-0.6B, $\tau = 1.5$

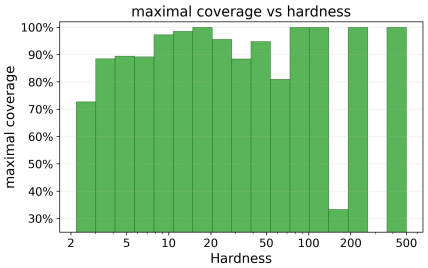


(b) Qwen3-0.6B, $\tau = 2.0$

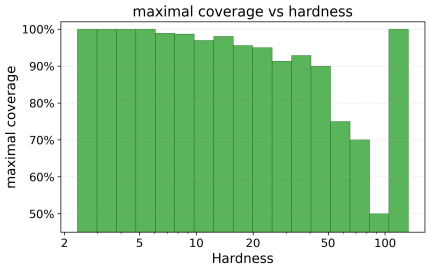


(c) Mistral 7B Instruct, $\tau = 2.0$

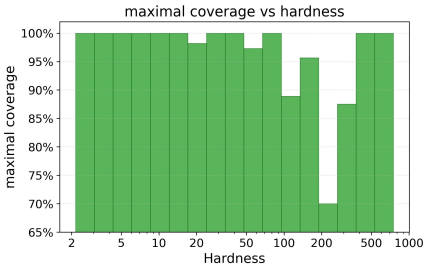
Figure 40: Per-question efficient range size against hardness, with the OLS fit on $\log_{10} H_q$, the counterpart of Figure 8. Here $|E_q|$ is the number of grid values in the efficient range, since the 10 and 5 values of these grids do not support the interpolated $t_R - t_L$.



(a) Qwen3-0.6B, $\tau = 1.5$



(b) Qwen3-0.6B, $\tau = 2.0$



(c) Mistral 7B Instruct, $\tau = 2.0$

Figure 41: Maximal coverage against hardness, the counterpart of Figure 11. Questions are placed in 17 log-uniform hardness bins, and the bar height is the largest share of the in-bin questions whose efficient ranges E_q hold one grid value. Empty bins have no bar, and the bins wholly above $H = 100$ hold 9, 2, and 68 questions in the three studies.



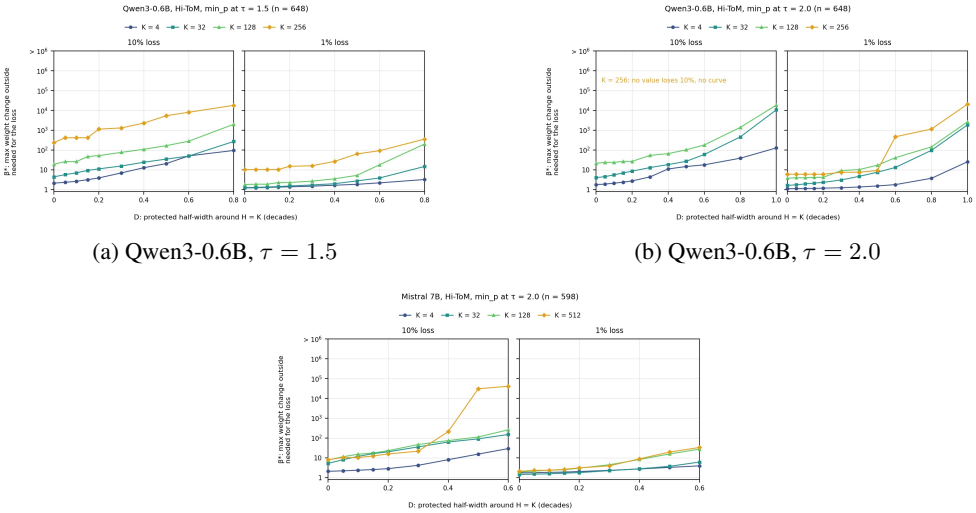
Figure 42: Distribution of $\mathrm{score}_B/\mathrm{score}_A$ as a function of $\mathrm{KL}(W \,\|\, \mathrm{uniform})$ for Qwen3-0.6B on Hi-ToM at $\tau = 1.5$ (648 included questions, $\nu = 500$, $K \in \{4, 32, 128, 256\}$, entropy ceiling $\ln(648) \approx 6.47$ nats), constructed as in Figure 17 with `min_p` in place of the temperature.



Figure 43: Distribution of $\mathrm{score}_B/\mathrm{score}_A$ as a function of $\mathrm{KL}(W \,\|\, \mathrm{uniform})$ for Qwen3-0.6B on Hi-ToM at $\tau = 2.0$ (648 included questions, $\nu = 400$, $K \in \{4, 32, 128, 256\}$, entropy ceiling $\ln(648) \approx 6.47$ nats), constructed as in Figure 17 with `min_p` in place of the temperature.



Figure 44: Distribution of $\text{score}_B/\text{score}_A$ as a function of $\mathrm{KL}(W \,\|\, \text{uniform})$ for Mistral 7B Instruct on Hi-ToM at $\tau = 2.0$ (598 included questions, $\nu = 1500$, $K \in \{4, 32, 128, 512\}$, entropy ceiling $\ln(598) \approx 6.39$ nats), constructed as in Figure 17 with `min_p` in place of the temperature.



(a) Qwen3-0.6B, $\tau = 1.5$

(b) Qwen3-0.6B, $\tau = 2.0$

(c) Mistral 7B Instruct, $\tau = 2.0$

Figure 45: β^* against D on Hi-ToM with `min_p` in place of the temperature, drawn as in Figure 29, for $K \in \{4, 32, 128, 256\}$ for Qwen3-0.6B and $K \in \{4, 32, 128, 512\}$ for Mistral 7B Instruct. Each panel shows D up to the last grid value before any of its curves exceeds 10^6: 0.8, 1.0, and 0.6 in (a), (b), and (c). Hardness is measured at each question's best value, and pass@K is the unbiased estimator of the sample counts. For Qwen3-0.6B at $\tau = 2.0$ and a 10% loss, no value loses 10% at $K = 256$, so C is empty and the panel has no curve for it. Every other curve has a reachable loss at $D = 0$.

J Expanded Proofs

In this section, we present the expanded proofs of several results from earlier sections and appendices. We begin with the proofs corresponding to Appendix B.1, which are included here to avoid unnecessary clutter.

J.1 Proofs of the Ideal Sampling Method Approximation Theorems

We start with the proof of Theorem B.1.

Proof - Theorem B.1. Let $\mathcal{Q}_{\text{good}}$ denote the set of questions for which Equations 20 and 21 hold. By the assumption,

$$\Pr_{q\sim\mathcal{Q}}\left(q \in \mathcal{Q}_{\text{good}}\right) \geq 1-\rho\,. \tag{54}$$

Fix arbitrary temperatures $\tau_1,\ldots,\tau_K \geq 0$. We define a corresponding set of discrete temperatures $\gamma_1,\ldots,\gamma_K \in \mathcal{G}$ as follows. If $\tau_k \leq T$, choose $\gamma_k \in \mathcal{G}$ such that $|\tau_k-\gamma_k| \leq \Delta$. This is possible by the definition of the grid in Equation 19. If $\tau_k > T$, set $\gamma_k = M\Delta$.

Now fix some $q \in \mathcal{Q}_{\text{good}}$. If $\tau_k \leq T$, then by Equation 20,

$$\bar{P}_q(\gamma_k) \leq \exp\left(L|\tau_k-\gamma_k|\right)\bar{P}_q(\tau_k) \leq \exp(L\Delta)\bar{P}_q(\tau_k)\,. \tag{55}$$

If $\tau_k > T$, then by Equation 21,

$$\bar{P}_q(\gamma_k) = \bar{P}_q(M\Delta) \leq \exp(\alpha)\bar{P}_q(\tau_k)\,. \tag{56}$$

Therefore, for every k, we have the common bound

$$\bar{P}_q(\gamma_k) \leq \exp(L\Delta+\alpha)\bar{P}_q(\tau_k)\,. \tag{57}$$

Multiplying over $k=1,\ldots,K$, we get

$$\prod_{k=1}^{K}\bar{P}_q(\gamma_k) \leq \exp\left(K(L\Delta+\alpha)\right)\prod_{k=1}^{K}\bar{P}_q(\tau_k)\,. \tag{58}$$

We can now average over $q\sim\mathcal{Q}$. Using the definition of $\bar{\mathcal{P}}$ from Equation 1, and splitting the expectation over $\mathcal{Q}_{\text{good}}$ and its complement, we obtain

$$\bar{\mathcal{P}}(\gamma_1,\ldots,\gamma_K) = \mathbb{E}_{q\sim\mathcal{Q}}\prod_{k=1}^{K}\bar{P}_q(\gamma_k)\,. \tag{59}$$

On the good set, we use the multiplicative bound above. On the complement of the good set, we use only the fact that $0 \leq \bar{P}_q(\tau) \leq 1$, and hence $\prod_{k=1}^{K}\bar{P}_q(\gamma_k) \leq 1$. Therefore,

$$\bar{\mathcal{P}}(\gamma_1,\ldots,\gamma_K) \leq \exp\left(K(L\Delta+\alpha)\right)\mathbb{E}_{q\sim\mathcal{Q}}\prod_{k=1}^{K}\bar{P}_q(\tau_k) + \Pr_{q\sim\mathcal{Q}}\left(q\notin\mathcal{Q}_{\text{good}}\right)\,. \tag{60}$$

Since $\Pr(q\notin\mathcal{Q}_{\text{good}}) \leq \rho$, this gives

$$\bar{\mathcal{P}}(\gamma_1,\ldots,\gamma_K) \leq \exp\left(K(L\Delta+\alpha)\right)\bar{\mathcal{P}}(\tau_1,\ldots,\tau_K) + \rho\,. \tag{61}$$

Since $\tau_1,\ldots,\tau_K$ were arbitrary, we may take the infimum over all continuous temperatures. Using the definitions of $\bar{\mathcal{P}}^\star$ and $\bar{\mathcal{P}}^\star_{\mathcal{G}}$ in Equation 22, we obtain

$$\bar{\mathcal{P}}^\star_{\mathcal{G}} \leq \exp\left(K(L\Delta+\alpha)\right)\bar{\mathcal{P}}^\star + \rho\,. \tag{62}$$

Rearranging gives

$$\bar{\mathcal{P}}^\star_{\mathcal{G}} - \bar{\mathcal{P}}^\star \leq \left(\exp\left(K(L\Delta+\alpha)\right)-1\right)\bar{\mathcal{P}}^\star + \rho\,. \tag{63}$$

Finally, since $\mathcal{G}\subseteq[0,\infty)$, the discrete optimization is over a restricted set of temperatures. Hence,

$$\bar{\mathcal{P}}^\star \leq \bar{\mathcal{P}}^\star_{\mathcal{G}}\,. \tag{64}$$

Combining the last two displays proves that

$$0 \leq \bar{\mathcal{P}}^\star_{\mathcal{G}} - \bar{\mathcal{P}}^\star \leq \left(e^{K(L\Delta+\alpha)}-1\right)\bar{\mathcal{P}}^\star + \rho\,, \tag{65}$$

as required. □

We can now present the proof of the other main result of this section, Theorem B.2, thereby completing all appendix proofs relevant to our analysis:

Proof - Theorem B.2. The proof follows the standard uniform convergence argument, but with a variance-sensitive bound. The main idea is to compare every grid assignment to the best grid assignment under the population objective, and then show that a bad population assignment is unlikely to look better on the sampled questions.

Let $\gamma = (\gamma_1, \ldots, \gamma_K)$ denote an arbitrary assignment of grid temperatures. For each sampled question q_i, define

$$X_i(\gamma) = \prod_{k=1}^{K} \bar{P}_{q_i}(\gamma_k) \,. \tag{66}$$

Since $0 \le \bar{P}_q(\tau) \le 1$, we have $0 \le X_i(\gamma) \le 1$. Moreover, by Equations 8 and 24,

$$\mathbb{E}_{q_i \sim \mathcal{Q}} X_i(\gamma) = \bar{\mathcal{P}}_{\mathcal{Q}}(\gamma_1, \ldots, \gamma_K), \qquad \frac{1}{n}\sum_{i=1}^{n} X_i(\gamma) = \bar{\mathcal{P}}_{\mathcal{Q}'}(\gamma_1, \ldots, \gamma_K) \,. \tag{67}$$

Let also $\gamma^\star = (\gamma_1^\star, \ldots, \gamma_K^\star)$ be an assignment minimizing the population objective over the grid. Thus, by Equation 22,

$$\bar{\mathcal{P}}_{\mathcal{Q}}(\gamma^\star) = \bar{\mathcal{P}}^\star_{\mathcal{G}} \,. \tag{68}$$

For any grid assignment γ, we can define its population excess over the best grid assignment by

$$R(\gamma) = \bar{\mathcal{P}}_{\mathcal{Q}}(\gamma) - \bar{\mathcal{P}}^\star_{\mathcal{G}} \,. \tag{69}$$

The goal is to show that the empirical minimizer cannot have a large value of $R(\gamma)$.

To do so, we compare γ directly to $\gamma^\star$. Define:

$$Y_i(\gamma) = X_i(\gamma) - X_i(\gamma^\star) \,. \tag{70}$$

Then:

$$\mathbb{E}_{q_i \sim \mathcal{Q}} Y_i(\gamma) = R(\gamma) \,. \tag{71}$$

Also, since $0 \le X_i(\gamma), X_i(\gamma^\star) \le 1$, we have:

$$Y_i(\gamma)^2 = \left(X_i(\gamma) - X_i(\gamma^\star)\right)^2 \le X_i(\gamma) + X_i(\gamma^\star) \,. \tag{72}$$

Therefore:

$$\mathrm{Var}\left(Y_i(\gamma)\right) \le \mathbb{E} Y_i(\gamma)^2 \le \bar{\mathcal{P}}_{\mathcal{Q}}(\gamma) + \bar{\mathcal{P}}_{\mathcal{Q}}(\gamma^\star) = R(\gamma) + 2\bar{\mathcal{P}}^\star_{\mathcal{G}} \,. \tag{73}$$

We can now apply Bernstein's inequality to the variables $Y_i(\gamma)$. For a fixed assignment γ, the event that γ looks at least as good as $\gamma^\star$ on the sampled questions is:

$$\bar{\mathcal{P}}_{\mathcal{Q}'}(\gamma) \le \bar{\mathcal{P}}_{\mathcal{Q}'}(\gamma^\star) \,, \tag{74}$$

or equivalently:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i(\gamma) \le 0 \,. \tag{75}$$

Since the expectation of this empirical average is $R(\gamma)$, Bernstein's inequality gives

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(\gamma) \le 0\right) \le \exp\left(-\frac{nR(\gamma)^2}{4\bar{\mathcal{P}}^\star_{\mathcal{G}} + \frac{10}{3}R(\gamma)}\right) \,. \tag{76}$$

This means that all that remains is to make this bound uniform over all possible grid assignments. Since the objective is invariant to the order of the K temperatures, it is enough to count unordered assignments. The number of multisets of size K from the $M+1$ grid points is

$$\binom{K+M}{M} \,. \tag{77}$$

Using a union bound over these assignments, and recalling the definition of Λ in Equation 27, we get that with probability at least $1-\rho'$, every assignment γ satisfying:

$$\frac{nR(\gamma)^2}{4\bar{\mathcal{P}}^\star_{\mathcal{G}} + \frac{10}{3}R(\gamma)} > \Lambda \tag{78}$$

has empirical value strictly larger than that of $\gamma^\star$.

Now apply this to the empirical minimizer $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_K)$ from Equation 25. Since $\hat{\gamma}$ minimizes the empirical objective, it satisfies:

$$\bar{\mathcal{P}}_{\mathcal{Q}'}(\hat{\gamma}) \leq \bar{\mathcal{P}}_{\mathcal{Q}'}(\gamma^\star)\ . \tag{79}$$

Therefore, on the high-probability event above, $\hat{\gamma}$ cannot satisfy the previous strict inequality. Hence:

$$\frac{nR(\hat{\gamma})^2}{4\bar{\mathcal{P}}^\star_{\mathcal{G}} + \frac{10}{3}R(\hat{\gamma})} \leq \Lambda\ . \tag{80}$$

Solving this quadratic inequality gives:

$$R(\hat{\gamma}) \leq \frac{5\Lambda}{3n} + \sqrt{\frac{4\bar{\mathcal{P}}^\star_{\mathcal{G}}\Lambda}{n} + \frac{25\Lambda^2}{9n^2}}\ . \tag{81}$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain:

$$R(\hat{\gamma}) \leq 2\sqrt{\frac{\bar{\mathcal{P}}^\star_{\mathcal{G}}\Lambda}{n}} + \frac{10\Lambda}{3n}\ . \tag{82}$$

Finally, by Equation 26, we have:

$$R(\hat{\gamma}) = \bar{\mathcal{P}}'_{\mathcal{Q}} - \bar{\mathcal{P}}^\star_{\mathcal{G}}\ . \tag{83}$$

Since $\bar{\mathcal{P}}^\star_{\mathcal{G}}$ is the minimum of the population objective over the grid, we also have:

$$0 \leq \bar{\mathcal{P}}'_{\mathcal{Q}} - \bar{\mathcal{P}}^\star_{\mathcal{G}}\ . \tag{84}$$

Combining the last two displays proves that, with probability at least $1-\rho'$,

$$0 \leq \bar{\mathcal{P}}'_{\mathcal{Q}} - \bar{\mathcal{P}}^\star_{\mathcal{G}} \leq 2\sqrt{\frac{\bar{\mathcal{P}}^\star_{\mathcal{G}}\Lambda}{n}} + \frac{10\Lambda}{3n}\ . \tag{85}$$

Which complete our proof. □

J.2 Lemma B.3 Proof

We will now prove Lemma B.3. The argument is a direct application of a standard idea from convex geometry. The key step is to move beyond viewing a temperature as a scalar and instead represent it by the vector it induces over the sampled questions.

For each candidate temperature τ_j, define its log-failure vector across the n questions by

$$a_j = \big(\log \bar{P}_{1j}, \ldots, \log \bar{P}_{nj}\big) \in \mathbb{R}^n. \tag{86}$$

Thus, for the finite optimization problem, each temperature corresponds to a point in $\mathbb{R}^n$, where the i-th coordinate records its log-failure probability on the i-th question.

Now consider a relaxed K-attempt strategy. Such a strategy assigns a total mass of K across the candidate temperatures. Equivalently, it produces a total log-failure profile z, obtained by averaging the temperature profiles and scaling by K. The set of all achievable vectors z is therefore exactly the scaled convex hull

$$K \cdot \mathrm{conv}\{a_1, \ldots, a_l\}. \tag{87}$$

This is the geometric core of the argument: the optimization is no longer over temperatures, but over points z obtained by mixing their log-failure profiles.

The objective depends on the strategy only through z. For $z = (z_1, \ldots, z_n)$, the average failure probability is

$$F(z) = \frac{1}{n} \sum_{i=1}^{n} \exp(z_i). \tag{88}$$

Hence, the relaxed problem becomes

$$\min_{z \in K \cdot \mathrm{conv}\{a_1, \ldots, a_l\}} F(z). \tag{89}$$

Let z^* be an optimal solution. Since F is convex and differentiable, first-order optimality implies that the gradient of F at z^* defines a supporting hyperplane to the feasible polytope. Equivalently, z^* lies on an exposed face of the polytope; see, for example, [56].

Moreover, the gradient of F is strictly positive in every coordinate and thus nonzero. Consequently, the supporting hyperplane is nondegenerate, and the exposed face containing z^* has dimension at most $n - 1$.

We now apply Carathéodory's theorem within this face [57, 58]. While the standard theorem guarantees a representation using at most $d + 1$ points in $\mathbb{R}^d$, here we operate within a set of dimension at most $n - 1$. Therefore, z^* can be expressed as a convex combination of at most n of the vectors $\{a_j\}$.

That is, there exist indices $j_1, \ldots, j_m$ with $m \leq n$, and nonnegative weights $\lambda_1, \ldots, \lambda_m$ satisfying

$$\sum_{r=1}^{m} \lambda_r = 1, \tag{90}$$

such that

$$z^* = \sum_{r=1}^{m} \lambda_r K a_{j_r}. \tag{91}$$

Translating this representation back into the language of temperature strategies, we obtain an alternative optimal strategy that achieves the same log-failure profile z^* using only the temperatures $\tau_{j_1}, \ldots, \tau_{j_m}$. Since the objective depends only on z^*, this strategy attains the same optimal value.

Therefore, there always exists an optimal strategy that uses at most n distinct temperatures, where n is the number of sampled questions.

This completes the proof.

J.3 Proofs of the Short-Sampling Effect Propositions

In this section we present the proofs corresponding to Appendix B.2.2, which are included here to avoid unnecessary clutter.

We start with the proof of the first lemma of the Appendix.

Proof - Proposition B.1. The proof is again only an algebraic extension of the symmetric two-question case. The only difference is that the two groups of questions no longer have the same weight. A fraction r of the questions behaves like the first question, and a fraction $1 - r$ behaves like the second question.

As before, using the compromise temperature assigns all K samples to the same temperature. Therefore, its failure probability remains exactly the same as in the unweighted case:

$$\bar{\mathcal{P}}_{\mathrm{comp}} = (1 - P_C)^K \ . \tag{92}$$

Now consider the two-expert strategy, where we allocate K' samples to the first expert temperature and $K - K'$ samples to the second expert temperature.

Since a fraction r of the questions belongs to the first group and a fraction $1 - r$ belongs to the second group, the weighted average failure probability is

$$\bar{\mathcal{P}}_{\mathrm{expert}}(K') = r \bar{P}_E^{K'} + (1 - r) \bar{P}_E^{K - K'} \ . \tag{93}$$

Then we can optimize this expression over K' in the continuous relaxation. The minimum is obtained when the two weighted contributions in Equation 93 are balanced. Indeed, differentiating with respect to K' gives

$$r\bar{P}_E^{K'} = (1-r)\bar{P}_E^{K-K'} \ . \tag{94}$$

This equation has a simple interpretation. At the optimum, neither side should dominate the total failure. If one group contributed much more to the weighted failure than the other, we could move a small amount of sampling mass toward that group and improve the strategy.

Substituting the balancing condition from Equation 94 back into Equation 93, we obtain

$$\bar{\mathcal{P}}^{\star}_{\text{expert}} = 2\sqrt{r(1-r)}\bar{P}_E^{\frac{K}{2}} \ . \tag{95}$$

This is the key point. In the original symmetric proposition, the expert failure term is $\bar{P}_E^{\frac{K}{2}}$. In the weighted setting, after optimizing the allocation between the two expert temperatures, the same term is multiplied by $2\sqrt{r(1-r)}$. Hence the weighted problem is equivalent to the original one after replacing $\bar{P}_E$ by an effective failure probability $\widetilde{\bar{P}}_E$ satisfying:

$$\widetilde{\bar{P}}_E^{\frac{K}{2}} = 2\sqrt{r(1-r)}\bar{P}_E^{\frac{K}{2}} \ . \tag{96}$$

Taking both sides to the power $\frac{2}{K}$ gives:

$$\widetilde{\bar{P}}_E = \sqrt[K]{4r(1-r)}\bar{P}_E \ . \tag{97}$$

This proves the exact failure-probability substitution in Equation 35.

It remains to translate the same statement into the first-order success-probability language. Using the standard low-probability approximation,

$$\bar{P}_E^{\frac{K}{2}} = (1-P_E)^{\frac{K}{2}} = e^{-P_E\frac{K}{2}}\left(1+O(KP_E^2)\right) \ , \tag{98}$$

Equation 95 becomes

$$\bar{\mathcal{P}}^{\star}_{\text{expert}} = 2\sqrt{r(1-r)}e^{-P_E\frac{K}{2}}\left(1+O(KP_E^2)\right) \ . \tag{99}$$

Since:

$$2\sqrt{r(1-r)} = e^{\frac{1}{2}\log(4r(1-r))} \ , \tag{100}$$

we may rewrite this as:

$$\bar{\mathcal{P}}^{\star}_{\text{expert}} = \exp\left(-\frac{K}{2}\left(P_E - \frac{1}{K}\log\left(4r(1-r)\right)\right)\right)\left(1+O(KP_E^2)\right) \ . \tag{101}$$

Thus, to first order, the weighted case has exactly the same form as Proposition 4.1, but with the effective substitution

$$P_E \mapsto P_E - \frac{1}{K}\log\left(4r(1-r)\right) \ . \tag{102}$$

This is precisely Equation 34.

Therefore, optimizing the allocation between the two expert temperatures reduces the weighted setting to the same short-sampling comparison as before, with the modified effective expert strength.

This proves the proposition.

□

Secondly, we will now present the two gaussians case proof, as the way this proof works will also be generalized to other more complicated cases.

Proof - Proposition B.3. The proof is again a direct comparison between the compromise temperature and the best possible two-temperature strategy. The only point that requires care is that the two expert temperatures should not be assumed to lie exactly at the Gaussian peaks. Instead, we allow them to move, and ask when moving away from the compromise temperature actually improves the objective.

By symmetry, we place the compromise temperature at 0, and write the two Gaussian peaks as $-a$ and a. Thus,

$$P_1(\tau) = A \exp\left(-\frac{(\tau + a)^2}{2\sigma^2}\right), \qquad P_2(\tau) = A \exp\left(-\frac{(\tau - a)^2}{2\sigma^2}\right) . \tag{103}$$

The peak probability is therefore:

$$P_E = A , \tag{104}$$

while the compromise probability is the value of either Gaussian at the intersection point:

$$P_C = A \exp\left(-\frac{a^2}{2\sigma^2}\right) . \tag{105}$$

Hence:

$$\frac{P_E}{P_C} = \exp\left(\frac{a^2}{2\sigma^2}\right) . \tag{106}$$

We now allow the two expert temperatures to move symmetrically around the compromise temperature. Write them as $-x$ and x. The case $x = 0$ corresponds to using the compromise temperature twice. Therefore, the two-expert strategy outperforms the compromise temperature exactly when some $x > 0$ gives a larger first-order success probability than $x = 0$.

For a fixed value of x, each question receives half of the samples at one temperature and half at the other. By symmetry, the effective first-order success probability is

$$S(x) = \frac{A}{2}\left(\exp\left(-\frac{(a - x)^2}{2\sigma^2}\right) + \exp\left(-\frac{(a + x)^2}{2\sigma^2}\right)\right) . \tag{107}$$

At the compromise temperature, this becomes:

$$S(0) = A \exp\left(-\frac{a^2}{2\sigma^2}\right) = P_C . \tag{108}$$

We now compare $S(x)$ with $S(0)$. Factoring out the compromise value gives

$$\frac{S(x)}{S(0)} = \exp\left(-\frac{x^2}{2\sigma^2}\right) \cosh\left(\frac{ax}{\sigma^2}\right) . \tag{109}$$

First suppose that $a \leq \sigma$. Using the standard inequality

$$\cosh(u) \leq \exp\left(\frac{u^2}{2}\right) , \tag{110}$$

Equation 109 gives

$$\frac{S(x)}{S(0)} \leq \exp\left(-\frac{x^2}{2\sigma^2} + \frac{a^2 x^2}{2\sigma^4}\right) = \exp\left(-\frac{x^2}{2\sigma^2}\left(1 - \frac{a^2}{\sigma^2}\right)\right) \leq 1 . \tag{111}$$

Thus, when $a \leq \sigma$, moving away from the compromise temperature cannot improve the objective. In this case, the best two-temperature strategy collapses to the compromise temperature.

Now suppose that $a > \sigma$. Expanding Equation 109 around $x = 0$, we obtain:

$$\frac{S(x)}{S(0)} = 1 + \frac{x^2}{2\sigma^2}\left(\frac{a^2}{\sigma^2} - 1\right) + O(x^4) . \tag{112}$$

Since $a > \sigma$, the coefficient of x^2 is positive. Therefore, for sufficiently small $x > 0$, we have:

$$S(x) > S(0) . \tag{113}$$

Hence, the optimized two-expert strategy strictly outperforms the compromise temperature exactly when:

$$a > \sigma . \tag{114}$$

It remains only to translate this condition into the ratio between the peak probability P_E and the compromise probability P_C. By Equation 106, the condition $a > \sigma$ is equivalent to:

$$\frac{P_E}{P_C} = \exp\left(\frac{a^2}{2\sigma^2}\right) > e^{1/2} = \sqrt{e} . \tag{115}$$

Equivalently,

$$P_E > \sqrt{e} P_C \simeq 1.65 P_C . \tag{116}$$

This proves the proposition. □

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 main claims in the abstract and introduction are aligned with the actual contributions: profile structure, hardness, efficient-temperature overlap, the ideal multi-temperature allocation, and the dulling effects are all developed in the main text and supported in the appendix.

   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: The paper includes a dedicated limitations discussion in Appendix A. It discusses empirical coverage, the gap between the theoretical assumptions and full real-world generality, the partial nature of the short-sampling theory, and the need for broader validation.

   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 formal results state their assumptions and are accompanied by proofs or derivations in the Appendix.

   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: [Yes]

   Justification: The paper provides the information needed to reproduce the main empirical conclusions: the pass@K formulation is defined in Section 2.2, the models and LiveCodeBench V1-V6 setup are stated in Section 2.1, and the appendix gives the temperature grids, sample counts, filtering rule, fitting procedure, optimization formulation, and pass@K tables. For the temperature-allocation comparison on MATH-500, Appendix C gives the dataset source, prompts, temperature grids, sample counts, token limits, and grading rules. For the comparison on Hi-ToM, Appendix C gives the same information, together with the first stage that selects each model's questions. Our code, at `https://github.com/avivorly/one-temperature`, implements the sampling, grading, pass@K estimation and temperature-allocation steps; see Appendix C.

   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: [Yes]

   Justification: The code is available at `https://github.com/avivorly/one-temperature`. It covers sampling with vLLM over a temperature grid, grading, the unbiased pass@K estimator, and the temperature-allocation methods with held-out evaluation, and its README gives the exact commands. All benchmarks and models are public; their sources, versions and licenses are listed under the licenses item.

   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: [Yes]

   Justification: The experimental setting is described in the main text and in Appendix C and Appendix D. The MATH-500 and Hi-ToM settings are in Appendix C. The `min_p` setting and results are in Appendix I.

   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: [Yes]

   Justification: The paper reports all of the uncertainty information and explains how it is computed.

   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: [Yes]

   Justification: Appendix C names the GPU types (NVIDIA L40S, A40, GeForce RTX 5090, and GeForce RTX 4090) and the `vLLM` inference engine, and gives the number of samples per question and temperature for each model and dataset, the GPU memory, and about 14,000 GPU-hours in total. The `min_p` sampling of Appendix I ran on NVIDIA L40S, GeForce RTX 5090, and GeForce RTX 4090 GPUs with `vLLM` 0.19.0, as reported there.

   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: The work studies existing models on three public benchmarks (LiveCodeBench, MATH-500, and Hi-ToM) and does not introduce human-subject experiments, private data, or deployment decisions about individuals.

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: [N/A]

Justification: This is foundational work: it analyses how the sampling temperature affects pass@K for existing public models and benchmarks, and it releases no new model, dataset or capability. Its practical effect is limited to choosing sampling temperatures more efficiently, which can lower the inference compute needed for a given pass@K. We see no direct negative societal impact beyond those of the models it studies.

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: Our work does not involve releasing any models or datasets that pose a risk of misuse.

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: [Yes]

    Justification: Each asset is cited and named with its creator, version, URL and license, and we use each only for research evaluation under that license. Models, all Apache-2.0 per their Hugging Face model cards: Mistral AI's Mistral-7B-Instruct-v0.3 [59], `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`; the Qwen team's Qwen3-8B and Qwen3-0.6B [60], `https://huggingface.co/Qwen/Qwen3-8B` and `https://huggingface.co/Qwen/Qwen3-0.6B`; and OpenAI's gpt-oss-20b [61], `https://huggingface.co/openai/gpt-oss-20b`. Benchmarks: LiveCodeBench [49], release_v6 (V1–V6) of `https://huggingface.co/datasets/livecodebench/code_generation_lite` (dataset card: `cc`), with code at `https://github.com/LiveCodeBench/LiveCodeBench` (MIT); MATH-500 [11, 31], `https://huggingface.co/datasets/HuggingFaceH4/MATH-500` at revision `6e4ed1a` (no license stated), which is the PRM800K test split of MATH, from `https://github.com/openai/prm800k` and `https://github.com/hendrycks/math` (both MIT); and Hi-ToM [50], the 1,200-question `Hi-ToM_data.json` of `https://github.com/ying-hui-he/Hi-ToM_dataset` (Apache-2.0). Inference engine: vLLM [62], `https://github.com/vllm-project/vllm` (Apache-2.0), version 0.19.0 for the runs of Appendix I. We scraped no data: LiveCodeBench states that it collected its problems from LeetCode, AtCoder and Codeforces, and we used its released files. We re-package and release none of these assets.

    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` 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: The paper does not introduce a new dataset.

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 research does not involve crowdsourcing experiments 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: The research does not involve crowdsourcing, study participants, or other human-subject research.

    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: While LLMs were used in preparing this work, their role was limited to editing, assisting with technical proofs, and supporting non-fundamental coding tasks. They were not used to formulate the research ideas, develop the main contributions, or discover the theorems. The final technical content and claims were produced, reviewed, and verified by the authors.

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.