CIS 6200: Emerging Theory of LLMs **Fall 2026**

Topic 5A: Sampling and Selection with Imperfect Rewards

Last updated: October 15, 2026 *Surbhi Goel*

Acknowledgements: *These notes draw on Huang et al. [6] for coverage and selection with imperfect scores, Schaeffer et al. [4] for scaling across prompts, and Chen et al. [7] for the coverage principle. The selection analysis is adapted here to mean absolute scoring error; the top-k guarantee is derived under the same coverage assumption.*

AI Disclosure: *These notes were drafted and revised with AI assistance. If you find an error, please let me know.*

Given a prompt, we can sample several responses from a language model, score them, and return the response with the highest score. This can improve accuracy without changing the model. But an incorrect response may receive a high score, and more samples can make such responses more likely to appear. How many responses should we sample, and how should we select among them when the score is imperfect?

The answer depends on the probability of generating correct responses and how selection amplifies scoring error.

1 What repeated sampling can achieve

Fix a prompt x and let $\mathcal{U}$ be its set of complete responses, including any chain of thought. A finite vocabulary and bounded response length make $\mathcal{U}$ finite. Write $q_0(y) = q_0(y \mid x)$ for the generator, suppressing x.

A **reward** $r(y) \in \{0, 1\}$ records correctness: it is 1 for a correct response and 0 otherwise. Let $G = \{y : r(y) = 1\}$ be the set of correct responses. For a response distribution q, its accuracy is

$$J(q) = \mathbb{E}_{Y \sim q}[r(Y)] = q(G).$$

Thus $1 - J(q)$ is the probability of returning an incorrect response.

The algorithm observes a **proxy score** $\widehat{r}(y) \in [0, 1]$, which may come from a learned reward model or an automated evaluator. It estimates correctness but can disagree with the true reward $r(y)$. Throughout the selection analysis, the generator and scorer remain fixed.

Best-of-N. Draw N independent responses and return one with largest score:

$$Y^{(1)}, \ldots, Y^{(N)} \overset{\text{iid}}{\sim} q_0, \qquad \widehat{Y} \in \operatorname*{arg\,max}_{y \in \{Y^{(1)}, \ldots, Y^{(N)}\}} \widehat{r}(y). \tag{1}$$

Ties are broken using a fixed ordering of responses. Write $q_N(y) = \mathbb{P}(\widehat{Y} = y)$ for the distribution of the returned response; $q_1 = q_0$. We study how $J(q_N)$ changes with N.

Experimental motivation. Cobbe et al. [1] showed that, with enough verifier training data, selecting the highest-scoring of 100 generated mathematics solutions improved on a single response. Figure 1 shows how accuracy changes with N for a fixed generator and a scorer that reads the complete solution and is trained to predict final-answer correctness [2].

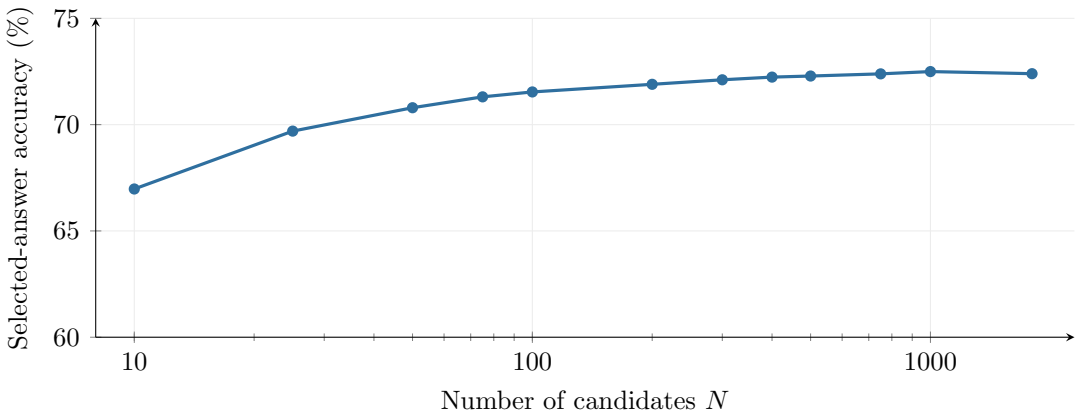

Figure 1: Best-of-N on 500 held-out MATH problems, using a fixed GPT-4-based generator and a scorer trained on final-answer correctness. Replotted from Lightman et al. [2, Figure 3]; other curves and subsampling variability bands are omitted.

1.1 Scaling across prompts

In Figure 1, accuracy initially rises and then changes little as N grows. More samples help only if they produce a correct answer on a previously unsolved prompt and the scorer selects it. To understand the limits of sampling, first suppose the scorer recognizes correct answers perfectly.

One prompt. Suppose $\widehat{r} = r$. If one attempt succeeds with probability $\alpha = q_0(G)$, all N attempts fail with probability $(1-\alpha)^N$. Thus

$$J(q_N) = \mathbb{P}(\widehat{Y} \in G) = 1 - (1-\alpha)^N \geq 1 - e^{-N\alpha}. \tag{2}$$

For $\alpha > 0$, a constant success probability requires on the order of $1/\alpha$ attempts. If $\alpha = 0$, sampling cannot produce a correct response. All logarithms below are natural.

Averaging across prompts. Let α_x be the probability that one response to prompt x is correct, and draw $X \sim \mu$ from the prompt distribution. Define *pass@N* as the probability that at least one of N attempts is correct, averaged over prompts. With a perfect scorer, this is also the selected-answer accuracy. The average failure probability is

$$F_N := 1 - \text{pass@}N = \mathbb{E}_X[(1-\alpha_X)^N]. \tag{3}$$

Prompts with large α_x are likely to be solved after few attempts. The remaining failures increasingly come from prompts with small α_x, where many more samples may be needed. Thus accuracy can appear to level off even with a perfect scorer.

Brown et al. [3] observed approximate power-law scaling of $-\log(\text{pass@}N)$ over the sampled budgets. Figure 2 contrasts individual prompts, whose curves bend downward, with the aggregate, which is roughly straight on log–log axes.

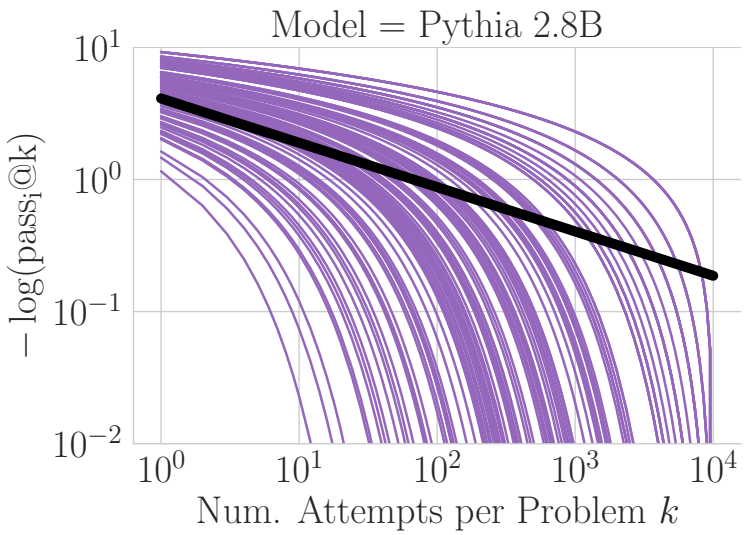

Figure 2: Pythia 2.8B on 128 MATH problems. Purple curves show $-\log(\text{pass}_i@k)$ for individual problems; black shows $-\log(\text{pass}@k)$, with averaging *before* taking the log. Both axes are logarithmic, and k is our N. Cropped from Schaeffer et al. [4, Figure 3]; pass rates are estimated from Brown et al.'s samples [3].

The figure plots negative log pass rate, whereas F_N is failure probability. They are related by

$$-\log(\text{pass}@N) = -\log(1 - F_N) = F_N + O(F_N^2) \qquad \text{as } F_N \to 0.$$

Thus they have the same asymptotic decay rate when failures become rare; they need not be close throughout the plotted range. We analyze F_N because it averages directly across prompts.

Why can this average decay as a power law when each prompt's failure decreases exponentially? Prompts with $N\alpha_x \gg 1$ are already likely to be solved; those with α_x of order $1/N$ or smaller still have a substantial chance of failure. If α_X has density $f(\alpha) \sim c\alpha^{\beta-1}$ as $\alpha \downarrow 0$, for $c, \beta > 0$, then the fraction of such hard prompts is of order $N^{-\beta}$, and [4, Section 3]

$$F_N = \int_0^1 (1-\alpha)^N f(\alpha)\, d\alpha = \Theta(N^{-\beta}).$$

Here $\sim$ means that the ratio tends to one. For example, the uniform distribution on $[0, 1]$ has $\beta = 1$ and $F_N = 1/(N + 1)$, so both failure probability and negative log pass rate decay as $1/N$. The distribution near zero determines how quickly additional samples resolve the remaining prompts.

This describes a population with success probabilities arbitrarily close to zero. On a fixed finite benchmark with all $\alpha_x > 0$, failure eventually decays exponentially. Prompts with $\alpha_x = 0$ leave a nonzero failure floor.

Figure 1 measures accuracy with an imperfect scorer, so its plateau can also reflect selection errors. The plot alone does not distinguish these errors from a shortage of correct candidates.

2 How scoring error affects selection

With a perfect scorer, adding candidates cannot reduce accuracy. Does this remain true if the scorer is accurate on most responses drawn from q_0? Fix the prompt again.

2.1 When more samples hurt

Consider the three-response generator in Table 1. The scorer underestimates the correct response and ranks a rare incorrect response above it.

Response	Base probability q_0	True reward r	Score $\widehat{r}$
Ordinary incorrect response	0.899	0	0
Correct response	0.100	1	0.8
Misleading incorrect response	0.001	0	1

Table 1: The generator and scorer for the three-response example.

The correct response is selected exactly when it appears and no misleading response appears. The probability of this event is the probability 0.999^N of seeing no misleading response, minus the probability 0.899^N that every response is ordinary and incorrect. Thus

$$J(q_N) = 0.999^N - 0.899^N.$$

The misleading response is selected whenever it appears, with probability $1 - 0.999^N$. Thus the expected proxy score is

$$\mathbb{E}_{Y\sim q_N}[\widehat{r}(Y)] = 0.8J(q_N) + 1 - 0.999^N.$$

These probabilities separate the sample sizes needed to encounter a correct response (about 10) and a misleading response (about 1000). Accuracy initially increases as the pool grows. For large N, the misleading response is selected with probability approaching one: accuracy tends to zero while the expected score tends to one. Figure 3 plots the two expectations above. This is an example of *reward-model overoptimization* [5].

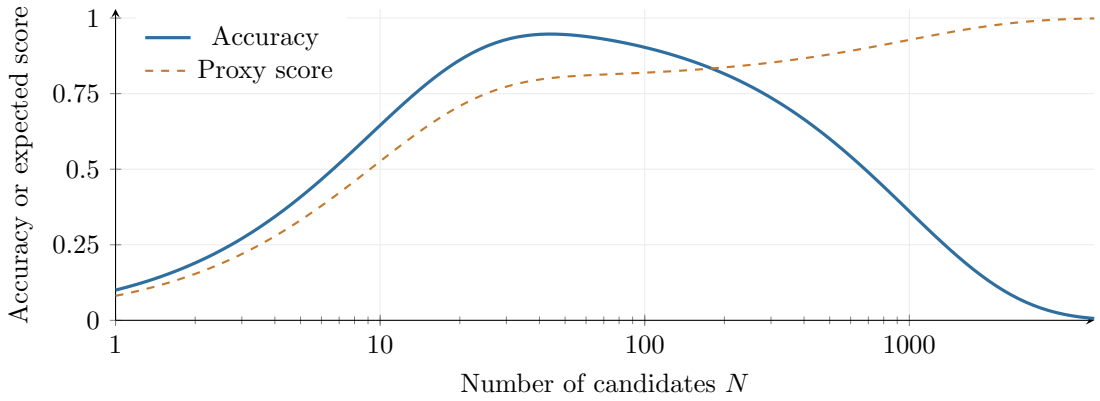

Figure 3: Exact accuracy and expected proxy score for Table 1, calculated from the displayed formulas.

3 Accuracy guarantees for sampling and selection

A selection rule can fail because the pool contains no correct response, or because it selects an incorrect response despite a correct candidate being available. We combine these two sources of failure, then ask how changing the selection rule affects them.

3.1 Coverage and scoring error

In Table 1, each draw is correct with probability 0.1. In general, the generator's coverage of correct responses is

$$\alpha = q_0(G) > 0. \tag{4}$$

The selector observes proxy scores, so it does not know which responses belong to G. We bound its failure probability $1 - J(q)$.

To quantify the scorer's error, write $e(y) = \widehat{r}(y) - r(y)$ and assume its mean absolute error under the generator is bounded by ε_r:

$$\mathbb{E}_{Y\sim q_0}|e(Y)| = \mathbb{E}_{Y\sim q_0}|\widehat{r}(Y) - r(Y)| \le \varepsilon_r. \tag{5}$$

This average allows large errors on rare responses. In Table 1, we can take $\varepsilon_r = 0.1 \cdot 0.2 + 0.001 \cdot 1 = 0.021$, yet the mean absolute error under q_N tends to 1 as selection concentrates on the misleading response. Our setup follows Huang et al., replacing their squared-error assumption with mean absolute error [6, Section 2].

3.2 What best-of-N guarantees

The following theorem is an L_1 adaptation of Huang et al.'s uniform coverage result [6, Section 3.2].

Theorem 3.1 (Best-of-N)**.** *Suppose $r : \mathcal{U} \to \{0, 1\}$ and $\widehat{r} : \mathcal{U} \to [0, 1]$ satisfy (5), and $q_0(G) = \alpha > 0$. For every integer $N \ge 1$,*

$$1 - J(q_N) \le \underbrace{e^{-N\alpha}}_{\textit{sampling}} + \underbrace{N\varepsilon_r}_{\textit{scoring}}. \tag{6}$$

Proof. Let $Y^\star$ be a response with largest true reward in the sampled pool, and let $\widehat{Y}$ be the response selected by proxy score.

Sampling. The best candidate is incorrect exactly when all N draws are incorrect. Hence

$$1 - \mathbb{E}[r(Y^\star)] = (1 - \alpha)^N \le e^{-N\alpha}.$$

Scoring. Since $r = \widehat{r} - e$, subtracting the rewards gives

$$\begin{aligned} r(Y^\star) - r(\widehat{Y}) &= \underbrace{\widehat{r}(Y^\star) - \widehat{r}(\widehat{Y})}_{\le 0} + e(\widehat{Y}) - e(Y^\star) \\ &\le e(\widehat{Y}) - e(Y^\star). \end{aligned}$$

The bracket is nonpositive because $\widehat{Y}$ maximizes proxy score. Let $I^\star, \widehat{I} \in \{1, \dots, N\}$ be the chosen indices, so $Y^\star = Y^{(I^\star)}$ and $\widehat{Y} = Y^{(\widehat{I})}$. For every realized pool,

$$\begin{aligned} e(\widehat{Y}) - e(Y^\star) &= \sum_{i=1}^{N} \big(\mathbf{1}\{\widehat{I} = i\} - \mathbf{1}\{I^\star = i\}\big) e(Y^{(i)}) \\ &\le \sum_{i=1}^{N} |e(Y^{(i)})|. \end{aligned}$$

Each coefficient belongs to $\{-1, 0, 1\}$; if the two indices coincide, all coefficients are zero. This inequality holds even though the indices depend on the sampled pool. Taking expectations therefore

requires only that each original draw has law q_0:

$$\mathbb{E}[r(Y^\star) - r(\widehat{Y})] \le \mathbb{E}\left[\sum_{i=1}^{N} |e(Y^{(i)})|\right] = N\mathbb{E}_{Y\sim q_0}|e(Y)| \le N\varepsilon_r.$$

Finally, $\mathbb{E}[r(\widehat{Y})] = J(q_N)$. Adding the sampling and scoring losses proves the bound on $1-J(q_N)$. □

Increasing N reduces the sampling term but increases the scoring bound. The three-response example shows that this second effect can dominate even when correct responses are plentiful.

3.3 The tradeoff in top-k selection

In that example, a large pool contains about 10% correct entries and 0.1% misleading entries. BoN lets one misleading entry determine the output. Instead, keep the highest-scoring 5% of the pool and choose uniformly among them. Misleading entries occupy only about $0.1/5 = 2\%$ of the retained set, so accuracy tends to 98% as the pool grows.

Retaining too few entries lets misleading responses dominate; retaining too many includes ordinary incorrect entries. The 5% choice is half the expected fraction of correct responses. This leaves room for sampling fluctuations while diluting the rare scoring errors.

In general, retain the k highest-scoring entries among N independent samples and return one uniformly at random. Write $q_{N,k}$ for its distribution. We retain duplicate *sample entries* and break ties by response ordering and then sample index. The choices $k = 1$ and $k = N$ recover BoN and q_0, respectively. Here top-k refers to complete responses, rather than token decoding.

Increasing k reduces the influence of any one scoring error, since each retained entry is chosen with probability $1/k$. But even with perfect scores, the retained set must include incorrect responses if the pool contains fewer than k correct responses.

Theorem 3.2 (Top-k selection)**.** *Suppose $r : \mathcal{U} \to \{0,1\}$ and $\widehat{r} : \mathcal{U} \to [0,1]$ satisfy (5), and $q_0(G) = \alpha > 0$. For integers $1 \le k \le N$, let $B \sim \mathrm{Bin}(N,\alpha)$ count the correct responses in the pool. Uniform selection among the k highest-scoring entries satisfies*

$$1 - J(q_{N,k}) \le \underbrace{\mathbb{P}(B < k)}_{\text{too few correct responses}} + \underbrace{\frac{N}{k}\varepsilon_r}_{\text{scoring loss}}.$$

In particular, choosing $k = \lceil N\alpha/2 \rceil$ gives

$$1 - J(q_{N,k}) \le \underbrace{e^{-N\alpha/8}}_{\text{sampling}} + \underbrace{\frac{2\varepsilon_r}{\alpha}}_{\text{scoring}}. \tag{7}$$

Proof. Let $V^\star$ and $\widehat{V}$ be the fractions of correct responses among the true and proxy top k entries. Uniform selection gives $\mathbb{E}[\widehat{V}] = J(q_{N,k})$.

Sampling. When $B \ge k$, the true top k entries are all correct, so $V^\star = 1$. Otherwise, $V^\star \ge 0$. Thus $1 - \mathbb{E}[V^\star] \le \mathbb{P}(B < k)$.

Scoring. Use the same argument as for BoN: the proxy top k have at least as large a total proxy score as the true top k. After substituting $r = \widehat{r} - e$, errors at shared indices cancel, and each

remaining index contributes at most its absolute scoring error. Dividing by k to compare fractions of correct entries and taking expectations gives

$$\mathbb{E}[V^\star - \widehat{V}] \le \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{N} |e(Y^{(i)})|\right] \le \frac{N}{k}\varepsilon_r.$$

The factor N comes from the N original draws; $1/k$ comes from uniform selection among the k retained entries. Adding the two losses proves the first claim.

For $k = \lceil N\alpha/2 \rceil$, $N/k \le 2/\alpha$, and a Chernoff bound gives $\mathbb{P}(B < k) = \mathbb{P}(B < N\alpha/2) \le e^{-N\alpha/8}$. □

Retaining a fixed fraction makes the scoring bound independent of N, at the cost of a slower sampling bound. The total bound decreases to $2\varepsilon_r/\alpha$; actual accuracy need not increase at every budget. To choose k, any known positive lower bound on α can replace α in the rule and guarantee.

Other ways to control overoptimization. Huang et al.'s *InferenceTimePessimism* limits concentration by penalizing departures from q_0. This discourages large probability increases based on overestimated scores. With fixed penalty strength, their regret bound does not deteriorate once the sample budget is sufficiently large. Their analysis uses squared scoring error. The paper also discusses KL penalties as another way to control concentration [6, Section 4].

3.4 Why an error floor remains

The scoring bound can be loose for a particular scorer: if its top k entries agree with the true top k, the selection loss is zero. With $k = \lceil N\alpha/2 \rceil$, however, our guarantee only reduces the scoring term to $2\varepsilon_r/\alpha$. Can another selector improve the dependence on ε_r/α once the sampling term is negligible?

Theorem 3.3 (A limit for every selector)**.** *Fix $0 < \alpha \le 1/2$ and $\varepsilon_r > 0$. For every rule using the generator and proxy scorer without true-reward feedback, there is an instance with binary rewards and proxy scores satisfying (5) and $q_0(G) = \alpha$ such that its output distribution q satisfies*

$$1 - J(q) \ge \frac{\varepsilon_r}{\alpha + 2\varepsilon_r}. \tag{8}$$

This holds even if the rule knows α and ε_r and uses arbitrarily many samples and proxy-score evaluations.

Proof. Consider $m \ge 2$ candidate responses $y_1, \ldots, y_m$ with equal generator probability and proxy score 1. Exactly one is incorrect. We need many candidates to make scoring error much smaller than coverage: the correct candidates together have probability α, while the single incorrect candidate must have probability at most ε_r. Each candidate has probability $\alpha/(m-1)$, so we choose the smallest suitable number, $m = 1 + \lceil \alpha/\varepsilon_r \rceil$.

In case j, the rewards are as follows; the second row describes each of the other $m-1$ candidates.

Response	q_0	$\widehat{r}$	r_j
y_j	$\alpha/(m-1)$	1	0
$y_i,\ i \ne j$	$\alpha/(m-1)$	1	1
Remaining response	$1 - m\alpha/(m-1)$	0	0

These probabilities are valid because $m \ge 2$ and $\alpha \le 1/2$.

The generator and proxy scores are the same in all m cases, so the selector has the same output distribution q in each. Every candidate y_i is incorrect in exactly one of the m cases; the remaining response is incorrect in all of them. Whatever response the selector returns, it therefore fails in at least a $1/m$ fraction of the cases. Averaging over its output preserves this lower bound.

Each case is a fixed reward assignment with its own failure probability. If every case had failure probability below $1/m$, their average would also be below $1/m$, contradicting the bound above. Hence at least one fixed case has failure probability at least $1/m$. Choose that case as the instance and use its reward function in $J(q)$. Since $m \le 2 + \alpha/\varepsilon_r$, we obtain

$$1 - J(q) \ge \frac{1}{m} \ge \frac{\varepsilon_r}{\alpha + 2\varepsilon_r}.$$

□

For $\varepsilon_r \le \alpha$, this lower bound has the same ε_r/α dependence as the fixed-fraction top-k guarantee, up to constants. The scorer's error is confined to one rare response, but more samples cannot tell the selector which high-scoring response is incorrect.

4 How maximum likelihood supplies coverage

Let $p(y \mid x)$ be the distribution that generates training responses, and let $q_0(y \mid x)$ be the model learned by maximum likelihood. Matching p requires reproducing its response probabilities. Sampling and selection need a weaker property: correct responses under p should retain enough probability under q_0 to be found. Chen et al. [7] study when MLE achieves this coverage.

The coverage profile. For prompts $X \sim \mu$, define

$$\mathrm{Cov}_M(p \| q_0) = \mathbb{P}_{X \sim \mu,\, Y \sim p(\cdot|X)}\left[\frac{p(Y \mid X)}{q_0(Y \mid X)} \ge M\right], \qquad M > 1,$$

with ratio $+\infty$ when $p(y \mid x) > 0 = q_0(y \mid x)$. This measures the data mass that q_0 underweights by a factor of at least M. A small value means that most responses drawn from p retain at least $1/M$ of their original probability under q_0; their probabilities need not match exactly. Here M is a probability-ratio threshold, distinct from the sampling budget N.

Coverage without identifying the exact model. Imagine 1,000 candidate models: 100 variants of each of 10 reference models. Every response probability of a variant is within a factor of 1.1 of its reference, for every prompt. If the true model p and the fitted model q_0 are in the same group, their probabilities differ by at most a factor $1.1^2 = 1.21$. The fitted model therefore has zero coverage error for $M > 1.21$, even if the data has not distinguished the exact variant.

For a finite class $\mathcal{P}$ of K models, let K_M count the smallest number of representatives needed to group models with sufficiently similar response probabilities. Larger M allows larger differences, so K_M can decrease.[1] Always $K_M \le K$. In the example, $K = 1{,}000$ but $K_M \le 10$ once $M^c \ge 1.1$, so each reference can represent all its variants. If the models have sufficiently different probabilities, K_M can equal K.

[1] Precisely, a representative q' accounts for q if $M^{-c} q'(y \mid x) \le q(y \mid x) \le M^c q'(y \mid x)$ for every prompt and response, where $c > 0$ is a suitable universal constant supplied by the theorem. Representatives need not belong to $\mathcal{P}$. This is the coarse covering number in Chen et al.'s theorem.

What maximum likelihood guarantees. The following finite-class specialization of Chen et al.'s theorem bounds severe underweighting when the data distribution belongs to the model class. We use it without proof [7, Theorem 4.1].

Theorem 4.1 (Coverage for finite-class MLE)**.** *Suppose $p \in \mathcal{P}$ and $|\mathcal{P}| = K < \infty$. Fit an exact maximum-likelihood model q_0 over $\mathcal{P}$ using n independent pairs $X_i \sim \mu$, $Y_i \sim p(\cdot \mid X_i)$. For each fixed $M \geq 8$ and $0 < \delta < 1$, with probability at least $1 - \delta$ over the data,*

$$\mathrm{Cov}_M(p\|q_0) \lesssim \underbrace{\frac{\log K}{n \log M}}_{\text{all candidate models}} + \underbrace{\frac{\log K_M + \log(1/\delta)}{n}}_{\text{groups of similar models}}. \tag{9}$$

Here $\lesssim$ hides an absolute constant.

The guarantee averages over prompts from μ; it does not give the same bound for every prompt.

What the bound gains. A usual finite-class MLE analysis gives a coverage bound of order $\log(K/\delta)/n$ [7, Appendix C]. The new bound divides the full-class term by $\log M$ and counts only K_M groups in the second term. In our example, that second count can be 10 instead of 1,000. When M is large and $K_M \ll K$, this certifies coverage with fewer training examples than the usual bound suggests. If $K_M = K$, the second term remains of the usual order. The estimator is still MLE; the sharper analysis uses groups of similar models to bound severe underweighting directly.

Why coverage is enough for selection. To connect this to the selection bounds, specialize to a single prompt: take μ concentrated on that prompt and suppress x. Suppose a response drawn from the training distribution p is correct with probability at least ρ, so $p(G) \geq \rho$. We want a lower bound on the learned model's probability $\alpha = q_0(G)$ of producing a correct response.

Assume $\mathrm{Cov}_M(p\|q_0) \leq \eta < \rho$, and let $E = \{y : p(y) > 0,\ q_0(y) \leq p(y)/M\}$ be the underweighted responses. Their total probability under p is at most η. Even if all of this mass belongs to correct responses, removing E leaves at least $\rho - \eta$ probability on correct responses under p. On each remaining response, $q_0(y) \geq p(y)/M$. Therefore

$$\alpha = q_0(G) \geq q_0(G \setminus E) \geq \frac{p(G \setminus E)}{M} \geq \frac{\rho - \eta}{M}. \tag{10}$$

This supplies the correctness probability needed to choose a sampling budget and a retained fraction for top-k. Larger M allows more probability distortion, but weakens this lower bound by the factor $1/M$. More training data reduces the bound on η at a fixed M.

References

[1] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, Ł. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. *Training Verifiers to Solve Math Word Problems.* arXiv:2110.14168, 2021. Full paper.

[2] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. *Let's Verify Step by Step.* arXiv:2305.20050, 2023. Full paper.

[3] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. *Large Language Monkeys: Scaling Inference Compute with Repeated Sampling.* arXiv:2407.21787, 2024. Full paper.

[4] R. Schaeffer, J. Kazdan, J. Hughes, J. Juravsky, S. Price, A. Lynch, E. Jones, R. Kirk, A. Mirhoseini, and S. Koyejo. *How Do Large Language Monkeys Get Their Power (Laws)?* ICML, PMLR 267:53132–53176, 2025. Full paper.

[5] L. Gao, J. Schulman, and J. Hilton. *Scaling Laws for Reward Model Overoptimization.* ICML, PMLR 202:10835–10866, 2023. Full paper.

[6] A. Huang, A. Block, Q. Liu, N. Jiang, A. Krishnamurthy, and D. J. Foster. *Is Best-of-N the Best of Them? Coverage, Scaling, and Optimality in Inference-Time Alignment.* ICML, PMLR 267:25075–25126, 2025. Full paper.

[7] F. Chen, A. Huang, N. Golowich, S. Malladi, A. Block, J. T. Ash, A. Krishnamurthy, and D. J. Foster. *The Coverage Principle: How Pre-Training Enables Post-Training.* ICLR, 2026. Full paper.