REGIME-DEPENDENT INFERENCE-TIME COMPUTE ALLOCATION FOR AGENTIC SYSTEMS: WHEN DEEPER REASONING WINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Longer reasoning, repeated sampling, revision, and communication have each been advocated as ways to improve language-model performance through their composition into agentic systems. These competing prescriptions leave a practical question: which use of compute is optimal for a given model, task, and compute budget? We organize this choice around the generation, preservation, and selection of correct answers, and compare 96 configurations of one model on four reasoning benchmarks. At a matched budget of 16,384 reasoning tokens, depth gives the highest observed accuracy in three harder groups defined retrospectively by short-run success. In the frequent-success group, voting reaches the observed ceiling at one-quarter of that expenditure. We then ask whether allocation can be prescribed before testing larger budgets on a new problem. A logistic rule trained on graded development outcomes at several depths chooses depth–width allocations using five short probes per new item. In an exploratory comparison on 200 unseen problems, it reaches 70.0% accuracy, versus 70.5% for fixed depth. The results support conditional allocation guidance and show how a calibrated rule can act before a new problem's scaling curve is measured.

1 INTRODUCTION

Empirical studies of test-time compute offer conflicting advice: use more agents, replace debate with voting, revise answers, or let one model reason longer (Kim et al., 2025; Dong & Shigida, 2026; Snell et al., 2025). Under a fixed budget, these choices compete for the same computation. Turning their reported successes into useful prescriptions requires knowing when each choice pays.

Existing work provides empirical recommendations conditioned on baseline performance (Snell et al., 2025; Kim et al., 2025) and mathematical descriptions of generation, verification, and cost (Singhi et al., 2025; Dong & Shigida, 2026; Sunkaraneni et al., 2026). A simple allocation formula, however, need not be cheap to use: its inputs must be measured or justified. Comparing performance at 1,024 and 2,048 reasoning tokens reveals the return to extra depth, but already pays for both runs. Learning across problems offers a way to reuse those measurements (Damani et al., 2025). The question then becomes:

When should a fixed inference budget be allocated to more attempts, more interaction, or deeper reasoning within an agentic system?

We study this question by separating three demands on computation: generating a correct answer, preserving it, and selecting it. For one model on four benchmarks, comparisons at 16,384 reasoning tokens favor depth in the three harder groups defined retrospectively by short-run success. In the frequent-success group, voting reaches the observed ceiling with one-quarter of the reasoning expenditure. On 200 unseen problems, depth achieves 70.5% accuracy, compared with 58.0% for a four-call allocation selected on development data.

We then fit a simple logistic rule on graded development outcomes at several depths. On a new problem, the rule uses five short, unlabeled probes to choose a depth–width allocation before observing any longer-run outcome for that problem. It predominantly selects depth and reaches 70.0% accuracy on the same 200 problems. This exploratory result gives an initial answer to the information

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.

question: calibration across representative problems can support allocation before a new problem's response to scaling is measured.

Building on model-relative allocation (Snell et al., 2025), we make three contributions:

1. **A common framework for competing designs.** Model-relative regimes and generation, preservation, and selection failures connect problem difficulty to the observed returns from depth, width, and revision.
2. **Controlled evidence for concentrating compute in depth.** Depth leads in three of four regimes, exceeds the estimated probability of obtaining any correct answer from equal-budget sampling at the reference allowance by 20.5 percentage points, and outperforms tested mixed allocations on 200 unseen problems. Frequent correctness at the reference reasoning allowance identifies when cheaper aggregation is sufficient.
3. **Supporting evidence from simple allocation learning.** A statistical model fitted on 96 development problems uses five diagnostic attempts per new item, each with 1,024 reasoning tokens. Its allocations predominantly favor deeper reasoning, nearly match fixed depth's accuracy, and outperform the tested LLM allocators. This supports the depth finding and suggests a route to economical allocation rules.

Takeaway. In the studied harder regimes, concentrating compute in deeper individual reasoning outperforms the tested allocations across agents and rounds. When solutions sampled with a smaller reasoning allowance are frequently correct, inexpensive aggregation through multi-agent voting can suffice.

2 Related Works

2.1 Agentic Design Choices and Competing Prescriptions

Self-consistency aggregates independent solutions (Wang et al., 2023), self-refinement revises an initial response (Madaan et al., 2023), and debate introduces peer feedback (Du et al., 2024). Their benefits are conditional: intrinsic self-correction can degrade reasoning without external feedback (Huang et al., 2024), and debate can fail to improve on self-consistency or derive most of its gain from voting (Smit et al., 2024; Choi et al., 2025). Peer exchange can destabilize correct answers (Bertalanič & Fortuna, 2026), yet help weaker models on difficult mathematics (Yang et al., 2025); some interacting configurations also improve the accuracy–compute frontier (Wunderlich et al., 2026). These findings motivate characterizing the model–problem regimes in which each strategy is useful, including when the same reasoning budget is better spent extending one trajectory than adding agents or revision rounds.

2.2 Agentic Systems as Inference-Time Compute Allocation

Test-time scaling frames these design choices as allocations of inference compute. Sampling coverage depends on per-problem success probabilities (Brown et al., 2024; Schaeffer et al., 2025), while more samples need not improve voting or filtering (Chen et al., 2024). Snell et al. (2025) establish that effective allocation among sampling, revision, and verifier-guided search depends on problem difficulty relative to the model. Budget forcing demonstrates the value of extending individual reasoning (Muennighoff et al., 2025). Agent studies extend this perspective to rollouts, revision, verification, and merging (Zhu et al., 2025), budgeted collaboration graphs (Wang et al., 2025), and model- and task-dependent coordination (Kim et al., 2025). Building on this foundation, we use enforced, matched reasoning-token budgets to compare depth, width, revision, and access to other agents' responses, characterizing regimes in which deeper individual reasoning gives greater returns than the evaluated orchestration strategies.

2.3 Generation, Selection, and Correctness Amplification

Oracle coverage (pass@k) measures whether a candidate pool contains a correct answer; practical selection must return one without oracle access. Coverage can rise after voting and reward-model

selection plateau (Brown et al., 2024), leaving a *generation–selection gap*. Weak-verifier ensembles address this gap (Saad-Falcon et al., 2025), whereas verifier false positives and frequent incorrect answers can limit sampling gains (Stroebl et al., 2026; Bay & Yearick, 2026). Work on solving versus verification and on proposal coverage versus reliable critics further distinguishes these uses of compute (Singhi et al., 2025; Sunkaraneni et al., 2026). This distinction motivates testing whether depth expands accessible correctness beyond sampling at the reference reasoning allowance, and whether reliable solutions at that allowance instead make aggregation sufficient.

3 PRELIMINARIES

Regimes describe a problem family relative to a model. Fix a model M, a problem distribution $\mathcal{D}$, and a reference generation protocol with reasoning allowance A. Let $p_{M,x}(A)$ be the probability of a correct response to problem x. Independent reference attempts estimate this probability and assign a descriptive regime $\widehat{r}_{M,A}(x)$. Each regime defines a subfamily

$$\mathcal{D}_r^{M,A} = \mathcal{D}\big(\cdot \mid \widehat{r}_{M,A}(x) = r\big). \tag{1}$$

This is a sampling-based approximation to model-relative difficulty, conditional on the chosen protocol. Finite observations leave uncertainty, especially near regime boundaries; no observed success does not imply that $p_{M,x}(A) = 0$. Our four strata are descriptive bins, not theoretical allocation thresholds.

Allocation trades depth against agents and rounds. An allocation π specifies agent count n, total rounds R, reasoning allowance a per call, communication, and final selection. For uniform call allowances, its reasoning cost is

$$C(\pi) = nRa \le B = kA. \tag{2}$$

Write $S_{\pi,x}$ for returned-answer correctness probability. The primary objective is to maximize $\mathbb{E}_{x\sim\mathcal{D}_r^{M,A}}[S_{\pi,x}]$ over the feasible allocations. At $B = kA$, depth uses $(n, R, a) = (1, 1, kA)$, width uses $(k, 1, A)$, and private revision uses $(1, k, A)$. A larger allowance within one call differs from restarting revision across calls. We assess correctness consistency and latency alongside accuracy, with accuracy taking priority; equal reasoning tokens do not equate all inference costs.

Correctness must be generated, preserved, and returned. Let H_{arc} be the probability that at least one recorded candidate is correct and H_{cur} the probability that at least one final selectable candidate is correct. If the output is selected from that final pool, then

$$1 - S = (1 - H_{\mathrm{arc}}) + (H_{\mathrm{arc}} - H_{\mathrm{cur}}) + (H_{\mathrm{cur}} - S). \tag{3}$$

The terms are generation, preservation, and selection failures. This accounting also applies to dependent agents. At kA, independent calls with A reasoning tokens each have oracle coverage $1 - [1 - p_{M,x}(A)]^k$. Depth can exceed even this oracle if $p_{M,x}(kA)$ is larger. Conversely, high coverage need not yield accurate voting when an incorrect answer attracts more support. Appendix B develops these comparisons; the regime label alone does not determine their outcome.

From probabilistic foundations to empirical regimes. Prior work offers a probabilistic basis for bridging the generation–selection gap. Singhi et al. (2025) study how to divide compute between generation and verification, while Sunkaraneni et al. (2026) formalize how proposal coverage, selection quality, and progress govern reasoning success. We build on this framing by asking whether model-relative regimes reveal consistent patterns in which allocations work best. Estimating how generation and selection respond to compute across many configurations can require extensive sampling. Our question is whether the patterns within regimes are strong enough to fit a simple model that infers near-optimal allocations from far fewer observations than a broad configuration sweep. Our primary goal is to characterize these regimes and establish when deeper individual reasoning is the most effective use of a fixed budget. The statistical allocator provides an initial test of whether limited calibration data and inexpensive diagnostic attempts can recover these allocation preferences.

4 METHODS

Characterize and compare. We evaluate Qwen3.8-27B on 96 problems drawn equally from AIME, HMMT, Omni-MATH, and GPQA-Diamond. Each has 25 independent reference attempts at $A = 1{,}024$ reasoning tokens. Among 94 fully scored sets of reference attempts, correct-draw counts $c = 0$, 1–5, 6–15, and 16–25 define no observed, rare, intermediate, and frequent success. We compare depth, width, private revision, and mixed allocations across 96 configurations with main-sweep budgets up to $25A$ (Figure 1). Matched private and communicating teams hold allocation and final voting fixed to isolate access to other agents' responses. Sampling and teams use plurality; single-agent procedures return their final answer. Enforced reasoning expenditure supports matched-budget comparisons, with other costs reported separately.

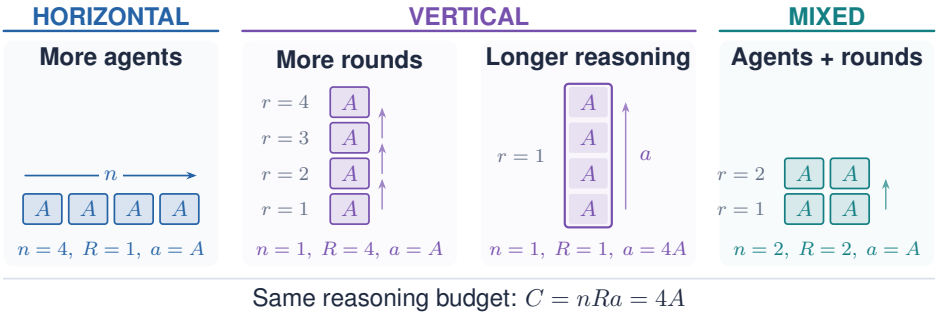


Figure 1: **Allocations at fixed reasoning cost.** All four spend $C = 4A$. Tiles denote A tokens; outlines enclose one agent in one round. Interaction and selection are separate choices.

Measure accuracy and repeated-run consistency. We compare final-answer accuracy with oracle coverage from sampling at A reasoning tokens per call and use Equation (3) to locate generation, preservation, and selection losses. A separate 48-problem repeated-run panel estimates correctness consistency, $K_x = S_x^2 + (1 - S_x)^2$, the probability that two independent runs are both correct or both incorrect. Because stable errors also increase K_x, we interpret it only alongside accuracy. A further 200 unseen problems test depth–width allocations; this panel has no fresh solver replications for estimating consistency.

Test simple allocation guidance. We fit a regularized logistic model to development outcomes labeled for correctness. For each new problem, it uses five diagnostic attempts with A reasoning tokens each, without their correctness labels, to choose among 15 one-round depth–width allocations up to $16A$. It prioritizes accuracy, then consistency within tolerance bands, then latency. We compare its choices with fixed depth and LLM allocators; protocols and cost accounting appear in Appendices A and E.

5 REGIMES AND THE RETURNS TO DEEPER REASONING

We first identify model-relative regimes, then compare allocations at fixed reasoning budgets and test whether the depth advantage persists on new problems.

5.1 SUCCESS AT THE REFERENCE ALLOWANCE SEPARATES GENERATION AND SELECTION FAILURES

With 25 attempts per problem, each using A reasoning tokens, estimated oracle coverage reaches 60.3%, but voting returns only 43.3% correct answers (averaged equally across benchmarks). Figure 2 separates this gap into four descriptive regimes. Problems with no observed success supply no correct candidate; rare-success problems have correct candidates, yet voting solves only 13.3% of them. Intermediate and frequent success yield 76.9% and 100% voting accuracy. Additional candidates can therefore leave either generation or selection as the limiting factor. Regime membership describes the model at this allowance, rather than intrinsic solvability or benchmark identity.

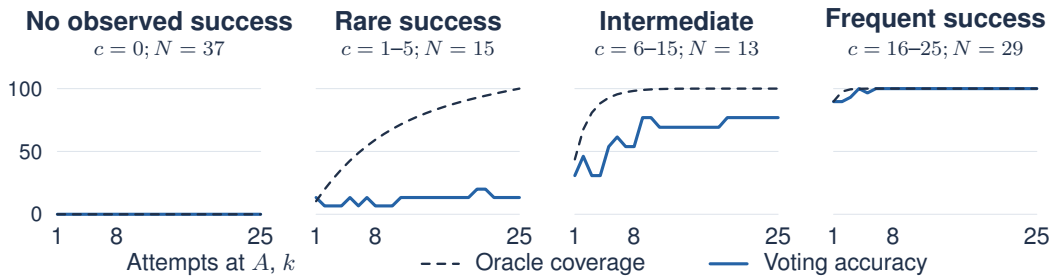


Figure 2: **The same sampling budget encounters different bottlenecks.** Coverage and accuracy (%) by correct answers c in a 25-attempt reference set; each draw uses 1,024 reasoning tokens. Dashed curves average oracle coverage over subsets; solid curves use observed voting prefixes. The reference sample sets also define the strata, so positive-stratum coverage at 25 is definitional.

5.2 At equal budgets, deeper reasoning wins in all three harder regimes

At $C = 16A = 16{,}384$ reasoning tokens, one deep call reaches 78.5% macro accuracy, versus 42.2% for voting, 52.8% for private revision, and 48.6% for debate (Figure 3). Depth leads in the no-, rare-, and intermediate-success regimes and solves $23/37$ problems with no correct response among the reference attempts. Its accuracy exceeds estimated equal-budget oracle coverage for sampling at A reasoning tokens per call (58.0%) by 20.5 percentage points. The advantage therefore extends beyond choosing among the candidates produced by those reference attempts.

In the frequent-success regime, voting already solves all 29 problems at $4A$, one-quarter of the $16A$ budget. At $16A$, every method ties. Intermediate-success rankings also reverse at some smaller budgets (Appendix D). **Allocate more reasoning tokens to one call when attempts at the reference allowance are unreliable; use inexpensive voting when they are already reliable.**

Allocation	No success	Rare	Intermediate	Frequent	Macro
16 calls at A + vote	0.0	13.3	69.2	100.0	42.2
16 private revision rounds	13.5	33.3	84.6	100.0	52.8
4 agents, 4 private rounds	5.4	20.0	69.2	100.0	45.5
4 agents, 4 debate rounds	10.8	33.3	61.5	100.0	48.6
One deep reasoning call	**62.2**	**60.0**	**100.0**	**100.0**	**78.5**

Depth: 78.5% accuracy; A-token sampling: 58.0% oracle coverage (+20.5 points)

Figure 3: **Depth leads in the three harder regimes; every method ties in frequent success.** Accuracy (%) on 94 fully graded problems at identical reasoning expenditure. Regime columns weight items equally; macro weights benchmarks equally. The oracle for sampling at A reasoning tokens per call is a diagnostic reference, not a deployable selector. Full uncertainty and budget sweeps are in Appendix D.

5.3 Depth also wins on unseen problems and mixed allocations

On 200 unseen problems, depth achieves 70.5% accuracy versus 65.0% for the strongest equal-budget split and 49.0% for voting over calls with A reasoning tokens each (Figure 4a). Against the four-call allocation selected on development data, depth gains 12.5 points (95% interval $[7.0, 18.4]$; Holm-adjusted $p = 0.0008$). Error analysis on the original problem set shows why more calls need not help: voting loses available correctness through selection, while revision loses earlier discoveries (Figure 4b). Confidence intervals for matched debate versus private revision comparisons include zero in every regime (Appendix D).

The separate repeated-run evaluation measures both accuracy and consistency: fixed depth attains the highest observed accuracy across that evaluation set (78.5%) with 88.6% correctness consistency

(Table 3). These pooled results show that high accuracy and consistency do not require additional agents; they do not establish a consistency advantage within every regime.

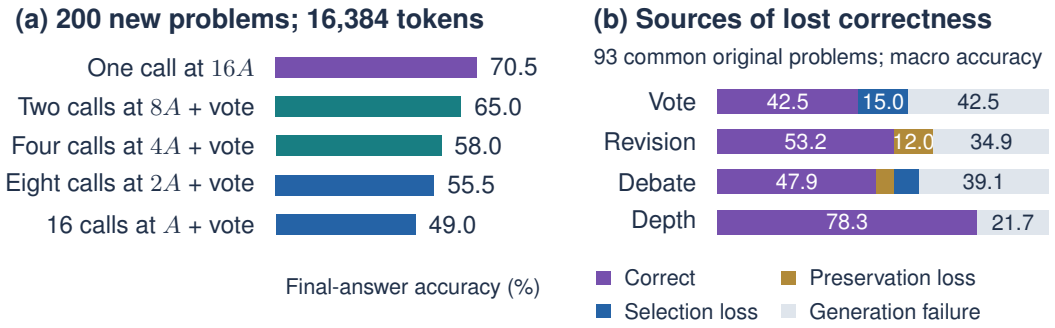


Figure 4: **Depth retains its advantage on unseen problems; additional answers can be lost during revision or selection. (a)** Independent depth–width allocations on unseen problems; benchmark weights are $(.10, .10, .50, .30)$ for AIME, HMMT, Omni-MATH, and GPQA. **(b)** Disjoint outcome components at $16A$ on the original common cohort. Single-call preservation and selection losses are zero by construction; detailed values are in Table 2.

6 Simple Allocation Learning Supports the Depth Finding

This supporting experiment tests whether a simple fitted rule recovers the preference for depth and whether inexpensive diagnostics can guide allocation within recurring problem families.

A small fitted model recovers a strong allocation. Regularized logistic regression, fitted on 96 development problems and 1,192 allocation outcomes labeled for correctness, uses allocation features and five diagnostic attempts at A reasoning tokens each, without their correctness labels, to choose among 15 depth–width options up to $16A$. These diagnostic attempts are ordinary model calls; fitting also requires responses labeled for correctness at several reasoning allowances. On 48 repeated-run problems, each evaluated with a predictor fitted without that problem's outcomes, the rule stays within one percentage point of fixed depth in accuracy and correctness consistency, and within 1% in latency (Figure 5a). It therefore recovers a near-best observed tradeoff using a simple decision rule.

Its choices reinforce the value of depth. On 200 new problems, the rule chooses depth on $151/200$ items and reaches 70.0% accuracy, versus fixed depth's 70.5%, with 7% lower mean modeled latency including probes. With the same set of solver allocations and an accuracy-first objective, LLM allocators reach 61.0% without probes and 59.3% with them. The statistical rule gains 9.0 points ($[5.4, 12.6]$) and 10.7 points ($[7.3, 14.2]$), respectively. The LLM choices are faster but miss the specified accuracy tolerance. The strongest inference is that calibration largely recovers the already effective deep allocation; greater orchestration complexity is unnecessary for that result.

Limited diagnostics suggest a practical research direction. A modest development panel and five diagnostic attempts at the reference reasoning allowance provide useful guidance on unseen items from the same problem families. This is directional evidence for learning economical allocation rules from representative evaluations. It does not establish how little calibration is sufficient, transfer to new models or families, or a globally optimal policy. The comparison is exploratory, probes add $5A$ to solver cost, and new-problem consistency is unmeasured. Thus, "near-best" refers to the observed accuracy–consistency–latency tradeoff among tested choices, not equal total compute or a certified optimum.

Allocation rule	(a) 48 repeated-run problems Accuracy (%)	Consistency (%)	Latency (s)	(b) 200 new problems Accuracy (%)	Latency (s)
Fixed deep call	78.5	88.6	277	70.5	290
Statistical allocator	77.7	87.7	279	70.0	270
LLM: problem only	67.0	78.5	155	61.0	158
LLM: five probes	69.4	82.7	202	59.3	178

Figure 5: **Statistical allocation recovers a near-best observed tradeoff.** **(a)** Each problem is excluded from fitting its predictor; diagnostic attempts are excluded from scoring. **(b)** Unseen problems with Figure 4a weights; consistency is unmeasured. Latency includes probes and LLM decisions. The set of solver allocations is shared; total costs differ. Comparisons are exploratory.

7 Discussion

The regime framework makes competing designs comparable under a common budget. At $16A$, deeper individual reasoning is the strongest tested allocation in the three harder regimes; frequent success identifies when less expensive voting is sufficient.

Regimes explain why apparently competing prescriptions can coexist. Sampling, voting, revision, and debate address different obstacles to correctness. In rare success, every complete set of reference attempts contains a correct answer, but voting returns correctness on only 13.3% of problems; in frequent success, voting reaches 100% (Figure 2). Regimes connect this difference to model-relative difficulty and the frequency of competing answers, extending work on sampling–selection gaps and difficulty-dependent allocation (Brown et al., 2024; Chen et al., 2024; Snell et al., 2025).

Depth improves both access to correctness and the answer returned. At $16A$, depth exceeds estimated oracle coverage for sampling at A reasoning tokens per call by 20.5 points and solves $23/37$ problems with no correct response generated with A reasoning tokens (Figure 3). Its advantage extends beyond selecting among the reference candidates. It also reaches 60.0% in rare success and 100% in intermediate success, where generation at the reference reasoning allowance already discovers correct answers that voting can miss. A generation–selection gap can therefore motivate longer reasoning as well as better selection. A single recorded answer has no separate selection step; its zero measured selection loss does not explain the advantage. The evidence for deeper reasoning is its higher final-answer accuracy, sustained against tested allocations across multiple calls on unseen problems.

More agents must improve on the reasoning they replace. Revision can discover correct answers and later lose them, while the matched comparisons show no consistent advantage from access to other agents' responses. Additional agents, rounds, and messages must therefore justify their opportunity cost against depth. The repeated-run results additionally show high accuracy and consistency without multiple agents. The prescription remains conditional. Private revision can lead at smaller budgets in intermediate success, and frequent-success voting reaches the highest observed accuracy at $4A$ with less sequential reasoning when calls run in parallel. Depth is an essential comparator, not a universal winner at every budget.

Simple allocation rules can recover the preference for depth. The statistical rule predominantly chooses depth and approaches its accuracy, while outperforming the tested LLM allocators. Fixed depth remains at least as accurate, showing that adaptive orchestration is unnecessary for the central gain. The experiment instead motivates testing whether a few diagnostic attempts and calibration reused across a problem family can retain accuracy while reducing computation or latency. Prospective comparisons must include both calibration and deployment costs.

8 LIMITATIONS

We study one 27B Qwen checkpoint, four reasoning benchmarks, and selected configurations; reliable budget enforcement constrained model coverage. Main sweeps stop at $25A$, so neither depth's eventual ceiling nor a global optimum over allocations is established. Regimes depend on finite reference samples, the model, and the reasoning protocol. The three-regime accuracy advantage is specific to the $16A$ budget; repeated-run consistency is evaluated on a separate selected panel, not established within every regime. Budgets match reasoning tokens rather than total FLOPs or deployment cost, since histories, peer messages, and answer-generation calls incur different costs.

Allocation learning covers one-round depth–width choices and needs offline calibration using correctness labels before five diagnostic attempts without labels can guide a new item. The expanded selector comparison was specified after evaluation outcomes were visible, so the comparison is exploratory. The 200-problem panel cannot estimate per-problem correctness consistency; the repeated-run panel is small and selected by baseline success. Modeled latency assumes parallel starts and excludes queueing, while diagnostic and decision overhead prevent an equal-total-budget interpretation. Broader prospective studies must establish transfer, calibration requirements, and net savings on real tasks.

9 CONCLUSION

In a controlled setting, we characterize how a fixed inference budget should be divided among deeper reasoning, more agents, and revision. Our regime-based framework makes this a measurable question about a model and a problem family. At $16{,}384$ reasoning tokens, deeper reasoning within one call achieves the highest observed accuracy in all three harder regimes and exceeds estimated equal-budget oracle coverage from sampling at A tokens per call by 20.5 points. Its advantage persists against tested allocations across multiple calls on unseen problems, while a separate repeated-run evaluation combines peak observed accuracy with high correctness consistency. When attempts at the reference allowance are frequently correct, voting reaches the highest observed accuracy with one-quarter of the budget. The statistical allocation experiment supports this central finding by predominantly selecting deeper reasoning and approaching its accuracy.

These findings bring order to agentic-system design by linking allocation choices to a model's behavior within defined problem subfamilies. Deeper individual reasoning becomes an essential baseline: additional agents and interaction must improve final-answer accuracy enough to justify the reasoning they displace. Practitioners with a recurring real-world problem family and a well-defined evaluation suite can apply this approach directly: characterize success at a reference reasoning allowance, compare deeper reasoning with alternative allocations under operational budgets, and validate the resulting choice on held-out tasks. This turns competing architectural prescriptions into testable decisions about accuracy, consistency, latency, and cost. The practical opportunity is substantial: in the studied harder regimes, better use of the existing model's reasoning budget outperforms the tested increases in agent count and interaction.

Future work should establish how these regimes and the advantage of deeper reasoning change across model families, training methods, and real-world tasks. Varying the reference reasoning allowance would show how regime labels and allocation preferences change; extending the compute range would identify where the gains from depth saturate and when independent attempts, verification, or communication become more effective. Studying how reasoning progresses within each regime could explain why additional depth recovers correctness and guide extensions to tasks involving tools and external feedback. Our preliminary allocation results also motivate a complementary direction: testing whether limited calibration data and easily generated diagnostic attempts can support near-optimal allocation decisions across broader settings. If validated with full cost accounting, such models could provide inexpensive inference-time orchestration grounded in measured regime behavior.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, to assist with conceptual framing and hypothesis refinement, feedback on experimental design, interpretation and presentation of results, and drafting and revision of mathematical exposition. We also used these tools for literature synthesis, manuscript drafting and editing, figure and LaTeX formatting, and citation checks. The language models evaluated as solvers and allocation decision makers are components of the experiments described in Section 4 and Appendices A and E. The human authors are responsible for validating all AI-assisted material and take full responsibility for the final text, claims, and artifacts in this paper.

REPRODUCIBILITY STATEMENT

Section 4 summarizes the experimental design. Appendix A specifies problem selection, allocation configurations, budget enforcement, grading, and cost accounting. Appendix B provides assumptions and proofs. Appendices C and D report additional comparisons and uncertainty, and Appendix E describes allocation prediction, evaluation splits, tuning, and statistical analysis.

REFERENCES

Yong Yi Bay and Kathleen A. Yearick. When more sampling hurts: The modal ceiling and correlation ceiling of test-time scaling, 2026. URL https://arxiv.org/abs/2606.28661.

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL https://arxiv.org/abs/2605.00914.

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

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/51173cf34c5faac9796a47dc2fdd3a71-Abstract-Conference.html.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

Mehul Damani, Idan Shenfeld, Andi Peng, Andreea Bobu, and Jacob Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ff414825df833edb8b1839e3d5d495e9-Abstract-Conference.html.

Yihe Dong and Boris Shigida. When independent sampling outperforms agentic reasoning. arXiv preprint arXiv:2605.08478, 2026. URL https://arxiv.org/abs/2605.08478.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Austin Feng, Marius Alonso, Ambroise Odonnat, Vasilii Feofanov, and Ievgen Redko. Optimal self-consistency for efficient reasoning with large language models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.12309.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01798.

Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. Towards a science of scaling agent systems. arXiv preprint arXiv:2512.08296, 2025. URL https://arxiv.org/abs/2512.08296v3. Version 3, revised April 2026.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.17651.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Jon Saad-Falcon, E. Kelly Buchanan, Mayee F. Chen, Tzu-Heng Huang, Brendan McLaughlin, Tanvir Bhathal, Shang Zhu, Ben Athiwaratkun, Frederic Sala, Scott Linderman, Azalia Mirhoseini, and Christopher Ré. Shrinking the generation-verification gap with weak verifiers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.18203.

Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (Laws)? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53132–53176. PMLR, 2025. URL https://proceedings.mlr.press/v267/schaeffer25a.html.

Nishad Singhi, Hritik Bansal, Arian Hosseini, Aditya Grover, Kai-Wei Chang, Marcus Rohrbach, and Anna Rohrbach. When to solve, when to verify: Compute-optimal problem solving and generative verification for LLM reasoning. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2504.01005.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905. PMLR, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

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. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/95591b7557edea46ae1d90b0e9e825c5-Abstract-Conference.html.

Varun Sunkaraneni, Pierfrancesco Beneventano, Riccardo Neumarker, Tomaso Poggio, and Tomer Galanti. Agentic systems as boosting weak reasoning models. arXiv preprint arXiv:2605.14163, 2026. URL https://arxiv.org/abs/2605.14163.

Fali Wang, Jihai Chen, Shuhua Yang, Runxue Bao, Tianxiang Zhao, Zhiwei Zhang, Xianfeng Tang, Hui Liu, Qi He, and Suhang Wang. Generalizing test-time compute-optimal scaling as an optimizable graph. arXiv preprint arXiv:2511.00086, 2025. URL https://arxiv.org/abs/2511.00086.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. 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://arxiv.org/abs/2203.11171.

Florian Valentin Wunderlich, Lars Benedikt Kaesberg, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Multi-agent reasoning improves compute efficiency: Pareto-optimal test-time scaling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1–14. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-srw.1. URL https://aclanthology.org/2026.acl-srw.1/.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. arXiv preprint arXiv:2505.22960, 2025. URL https://arxiv.org/abs/2505.22960.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, Changwang Zhang, Chenghua Lin, Jun Wang, Ge Zhang, and Wangchunshu Zhou. Scaling test-time compute for LLM agents. arXiv preprint arXiv:2506.12928, 2025. URL https://arxiv.org/abs/2506.12928.

A Extended methods

This appendix supplies the detailed design and execution procedures summarized in Section 4.

A.1 Reference setting and problem selection

The main panel contains 96 problems, with 24 from each of four benchmark families: AIME 2025/2026, HMMT February 2025/2026, Omni-MATH at difficulty at least seven, and GPQA-Diamond. Problems follow a frozen stratified sampling procedure. Prose-only or unparseable Omni-MATH references are excluded before generation. Eight disjoint development problems per family support calibration, with a pre-specified acceptance range of 0.15–0.85 for accuracy at the base reasoning allowance; all four families satisfy this criterion.

We evaluate one model, Qwen3.8-27B (`qwen/qwen3.8-27b`), using a reference allowance of $A = 1{,}024$ reasoning tokens. Each problem has 25 independent reference samples and one realization at each deeper allowance; the depth-one baseline reuses the first reference sample. These sample sets characterize model-relative sampling behavior at A, rather than intrinsic problem difficulty or impossibility.

A.2 Compute allocation and interaction

A uniform allocation (n, R, a) uses n agents, R total rounds, and a reasoning tokens per agent per round. We write π for the complete procedure, including these coordinates, the interaction protocol, and final selection. Its reasoning budget and equivalent number of reference attempts are

$$C = nRa = kA, \qquad k = \frac{nRa}{A}.$$

The matched reference spends C on k independent attempts of A tokens each. This matches reasoning-token expenditure: context processing, answer generation, and coordination can differ across allocations, so equal C does not imply equal total inference cost. Rounds include initialization, making four revisions five total rounds.

At budget kA, the basic scaling directions are independent width $(k, 1, A)$, private revision $(1, k, A)$, and within-call depth $(1, 1, kA)$. Mixed allocations distribute compute across agents and rounds: $(4, 4, A)$ and $(5, 5, A)$ spend $16A$ and $25A$. Figure 1 illustrates the coordinates at a common budget.

Table 1 specifies the 96 configurations. Private agents revise their own answers without peer exposure; communicating teams exchange responses between rounds. Their matched comparison holds the allocation and final voting rule fixed. Plurality returns the most frequent answer class, which need not have an absolute majority. The $R = 1$ team configurations are initialization-only baselines.

Family	(n, R, a)	Settings (count)	Final answer rule
Independent sampling	$(k, 1, A)$	$k = 1, \ldots, 25$ (25)	Plurality
Private revision	$(1, k, A)$	$k = 1, \ldots, 25$ (25)	Last answer
Private teams	(n, R, A)	$n = 2, \ldots, 5;\ R = 1, \ldots, 5$ (20)	Final-round plurality
Communicating teams	(n, R, A)	$n = 2, \ldots, 5;\ R = 1, \ldots, 5$ (20)	Final-round plurality
Within-call depth	$(1, 1, mA)$	$m \in \{1, 2, 4, 8, 16, 25\}$ (6)	Single answer

Table 1: **Allocation and final-answer rules.** Counts include initialization-only configurations; shared baselines do not constitute independent repetitions. Interaction and final selection are specified separately from the allocation coordinates.

Principal matched-budget comparisons use $C/A \in \{2, 4, 8, 16, 25\}$. After completed round t, a uniform team has spent $C_t = nta$; its matched sampling reference therefore uses $k_t = nta/A$ attempts. Comparing rounds at their cumulative expenditure charges revision and communication for reasoning that could instead produce fresh candidates.

A.3 Sampling diagnostics and error decomposition

Final-answer accuracy S is the fraction of correct final answers. Current coverage H_{cur} is the probability that the selection pool contains a correct answer: all sampled candidates, final-round

team answers, or the last private-chain answer. Archive coverage H_{arc} considers every generated candidate, independently of whether the final readout can return it. On the same scored problems, selection restricted to the current pool gives

$$1 - S = (1 - H_{\mathrm{arc}}) + (H_{\mathrm{arc}} - H_{\mathrm{cur}}) + (H_{\mathrm{cur}} - S).$$

The terms distinguish generation failure, loss of generated correctness from the selectable pool, and selection failure. Archive coverage records whether correctness was ever generated; it does not substitute an archive-based selector for the native readout. For one call, both coverages equal accuracy; for independent sampling, archive and current coverage coincide. A private chain has $H_{\mathrm{cur}} = S$, so losing an earlier correct answer appears as preservation failure. The decomposition describes outcomes under these specified pools; revision and communication can affect both generation and final availability.

For a fully graded reference sample set with c_x correct answers on problem x, estimated oracle coverage over $k \leq 25$ attempts is

$$\widehat{P}_{k,x}(A) = 1 - \frac{\binom{25-c_x}{k}}{\binom{25}{k}}. \tag{4}$$

This estimator averages over candidate subsets. For an observed k-sample prefix, its coverage and voting accuracy instead describe the same realized pool. At $C = kA$, define the oracle coverage for sampling at A reasoning tokens per call and signed accuracy gap

$$O(C) = \mathrm{pass@}k(A), \qquad D_\pi(C) = O(C) - S_\pi(C). \tag{5}$$

Positive D_π indicates a shortfall relative to the reference oracle; negative D_π indicates exceeding it through another allocation. This cross-allocation comparison is distinct from the within-pool selection gap $H_{\mathrm{cur}} - S$. At round t, $D_{\pi,t} = \mathrm{pass@}k_t(A) - S_{\pi,t}$.

We estimate baseline success as $\widehat{p}_A(x) = c_x/25$ and stratify the 94 fully graded banks into no, rare, intermediate, and frequent success: $c_x = 0$, 1–5, 6–15, and 16–25, respectively. The same banks define strata and sampling baselines, making these descriptive success regimes. Coverage and voting jointly characterize their bottlenecks: equal success frequencies can yield different selection gaps when incorrect answers concentrate on different alternatives. Zero observed success does not establish zero underlying success probability, and the strata alone do not prescribe an allocation. A prospective allocation rule requires separate calibration and held-out evaluation.

A.4 Execution and analysis

Calls use OpenRouter with temperature 1.0, top-p 0.95, top-k 20, and provider fallbacks and response caching disabled. Provider assignment is fixed within each problem across configurations: Mancer serves 58 problems and Parasail 38. To compare consumed reasoning rather than unused allowances, each generation has separate THINK and ANSWER calls. THINK applies logit bias -100 to the end-of-reasoning token and two end-of-sequence tokens, with its output limit equal to the assigned allowance. ANSWER receives the original prompt and reasoning, closes the reasoning block, and produces a visible answer under a 4,096-token cap. Provider checks verify bias support and exact-length generation; prematurely terminated THINK calls are retried. All 13,344 accepted main-run generations consume their assigned reasoning allowance. The comparison therefore concerns this enforced generation protocol.

Math-Verify grades mathematical answers; GPQA uses exact option matching after a fixed per-problem option shuffle. Invalid final answers count as incorrect. Grading timeouts remain unknown and are excluded from scored comparisons. Main sampling and allocation comparisons use the 94 fully graded reference sample sets. The joint error decomposition at $16A$ (Table 2) additionally requires fully graded current and archive pools for every compared procedure, leaving 93 common problems (24 AIME, 22 HMMT, 23 Omni-MATH, and 24 GPQA). It uses archive coverage and the native current-pool answer, never the answer selected from the archive. Macro averages weight benchmark families equally; regime-level results weight problems equally. Accuracy bars in the time comparison use 95% Wilson intervals over problems. Peer-minus-private differences use 95% paired item-bootstrap intervals from 10,000 resamples; these intervals describe variation across problems rather than repeated independent runs.

The time comparison uses modeled serial reasoning length rather than measured wall-clock duration. Assuming sufficient parallel capacity, latency proportional to reasoning length, and negligible combination overhead, normalized time is

$$\ell(\pi) = \frac{Ra}{A} = \frac{C}{nA}. \tag{6}$$

At $C = 16A$, this equals one for independent sampling, four for four-agent teams, and sixteen for either single-agent allocation. It distinguishes sequential dependence at equal reasoning budgets without equating the proxy with total execution latency.

B PROOFS AND FURTHER THEORETICAL RESULTS

This appendix develops the conditional comparisons introduced in Section 3 and supplies their assumptions, proofs, and extensions. All probabilities condition on a fixed problem, model, and response protocol. Population-level quantities average the resulting per-problem expressions; in general, inserting a benchmark-average success rate into a per-problem product is invalid. A candidate classified as incorrect includes an invalid output. The independent voting analysis assumes a fixed finite answer partition with one correct class; the decomposition and interaction results require only a correctness event. The grouping is an analytical convention and does not grant the deployed selector access to correctness labels.

The probability and concentration arguments are standard. Their role here is to distinguish allocation-dependent discovery, preservation, and selection, and to state the conditions under which each comparison applies. Counterfactual predictions based on simplified transition models additionally require the stated invariance assumptions. None of these assumptions follows from high empirical coverage alone.

B.1 ACCURACY REQUIRES GENERATION, PRESERVATION, AND SELECTION

Let $L_{\mathrm{pre}} = H_{\mathrm{arc}} - H_{\mathrm{cur}}$ and $L_{\mathrm{sel}} = H_{\mathrm{cur}} - S$ denote the preservation and selection losses defined in Preliminaries. For any two allocations, Equation (3) gives

$$\begin{aligned} S_\pi - S_\sigma = H_{\mathrm{arc},\pi} - H_{\mathrm{arc},\sigma} \\ - (L_{\mathrm{pre},\pi} - L_{\mathrm{pre},\sigma}) - (L_{\mathrm{sel},\pi} - L_{\mathrm{sel},\sigma}). \end{aligned} \tag{7}$$

This identity allows dependent and communicating agents, provided each output belongs to its final candidate pool. It explains why revision can discover more correct answers yet return fewer, and why communication must be judged by its net accuracy gain. A single recorded answer has zero pool-relative preservation and selection losses by construction; this says nothing about intermediate reasoning. Section 5.3 applies this accounting to the observed allocations, without treating its components as independent causal effects.

B.2 WHEN DEPTH EXCEEDS ORACLE COVERAGE AT THE REFERENCE ALLOWANCE

Let $p(a)$ be a fresh call's correctness probability at reasoning allowance a. At budget $C = kA$, k independent calls with A reasoning tokens each have oracle coverage $1 - [1 - p(A)]^k$. Thus one depth-kA call outperforms even perfect selection from that pool exactly when

$$p(kA) > 1 - [1 - p(A)]^k \quad \Longleftrightarrow \quad h(kA) > kh(A), \tag{8}$$

where $h(a) = -\log[1 - p(a)]$ is log-failure reduction (Dong & Shigida, 2026). Depth must remove more generation failure than the fresh attempts it replaces. This criterion motivates the oracle comparison in Section 5.2: exceeding coverage at the reference reasoning allowance demonstrates an advantage beyond selecting among solutions generated at the reference reasoning allowance. Beating voting alone requires less. The criterion depends on the response to depth, not just success at the reference reasoning allowance, and does not imply that recovered solutions were impossible under generation at the reference reasoning allowance. Appendix B.6 extends it to mixed depths.

B.3 WHEN RELIABLE ANSWERS AT THE REFERENCE ALLOWANCE MAKE VOTING SUFFICIENT

For independent draws at a fixed allowance, assume a finite answer partition with one correct class, matching the selector's grouping. As the number of draws grows, plurality accuracy approaches one when correctness is the unique most probable class. If an incorrect class is strictly more probable, voting accuracy instead approaches zero, even as coverage approaches one whenever $p(a) > 0$ (Feng et al., 2026; Bay & Yearick, 2026). Coverage therefore measures discovery; reliable voting additionally requires favorable support relative to competing errors. This distinction explains why rare and frequent success at the reference reasoning allowance can give different aggregation returns (Section 5.1). Where voting already achieves comparable accuracy, parallel sampling can reduce sequential reasoning (Section 5.2).

The four observed regimes are descriptive, not proven allocation boundaries. Identical diagnostic attempts at the reference reasoning allowance can accompany opposite intervention responses; prospective allocation therefore requires calibrated response information and accounting for probe costs (Appendix B.11). The theory supplies conditional explanations for the measured advantages, not a universal guarantee that depth or aggregation wins.

B.4 Comparing allocations through their error components

For a policy π, define the preservation and selection losses by

$$L_{\mathrm{pre},\pi} = H_{\mathrm{arc},\pi} - H_{\mathrm{cur},\pi}, \qquad L_{\mathrm{sel},\pi} = H_{\mathrm{cur},\pi} - S_\pi.$$

Both are nonnegative because the final selector returns a candidate from its current pool, which is contained in the generated archive. Hence $S_\pi = H_{\mathrm{arc},\pi} - L_{\mathrm{pre},\pi} - L_{\mathrm{sel},\pi}$. Subtracting this equality for policies π and σ proves Equation (7). The equality also holds after averaging over the same problems with the same weights. It attributes the observed accuracy difference to differences in archive coverage and the two pool-relative losses; it does not identify the effect of changing one mechanism while holding the others fixed.

For independent sampling, $L_{\mathrm{pre}} = 0$. For an allocation that generates only one recorded candidate, both losses are zero by construction. Consequently, depth's accuracy advantage over independent voting equals its archive-coverage advantage plus the independent pool's selection loss. Intermediate hypotheses within a reasoning trace are not separately graded candidates in this accounting, so zero pool-relative loss does not establish perfect preservation or selection inside a long call.

B.5 Discovery, availability, and selection

We fix a problem and model throughout; probabilities concern inference randomness. A budget-feasible policy generates at most T batches, each containing one or more candidates. Actions may depend on observable history, and outputs need not be independent. Early stopping is padded with empty batches. Any newly synthesized candidate answer, including one produced during verification, counts as a generation. The final selector returns a member of its terminal candidate pool.

Let τ be the first batch containing a correct candidate, with $\tau = \infty$ if none does. Define its first-discovery hazards and cohort masses by

$$d_t = \Pr(\tau = t \mid \tau \geq t), \qquad w_t = d_t \prod_{u<t}(1 - d_u).$$

Let r_t be terminal-pool coverage conditional on $\tau = t$, and let η_t be the probability of correct selection conditional on that coverage and $\tau = t$. These quantities are properties of the policy's trajectory distribution; computing the decomposition does not require the policy to observe correctness during execution.

Proposition B.1 (Discovery, availability, and selection)**.** *For any adaptive policy satisfying the specified pool convention, including dependent and communicating agents,*

$$\begin{aligned} H_{\mathrm{arc}} &= 1 - \prod_{t=1}^{T}(1 - d_t) = \sum_{t=1}^{T} w_t, \\ H_{\mathrm{cur}} &= \sum_{t=1}^{T} w_t r_t, \qquad\qquad S = \sum_{t=1}^{T} w_t r_t \eta_t. \end{aligned} \tag{9}$$

Proof. Write C_T for terminal-pool coverage and Y_T for correctness of the returned answer. Thus

$$\{Y_T = 1\} \subseteq \{C_T = 1\} \subseteq \{\tau \leq T\}.$$

For $d_t = \Pr(\tau = t \mid \tau \geq t)$, the probability chain rule gives

$$\Pr(\tau > t) = \prod_{u \leq t}(1 - d_u), \tag{10}$$

$$\Pr(\tau = t) = d_t \prod_{u<t}(1 - d_u) = w_t. \tag{11}$$

On null conditioning events, define conditional probabilities arbitrarily; the corresponding w_t is zero. Since the events $\{\tau = t\}$ partition archive success, total probability yields

$$H_{\mathrm{arc}} = \sum_{t=1}^{T} w_t, \qquad H_{\mathrm{cur}} = \sum_{t=1}^{T} w_t r_t, \qquad S = \sum_{t=1}^{T} w_t r_t \eta_t, \tag{12}$$

where $r_t = \Pr(C_T = 1 \mid \tau = t)$ and $\eta_t = \Pr(Y_T = 1 \mid C_T = 1, \tau = t)$. In particular,

$$1 - S = \prod_{t=1}^{T}(1 - d_t) + \sum_{t=1}^{T} w_t(1 - r_t) + \sum_{t=1}^{T} w_t r_t(1 - \eta_t). \tag{13}$$

□

These are probability identities, requiring neither a Markov approximation nor access to correctness during execution. The hazards condition on *no earlier discovery*; unconditional per-round correctness rates cannot replace them. Final availability r_t includes paths that lose correctness and subsequently rediscover it. For n independent initial calls, $d_1 = 1 - [1 - p(a)]^n$. Providing the full archive to the final selector makes $r_t = 1$ for every positive-mass discovery cohort, but does not make $\eta_t = 1$.

Corollary B.2 (The discovery opportunity cost of interaction)**.** *Suppose a team uses n calls of allowance a in each of R total rounds, with independent initial answers of success probability $p(a) < 1$. For its round-level first-discovery hazards d_t,*

$$H_{\mathrm{arc}} = 1 - [1 - p(a)]^n \prod_{t=2}^{R}(1 - d_t). \tag{14}$$

At the same reasoning expenditure $C = nRa$, its archive coverage exceeds that of nR independent calls exactly when

$$\sum_{t=2}^{R} -\log(1 - d_t) > n(R - 1)h(a), \tag{15}$$

where $h(a) = -\log[1 - p(a)]$.

Proof. Initialization fails with probability $[1 - p(a)]^n$. The chain rule for all remaining first-discovery events proves Equation (14). The independent reference fails with probability $[1 - p(a)]^{nR}$. Cancel the positive initialization-failure factor and take negative logarithms. A hazard $d_t = 1$ gives infinite log-failure reduction and perfect archive coverage; $R = 1$ gives equal allocations. If $p(a) = 1$, both archives are already perfectly covered and strict improvement is impossible. □

The right side of Equation (15) is the discovery supplied by the fresh calls displaced by interaction. Exceeding it establishes greater generation coverage, but does not remove the availability and selection losses in Proposition B.1.

Proposition B.3 (A sharp preservation guarantee)**.** *Conditional on $\tau = t$, enumerate all subsequent pool-update stages by $j = 1, \ldots, m_t$, including pruning immediately after batch t is generated. Assume a finite deterministic stage bound and pad early stopping with identity updates so that m_t is deterministic. Suppose the probability of losing all available correct candidates at stage j, conditional on uninterrupted availability up to that stage, is at most $\bar{e}_{t,j}$. Suppose terminal selection succeeds with probability at least $\underline{\eta}_t$ conditional on uninterrupted availability. Then*

$$S \geq \sum_{t=1}^{T} w_t \underline{\eta}_t \prod_{j=1}^{m_t}(1 - \bar{e}_{t,j}). \tag{16}$$

The bound is sharp given these constraints.

Proof. For bookkeeping, the pool immediately after generation includes the newly generated batch before any discard; hence it is covered when $\tau = t$. Its immediate pruning is the first update, rather than an unaccounted loss. Let $E_{t,j}$ denote continued coverage through update j. Repeated conditioning gives

$$\Pr(E_{t,m_t} \mid \tau = t) = \prod_{j=1}^{m_t} \Pr(E_{t,j} \mid E_{t,j-1}, \tau = t) \geq \prod_{j=1}^{m_t} (1 - \bar{e}_{t,j}).$$

Restrict the success probability to these surviving paths and apply the assumed selection bound. Summing over first-discovery cohorts proves the claim. Equality is attainable by a process with states "never discovered," "available," and "permanently lost": use the specified d_t for first discovery, lose availability with probability $\bar{e}_{t,j}$ at each subsequent update, and never rediscover after a loss. Keep an incorrect decoy available, and return the correct candidate with probability $\underline{\eta}_t$ on surviving paths. Each inequality is then an equality. □

An archive supplied to the selector ensures $r_t = 1$, but does not ensure $\eta_t = 1$. Conversely, marginal extinction estimates do not by themselves establish the conditional bounds above. If $d_t \geq \underline{d}_t$ and $r_t \eta_t \geq \underline{\kappa}$ uniformly, a simpler certificate is

$$S \geq \underline{\kappa} \left[1 - \prod_{t=1}^{T} (1 - \underline{d}_t) \right]. \tag{17}$$

Uniform recovery is essential to this substitution: increasing early discovery can expose more correct candidates to subsequent destructive updates.

B.6 Independent allocation and selection

All probabilities in this subsection condition on one problem and model. Calls are independent under this conditioning; answer classes are fixed across depths. Write $*$ for the correct class and $j \in \{1, \ldots, J\}$ for the finitely many incorrect classes. The voting arguments use standard concentration methods underlying self-consistency analyses (Feng et al., 2026; Bay & Yearick, 2026); we make their dependence on heterogeneous reasoning allowances explicit. The single-correct-class assumption applies to an implemented voting rule only when its partition merges equivalent correct answers. If correct outputs occupy several vote classes, their aggregate correctness probability cannot be substituted for one class probability in the margins below.

Consider allowances $a_1, \ldots, a_n$ with $\sum_i a_i \leq C$. Put $p_i = p(a_i)$, $q_{ij} = q_j(a_i)$, $h(a) = -\log[1 - p(a)]$, and $\Delta_j = \sum_i (p_i - q_{ij})$. The log-failure-per-cost criterion connects to prior analysis of independent allocation (Dong & Shigida, 2026); the results below additionally distinguish generation from selection over the competing answer classes.

Proposition B.4 (Generation and selection under an independent compute split)**.** *For independent calls whose outputs are all retained in the final pool,*

$$H_{\mathrm{arc}} = H_{\mathrm{cur}} = 1 - \exp\left[-\sum_{i=1}^{n} h(a_i) \right]. \tag{18}$$

If $\Delta_j > 0$ for every incorrect class, plurality selection with any tie-breaking rule satisfies

$$1 - S_{\mathrm{vote}} \leq \sum_{j=1}^{J} \exp\left(-\frac{\Delta_j^2}{2n} \right). \tag{19}$$

Proof. For independent answers $Y_1, \ldots, Y_n$, the probability of generating no correct candidate is

$$1 - H = \prod_{i=1}^{n} (1 - p_i) = \exp\left(-\sum_{i=1}^{n} h(a_i) \right).$$

This remains valid when some $p_i = 1$, with $h(a_i) = +\infty$ and $e^{-\infty} = 0$; $p_i = 0$ contributes zero hazard. For each incorrect class define

$$X_{ij} = \mathbf{1}\{Y_i = *\} - \mathbf{1}\{Y_i = j\}, \qquad \mathbb{E}X_{ij} = p_i - q_{ij}.$$

These variables are independent over i, lie in $[-1, 1]$, and sum to the difference between the correct and class-j vote counts. An incorrect plurality requires $\sum_i X_{ij} \le 0$ for at least one j, including any tie resolved against the correct answer. If $\Delta_j = \sum_i (p_i - q_{ij}) > 0$, Hoeffding's inequality gives

$$\Pr\left(\sum_i X_{ij} \le 0\right) \le \exp\left(-\frac{2\Delta_j^2}{\sum_i (1-(-1))^2}\right) = e^{-\Delta_j^2/(2n)}.$$

A union bound proves the selection statement for every tie rule. No independence across different incorrect classes is required. □

Proposition B.5 (Exact independent generation allocation)**.** *Let $\mathcal{D}$ be a finite set of positive integer allowances with $p(a) < 1$, and let C be an integer budget. Set $F(0) = 0$ and, for $b \ge 1$,*

$$F(b) = \max(\{0\} \cup \{h(a) + F(b-a) : a \in \mathcal{D},\ a \le b\}). \tag{20}$$

The greatest independent coverage is $1 - e^{-F(C)}$. If at least one allowance is affordable, let a_* maximize $h(a)/a$ over affordable allowances, write $\rho_C = h(a_*)/a_*$ and $r = C - a_* \lfloor C/a_* \rfloor$. Then

$$(C - r)\rho_C \le F(C) \le C\rho_C. \tag{21}$$

Consequently, repeating a_* is optimal when a_* divides C; otherwise mixed allowances can improve on uniform repetition.

Proof. Maximizing coverage is equivalent to maximizing the sum of hazards. Every nonempty allocation consists of a first call a and an allocation with budget at most $b - a$. Replacing that remainder by its optimum proves (20) in both directions; the zero option permits unused budget. Every feasible call satisfies $h(a) \le a\rho_C$, proving the upper bound after summation. Taking $\lfloor C/a_* \rfloor$ calls attains $(C - r)\rho_C$. If no call is affordable, $F(C) = 0$. If an affordable call has $p(a) = 1$, maximum coverage is already one, so the finite-hazard optimization is unnecessary. □

Proposition B.6 (Depth crossover and curvature)**.** *For $C = kA$ with integer k, one depth-C call has greater coverage than k depth-A calls exactly when $h(C) > kh(A)$. With a fixed number n of calls and fixed total allowance C, concavity of h makes equal allowances optimal whenever C/n is allowed. If h is convex on $[0, C]$ and $h(0) = 0$, concentrating the budget in one call maximizes coverage. Conversely, subadditivity of h on the grid of A multiples makes k depth-A calls optimal among allocations on that grid. In the differentiable relaxation with $n = C/a$, increasing a improves coverage exactly when*

$$ah'(a) > h(a). \tag{22}$$

Proof. The crossover follows by comparing summed hazards. Jensen's inequality gives $\sum_i h(a_i) \le nh(C/n)$ for concave h and total allowance C. Convexity and $h(0) = 0$ give $h(a) \le ah(a + b)/(a + b)$ and the analogous bound for b, hence $h(a) + h(b) \le h(a + b)$; iterating merges all calls. Subadditivity instead permits repeated splitting into A units. Finally, relaxed coverage is $1 - \exp[-Ch(a)/a]$, whose derivative has the sign of $ah'(a) - h(a)$. □

The relevant curvature is that of log failure, not accuracy:

$$h''(a) = \frac{p''(a)}{1 - p(a)} + \frac{p'(a)^2}{(1 - p(a))^2}.$$

For example, $p(a) = 1 - e^{-a^2}$ is concave for $a > 1/\sqrt{2}$, but $h(a) = a^2$ favors concentration. These are conditional properties of response curves, not assumptions imposed on our empirical model.

Proposition B.7 (Chernoff voting certificate and wrong-mode gap)**.** *For heterogeneous independent calls, define*

$$K_j = \sup_{\lambda \ge 0} \sum_{i=1}^{n} -\log\left(1 - p_i - q_{ij} + p_i e^{-\lambda} + q_{ij} e^{\lambda}\right). \tag{23}$$

Then $1 - S_{\text{vote}} \le \sum_j e^{-K_j}$. At a uniform depth with $p > q_j$, $K_j = nI_j$, where

$$I_j = -\log\left[1 - (\sqrt{p} - \sqrt{q_j})^2\right]. \tag{24}$$

If instead $0 < p < q_j$ for some class, $S_{\text{vote}} \le e^{-nI_j}$ and

$$H - S_{\text{vote}} \ge 1 - e^{-nh} - e^{-nI_j}. \tag{25}$$

Proof. For $\lambda \geq 0$, Markov's inequality and independence give

$$\Pr\left(\sum_i X_{ij} \leq 0\right) \leq \mathbb{E}e^{-\lambda \sum_i X_{ij}} = \prod_i (1 - p_i - q_{ij} + p_i e^{-\lambda} + q_{ij} e^{\lambda}).$$

Optimize over one common λ per class and apply the union bound. For identical calls with $p > q_j > 0$, the minimizing value is $\lambda = \frac{1}{2}\log(p/q_j)$, yielding (24); $q_j = 0$ follows by taking $\lambda \to \infty$. When $q_j > p$, correct plurality requires $\sum_i X_{ij} \geq 0$. Applying the same argument to $-X_{ij}$ gives the reversed bound. Subtract it from $H = 1 - e^{-nh}$ to obtain (25). □

Thus $n \geq \max\{\log(2/\epsilon)/h, \log(2/\epsilon)/I_j\}$ suffices for a gap of at least $1 - \epsilon$ in the strict wrong-mode regime. At equal probabilities these exponential bounds are uninformative. The uniform correct-mode certificate scales with $I(a)/a$, where $I(a) = \min_j I_j(a)$, whereas generation scales with $h(a)/a$.

Proposition B.8 (When heterogeneous depths make voting reliable)**.** *For a finite allowance menu, define*

$$v(a) = (p(a) - q_1(a), \ldots, p(a) - q_J(a)).$$

A finite allocation with every $\Delta_j > 0$ exists exactly when the convex hull of $\{v(a) : a \in \mathcal{D}\}$ intersects the strictly positive orthant. Repeating such an allocation independently m times gives

$$1 - S_{\text{vote}} \leq \sum_j \exp\left(-\frac{m\Delta_j^2}{2n}\right). \tag{26}$$

For budget fractions $w_a \geq 0$ summing to one, with integer counts $n_a = Cw_a/a$, set

$$\nu = \sum_a \frac{w_a}{a}, \qquad \kappa_j = \sum_a \frac{w_a[p(a) - q_j(a)]}{a}.$$

If all $\kappa_j > 0$, then $1 - S_{\text{vote}} \leq \sum_j e^{-C\kappa_j^2/(2\nu)}$.

Proof. Normalizing the call counts expresses $(\Delta_1, \ldots, \Delta_J)/n$ as a convex combination of the $v(a)$. Conversely, a strictly positive convex combination remains positive after a sufficiently small rational approximation of its weights. Clearing denominators produces finite integer call counts. Repetition multiplies the margins by m and the call count by m, proving (26) by Proposition B.4. Substituting $n = C\nu$ and $\Delta_j = C\kappa_j$ proves the budget-fraction statement. □

For a concrete counterexample, two depths may have distributions $(0.4, 0.5, 0.1)$ and $(0.4, 0.1, 0.5)$ over (correct, wrong 1, wrong 2). Each depth alone converges to a wrong plurality. Equal call counts instead average to $(0.4, 0.3, 0.3)$, giving reliable voting at cost $a_1 + a_2$ per pair. This constructed example isolates a possibility, not an empirical finding. In the divisible relaxation, generation hazard per cost is the linear objective $\sum_a w_a h(a)/a$ and admits a pure-depth maximizer; the selection certificate depends on the entire error profile and can favor mixing. Zero-margin boundary cases require separate treatment and are outside this strict-margin characterization.

B.7 Verification and the evidence needed for rare correctness

Imperfect verification can limit resampling (Stroebl et al., 2026); proposal coverage alone does not establish identifiability (Sunkaraneni et al., 2026). The following Bayesian bound specifies one restricted mechanism. It applies to selection from candidate-local evidence, not to selectors that obtain additional correctness information from cross-candidate relations or new reasoning. The finite likelihood-ratio assumption must hold for all evidence actually exposed to the selector, such as a restricted score channel. It need not hold for full answer contents: for a fixed graded problem, correctness is a deterministic function of a complete answer, so conditioning on that answer can separate correct and incorrect evidence laws.

Proposition B.9 (Evidence barrier for candidate-local selection)**.** *Let (Z_i, E_i) be independent pairs, with $Z_i \sim \mathrm{Bernoulli}(p_i)$ and $0 < p_i < 1$. Here Z_i records correctness and E_i is all evidence exposed to the selector about candidate i. Suppose the conditional evidence laws satisfy*

$$P_{i,1} \ll P_{i,0}, \qquad \frac{dP_{i,1}}{dP_{i,0}} \le L_i < \infty.$$

Any selector returning an index using only $(E_1, \dots, E_n)$ and independent randomization has accuracy

$$S \le \min\left\{1 - \prod_i (1-p_i),\ \max_i \frac{p_i L_i}{1 - p_i + p_i L_i}\right\}. \tag{27}$$

For common $p_i = p$ and $L_i = L$, achieving $S \ge 1 - \epsilon$ requires

$$L \ge \frac{(1-\epsilon)(1-p)}{\epsilon p}. \tag{28}$$

Proof. Write $\ell_i(E_i) = dP_{i,1}/dP_{i,0}(E_i)$. Independence across candidate pairs and Bayes' rule give

$$\Pr(Z_i = 1 \mid E_1, \dots, E_n) = \frac{p_i \ell_i(E_i)}{1 - p_i + p_i \ell_i(E_i)} \le \frac{p_i L_i}{1 - p_i + p_i L_i}.$$

Conditional on the observed evidence, the selected candidate's correctness probability is a convex combination of these posterior probabilities, including any independent randomized tie-breaking. It cannot exceed their largest upper bound. Averaging proves the second term of (27); success also requires at least one correct candidate, proving the first. Rearranging the common-p, L bound gives (28). □

For fixed target error, the necessary log-likelihood evidence grows as $\log(1/p)$ when $p \to 0$. Increasing depth can therefore help both by raising coverage and by reducing the evidence needed to identify a correct candidate. This conclusion does not infer verification ability from generation ability; the likelihood bounds are additional assumptions requiring separate measurement.

B.8 Revision and interaction at a fixed budget

Let the current pool have coverage probability G. An interaction uses n calls of allowance a. Write γ for coverage after interaction conditional on current absence, and e for absence after interaction conditional on current presence. Here γ includes recovery after a correct answer was lost earlier, unlike the first-discovery hazard d_t. An alternative spends the same reasoning budget on n fresh calls, retains the old pool, and selects from the enlarged pool. Assume these calls are mutually independent and independent of the old pool; their coverage is $g = 1 - [1 - p(a)]^n$.

Proposition B.10 (Interaction versus fresh exploration)**.** *Under the stated pool and independence assumptions,*

$$H_{\mathrm{int}} - H_{\mathrm{fresh}} = (1 - G)(\gamma - g) - Ge. \tag{29}$$

Let η_{int} and η_{fresh} denote each alternative's selection reliability conditional on its resulting pool being covered. Interaction improves delivered accuracy exactly when

$$\eta_{\mathrm{int}}[G + (1-G)\gamma - Ge] > \eta_{\mathrm{fresh}}[G + (1-G)g]. \tag{30}$$

Proof. Let E indicate that the existing candidate pool contains a correct answer, with $\Pr(E = 1) = G$. Write E' for coverage after the interaction. By definition,

$$\Pr(E' = 1 \mid E = 0) = \gamma, \qquad \Pr(E' = 0 \mid E = 1) = e.$$

Total probability therefore gives

$$H_{\mathrm{int}} = G(1-e) + (1-G)\gamma = G + (1-G)\gamma - Ge. \tag{31}$$

The alternative retains the existing pool and adds n fresh, independent calls, each succeeding with probability $p(a)$, independently of the pool. Their coverage is $g = 1 - [1-p(a)]^n$. Its enlarged pool fails precisely when both the old pool and all fresh calls fail, so

$$H_{\mathrm{fresh}} = 1 - (1-G)(1-g) = G + (1-G)g.$$

Subtracting proves

$$H_{\text{int}} - H_{\text{fresh}} = (1 - G)(\gamma - g) - Ge. \tag{32}$$

For either method, selection restricted to its available pool succeeds with probability equal to coverage times conditional selection reliability. Consequently,

$$\begin{aligned} S_{\text{int}} &= \eta_{\text{int}}[G + (1 - G)\gamma - Ge], \\ S_{\text{fresh}} &= \eta_{\text{fresh}}[G + (1 - G)g]. \end{aligned} \tag{33}$$

This argument requires neither independence among the interacting agents nor a Markov description of their histories. □

For $G < 1$, interaction improves coverage exactly when $\gamma - g > Ge/(1 - G)$. At $G = 0$, only relative discovery matters; at $G = 1$, interaction cannot improve coverage and decreases it whenever $e > 0$. Conditional probabilities on null events may be assigned arbitrarily, because their coefficients vanish. Likewise, selection reliability on a zero-coverage pool may be set to zero. A coverage disadvantage need not imply lower delivered accuracy: the appropriate comparison is Equation (33).

Both alternatives use na additional reasoning tokens. The fresh-sampling reference explicitly retains the old pool. Input processing, storage, and selection over different pool sizes can have different costs; the statement compares the paper's reasoning resource and does not equate total inference cost. If an interaction returns a newly generated answer, that answer belongs to its generated pool before this accounting applies.

Corollary B.11 (Selection gains must compensate for extinction)**.** *Suppose an additional interaction round cannot recover correctness from an all-wrong current pool, so $\gamma = 0$. If current coverage $G > 0$, the round increases delivered accuracy exactly when*

$$(1 - e)\eta_{\text{new}} > \eta_{\text{old}}. \tag{34}$$

In particular, a perfect current selector cannot benefit from this round.

Proof. The old and new accuracies are $G\eta_{\text{old}}$ and $G(1 - e)\eta_{\text{new}}$. Divide their strict comparison by G. Since $(1 - e)\eta_{\text{new}} \leq 1$, an old reliability of one cannot be improved. If $G = 0$, both accuracies are zero. □

Copying or reranking only currently available answers is sufficient for $\gamma = 0$, because it cannot make an all-wrong pool correct. Literal copying is not necessary: any channel with an absorbing all-wrong coverage state satisfies the corollary. Conversely, debate that generates new answers need not satisfy it. Absence of first discoveries in observed runs also does not establish $\gamma = 0$, which additionally rules out recovery after an earlier correct answer was lost. Along a common run prefix with this property at every round,

$$S_R = G_1\eta_R \prod_{t=2}^{R}(1 - e_t).$$

Thus improved agreement or conditional selection can coexist with declining coverage; its accuracy gain must exceed the accumulated extinction loss.

B.9 An explicit model of private revision

One-step correction and corruption. Let q_r be the correctness probability before revision, and define $\alpha_r = \Pr(\text{correct next} \mid \text{wrong now})$ and $\beta_r = \Pr(\text{wrong next} \mid \text{correct now})$ for the actual entering history distribution. Total probability gives

$$q_{r+1} = (1 - \beta_r)q_r + \alpha_r(1 - q_r).$$

Subtracting q_r gives the one-step correction–corruption balance

$$q_{r+1} - q_r = \alpha_r(1 - q_r) - \beta_r q_r. \tag{35}$$

Another round improves correctness exactly when correction outweighs corruption. No stationarity or Markov assumption is needed for this identity. Reusing the same rates at other rounds or under different preceding actions requires additional assumptions; the following model states them explicitly.

A homogeneous model. Fix allowance a and suppress the problem index. Consider n independent chains. Each initial answer succeeds with probability $p = p(a)$. Within each chain, correctness is a homogeneous two-state Markov process with correction probability $\alpha = \alpha(a)$ and corruption probability $\beta = \beta(a)$:

$$\Pr(\text{correct next} \mid \text{wrong now}) = \alpha, \qquad \Pr(\text{wrong next} \mid \text{correct now}) = \beta.$$

These are explicit modeling assumptions, not consequences of baseline sampling. They exclude dependence on earlier history beyond current correctness and require independence between chains.

Proposition B.12 (Depth, width, and private revision)**.** *If $\alpha+\beta > 0$, the final correctness probability of one chain is*

$$q_R = \frac{\alpha}{\alpha+\beta} + \left(p - \frac{\alpha}{\alpha+\beta}\right)(1-\alpha-\beta)^{R-1}. \tag{36}$$

If $\alpha = \beta = 0$, then $q_R = p$. Current and archive coverage are

$$\begin{aligned} H_{\mathrm{cur}}(n,R,a) &= 1-(1-q_R)^n, \\ H_{\mathrm{arc}}(n,R,a) &= 1-\left[(1-p)(1-\alpha)^{R-1}\right]^n. \end{aligned} \tag{37}$$

For $R > 1$, $n \geq 1$, and $p, \alpha < 1$, archive coverage is strictly greater than that of nR independent calls of allowance a if and only if $\alpha > p$.

Proof. One chain obeys $q_{r+1} = (1-\beta)q_r + \alpha(1-q_r) = \alpha + (1-\alpha-\beta)q_r$, with $q_1 = p$. When $\alpha + \beta > 0$, subtract the fixed point $\alpha/(\alpha+\beta)$ and iterate, proving Equation (36). The zero-transition case is immediate. The n final answers are independent Bernoulli variables with parameter q_R, which gives current coverage.

A chain contains no correct answer anywhere exactly when initialization fails and all $R-1$ subsequent wrong-to-next transitions stay wrong. By the Markov assumption this event has probability $(1-p)(1-\alpha)^{R-1}$. Independence across chains proves the archive formula. The equal-cost independent reference has failure probability $(1-p)^{nR}$. When $p, \alpha < 1$, comparing these positive probabilities, taking nth roots, and canceling $1-p$ gives $(1-\alpha)^{R-1} < (1-p)^{R-1}$, equivalent to $\alpha > p$. □

If $p = 1$, archive coverage is already one. If $p < 1$ and $\alpha = 1$, any $R > 1$ gives perfect archive coverage. The exponent-zero case $R = 1$ means initialization alone. Corruption β affects current coverage but not whether correctness was ever generated. When $\alpha + \beta \leq 1$, q_R approaches its fixed point monotonically; when $\alpha + \beta > 1$, it can oscillate. Thus more revision need not monotonically improve even an individual chain.

Write $h_\alpha(a) = -\log[1-\alpha(a)]$, alongside the fresh-call quantity $h(a) = -\log[1-p(a)]$. The constraint $C = nRa$ gives

$$H_{\mathrm{arc}} = 1 - \exp\{-CJ(R,a)\}, \qquad J(R,a) = \frac{h(a) + (R-1)h_\alpha(a)}{Ra}. \tag{38}$$

For finite $h(a), h_\alpha(a)$, rewrite the rate as

$$J(R,a) = \frac{1}{a}\left[h_\alpha(a) + \frac{h(a) - h_\alpha(a)}{R}\right].$$

At fixed a, archive-optimal revision therefore lies at an endpoint of the feasible set: its smallest R when $\alpha < p$, its largest R when $\alpha > p$, and any R when $\alpha = p$. This comparison is exact among allocations for which $n = C/(Ra)$ is a positive integer; a continuous relaxation gives the same monotonicity. An interior optimum for archive coverage requires departures from this homogeneous model, additional constraints, or changing a. Final-answer accuracy can have an interior optimum because corruption and selection introduce further tradeoffs.

The proposition predicts coverage, not multiclass voting accuracy. With only one incorrect answer class, final plurality accuracy is the upper-tail probability of $\mathrm{Bin}(n, q_R)$, with the prescribed tie rule. With several incorrect answer classes, their probabilities are additionally required. Fitting the transition rates and testing unseen rounds makes this specialization falsifiable; its failure would identify a limitation of the homogeneous model rather than of the general generation–selection decomposition.

B.10 NONSTATIONARY INTERACTIONS AND DOWNSTREAM EFFECTS

For an arbitrary protocol, let E_t indicate current coverage and define its actual marginal conditional rates $\gamma_t = \Pr(E_t = 1 \mid E_{t-1} = 0)$ and $e_t = \Pr(E_t = 0 \mid E_{t-1} = 1)$. With $\lambda_t = 1 - \gamma_t - e_t$, total probability gives

$$G_t = \gamma_t + \lambda_t G_{t-1}, \qquad G_R = G_1 \prod_{t=2}^{R} \lambda_t + \sum_{t=2}^{R} \gamma_t \prod_{u=t+1}^{R} \lambda_u. \tag{39}$$

Empty products equal one. No Markov assumption is needed for these descriptive identities: the rates incorporate the histories encountered by that protocol.

For two protocols $+$ and $-$ starting with equal coverage, write $\Delta\gamma_t = \gamma_t^+ - \gamma_t^-$ and $\Delta e_t = e_t^+ - e_t^-$. Subtracting their recurrences gives

$$\Delta G_t = \lambda_t^+ \Delta G_{t-1} + (1 - G_{t-1}^-)\Delta\gamma_t - G_{t-1}^- \Delta e_t.$$

Induction from $\Delta G_1 = 0$ yields the exact decomposition

$$\Delta G_R = \sum_{t=2}^{R} \left[(1 - G_{t-1}^-)\Delta\gamma_t - G_{t-1}^- \Delta e_t\right] \prod_{u=t+1}^{R} \lambda_u^+. \tag{40}$$

This separates changes in discovery and extinction at each round, weighted by the later transition slopes. It is an algebraic comparison of actual trajectory distributions, not an identification of intervention effects.

For a counterfactual interpretation, assume the continuation's coverage transition law depends only on current coverage and is invariant to the earlier action and history distribution. If two local actions create a coverage difference ΔG_t and share this continuation, subtracting the common subsequent recurrences gives

$$\Delta G_R = \Delta G_t \prod_{u=t+1}^{R} \lambda_u.$$

If terminal selection reliability is also invariant and equals η_R, the final accuracy difference is $\eta_R \Delta G_t \prod_{u=t+1}^{R} \lambda_u$. When $\gamma_u + e_u \le 1$, each factor lies in $[0, 1]$ and later rounds attenuate coverage differences. Otherwise a factor can be negative and reverse their ordering. These slopes are not pure survival probabilities: rapid recovery of uncovered teams also reduces differences. Without the invariance assumptions, rates estimated under one allocation cannot be reused as counterfactual transition kernels under another.

B.11 SCHEDULING AND LIMITS OF ALLOCATION DIAGNOSTICS

An explicit state abstraction. For an action a, let g_a be the probability of current-pool coverage afterward given its absence before, and e_a the probability of losing current coverage given its presence. At a particular entering distribution, total probability gives

$$T_a(G) = g_a + (1 - g_a - e_a)G. \tag{41}$$

Using this same map under different preceding actions is an additional assumption: conditional on the covered/uncovered state, the action's transition kernel must be invariant to history and ordering. Pooled empirical transition rates need not satisfy this requirement.

Proposition B.13 (Ordering a fixed set of interactions)**.** *Consider a fixed multiset of actions with invariant transition kernels as above, $0 \le g_a + e_a \le 1$, and costs unchanged by ordering. Suppose conditional terminal selection reliability is the same η for every schedule. Then ordering nonidentity actions by nondecreasing*

$$\rho_a = \frac{g_a}{g_a + e_a} \tag{42}$$

maximizes terminal coverage and delivered accuracy. Actions with $g_a = e_a = 0$ may occur anywhere.

Proof. Write $\lambda_a = 1 - g_a - e_a \geq 0$. Direct expansion gives the adjacent-exchange identity

$$T_b(T_a(G)) - T_a(T_b(G)) = g_b e_a - g_a e_b \tag{43}$$

$$= (g_a + e_a)(g_b + e_b)(\rho_b - \rho_a). \tag{44}$$

The entering G cancels. A common suffix multiplies this difference by the product of its λ values, which is nonnegative. Therefore swapping an inverted adjacent pair cannot reduce terminal coverage. Successive swaps produce the stated order; multiplication by the common $\eta \geq 0$ preserves optimality. □

This is a conditional scheduling result, not a universal ordering of debate and revision. It predicts that actions with a more favorable discovery-to-extinction balance belong later, provided the state abstraction holds. The suffix product quantifies how later rounds attenuate the value of earlier choices. When every action preserves an archive, $e_a = 0$ and the ordering effect disappears.

When another action helps. For unchanged selection reliability,

$$T_a(G) - G = (g_a + e_a)(\rho_a - G). \tag{45}$$

Thus another action increases coverage exactly below its equilibrium, except for the identity action. If selection reliability changes from η to η', the exact accuracy condition is instead

$$\eta' \big[G(1 - e_a) + (1 - G)g_a\big] > \eta G. \tag{46}$$

For two equal-cost actions followed by a common suffix with nonnegative slopes and common terminal selection,

$$T_a(G) - T_b(G) = (g_a - g_b)(1 - G) - (e_a - e_b)G. \tag{47}$$

If both differences are positive, the more exploratory action a gives higher one-step coverage when $G < (g_a - g_b)/(g_a - g_b + e_a - e_b)$; a zero suffix slope or $\eta = 0$ makes final accuracy equal. Greater discovery and lower extinction imply weak dominance at every G. These are equal-cost comparisons; dividing gains by unequal costs does not prove global budget optimality.

Proposition B.14 (Baseline diagnostics do not identify unseen interventions)**.** *Without restrictions linking baseline and intervention response laws, no rule based only on baseline observations can uniformly identify the best intervention. Specifically, for a menu of two interventions that each consume the remaining budget, there are two systems with identical laws of all baseline observations and opposite optimal interventions. Every such rule has worst-case accuracy regret at least* $1/2$ *over these systems.*

Proof. Give both systems the same baseline law, including prompts, outputs, traces, and any observed baseline correctness labels. In system 0, intervention D succeeds surely and I fails surely; in system 1, swap these outcomes. No assumption on the baseline law restricts these unseen response kernels. Let α be the rule's probability of choosing D, averaged over the common diagnostic law and its own randomization. Optimal accuracy is one in each system, while the rule's regrets are $1 - \alpha$ and α. Their maximum is at least $1/2$. The result holds for arbitrarily many baseline probes, and therefore also when only sampling and voting rates are available. □

This standard indistinguishability argument concerns a specified intervention menu, not arbitrary computations or an information-theoretic limit on reasoning. It establishes that prediction across scaling directions requires calibrated response information or structural assumptions. Moreover, unlabeled deployment samples reveal agreement, not their true correctness probabilities.

A calibrated finite menu. For completeness, the standard uniform-error argument provides a modest positive guarantee. For feasible allocations $a \in \mathcal{A}$, suppose predicted delivered accuracies satisfy $\sup_a |\widehat{S}_a - S_a| \leq \epsilon$. If $\widehat{a} \in \arg\max_a \widehat{S}_a$ and $a^* \in \arg\max_a S_a$, then

$$S_{a^*} - S_{\widehat{a}} \leq (S_{a^*} - \widehat{S}_{a^*}) + (\widehat{S}_{\widehat{a}} - S_{\widehat{a}}) \leq 2\epsilon. \tag{48}$$

The first inequality uses estimated optimality. Obtaining the uniform prediction bound, including under held-out tasks, is a separate statistical problem; it does not follow from a small unlabeled diagnostic sample. Diagnostic costs must also be subtracted before defining the feasible menu.

B.12 CORRESPONDENCE WITH THE EMPIRICAL COMPARISONS

The empirical decomposition in Table 2 uses the same 93 problems, benchmark weights, and native readouts for all procedures. It therefore permits the gain identity in Equation (7) to be applied directly. At $16A$, deep reasoning exceeds independent voting by $78.3 - 42.5 = 35.8$ percentage points. Its archive-coverage advantage contributes $78.3 - 57.5 = 20.8$ points, and the independent pool's selection loss contributes 15.0 points. Thus the rounded entries give

$$35.8 = 20.8 + 15.0 \quad \text{percentage points.}$$

This comparison separates expanded observed coverage from correctness available to sampling but not returned by voting. It uses realized current-pool coverage on the common cohort; the 20.5-point comparison with estimated oracle coverage for sampling at A reasoning tokens per call elsewhere uses a different reference and cohort.

The same accounting exposes the tradeoff in repeated revision: its 65.1% archive coverage falls to 53.2% final accuracy, corresponding to a 12.0-point preservation loss computed before rounding. Peer and private teams likewise trade off the components: peer teams have approximately 3.1 points less archive coverage, a similar preservation loss, and a 5.5-point smaller selection loss. These changes account for their 2.2-point delivered-accuracy difference before rounding. They neither establish a statistically significant communication benefit nor identify a causal effect on any one component while holding the others fixed.

These measurements illustrate the general decomposition without fitting the simplified transition models. The reported experiments do not establish stationary correction and corruption rates, invariant scheduling kernels, per-problem depth-response curvature, or the bounded-likelihood evidence assumption. Likewise, observed success strata do not establish the voting-margin conditions or prescribe an intervention: those require the relevant response laws, as Proposition B.14 makes explicit. All empirical correspondences concern the forced reasoning protocol, its recorded candidate pools, and the evaluated budget range.

C ADDITIONAL EMPIRICAL COMPARISONS

The following displays retain the benchmark-level comparisons and uncertainty underlying the regime-centered Results section. Figure 6 separates estimated oracle coverage from realized prefix coverage and voting. Figure 7 gives the benchmark-level $16A$ comparison, and Figure 8 retains the regime-level Wilson intervals for all five allocations.

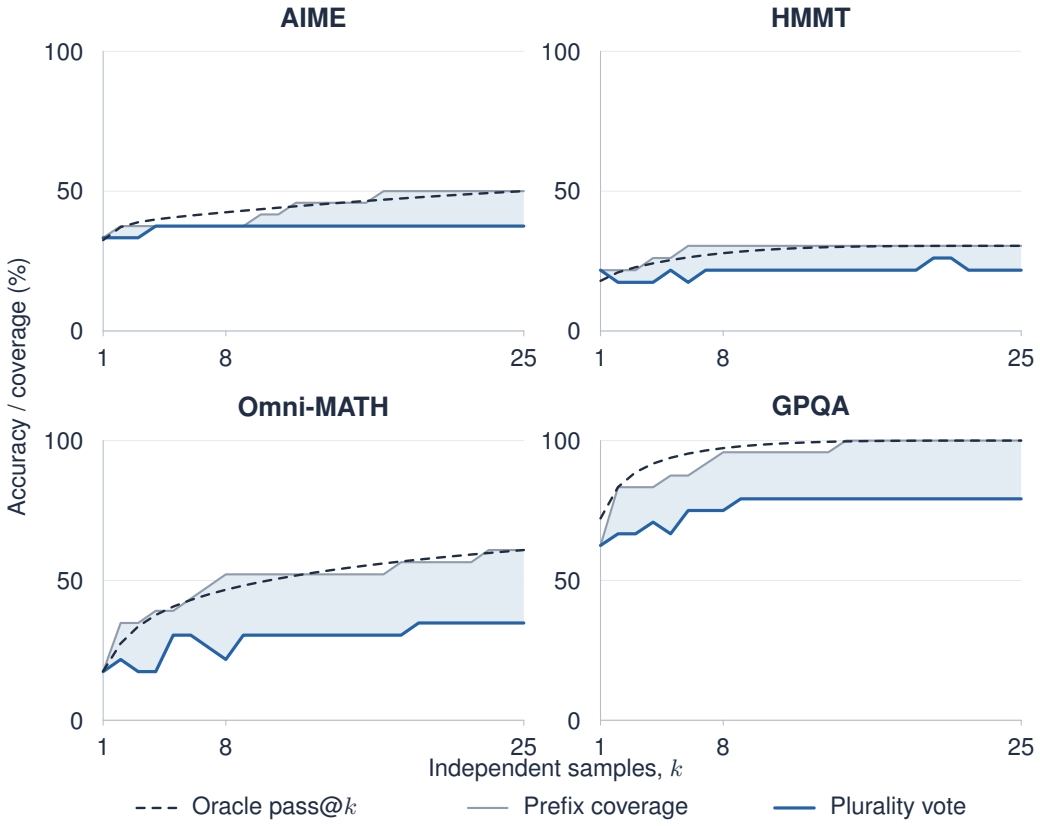


Figure 6: **Generation and selection across benchmarks.** Dashed curves estimate oracle pass@k from 25 draws per problem. Gray and blue curves show observed prefix coverage and plurality-vote accuracy; shading marks their selection gap. Benchmarks comprise AIME 2025/26 ($N = 24$), HMMT February 2025/26 ($N = 23$), Omni-MATH at difficulty ≥ 7 ($N = 23$), and GPQA-Diamond ($N = 24$). Each draw uses $A = 1{,}024$ reasoning tokens. All 94 sample sets are fully graded.

Configuration	$(n, R, a/A)$	AIME	HMMT	Omni	GPQA	Macro
Sampling + vote	$(16, 1, 1)$	37.5	21.7	30.4	79.2	42.2
Private revision	$(1, 16, 1)$	50.0	30.4	39.1	**91.7**	52.8
Private team	$(4, 4, 1)$	41.7	21.7	43.5	75.0	45.5
Peer communication	$(4, 4, 1)$	50.0	39.1	30.4	75.0	48.6
Deep reasoning	$(1, 1, 16)$	**83.3**	**78.3**	**60.9**	**91.7**	**78.5**

Accuracy (%)

Figure 7: **Compute allocation and combination at a fixed reasoning budget.** Final-answer accuracy for $C = 16A$ reasoning tokens, where $A = 1{,}024$ and $(n, R, a/A)$ denotes agents, sequential rounds, and reasoning allowance per call. Private and peer teams have the same allocation; peer teams additionally exchange responses between rounds. Sampling and teams use their native voting readout. Deep reasoning achieves the highest benchmark-macro accuracy. Bold values identify the highest accuracy in each column. Results use 94 problems with fully graded sample sets; Macro gives each benchmark equal weight.

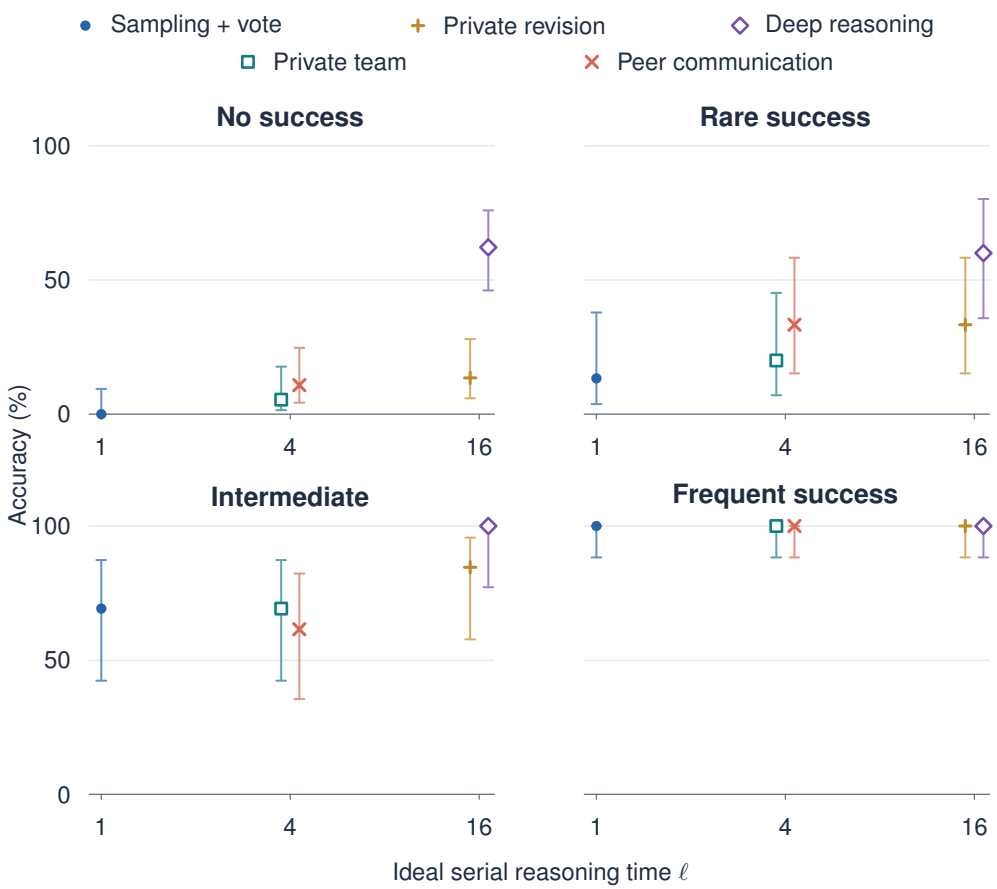


Figure 8: **Accuracy and time across sampling regimes.** All configurations use $C = 16A$. With parallel execution and negligible combination overhead, serial reasoning time $\ell = Ra/A$ equals 1 for sampling, 4 for teams, and 16 for revision or deep reasoning. Bars show 95% Wilson intervals. Regimes contain 0, 1–5, 6–15, or 16–25 correct answers in 25 baseline draws ($N = 37, 15, 13, 29$, respectively). Equal-time points are separated horizontally for visibility. Deep reasoning leads in the first three regimes; parallel sampling reaches the frequent-success ceiling at the lowest time.

D Detailed regime and allocation results

D.1 Distinct sampling regimes reveal different constraints on correctness

We first characterize what sampling at the reference reasoning allowance generates and what voting returns. Mean single-draw correctness is 35.0% under benchmark-macro averaging. With 25 draws, oracle coverage reaches 60.3%, while voting reaches 43.3%: a 17.0 percentage-point gap between available and delivered correctness. Benchmark-specific curves appear in Appendix C.

These averages combine distinct generation–selection profiles. We stratify the 94 fully graded banks by the number of correct answers c among 25 draws: *no observed success* ($c = 0$; $N = 37$), *rare success* ($c = 1$–5; $N = 15$), *intermediate success* ($c = 6$–15; $N = 13$), and *frequent success* ($c = 16$–25; $N = 29$). Figure 9 characterizes coverage and voting within these strata.

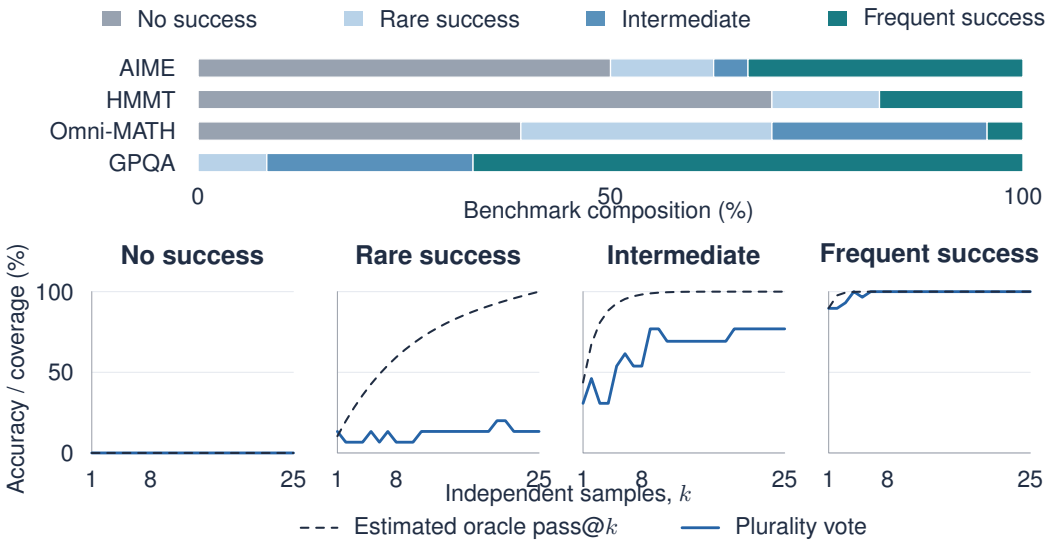


Figure 9: **Distinct generation–selection profiles under sampling at the reference reasoning allowance.** Upper bars show benchmark composition. Panels compare estimated oracle coverage (subset averaging) with voting on observed prefixes; at $k = 25$, both use the complete sample set. Strata contain 0, 1–5, 6–15, or 16–25 correct answers in 25 draws ($N = 37, 15, 13, 29$). The same banks define groups and curves; positive-stratum coverage at 25 is therefore definitional. Each draw uses $A = 1{,}024$ reasoning tokens.

In the no-success stratum, the observed bank supplies no solution for a selector to recover. Every positive-success bank contains a correct answer by construction, yet voting over all 25 candidates succeeds on only 13.3% of rare-success problems, compared with 76.9% in intermediate success and 100% in frequent success. Thus, rare success combines occasional generation with poor selection, intermediate success converts more available correctness into answers, and frequent success gives voting strong support for correctness. These profiles coexist within benchmarks: AIME contains both 12 no-success and eight frequent-success problems.

The regimes provide the foundation for the allocation comparisons below: additional compute may need to discover correctness, make available correctness determine the output, or achieve comparable accuracy with less sequential dependence. They describe the measured reference process, rather than intrinsic difficulty or proven allocation boundaries. The same banks define the strata and their sampling curves; zero observed successes does not imply zero success probability.

D.2 Compute allocation changes accessible and delivered correctness

We next hold the reasoning budget fixed at $C = 16A = 16{,}384$ tokens and compare allocations within these regimes. A single $16A$ call achieves 78.5% benchmark-macro accuracy, versus 42.2%

for 16 independent calls with A reasoning tokens each with voting and 52.8% for 16 rounds of private revision. Its accuracy also exceeds estimated oracle coverage for sampling at A reasoning tokens per call (58.0%) by 20.5 percentage points. Figure 10 separates these aggregate outcomes from their regime-specific returns.

Allocation	$(n, R, a/A)$	No success $N = 37$	Rare $N = 15$	Intermed. $N = 13$	Frequent $N = 29$	Benchmark macro
Sampling + vote	$(16, 1, 1)$	0.0	13.3	69.2	**100.0**	42.2
Private revision	$(1, 16, 1)$	13.5	33.3	84.6	**100.0**	52.8
Private team	$(4, 4, 1)$	5.4	20.0	69.2	**100.0**	45.5
Peer communication	$(4, 4, 1)$	10.8	33.3	61.5	**100.0**	48.6
Deep reasoning	$(1, 1, 16)$	**62.2**	**60.0**	**100.0**	**100.0**	**78.5**

Reference at $16A$: oracle coverage from A-token calls: 58.0% (equal benchmark weights)

Figure 10: **Allocation changes delivered correctness at the same reasoning budget.** All procedures spend $C = 16A = 16{,}384$ reasoning tokens. Regime columns weight problems equally; benchmark-macro weights the four benchmark families equally and is not the mean of the regime columns. Values are accuracy (%) on 94 fully graded reference sample sets; bold marks column maxima. The oracle is a diagnostic reference, not an executable selector. Appendix C gives benchmark-level results and regime-level uncertainty.

Depth expands observed correctness beyond the reference allocation. The longer call solves 23 of the 37 problems whose reference sample sets contain no correct answer, compared with five for private revision. Sampling prefixes inherit the full sample set's absence of correct candidates. The deeper call therefore generates solutions absent from the observed set of responses generated with A reasoning tokens each, rather than selecting them from that sample set. This illustrates the allocation-specific oracle comparison in Section B.2: concentrating compute can outperform the opportunities supplied by equal-cost attempts at the reference reasoning allowance. It does not establish that those solutions have zero probability under generation at the reference reasoning allowance.

The value of depth depends on the regime. At $16A$, depth achieves 60.0% accuracy in rare success and 100% in intermediate success, versus 13.3% and 69.2% for independent voting. It has the highest observed delivered accuracy among the compared allocations in the first three regimes. In frequent success, all five procedures solve all 29 problems, leaving sequential time to distinguish them (Section 5.2). The three-regime accuracy lead and the aggregate excess over the oracle for sampling at A reasoning tokens per call are distinct findings: depth need not exceed oracle coverage within each regime to improve the returned answer.

D.3 Generation, preservation, and selection explain allocation differences

We now apply Equation (3) to distinguish correctness never generated, generated but no longer selectable, and selectable but not returned. Table 2 uses 93 common problems with fully graded archive and current pools for every procedure; one HMMT problem is excluded because its private-team archive coverage remains unresolved. Its rates therefore differ slightly from the preceding 94-problem comparison.

	Correctness (%)			Error components (pp)		
Procedure	H_{arc}	H_{cur}	S	Generation	Preservation	Selection
Sampling	57.5	57.5	42.5	42.5	0.0	15.0
Private revision	65.1	53.2	53.2	34.9	12.0	0.0
Private team	64.0	58.7	45.7	36.0	5.3	13.0
Peer team	60.9	55.4	47.9	39.1	5.4	7.5
Deep reasoning	78.3	78.3	78.3	21.7	0.0	0.0

Table 2: **Generation, preservation, and selection at $16A$.** Sampling uses 16 independent calls; revision uses 16 successive calls; teams use four agents and four rounds; depth uses one long call. Rates are benchmark-macro averages on 93 common, fully graded problems. Error components follow Equation (3) and sum to $1 - S$ before rounding; S uses the native current-pool readout.

Discovery and selection. On this common cohort, depth exceeds sampling with voting by 35.8 percentage points. The gain identity in Section B.1 separates that difference into

$$\underbrace{78.3 - 42.5}_{\text{accuracy gain } 35.8} = \underbrace{78.3 - 57.5}_{\text{coverage gain } 20.8} + \underbrace{57.5 - 42.5}_{\text{selection shortfall } 15.0} .$$

This uses realized coverage on 93 problems; the 20.5-point headline instead uses estimated oracle coverage on 94. A single output has zero preservation and selection losses by construction, without characterizing selection inside its reasoning.

Preserving revision gains. Private revision raises archive coverage to 65.1%, above sampling's 57.5%, but achieves only 53.2% benchmark-macro final accuracy. Approximately 12.0 points are lost before the final answer. As Section B.1 distinguishes, further computation can add correct discoveries without preserving them in the final pool. These losses instantiate the decomposition without fitting the stationary correction–corruption model in Appendix B.9.

The net value of peer exposure. Compared with private teams, peer teams have lower archive coverage (60.9% versus 64.0%), similar preservation loss, and a smaller selection loss (7.5 versus 13.0 points). These components account descriptively for their 2.2-point accuracy difference on the common cohort. Separately, the matched peer–private comparison on the 94-problem panel holds $(n, R, a) = (4, 4, A)$ and the final voting rule fixed. Its regime-specific paired intervals all include zero (Figure 11). Thus, a smaller selection-loss term does not establish a reliable communication advantage; interaction must be judged by its net effect on delivered correctness.

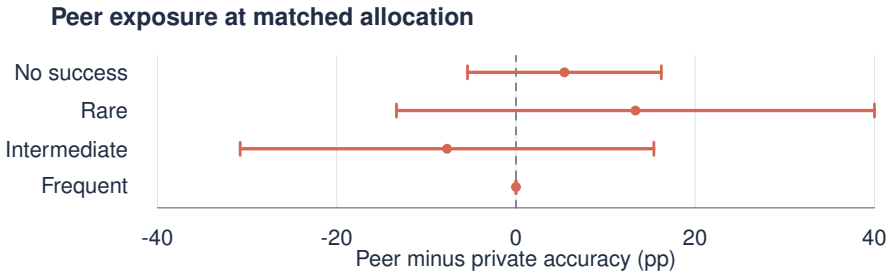
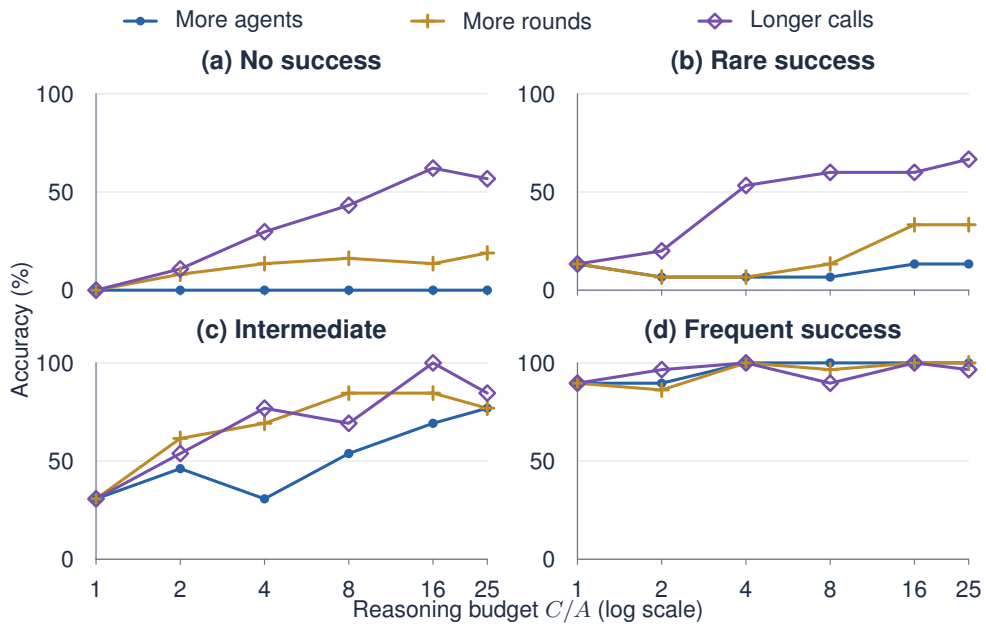


Figure 11: **Peer exposure has no consistent detected accuracy advantage.** Peer minus private-team accuracy at $(n, R, a/A) = (4, 4, 1)$, with identical final voting rules. Bars are 95% paired item-bootstrap intervals from 10,000 resamples. The frequent-stratum interval is degenerate because all observed paired outcomes agree; it does not establish population equivalence.

D.4 Scaling returns depend on regime and budget

The fixed-budget comparison does not imply a universal ordering as compute grows. Figure 12 traces the separate returns to more independent agents, more private revision rounds, and longer calls within each sampling regime.



Frequent-success stratum at $16A$: modeled sequential time			
	Sampling + vote	Four-agent teams	Revision / depth
Accuracy	$29/29$	$29/29$	$29/29$
$\ell = Ra/A$	1	4	16

Figure 12: **Scaling returns vary with regime and budget.** Curves increase compute through independent agents with voting $(n, 1, A)$, private revision $(1, R, A)$, or within-call reasoning $(1, 1, a)$. Panels use the same four descriptive strata as Figure 9. Each deeper allowance has one realization per problem, so small rank reversals describe observed outcomes. The inset compares all five procedures at $16A$: equal observed accuracy in frequent success accompanies different modeled serial reasoning lengths under sufficient parallel capacity and negligible combination overhead. These times are not measured wall-clock latency.

In the no-success and rare-success strata, longer calls yield larger gains than the evaluated sampling and revision alternatives. The intermediate stratum shows budget-dependent rank reversals: private revision solves one more problem than depth at $2A$ and two more at $8A$, out of 13, whereas depth leads at $16A$. These are observed comparisons with one realization per deeper allowance, not estimated crossover thresholds or monotonic scaling laws.

In frequent success, independent voting already solves all 29 problems at $4A$. At $16A$, all five procedures solve those same problems, but their execution structures differ. Under the parallelism assumptions of Equation (6), modeled serial reasoning length is $\ell = 1$ for sampling, 4 for four-agent teams, and 16 for either single-agent allocation. Width therefore attains the same observed accuracy with the shortest modeled time; this is not a measured wall-clock speedup.

Together, the results identify distinct roles for the scaling axes under the enforced reasoning budget: depth can expand observed generation, revision can add discoveries that require preservation, and width can exploit already reliable aggregation with shorter sequential dependence. Their value depends on the regime and budget. As Section B.3 makes explicit, turning these descriptive comparisons into a prospective allocation rule additionally requires calibrated intervention responses.

E ALLOCATION PREDICTION AND SELECTOR EVALUATION

E.1 PANELS, DIAGNOSTIC INFORMATION, AND RESPONSE FITTING

The extension retains the model, sampling parameters, provider endpoints, and exact-length THINK protocol of Section A.4. The 200-problem evaluation is sampled from eligible benchmark problems after excluding the original main and development panels and exact or reviewed near duplicates. Its benchmark sizes are 20/20/100/60; provider assignment is balanced within each benchmark, giving 100 problems per provider. The repeated-run panel contains 12 problems per benchmark, selected from the 94 fully graded original reference sample sets by benchmark and success stratum at the reference reasoning allowance. Its unweighted results describe this selected panel, rather than estimate the full original panel or benchmark populations.

Each problem has five diagnostic draws at A. Development diagnostics use slots in the original reference sample set 20–24; the probes on new problems form a separate draw partition. Correctness labels are removed before constructing diagnostic features. Features are the five sorted answer-class frequencies, the largest-minus-second-largest frequency, the number of classes divided by five, entropy normalized by $\log 5$, invalid-output fraction, mean $\log(1+\text{answer characters})$, and the fraction of answers reaching the 4,096-token cap. Answer equivalence uses the experiment's mathematical-equivalence grouping. No new problem's correctness label is an input. A provider-normalized latency feature was considered and dropped by development cross-validation.

There are 1,192 development pool targets. Every development problem contributes available pools at 1×1, 2×1, 1×2, 4×1, 2×2, 1×4, 16×1, 8×2, 2×8, 1×16, and 1×8. The repeated-run panel adds 2×16, 4×8, and 8×4 training pools. Thus, development calibration includes additional depth replicates and some $32A$ allocations; it is not a five-example fit. The cells 4×4, 8×1, 4×2, and 2×4 are withheld from all fitting and tuning. Plurality targets average uniformly over tied winning classes; evaluation uses seeded realized ties. Unknown grades are excluded and invalid outputs count as wrong.

For all three regression variants, the delivered-success predictor is

$$\widehat{S}_x(c) = \operatorname{logit}^{-1}\Big(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{\phi}(x,c)\Big). \tag{49}$$

The population regression uses benchmark indicators and $(\log n, \log d, R, nd/16)$, with $R = 1$ throughout. The additive regression adds provider and the 11 diagnostic features. The interaction regression, used as the statistical allocator in Figure 5, adds all squares and pairwise products of the 14 continuous features: $\log n$, $\log d$, $nd/16$, and the diagnostic features. Columns are standardized, the intercept is unpenalized, and ridge strength is chosen from $\{0.3, 1, 3, 10, 30, 100\}$ by five-fold cross-validation grouped by problem. Ties favor stronger regularization. The delivered-outcome fits choose ridge 30 for population regression and 100 for additive regression and interaction regression. The additive regression form changes probability levels with diagnostic features but leaves its accuracy-only allocation ranking invariant across problems; the interaction regression allows allocation–diagnostic interactions. The additive regression also adds provider information, so its contrast with population regression does not isolate the contribution of diagnostic samples alone.

The coverage–selection predictor fits a logistic per-call response model with benchmark–depth indicators, diagnostic features and their interactions with log depth, and a Gaussian problem intercept. It estimates coverage by averaging $1-(1-p_x(d))^n$ over 2,000 simulations, with a shared latent problem effect across calls. A separate ridge logistic model estimates selection conditional on coverage, using out-of-fold response predictions on covered multi-call training pools; single calls have selection factor one. The coverage–selection predictor multiplies the two estimates. Its response fit uses 4,160 graded development draws. Removing answer-concentration features gives ablations of the additive regression and coverage–selection predictor without agreement information. Those ablations do not establish a distinct predictive benefit from concentration features in this experiment.

The original nine-option predictions were frozen before release of labels for the 200 new problems. For the later comparison of 15 allocations, fitting on the development panel reproduces those predictions exactly on all 10,800 overlapping predictor–problem–allocation cells. On the repeated-run panel, the entire fitting and tuning pipeline is repeated inside five outer problem folds, so a scored problem contributes no outcomes to its predictor's training. The common random seed for splits,

folds, subset sampling, and bootstrap analysis is 20260925. Model generation itself has no fixed API seed.

E.2 Selection objective, LLM comparison, and cost accounting

For each problem, let $\widehat{K}_x(c) = \widehat{S}_x(c)^2 + [1 - \widehat{S}_x(c)]^2$. The statistical selector constructs

$$\mathcal{A}_{\mathrm{acc}}(x) = \{c \in \mathcal{A} : \widehat{S}_x(c) \geq \max_b \widehat{S}_x(b) - \delta\}, \tag{50}$$

$$\mathcal{A}_{\mathrm{stable}}(x) = \{c \in \mathcal{A}_{\mathrm{acc}}(x) : \widehat{K}_x(c) \geq \max_{b \in \mathcal{A}_{\mathrm{acc}}(x)} \widehat{K}_x(b) - \delta\}, \qquad \delta = 0.02, \tag{51}$$

then minimizes estimated latency on $\mathcal{A}_{\mathrm{stable}}(x)$, breaking ties by fewer reasoning tokens. The latency estimates are mean option latencies on the repeated-run panel within the problem's provider, leaving the item out on the repeated-run panel. This rule does not compute a Pareto frontier. Correctness consistency is a function of accuracy, not an additional per-problem observation; below $S = 0.5$, it can favor a less accurate option inside the tolerance band. Population consistency averages per-problem K_x, not K evaluated at population accuracy.

The agreement heuristic uses the diagnostic plurality share f_1. It chooses an allocation X when $f_1 \geq \tau$, and 1×16 otherwise. Development selection searches $\tau \in \{0.6, 0.8, 1.0\}$ and X in the menu, using the same ordered population objective. Its choices on the repeated-run panel are cross-fitted. The agreement rule fitted for the 200 new problems chooses 4×1 when at least four diagnostics agree, and 1×16 otherwise.

The two LLM allocators use the solver model and the problem's provider, with 1,024 forced reasoning tokens per allocation decision and the same sampling parameters. Their prompt explains independent parallel calls, plurality with uniform ties, and all 15 allocation options with typical times from the repeated-run panel rounded to 10 s. It requests accuracy first, treating options within two percentage points as tied; then stable correctness, defined as a probability closest to zero or one; then speed. The probe-informed LLM allocator sees the five diagnostic final answers marked as having unknown correctness. Each arm makes five independent decisions per item with separately shuffled option order. The prompt specifies the accuracy tolerance numerically but does not give a numerical consistency tolerance; both arms retain an accuracy-oriented system message. These comparisons therefore concern the implemented prompt and model, not all possible LLM allocation strategies.

One outcome-blind retry follows an invalid decision; a second invalid result falls back to 1×16. Eight of 2,480 final decisions require this fallback. The five decisions are evaluated separately and their scores averaged, rather than voting over allocation choices. Each selects an already generated solver pool, so repeated choices share that pool's outcome. On the 200-problem evaluation, one probe-informed decision selects a pool with unknown correctness; it is omitted from that item's accuracy average and retained for latency. The original accuracy-only LLM arms use a smaller five-option menu and are not the primary comparison here.

Measured call duration includes THINK, accepted ANSWER, and failed ANSWER attempts, but excludes dispatcher queueing, backoff sleeps, and earlier failed THINK attempts. Solver latency is the maximum member duration under simultaneous starts. Diagnostics add their maximum duration before selection; LLM decision time includes retries and precedes the solver. Reported latency includes these overheads, but not offline model fitting or local statistical prediction time, and depends on provider load during collection. A diagnostic statistical selector costs $5A + ndA$ in reasoning tokens; the problem-only LLM allocator costs $A + ndA$ and the probe-informed LLM allocator costs $6A + ndA$ per use, before retries. Five LLM decisions are evaluation replication, not a five-decision deployment requirement. These are common solver-menu comparisons, not equal-total-cost comparisons after overhead.

Prediction quality and ablations. The additive regression's frozen predictions achieve mixture-weighted Brier score 0.179 on the 200 new problems, versus 0.202 for population regression (difference -0.0226, 95% percentile interval $[-0.0441, -0.0002]$). The interaction regression scores 0.184; its difference from additive regression, 0.0049 ($[-0.0008, 0.0109]$), does not establish a predictive improvement. Interactions instead allow allocation preferences to vary with

probe features. The coverage–selection predictor incurs 0.0091 higher loss than additive regression ($[0.0010, 0.0177]$; pre-registered comparison, Holm-adjusted $p = 0.034$). Thus, the theoretical decomposition does not itself ensure better predictive accuracy.

E.3 Scoring, uncertainty, and interpretation

The repeated-run panel's scoring banks contain 36, 16, 12, 10, and 9 draws per problem at $A, 2A, 4A, 8A, 16A$, respectively, before excluding unknown grades. The five diagnostic draws are absent from the reference-allowance sample set used for scoring. Single-call estimates average bank correctness; multi-call estimates average tie-corrected plurality over all subsets when there are at most 5,000, and otherwise over 5,000 seeded subsets. The same subsets estimate maximum call latency. For LLM arms, per-item system accuracy averages the five selected options' bank accuracies; correctness consistency is computed from that system accuracy, incorporating allocator randomness. The resulting estimates are conditional on finite banks, including only nine deep draws per problem.

The 200-problem evaluation has one realized solver pool per item and configuration. Averaging five LLM decisions estimates allocation randomness conditional on these pools; it does not provide five fresh solver replications. Consequently, the 200-problem evaluation cannot estimate per-problem correctness consistency. Population selection on the 200 new problems uses accuracy and then latency, whereas the repeated-run panel uses all three criteria. The point-estimate population ordering is descriptive and changes under resampling; no selector wins more than half the bootstrap replicates.

Table 3 reports representative selectors over the fifteen allocations and the fixed baseline. The 200-problem evaluation mixture weights are $(0.10, 0.10, 0.50, 0.30)$; macro weights are equal. Every paired bootstrap resamples whole problems within benchmarks, carrying their allocations and decisions together, with 10,000 replicates. Selector-comparison intervals are percentile intervals and are exploratory, without multiplicity correction. The pre-registered Brier comparison between the coverage–selection predictor and additive regression instead uses a studentized interval and Holm adjustment across the two core primary tests. Brier averages squared prediction errors over each item's known $16A$ configurations, then aggregates problems with the stated weights. The Brier interval comparing additive and population regression is an exploratory percentile contrast on the frozen predictions.

	48 repeated problems			200 new problems		
Selector	Accuracy (%)	Consistency (%)	Latency (s)	Acc. mix (%)	Acc. macro (%)	Lat. mix (s)
Fixed depth / population	78.5	88.6	277	70.5	79.4	290
Additive regression	78.5	88.6	320	70.5	79.4	319
Coverage–selection predictor	76.3	90.2	266	67.5	77.6	257
Statistical allocator	77.7	87.7	279	70.0	79.0	270
Agreement heuristic	75.1	87.8	262	68.5	77.9	244
LLM: problem only	67.0	78.5	155	61.0	66.4	158
LLM: five probes	69.4	82.7	202	59.3	66.6	178

Table 3: **Allocation quality and cost on both scoring panels.** Point estimates for the 15-allocation comparison. Accuracy and correctness consistency are percentages; latency includes applicable diagnostic and decision overhead in seconds. The fixed baseline and population regression have identical choices and scores. The 200-problem evaluation mixture and macro columns use different declared benchmark weights, rather than different cohorts. Consistency is not estimated on the 200 new problems. Paired statistical-allocator–LLM accuracy intervals are reported in Section 6.

Meaning of the observed near-best tradeoff. The three-criterion comparison includes 33 deployable strategies, excluding oracle references. On the 48-problem panel, the best observed accuracy is 78.47%; the statistical allocator's 77.73% lies within the two-point band. Among accuracy-admissible strategies, the highest observed consistency is 89.49%; the allocator's 87.69% also lies within the two-point band. Its latency, 279.16 s, is 0.94% above the fastest admissible strategy's 276.55 s. These point estimates motivate "near-best observed tradeoff." They do not certify equivalence: paired 95% intervals against fixed depth are $[-3.26, 1.25]$ points for accuracy, $[-3.50, 1.31]$ points for consistency, and $[-19.9, 24.9]$ s for latency. On the 200 new problems, the statistical al-

locator's accuracy difference from fixed depth is -0.5 points ($[-1.5, 0.0]$). The observed gain is a useful allocation tradeoff, rather than improved maximum accuracy.

The results separate prediction quality from adaptive value. In the original accuracy-only $16A$ policy readout, the additive regression selects 1×16 on all 200 new problems. Under the later lexicographic rule, it selects depth on 189 and 2×8 on 11; both choices yield the same scored accuracy as fixed depth on this panel. Regression interactions permit more varied choices and reduce latency, with a small observed accuracy loss. Across all evaluated statistical and heuristic selectors, no accuracy gain over fixed depth is established. Outcome-based selection of the repeated-run panel, finite scoring banks, post-release comparison design, and provider-dependent latency limit stronger claims of per-problem optimality or deployment savings.

E.4 INDEPENDENT HOLDOUT TEST OF DEPTH AGAINST MIXED ALLOCATIONS

The 200 new problems are disjoint from the original main and development panels, with 20 AIME, 20 HMMT, 100 Omni-MATH and 60 GPQA problems. The pre-specified evaluation weights are $(0.10, 0.10, 0.50, 0.30)$. The matched $16A$ menu uses one $16A$ call, two $8A$ calls, four $4A$ calls, eight $2A$ calls, or sixteen A calls; multi-call outputs use plurality. The four-call allocation was locked from development results before scoring the new problems. Its accuracy is 58.0%, compared with 70.5% for depth; the paired studentized interval for depth minus the locked alternative is $[7.03, 18.43]$ percentage points, with Holm-adjusted $p = 0.0008$ across the two core primary tests. The strongest observed mixed option is two $8A$ calls at 65.0%; the remaining 8×2 and 16×1 options achieve 55.5% and 49.0%. This extension tests one-round independent allocations; the revision and debate evidence comes from the original panel. Fresh-problem labels were hidden during the original predictor fitting. The later expanded selector comparison is exploratory and was specified after these outcomes were visible.