FROM COVERAGE TO ACCURACY: MAKING INDEPENDENT SAMPLING COMPETITIVE WITH GUIDED SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Verifier-guided search can raise accuracy while narrowing generation onto a few reasoning paths. We ask how that narrowing changes the answer a method returns, separating coverage (a correct answer is in the pool) from selection (the decision rule returns it). On GSM8K and MATH, independent sampling covers correct answers about as often as guided search, yet turns less of that coverage into correct votes: on MATH it leads guided coverage by 6 to 11 points while its vote trails by 12 to 14. Guidance raises the share of correct candidates, and it makes wrong answers agree more often. Conditional comparisons and controlled answer-group interventions show the agreement can outweigh the enrichment for a correct minority, even while guidance lifts the vote overall. This points to improving selection as an alternative to spending compute on guided generation. We introduce certified selective verification, which returns the exact weighted-voting winner for a fixed bounded score vector while scoring only the candidates needed to decide it. In live end-to-end runs it reproduces every full-verification decision using 70 to 83% fewer verifier tokens on GSM8K, 36 to 62% on MATH. Sweeping the pool size gives certified independent sampling competitive accuracy–compute tradeoffs: it matches REBASE's best observed GSM8K accuracy at about a third less compute, and reaches higher observed accuracy on MATH (62.2% versus 60.3%) at 27% less compute.

1 INTRODUCTION

Guided search has a familiar tension: concentrating on promising reasoning paths also makes its candidates more alike. Once a prefix receives a high verifier score, many later candidates descend from it. Beeching et al. (2024) and Giannone et al. (2026) document this loss of diversity in verifier-guided beam search and particle resampling, and propose methods that preserve separate subtrees or delay concentration. Yet guided search can outperform independent sampling in reported accuracy–compute comparisons (Wu et al., 2025; Beeching et al., 2024). What does guidance gain when it gives up diversity?

Final accuracy alone cannot answer this. A method may find a correct answer on more questions, or it may make that answer easier to select on the questions it already covers. Repeated descendants can hand a correct answer a decisive vote, and they can hand the same wrong answer enough votes to beat a correct minority. Agreement helps only when the candidates agree on a correct answer.

We separate these effects on GSM8K and MATH with independent sampling, two controlled guidance rules, and REBASE (Wu et al., 2025). Independent sampling reaches comparable or higher coverage, the probability that at least one candidate is correct, but guided search still wins more votes. On MATH, REBASE improves ordinary voting by 12.0 points over independent sampling while covering 7.3 points fewer questions. That advantage comes from the composition of the population, not from greater coverage.

Guidance enriches the pool with correct candidates, which pushes many populations toward a decisive correct vote. It also makes wrong answers agree, mostly among descendants of a shared prefix, and at a fixed correct count that agreement can sink a correct minority. Enrichment outweighs the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

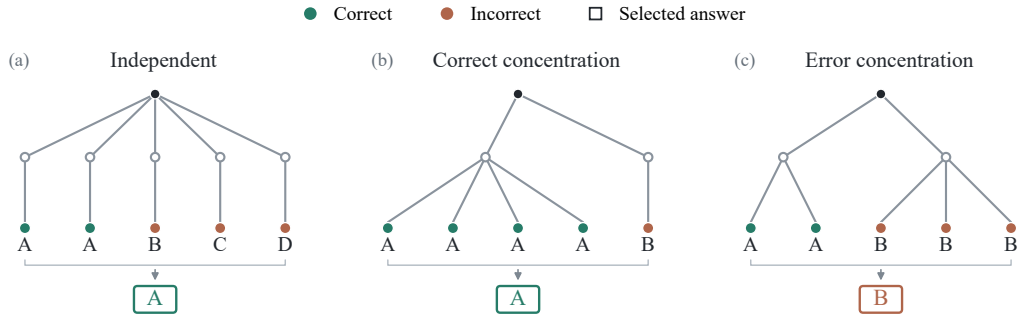

Figure 1: **Answer concentration can help or hurt voting.** Shared prefixes can concentrate correct support or repeated errors. Panels (a) and (c) contain the same number of correct answers, yet select different winners. Illustrative examples.

penalty in the overall vote. Splitting only the wrong-answer groups ties the same competition to verification: a method rules out smaller rivals with fewer score queries.

Independent sampling already supplies the alternatives; the difficulty is selecting among them without running a verifier through the search. We first separate coverage from selection and locate where guided search's vote advantage comes from, splitting it into the correct candidates guidance adds and the agreement it creates among errors (Sections 3–4). That analysis moves the verifier off the search and onto the final vote. Prior selectors either reweight votes by a verifier score (Li et al., 2023) or stop sampling once the majority answer is statistically settled (Cordero-Encinar & Duncan, 2025). We instead certify the exact weighted winner of a fixed candidate pool: bounded scores give a deterministic guarantee for any score vector, and we score candidates only until no unscored one can change the outcome (Section 5). The same error clustering that costs a correct vote also sets the price of this certificate (Section 6). End to end, certified selection reproduces every full-verification decision while cutting verification work, and sweeping the pool size gives competitive accuracy–compute trade-offs against REBASE, matching its best GSM8K accuracy at lower compute and reaching higher observed accuracy on MATH (Section 7). Independent sampling already finds the answers; the work is in selecting them, and in verifying only as far as the selection needs.

2 Generation and Selection

We separate the generation of candidate solutions from the rule that selects their final answer. Every method uses the same policy, Qwen2.5-1.5B-Instruct (Qwen et al., 2024), and the same process reward model, Qwen2.5-Math-PRM-7B (Zhang et al., 2025), on GSM8K (Cobbe et al., 2021) and MATH-500 (Lightman et al., 2023), a subset of MATH (Hendrycks et al., 2021). A prefix is a partially generated solution, and the PRM scores its promise. Table 1 sets out how those scores shape generation. Throughout, *guided search* names the family of verifier-guided methods and *soft selection* our stochastic retention rule. REBASE follows Wu et al. (2025); soft and hard selection vary the selection pressure under our control.

Table 1: **Four ways to generate a candidate population.** Soft and hard selection score an expanded pool before retaining candidates. REBASE uses parent scores to allocate the next expansions.

Method	Generation rule	Verifier in generation?
Independent	Extend separate chains without competing for retention.	No
Soft	Expand each active prefix equally; sample survivors without replacement, favoring higher scores.	Yes, for retention
Hard	Expand each active prefix equally; retain the highest-scoring candidates.	Yes, for retention
REBASE	Allocate more continuations to higher-scoring prefixes, subject to a fixed total width.	Yes, for allocation

Budgets. $C = dmk$, for depth d, width m, and branching k: the nominal number of prefix expansions it may make. Independent sampling has $k = 1$; the guided rules have $k = 2$. At depth 12, $C = 240$ gives independent sampling $m = 20$ chains, soft and hard selection $m = 10$ with $k = 2$, and REBASE width 20, so every method does the same nominal generation work. Completed solutions stop expanding, so realized work can fall below C. The saved comparisons (Sections 3–6) fix C=240 and add two larger independent budgets, C=288 and 336, which raise the pool by 20% and 40% (24 and 28 chains) to isolate the effect of more candidates. The live frontier (Section 7) sweeps the two axes directly: the independent candidate count N, with C = depth $\times$ N, and the REBASE width w, and it reports realized compute rather than nominal C. Exact retention and allocation rules, stopping conditions, and per-method budgets appear in Appendix A.1. All main-text runs use depth 12; a depth-24 control on MATH shows the same vote gap (Appendix E.3).

Selecting an answer. *Ordinary voting* returns the most frequent parsed answer, even without an absolute majority (Wang et al., 2023). *Verifier-weighted voting* sums candidate weights within each answer group and returns the largest sum (Li et al., 2023). *Certified selective verification* queries only enough candidate scores to fix that same weighted winner for a given score vector (Section 5); it generates and parses every candidate and varies only how many receive a score. Independent generation can therefore use a verifier for the final selection without steering the chains. Diagnostic runs score intermediate independent prefixes for analysis, and those scores never change generation.

Coverage and accuracy. Coverage records whether any final candidate is correct. Answer accuracy records whether the selection rule returns a correct answer. The gap between them is how often a correct answer was available but not selected. Only parsed answers cast votes. In the conditional analyses we drop unparsed candidates and compare exactly ten parsed answers, subsampling independent and REBASE populations when needed, then measure voting success at matched numbers of correct answers. Overall accuracy and coverage always include every search.

Evaluation cohorts. The main paired presentation uses 200 GSM8K test questions and all 500 MATH-500 questions, each with two seeds, and a separate 500-question GSM8K cohort corroborates the pattern. Selectors are tuned on disjoint development questions (Appendix A.2).

3 Independent sampling covers more than its vote recovers

A method can hold a correct answer in its pool and still return a wrong one. Independent sampling matches guided search at the first and falls behind at the second: its coverage is comparable to guided search on GSM8K and higher on MATH, while ordinary voting recovers less of that coverage (Figure 2). On GSM8K coverage sits near 94% throughout, yet voting climbs from 80.75% under independent sampling to 84.50% under soft selection and 88.25% under hard. MATH widens the split: independent coverage leads soft and hard selection by 6.3 and 10.7 points while its voting trails them by 14.0 and 11.7. REBASE sits on the same pattern: 94.50% coverage and 89.25% vote accuracy on GSM8K, 64.8% and 58.5% on MATH.

A 40% larger independent pool raises coverage to 96.25% on GSM8K and 75.0% on MATH, but majority voting barely moves, to 80.75% and 48.5%. The extra correct candidates do not turn into correct decisions on their own: the larger pools leave roughly 15 points of headroom on GSM8K and 27 on MATH.

A verifier reclaims part of that headroom. Verifier-weighted voting, with its weighting rule fixed on a separate development split of questions disjoint from the evaluation set, lifts the larger independent pools to 85.25% on GSM8K and 52.4% on MATH, though the guided populations still lead: hard selection reaches 89.0% on GSM8K and soft selection 60.8% on MATH. Tuned selectors are chosen per method and dataset on development data (Appendix A.2); coverage and majority vote do not depend on them.

The same effect appears on a second, larger GSM8K set run separately from the 200-question diagnostic cohort: soft selection improves majority voting by 11.3 and 10.5 points at two budgets while coverage barely moves (-1.6 and -1.8). Appendix B.1 gives the full comparisons and uncertainty estimates.

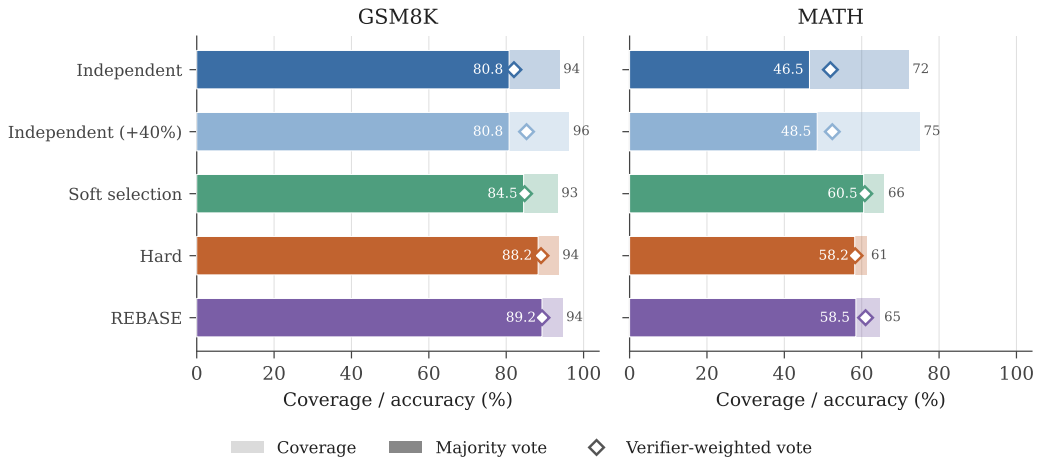

Figure 2: **Independent sampling finds answers that voting fails to recover.** Pale bars show coverage, solid bars majority-vote accuracy, and diamonds verifier-weighted voting tuned on development data. The larger independent budget is C=336 (28 chains). Guided methods select more accurately despite comparable or lower coverage. The cohorts contain 200 GSM8K and 500 MATH questions, each with two seeds.

4 Guidance concentrates correct answers and errors

Coverage records whether a correct answer exists. Voting also depends on how strongly that answer is represented and how its competitors are distributed. Two correct votes beat three different wrong answers but lose to three copies of a single wrong answer. Guidance changes both sides of that comparison: it adds correct votes, and it lines the wrong ones up.

Correct-answer enrichment. Among parsed candidates, the correct fraction rises from 45.4% under independent sampling to 72.0% under soft selection and 80.7% under hard selection on GSM8K. On MATH it rises from 25.9% to 68.8% and 79.2%. REBASE also enriches correct support, reaching 72.2% on GSM8K and 72.3% on MATH. These are proportions among actual voters, and they explain why a guided population can be easier to select from even when it covers fewer questions.

Error clustering. Guidance does the same to wrong answers. On MATH, two parsed wrong candidates agree with probability 4.6% under independent sampling, 22.8% under soft selection, and 35.4% under hard; GSM8K shows the same ordering (Figure 3). The agreement concentrates within shared-prefix families, where descendants of one prefix repeat its error, and REBASE clusters wrong answers too, at 33.4% on GSM8K and 27.5% on MATH. Appendix B.2 gives the within- and across-family rates.

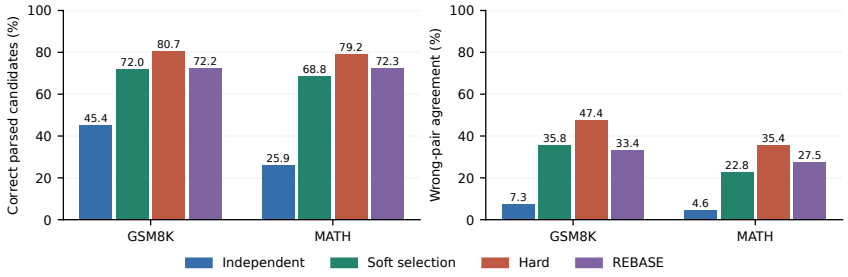

Figure 3: **Guidance concentrates both correct answers and errors.** Left: correct fraction among parsed candidates. Right: agreement among parsed wrong-answer pairs. Both effects strengthen under hard selection and appear on both datasets.

A correct minority faces stronger competition. Table 2 holds the voting population at ten parsed answers and compares each minority count on its own. With three correct voters, independent sampling returns the correct answer 88.9% of the time; soft selection, hard selection, and REBASE return it 35.7%, 12.5%, and 47.7%. The correct support is the same, and its competition differs. Independent errors split across several answers, while a repeated wrong answer can outvote the correct group.

Table 2: **The same correct count can win fewer votes after guidance.** GSM8K vote accuracy (%) among ten parsed voters, for every minority count. Independent and REBASE populations are uniformly subsampled to ten. Each original search has total weight one. Question support and uncertainty are reported in Appendix B.3.

Correct voters (of 10)	Independent	Soft	Hard	REBASE
1	2.4	0.0	0.0	0.0
2	55.0	0.0	0.0	6.6
3	88.9	35.7	12.5	47.7
4	98.7	75.0	50.0	83.8

REBASE gives a check beyond the controlled guidance rules: at three correct voters, its gap from independent sampling is -41.3 points (paired 95% interval $[-55.4, -25.6]$). The table describes conditional populations, not a comparison of identical questions within each count. The controlled answer-group intervention in Section 6 isolates grouping while holding candidates and scores fixed.

Enrichment outweighs the competition penalty. For a population with n parsed answers, let J be its correct count, $\pi(j) = \Pr(J = j)$ its count distribution, and $v(j)$ its conditional vote accuracy, so that

$$A = \sum_{j=0}^{n} \pi(j)v(j). \tag{1}$$

Guidance can raise A by shifting mass toward larger J while lowering $v(j)$ at minority counts. The symmetric two-factor decomposition (Shorrocks, 2013) separates the contributions at a fixed parsed population size:

$$A_g - A_i = \underbrace{\sum_j [\pi_g(j) - \pi_i(j)] \, \tfrac{v_g(j)+v_i(j)}{2}}_{\text{enrichment}} + \underbrace{\sum_j [v_g(j) - v_i(j)] \, \tfrac{\pi_g(j)+\pi_i(j)}{2}}_{\text{conditional voting}}. \tag{2}$$

On GSM8K, soft selection's 10.87-point gain over independent ten-answer subsets splits into $+18.63$ from count enrichment and -7.76 from conditional voting. Hard selection gives $+24.40$ and -9.78, a net 14.62. REBASE gives $+20.79$ and -5.93, a net 14.86. Enrichment outweighs the conditional penalty in every case. The decomposition tracks population changes and does not require the same questions in every correct-count stratum.

This accounts for the gap in Section 3. Guided search wins more votes by building stronger correct coalitions on the questions both methods cover, while independent sampling keeps the broader coverage that a better decision rule can exploit.

5 A CERTIFICATE FOR THE WEIGHTED WINNER

Independent sampling already supplies the candidates, so the verifier does more good choosing among them than guiding the search. Selection needs a score only until one answer moves beyond the reach of the rest.

Group candidates by their parsed answer. A candidate with fixed verifier score $s_i \in [0, 1]$ carries weight $w_i = \exp(\beta(s_i - 1))$, with $\beta = 8$ in our experiments, and weighted voting returns the answer with the largest total weight. This rule is fixed across datasets and is not the tuned selector of any method on MATH (Appendix A.2). Suppose answer a has summed weight K_a over its scored candidates and u_a candidates still unscored. Its final weight lies between

$$L_a = K_a + u_a e^{-\beta}, \qquad U_a = K_a + u_a. \tag{3}$$

We score one candidate at a time and stop as soon as $L_a > \max_{b \neq a} U_b$. Every unscored weight lies in $[e^{-\beta}, 1]$, so once this holds no assignment of the remaining scores can change the winner; until it holds, we score a candidate from the unresolved group with the largest upper bound. The rule groups every candidate but scores only some, and it returns the answer that full weighted voting would choose, for any fixed set of scores and regardless of how the candidates were produced or how the verifier is calibrated.

If one answer has reached weight 3.2 and no other can exceed 2.1, the winner is settled and the remaining candidates go unscored. The rule acts on a margin that is already decided, not on a guess that the leading answer is probably right.

A rival's upper bound depends on how its unscored candidates group together. Split a rival into smaller groups and its unscored count divides among them, so its largest upper bound U_b can only fall, while the winner's lower bound L_a stays fixed. The stopping test $L_a > \max_{b \neq a} U_b$ then still holds: a winner certified before the split stays certified, with no extra scoring. Clustered errors therefore cost the most to rule out, and *splitting competing groups cannot increase the cost of certifying a fixed winner*. Algorithm 2 in Appendix C states the scheduler, and the splitting lemma there bounds the minimum certificate size.

6 Clustered errors raise the cost of certification

Guidance makes wrong answers agree (Section 4). At the vote, this lets a repeated wrong answer beat a correct minority; at the certificate, it adds work, because a losing group's unscored candidates still have to be ruled out before the winner is safe. Splitting wrong-answer groups in the saved populations isolates this second cost, holding every candidate score and every correct group fixed. The split relies on correctness labels, so it measures grouping rather than a selector one could deploy. As the wrong groups shrink, ordinary voting rises at unchanged coverage and the same correct winner certifies with fewer scores (Figure 4): mean queries for independent populations drop from 4.05 to 2.26 on GSM8K and from 5.61 to 3.64 on MATH, cuts of 44% and 35%, and REBASE drops the same way, from 2.62 to 1.62 and from 2.99 to 2.65. The splitting lemma of Section 5 bounds the minimum certificate size; in this replay the scheduler's queries fall too, since a smaller rival carries a lower upper bound and clears with fewer scores.

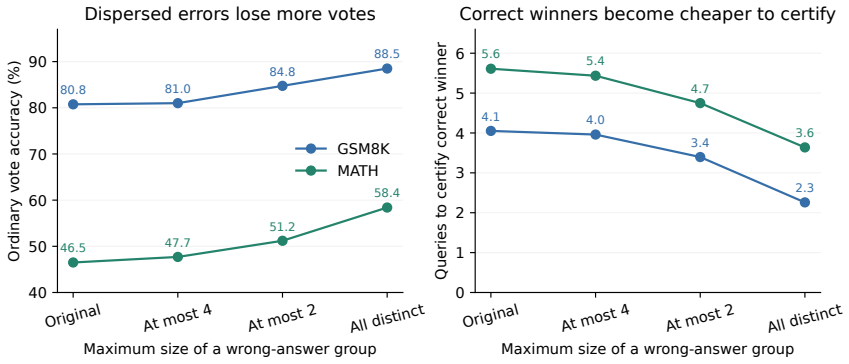

Figure 4: **Error clustering links voting failures to verification work.** Controlled replay of independent populations: only wrong-answer labels are split. Left: ordinary vote accuracy across all searches. Right: queries on the same searches whose original weighted winner is correct (328 GSM8K, 514 MATH); that winner and all scores stay fixed. Splitting requires known correctness and serves only as a diagnostic.

On the unmodified saved populations, Figure 5 plots the share of candidates scored against how concentrated the answers are and against the final weighted-vote margin. When independent sampling's largest answer group holds under a fifth of the parsed candidates, the certificate scores 67% of them on GSM8K and 85% on MATH; when it holds at least four fifths, 10% and 8%. A clear winner settles with little scoring, and close competitors keep more candidates in play. We read the margin after full scoring to account for the cost, and the certificate itself never sees it. Guided populations

often need fewer terminal replay queries, but they obtain those scores during generation. Independent sampling does no generation-time verification and pays only for selective final scoring, so its total verifier work is lower.

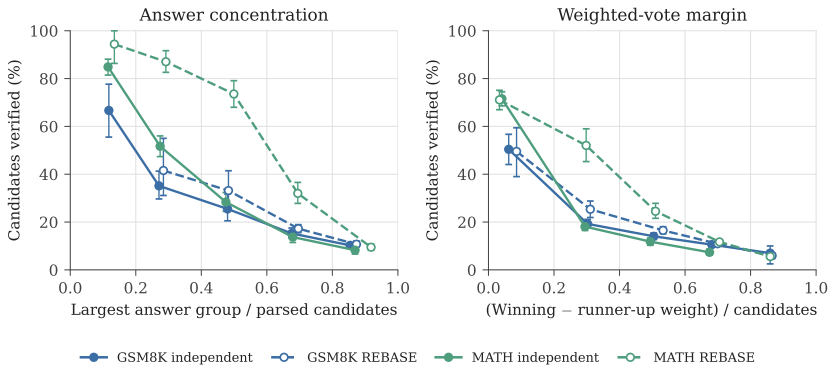

Figure 5: **Clearer decisions require less verification.** Saved-score replay on independent (solid) and REBASE (dashed) populations from GSM8K and MATH, including correct and incorrect winners. Points average searches within fixed-width bins; bars are 95% question-bootstrap intervals. Both axes describing support use the number of parsed candidates as denominator. Appendix C reports bin support. For REBASE these are terminal replay queries as generation has already used the verifier.

7 Independent sampling at lower compute

Independent generation never calls the verifier, and certification calls it only until the winner is fixed. We run the full pipeline end to end to measure what this saves, with fresh generation and separate full and selective scoring. Every weighted curve (certified, full verification, REBASE weighted) on Figure 6 selects by last-step score with weight $\exp(8(s-1))$ and the same tie rule; majority and REBASE vote use ordinary voting with that tie rule.

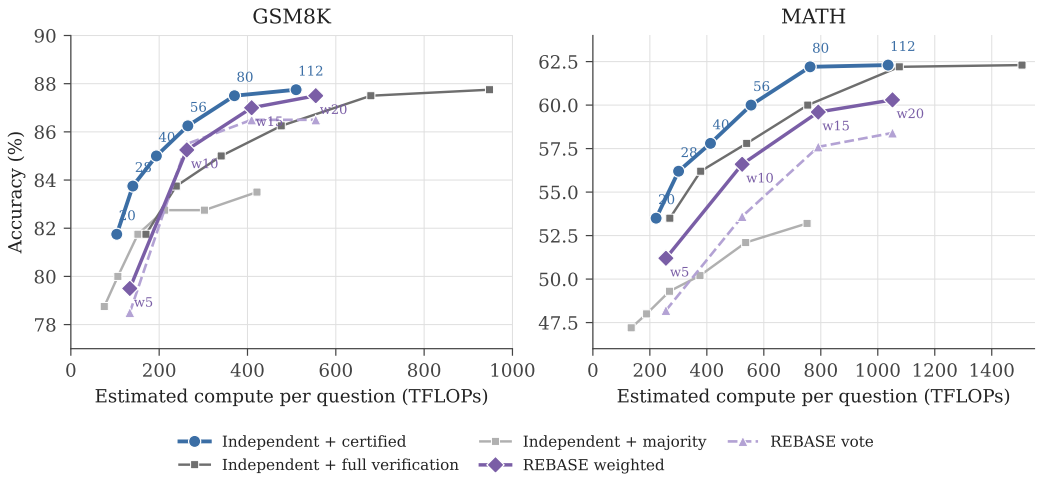

Figure 6: **Independent sampling with certified selection matches or improves on REBASE's accuracy–compute trade-off at the tested operating points.** End-to-end evaluation on GSM8K (left, 200 questions) and MATH (right, 500 questions), two seeds. Independent + certified sweeps $N = 20$–112; REBASE sweeps width 5–20 (point labels). On GSM8K the certified and REBASE accuracies overlap within noise while certified reaches them at lower compute; on MATH certified reaches higher observed accuracy at lower compute.

Certification changes no decisions, only their cost. Across 200 GSM8K questions, two seeds, and six candidate budgets from $N{=}20$ to 112, all 2,400 selected answers match full verification, while verifier input tokens fall by 70–83% and estimated compute by 39–46% at the same budget and accuracy. MATH agrees on all 6,000 comparisons too, with verifier input down 36–62% and compute down 18–31%; the smaller saving follows from weaker consensus on MATH, which leaves more answer groups contested. The compute estimate sums policy input and output and verifier input tokens weighted by model size, with the accounting in Appendix A.4.

Sweeping both methods traces two accuracy–compute frontiers (Figure 6); we compare methods on the same questions, and Appendix D.2 reports the paired intervals on their differences. Certified independent sampling gives competitive accuracy–compute trade-offs on both datasets. On GSM8K the two frontiers reach the same accuracies, and certified reaches them for less compute: it matches REBASE's best 87.5% at $N{=}80$ for 370 TFLOPs against REBASE's 555, about a third less, with the accuracy differences inside the question-bootstrap intervals. On MATH the margin is larger and favors certified selection: $N{=}80$ reaches 62.2% at 762 TFLOPs, above REBASE's best 60.3% at 1,052. Returns flatten past $N{=}80$, where raising the pool to 112 adds at most 0.25 points while spending more.

Harder problems. Partitioning the MATH questions by level (Levels 1–3 easy, Levels 4–5 hard, Appendix D.3) sharpens the separation on harder questions. At $N{=}80$, independent coverage falls from 97.3% on the easy slice to 75.2% on the hard one, while the gap between coverage and ordinary vote widens from 22.1 to 44.1 points; on Level 5, coverage is 67.2% against 18.7% vote accuracy. Adding candidates raises coverage on the hard slice far faster than certified accuracy: from $N{=}20$ to 112, Level-5 coverage climbs from 43.7% to 71.6% while certified accuracy moves only from 21.6% to 26.9%. Certification keeps its guarantee on every slice and still saves compute, 23.5% over full verification at $N{=}112$ on Level 5, though less than on easy questions. Against the REBASE budget curve, the certified advantage is clear on the easy slice ($+5.0$ points at $N{=}80$ over width-15 weighting, at 8.2% less compute) and competitive on the hard slice ($+0.4$ points, interval $[-3.4, 4.2]$).

Table 3: **MATH by difficulty at $N{=}80$.** Coverage falls with difficulty, while ordinary vote accuracy falls faster, widening the gap; certification matches full verification and saves less as consensus weakens. All values from the end-to-end runs; full budgets and intervals in Appendix D.3.

Slice	Coverage	Ord. vote	Certified	Gap (pp)	Save vs full
Levels 1–3	97.3	75.2	85.7	22.1	37.0%
Levels 4–5	75.2	31.1	40.8	44.1	23.9%
Level 5	67.2	18.7	27.6	48.5	20.2%

8 Related work

Coverage and inference-time allocation. Self-consistency returns the majority answer across sampled reasoning paths (Wang et al., 2023), and best-of-n ranks samples by a reward or verifier score. Brown et al. (2024) separate two quantities that scaling treats as one: coverage, the chance that some sample is correct, keeps rising with more samples long after majority voting and reward-based selection plateau. That gap, between what a pool contains and what a rule extracts, is what we study. Compute-optimal scaling work asks how to divide a fixed budget among sampling width, search, and verification, and reports that verifier-guided search can beat independent sampling per unit compute (Snell et al., 2025; Wu et al., 2025). Liu et al. (2025) find that which method wins depends on the policy model, verifier, and problem difficulty. These comparisons rank methods by final-answer accuracy at a given compute, and some also report coverage, but they do not analyze the population each method hands to the selector. We measure the population each method hands to the selector and decompose the guided-search advantage into correct-answer enrichment and agreement among surviving errors. That advantage comes from reshaping the vote so the allocation question becomes where to spend verification.

Dependence and verification. Once the verifier favors a prefix, most later candidates descend from it and repeat its errors. Diverse verifier tree search keeps separate subtrees alive to slow that

collapse (Beeching et al., 2024). DORA groups candidates by reasoning direction and matches compute to direction quality, so a direction earns no extra budget just by holding more candidates (Wang et al., 2025); we share its grouping premise. A parallel line asks how the verifier itself misleads. Imperfect verifiers cap the returns from resampling, since false positives accumulate however many samples you draw (Stroebl et al., 2026), and taking the maximum over many noisy scores overestimates quality, so a wider pool inflates confidence in wrong answers (Dalal et al., 2026). All of this rests on process reward models for the step-level signal (Lightman et al., 2023; Zhang et al., 2025). Beeching et al. (2024) and Wang et al. (2025) fix the collapse at generation time; Stroebl et al. (2026) and Dalal et al. (2026) measure the accuracy it costs. We price it in the vote instead: we trace a shared-prefix genealogy that ties dependence to the final answer histogram and to voting accuracy at matched correct counts, and we show that clustered errors take votes from the correct answer and, through the certificate of Section 5, raise the cost of certifying the winner.

Efficient answer selection. Several methods lower the cost of self-consistency. Confidence-Informed Self-Consistency reweights votes by model confidence (Taubenfeld et al., 2025), and VecCISC clusters reasoning traces to sharpen confidence-weighted answer selection (Petullo et al., 2026). Adaptive-sampling variants stop drawing new samples once the leading answer looks stable (Aggarwal et al., 2023). Certified Self-Consistency sits closest to our rule: it applies a statistical stopping test to decide when enough samples identify the mode of the model's answer distribution (Cordero-Encinar & Duncan, 2025). Its target is distributional and its guarantee is probabilistic. Ours is neither. We certify the exact weighted winner of a fixed, finite candidate pool: bounded verifier weights $w_i \in [e^{-\beta}, 1]$ bound every unqueried group, and we stop once no completion of the unseen scores can move the outcome, so the certificate is deterministic and holds for any score vector, whatever generated the candidates and however the verifier is calibrated. These methods take the pool as given. We connect exact selection back to how the pool is built: independent sampling supplies the coverage, and certification prices the decision among the alternatives it provides.

9 Conclusion

Independent sampling finds correct answers about as often as guided search but selects them less well. Guidance reshapes the selection problem by concentrating both correct answers and surviving errors, and separating those effects accounts for its voting gains. The same error clustering that costs a correct vote also raises the cost of certifying a winner. This points to improving selection, rather than guidance, as where the verifier helps most. On GSM8K and MATH this reaches guided-search voting accuracy at lower compute: about a third less than REBASE on GSM8K, and 27% less on MATH with higher observed accuracy.

Limitations. The study uses one policy and one process reward model from a single family, on two datasets. It does span a range of difficulty: stratifying MATH by level, the coverage–selection gap holds from Levels 1–3 down to Level 5, where coverage is 44% and ordinary voting 14% and certification stays exact throughout (Section 7). We do not reach tasks where a correct sample is rarer still, or other model families and scales. The compute figures are a token and model-size proxy rather than measured FLOPs. We score with a process reward model to keep the verifier identical across methods, but certified selection needs only one bounded score per completed answer; an outcome reward model fits that setting directly and could improve both selection quality and cost.

Reproducibility Statement

Code, data splits, and saved results are available in an anonymous repository. A script recomputes the reported results from the saved outputs without GPUs, and the README explains how to rerun the experiments. Models and datasets are public. Appendix A gives the search rules, datasets, selectors, bootstrap procedure, and compute accounting.

Ethics statement

The goal of this work is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

AI Use Statement

We used generative AI tools, under our direction, in three supporting roles: polishing writing (grammar, phrasing, and clarity edits), generating the illustrative schematic in Figure 1 with a generative image model, and helping write code for the experiments and figures. The research questions, method, experimental design, analysis, and all reported results are our own. We reviewed and verified every AI-assisted output (text, figure, and code) and take full responsibility for the paper's content.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning with llms, 2023.

Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models. Hugging Face blog, 2024. URL https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute.

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, 2024. URL https://arxiv.org/abs/2407.21787.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Paula Cordero-Encinar and Andrew B. Duncan. Certified self-consistency: Statistical guarantees and test-time training for reliable reasoning in llms. *arXiv preprint arXiv:2510.17472*, 2025.

Gal Dalal, Assaf Hallak, Gal Chechik, and Yftah Ziser. More test-time compute can hurt: Overestimation bias in LLM beam search, 2026. URL https://arxiv.org/abs/2603.15377.

Giorgio Giannone, Guangxuan Xu, Nikhil Shivakumar Nayak, Rohan Mahesh Awhad, Shivchander Sudalairaj, Kai Xu, and Akash Srivastava. Mitigating premature exploitation in particle-based Monte Carlo for inference-time scaling, 2026. URL https://arxiv.org/abs/2510.05825.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.291. URL https://aclanthology.org/2023.acl-long.291/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1B LLM surpass 405B LLM? rethinking compute-optimal test-time scaling, 2025. URL https://arxiv.org/abs/2502.06703.

James Petullo, Sonny George, Dylan Cashman, and Nianwen Xue. VecCISC: Improving confidence-informed self-consistency with reasoning trace clustering and candidate answer selection. *arXiv preprint arXiv:2605.08070*, 2026.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Anthony Shorrocks. Decomposition procedures for distributional analysis: a unified framework based on the shapley value. *The Journal of Economic Inequality*, 11:99–126, 2013. doi: 10.1007/s10888-011-9214-z.

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. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The limits of inference scaling through resampling, 2026. URL https://arxiv.org/abs/2411.17501v3. Version 3, 26 March 2026; first posted November 2024.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. doi: 10.18653/v1/2025.findings-acl.1030.

Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Every rollout counts: Optimal resource allocation for efficient test-time scaling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/940b66bcdb34fb9ae47d85de838c372d-Abstract-Conference.html.

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. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8c3caae2f725c8e2a55ecd600563d172-Abstract-Conference.html.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

A SETUP AND PROTOCOLS

A.1 SEARCH RULES AND BUDGETS

The policy is Qwen2.5-1.5B-Instruct (Qwen et al., 2024) and the verifier Qwen2.5-Math-PRM-7B (Zhang et al., 2025). Generation uses temperature 1, a 160-token step limit, and depth 12 (depth 24 only in Appendix E.3). The budget is $C = d\,m\,k$ for depth d, retained width m, and branching k. Independent sampling uses k=1, soft and hard selection k=2; at depth 12, C=240 gives 20 independent chains, width 10 for soft and hard, and width 20 for REBASE.

Algorithm 1 Candidate generation ($C = d\,m\,k$)

Require: policy, depth d, retained width m, branching k, method M

initialise the active set A: m chains for independent (k=1); m first steps from the prompt for soft, hard, and REBASE, each scored and stored
for $t = 1$ to d **do**
  move completed prefixes to the output; they keep their slots and stop expanding
  $A \leftarrow$ active incomplete prefixes; $w \leftarrow m -$ (completed count) ▷ remaining width
  stop if $A = \varnothing$ or $w = 0$
  if $M =$ independent **then**
    extend each $a \in A$ by one step
  else if $M \in \{\text{soft}, \text{hard}\}$ **then**
    expand each $a \in A$ into k children; score and store each child
    keep the top w: soft by $\lambda S_i + G_i$ (λ=4, Gumbel G_i), hard by S_i
  else if $M =$ REBASE **then**
    allocate w offspring across A by $\mathrm{softmax}(S/\tau)$ on the stored parent scores (τ=0.1, descending-score sequential ceiling), pruning parents given none
    generate the allocated children; score and store each child
  end if
end for
return completed and depth-limited leaves

Algorithm 1 states the four rules. Two retained indices can parse to the same answer. Completed solutions leave the active set but keep their slots, so the retained width shrinks over depth and realized expansions fall below C.

A.2 DATASETS, COHORTS, AND SELECTORS

We evaluate on 200 GSM8K questions (Cobbe et al., 2021) and the 500 questions of MATH-500 (Lightman et al., 2023), a subset of MATH (Hendrycks et al., 2021), each with two seeds, plus a separate 500-question GSM8K cohort (Appendix B.1) for replication. MATH selectors are tuned on 100 problems sampled from the MATH training split, excluding any exact match to an evaluation question; GSM8K selectors use a disjoint development split. Selectors are frozen and never retuned. Grading uses saved correctness labels; MATH keeps distinct correct strings as separate voting groups. Full-run accuracy and coverage keep every search. The conditional and candidate-level analyses drop unparsed candidates and use only searches with at least ten parsed answers; for larger pools we draw 500 uniform ten-answer subsets that keep candidate order, and each eligible search carries total weight one.

The tuned selector picks among ordinary voting, best-score selection, and weighted voting, $\sum \exp(\beta(s_i - \max_j s_j))$ within an answer, over score aggregations (minimum, mean, last step) and $\beta \in \{0, 0.5, 1, 2, 4, 8, 16\}$. The frozen choices differ by dataset. On GSM8K they are last-step β=8 for independent sampling, minimum β=0.5 for soft, and minimum β=2 for hard; on MATH they are mean β=2 for independent, mean β=16 for soft, and minimum β=2 for hard. Every frozen winner is a weighted selector. Certified selection (Section 5) targets the last-step β=8 winner on both datasets.

A.3 UNCERTAINTY

All intervals are 95% percentile question bootstraps that resample questions with both seeds kept together and average the seeds within a question. Unless noted the draw count is 10,000; it is 20,000 for the confirmation cohort (Appendix B.1), 5,000 for the matched-count voting gaps, the splitting savings, and the depth-24 contrasts, and 3,000 for the answer-concentration and vote-margin curves. Paired contrasts reuse the draw weights across methods, and savings recompute 1 − cost ratio per draw. Conditional rates and pair-agreement intervals pool the numerator and denominator across the resampled questions. Intervals are pointwise and unadjusted.

A.4 COMPUTE ACCOUNTING

We estimate compute at $2N$ FLOPs per token for an N-parameter model, $F = 2\,[\,1.5 \times 10^9(\text{policy in} + \text{out}) + 7 \times 10^9(\text{PRM in})\,]$, in TFLOPs after dividing by 10^{12}. It omits attention and cache-reuse effects. Because all methods share the policy and verifier and count tokens identically, we report savings as ratios. The saved diagnostic runs lack verifier input-token counts, so their verifier work is reported as scored-candidate counts.

B THE COVERAGE–SELECTION GAP

B.1 GSM8K CONFIRMATION COHORT

The confirmation cohort is a separate 500-question GSM8K set with two seeds and nine conditions. It reproduces the main pattern on independent data (Table 4): guided selection lifts the vote well above independent sampling at slightly lower coverage. Soft selection improves ordinary voting over independent sampling by 11.3 points ($[8.7, 14.0]$) at C=240 and 10.5 ($[7.8, 13.2]$) at C=480, while coverage moves by -1.6 ($[-3.4, 0.1]$) and -1.8 ($[-3.4, -0.3]$); hard selection gives $+10.6$ and $+10.2$ on the vote at -1.5 and -2.8 coverage.

Table 4: GSM8K confirmation cohort, 500 questions and two seeds. Conversion is ordinary vote accuracy over coverage.

Method	C	Coverage	Vote	Conversion
Independent	240	92.5	74.4	80.4
Unguided (k=2)	240	82.4	62.4	75.7
Soft (k=2)	240	90.9	85.7	94.3
Hard (k=2)	240	91.0	85.0	93.4
Soft (k=4)	240	88.5	81.6	92.2
Hard (k=4)	240	86.9	84.4	97.1
Independent	480	95.9	76.4	79.7
Soft (k=2)	480	94.1	86.9	92.3
Hard (k=2)	480	93.1	86.6	93.0

B.2 ERROR CLUSTERING WITHIN SHARED-PREFIX FAMILIES

Wrong-answer agreement concentrates among descendants of a shared depth-one prefix (Table 5). Guided methods agree far more within a family than across families, while independent sampling has no within-family pairs.

Table 5: Wrong-pair agreement (%) within and across shared-prefix families, with 95% intervals where reported.

Dataset	Method	Within	Across
GSM8K	Soft	54.0 [45.2, 63.1]	25.9 [17.0, 34.6]
GSM8K	Hard	58.7 [48.8, 68.9]	36.6 [25.2, 48.0]
GSM8K	REBASE	38.8	29.0
GSM8K	Independent	—	7.3 [5.8, 9.0]
MATH	Soft	50.6	8.7
MATH	Hard	58.7	17.3
MATH	REBASE	38.4	18.0
MATH	Independent	—	4.6

B.3 Conditional voting and the decomposition

Table 6 reports the GSM8K vote-accuracy gaps to independent sampling at matched correct counts: guidance loses ground at every minority count, most where independent sampling still wins. Table 7 splits the ten-answer voting gain by Equation 2; correct-count enrichment outweighs the conditional-voting loss in every case. MATH minority strata are sparse, often fewer than ten questions per count, so we do not tabulate them.

Table 6: GSM8K vote-accuracy gaps to independent sampling at matched correct counts (percentage points), with 95% intervals and contributing questions.

Method	Count	Gap	95% interval	Questions
Soft	1	−2.4	[−4.0, −1.2]	16
Soft	2	−55.0	[−61.5, −48.1]	6
Soft	3	−53.2	[−78.8, −22.8]	12
Soft	4	−23.7	[−58.8, 1.4]	8
Hard	3	−76.4	[−91.2, −47.7]	8
Hard	4	−48.7	[−98.7, 1.3]	5
REBASE	2	−48.4	[−56.4, −39.0]	37
REBASE	3	−41.3	[−55.4, −25.6]	52
REBASE	4	−14.9	[−22.6, −8.1]	70

Table 7: Two-factor decomposition of the GSM8K ten-answer voting gain over independent sampling (percentage points).

Method	Count enrichment	Conditional voting	Net
Soft	+18.6	−7.8	+10.9
Hard	+24.4	−9.8	+14.6
REBASE	+20.8	−5.9	+14.9

C Certification cost

C.1 The scheduler and the splitting lemma

Algorithm 2 scores candidates one at a time and stops once one answer's lower bound clears every rival's upper bound.

Algorithm 2 Certified selective verification

Require: parsed answers partitioned into groups $\{G_a\}$; verifier on demand; weight $w(s) = \exp(\beta(s-1))$, $s \in [0,1]$, $\beta \geq 0$; margin $\epsilon = 10^{-12}$

if no parsed answers **then return** $\varnothing$
end if
if $\beta = 0$ **then return** $\arg\max_a |G_a|$ ▷ weights are 1; weighted vote = ordinary vote
end if
$K_a \leftarrow 0$, $u_a \leftarrow |G_a|$ for all a ▷ scored weight, unscored count
loop
  $L_a \leftarrow K_a + u_a e^{-\beta}$, $U_a \leftarrow K_a + u_a$ ▷ bounds, all a
  $a^\star \leftarrow \arg\max_a L_a$ ▷ ties: larger K_a, then group order
  if $L_{a^\star} > \max_{b \neq a^\star} U_b + \epsilon$ **then return** $a^\star$
  end if
  $b \leftarrow \arg\max_{a:\, u_a > 0} U_a$ ▷ ties: larger u_a, then group order
  score the first unscored $j \in G_b$: $K_b \mathrel{+}= w(s_j)$, $u_b \mathrel{-}= 1$
  if $\sum_a u_a = 0$ **then return** $\arg\max_a K_a$ ▷ full scoring; same rescaled vote/tie rule as full verification
  end if
end loop

Splitting a rival cannot raise the cost of certifying a fixed winner. Fix the scores and a queried set Q; for a group g write $L_g = K_g + u_g e^{-\beta}$ and $U_g = K_g + u_g$, with K_g the scored weight and u_g the unscored count. Partition a rival b into disjoint subgroups $b_1, \ldots, b_k$. Their weights and unscored counts sum to b's and all terms are nonnegative, so $\max_r U_{b_r} \leq \sum_r (K_{b_r} + u_{b_r}) = U_b$, while the target's L_a is unchanged. Hence $L_a > \max_{b \neq a} U_b$ before the split implies the same after: any Q that certifies a still does, so the minimum certifying $|Q|$ cannot rise. The bound is on the best queried set, not on Algorithm 2's particular order.

The controlled diagnostic splits each wrong-answer group in encounter order into groups of at most four, two, or one, holding scores and the correct group fixed (Table 8). As the groups shrink, ordinary voting rises and the same correct winner certifies with fewer scores. The Queries column reports the mean scores Algorithm 2 uses; the splitting lemma bounds the minimum certifying set. Across 8,400 saved records and 32,512 non-empty split populations, the selective winner matched the full weighted winner every time.

Table 8: Controlled wrong-group splitting, original → all-distinct. Vote is ordinary vote accuracy; queries are mean scores to certify a correct winner. Queries saved carries a 95% interval.

Dataset	Method	Vote	Queries	Queries saved
GSM8K	Independent	80.8 → 88.5	4.05 → 2.26	1.79 [1.53, 2.07]
GSM8K	Soft	84.5 → 89.0	1.56 → 1.14	0.42 [0.31, 0.54]
GSM8K	Hard	88.2 → 92.0	0.94 → 0.62	0.32 [0.21, 0.44]
GSM8K	REBASE	89.2 → 93.5	2.62 → 1.62	1.00 [0.84, 1.18]
MATH	Independent	46.5 → 58.4	5.61 → 3.64	1.97 [1.69, 2.27]
MATH	Soft	60.5 → 62.9	1.52 → 1.34	0.18 [0.12, 0.25]
MATH	Hard	58.2 → 60.0	0.92 → 0.81	0.11 [0.06, 0.16]
MATH	REBASE	58.5 → 60.5	2.99 → 2.65	0.34 [0.25, 0.45]

C.2 Verification structure

Figure 5 uses every independent and REBASE C=240 population with a parsed answer. Concentration is the largest answer group over the parsed count; margin is the winning minus runner-up weighted sum, with $\exp(\beta(s_i - 1))$ weights, over the same count. Points are bin means with 3,000-draw intervals; Table 9 gives the searches per bin. Verified fraction falls with both quantities: the Pearson correlations are -0.56 and -0.64 for GSM8K independent, -0.61 and -0.76 for GSM8K REBASE, -0.68 and -0.70 for MATH independent, and -0.81 and -0.69 for MATH REBASE.

Table 9: Searches per bin in Figure 5.

Dataset	Method	Axis	$[0, .2)$	$[.2, .4)$	$[.4, .6)$	$[.6, .8)$	$[.8, 1]$
GSM8K	Indep.	Concentration	45	109	83	104	59
GSM8K	Indep.	Margin	145	97	89	64	5
GSM8K	REBASE	Concentration	—	16	38	89	257
GSM8K	REBASE	Margin	33	29	83	192	63
MATH	Indep.	Concentration	435	287	146	92	40
MATH	Indep.	Margin	754	130	78	36	2
MATH	REBASE	Concentration	26	152	178	183	417
MATH	REBASE	Margin	377	97	141	225	116

D LIVE FRONTIER AND BUDGETS

D.1 PROTOCOL AND AGREEMENT

The live runs use the same policy and verifier, temperature 1, depth 12, and step limit 160, with independent pools $N \in \{20, 28, 40, 56, 80, 112\}$ as nested prefixes and REBASE at widths 5–20. Full and certified verification make separate verifier calls on the same candidates; we audit answer agreement. Certified and full selected the same answer on every comparison (Table 10); MATH shows score drift in 113 comparisons, with no answer mismatches. Relative to the saved runs, GSM8K coverage is unchanged and ordinary-vote accuracy is 2–3 points lower for both independent sampling (80.75 $\rightarrow$ 78.75) and REBASE (89.25 $\rightarrow$ 86.5), within the question-bootstrap intervals; MATH agrees within one point. All comparisons are made within a single set of runs.

Table 10: Certified-versus-full agreement across the six budgets.

Dataset	Comparisons	Answer mismatches	Score-drift cases	Max score Δ
GSM8K	2,400	0	0	0
MATH	6,000	0	113	0.0098

D.2 BUDGET SWEEP AND THE REBASE FRONTIER

Tables 11 and 12 give the independent accuracy–compute curve; certified and full verification return the same answers, so their accuracy is shared and only the cost differs. Table 13 gives REBASE across widths. Certified selection gives competitive trade-offs against REBASE on both datasets (Table 14). Accuracy gains flatten past N=80; raising the pool to 112 adds 0.25 points on GSM8K and 0.1 on MATH.

Table 11: GSM8K independent curve. Accuracy is certified (= full); TFLOPs and saving are certified against full verification.

N	Coverage	Ordinary	Accuracy	TFLOPs	Save vs full
20	94.25	78.75	81.75	103.9	38.9%
28	95.25	80.00	83.75	140.3	41.2%
40	96.25	81.75	85.00	193.7	43.1%
56	97.25	82.75	86.25	264.9	44.4%
80	98.25	82.75	87.50	370.5	45.5%
112	98.75	83.50	87.75	510.1	46.2%

Table 12: MATH independent curve.

N	Coverage	Ordinary	Accuracy	TFLOPs	Save vs full
20	71.9	47.2	53.5	221.9	17.8%
28	76.4	48.0	56.2	300.9	20.5%
40	80.0	49.3	57.8	412.8	23.5%
56	83.7	50.2	60.0	555.2	26.4%
80	85.7	52.1	62.2	762.5	29.1%
112	87.3	53.2	62.3	1036.2	31.2%

Table 13: REBASE across widths, both selectors.

Dataset	Width	Coverage	Vote	Weighted	TFLOPs
GSM8K	5	84.5	78.5	79.5	133.5
GSM8K	10	91.25	85.5	85.25	262.6
GSM8K	15	92.5	86.5	87.0	409.4
GSM8K	20	94.5	86.5	87.5	554.6
MATH	5	53.4	48.2	51.2	256.4
MATH	10	59.4	53.6	56.6	524.0
MATH	15	63.0	57.6	59.6	791.1
MATH	20	63.9	58.4	60.3	1051.6

Table 14: Certified selection against REBASE weighted voting at nearby compute budgets. Positive accuracy favors certified; positive saving means certified is cheaper.

Dataset	Certified N / width	Accuracy diff. (pp)	Compute saving
GSM8K	56 / 10	$+1.0\ [-2.3, 4.3]$	$-0.9\%\ [-4.8, 2.9]$
GSM8K	80 / 15	$+0.5\ [-3.0, 3.8]$	$9.5\%\ [6.2, 12.6]$
GSM8K	112 / 20	$+0.3\ [-3.5, 4.0]$	$8.0\%\ [4.8, 10.9]$
MATH	56 / 10	$+3.4\ [0.7, 6.1]$	$-6.0\%\ [-8.7, -3.2]$
MATH	80 / 15	$+2.6\ [0.0, 5.3]$	$3.6\%\ [1.3, 5.9]$
MATH	112 / 20	$+2.0\ [-0.8, 4.8]$	$1.5\%\ [-0.7, 3.6]$

D.3 MATH DIFFICULTY

Stratifying MATH by level shows the gap widening on harder questions and the selection bottleneck persisting. Coverage falls with difficulty and the vote falls faster, so the coverage-to-vote gap grows (Table 15). More candidates raise coverage on every slice but do not move certified accuracy on the hard slice: Level-5 coverage climbs from 67.2% at $N{=}80$ to 71.6% at $N{=}112$ while certified accuracy holds near 27%. Certification saves less as consensus weakens yet still cuts compute on the hardest questions, 23.5% over full verification at $N{=}112$ on Level 5. The certified frontier sits above REBASE on the easy slice and meets it on the harder ones (Figure 7): certified $N{=}80$ beats width-15 weighted REBASE by 5.0 points $[1.3, 9.0]$ on Levels 1–3 but by 0.4 $[-3.4, 4.2]$ on Levels 4–5, and REBASE edges ahead at the largest Level-5 budget. The saved depth-24 search widens the same gap.

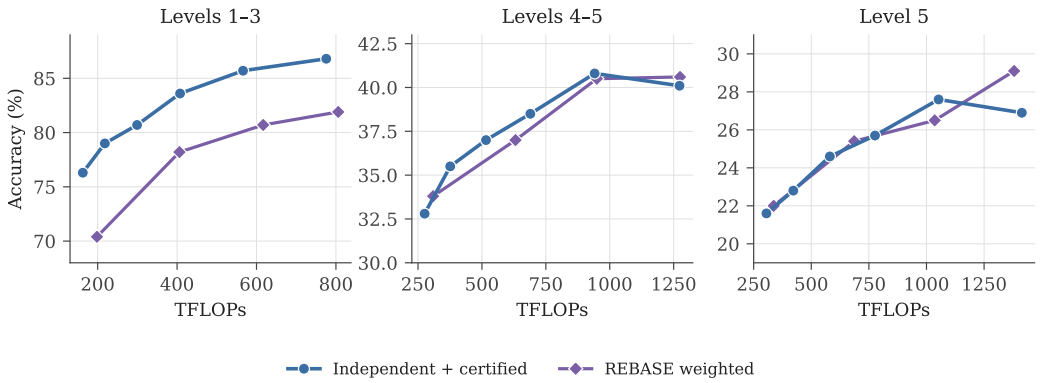

Figure 7: MATH accuracy–compute frontiers by difficulty. Independent + certified sweeps N=20–112; REBASE weighted sweeps width 5–20. The certified curve sits above REBASE on the easy slice and overlaps it on the harder slices, where REBASE edges ahead at the largest budget on Level 5.

Table 15: Independent sampling on MATH by difficulty and budget. Accuracy is certified (= full); saving is against full verification.

Slice	N	Coverage	Ordinary	Certified	TFLOPs	Save vs full
Levels 1–3	20	91.4	70.4	76.3	162.6	28.4%
Levels 1–3	40	94.5	72.9	80.7	299.3	33.7%
Levels 1–3	80	97.3	75.2	85.7	566.1	37.0%
Levels 1–3	112	97.7	76.1	86.8	775.6	38.3%
Levels 4–5	20	54.2	26.1	32.8	275.8	10.8%
Levels 4–5	40	66.8	27.9	37.0	515.9	16.8%
Levels 4–5	80	75.2	31.1	40.8	940.9	23.9%
Levels 4–5	112	77.9	32.4	40.1	1272.9	26.5%
Level 5	20	43.7	14.2	21.6	305.4	7.3%
Level 5	40	57.8	19.0	24.6	580.7	12.5%
Level 5	80	67.2	18.7	27.6	1052.6	20.2%
Level 5	112	71.6	21.6	26.9	1413.9	23.5%

Table 16: REBASE on MATH by difficulty and width.

Slice	Width	Coverage	Vote	Weighted	TFLOPs
Levels 1–3	5	72.5	67.6	70.4	198.3
Levels 1–3	15	81.9	77.9	80.7	616.9
Levels 1–3	20	84.5	80.5	81.9	805.8
Levels 4–5	5	36.1	30.5	33.8	309.2
Levels 4–5	15	45.8	39.1	40.5	949.5
Levels 4–5	20	45.2	38.4	40.6	1275.0
Level 5	5	24.3	20.5	22.0	337.0
Level 5	15	31.3	25.0	26.5	1035.8
Level 5	20	32.8	26.5	29.1	1380.4

E ROBUSTNESS AND CONTROLS

E.1 SCORE AGGREGATION

On GSM8K the frozen weighted selector aggregates PRM scores by last step for independent sampling and by minimum for the guided rules; the MATH choices differ (Appendix A.2). Varying both choices on the diagnostic cohort (Table 17), weighted voting moves at most 1 point between minimum and last-step aggregation, while best-of-n selection swings 3 to 9 points. Aggregation matters for single-candidate selection, not for weighted voting, so the selector choice does not drive the main results.

Table 17: GSM8K accuracy under minimum and last-step aggregation, for weighted voting and best-of-n selection (C=240).

	Weighted (min)	Weighted (last)	Best (min)	Best (last)
Independent	82.25	82.00	72.75	75.75
Soft	84.75	85.75	74.75	83.75
Hard	89.00	89.00	78.75	84.75

E.2 TIE-BREAKING

Ordinary voting breaks ties by candidate order. Averaging over tied answers instead keeps the ranking (Table 18): soft and hard selection still beat independent sampling, by 4.6 $[1.1, 8.3]$ and 8.7 $[5.0, 12.8]$ points.

Table 18: GSM8K accuracy under uniform tie-breaking (C=240). Terminal and unparsed are candidate fractions.

Method	Uniform vote	Terminal	Unparsed
Independent	79.9	98.2	0.01
Soft	84.5	95.2	0.00
Hard	88.6	94.3	0.00

E.3 DEPTH-24 CONTROL

Doubling depth to 24 on MATH raises guided coverage and voting more than independent, and the coverage-selection gap persists (Table 19). At matched width, depth 24 minus depth 12 changes independent coverage by 0.0 $[-2.1, 2.2]$ and vote by -0.8 $[-2.8, 1.3]$, against soft coverage $+7.7$ and vote $+3.6$, and hard coverage $+7.9$ and vote $+6.5$. Independent sampling still covers far more than it votes.

Table 19: MATH at depth 24. Independent widens the budget by adding chains (C=480, 576, 672); soft and hard hold width 10.

Method	C	Coverage	Vote	Tuned
Soft	480	73.5	64.1	66.3
Hard	480	69.3	64.7	65.3
Independent	480	72.1	45.7	53.7
Independent	576	75.2	48.9	53.5
Independent	672	76.4	49.7	54.5