PARETO FRONTIER OF LLM TEST-TIME COMPUTE: ADAPTIVE ROLLOUTS AND VOTING UNDER TOKEN BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models improve their test-time accuracy by generating multiple rollouts and aggregating them to produce a final answer by voting. As each additional rollout incurs token costs, test-time compute entails a trade-off between accuracy and inference cost. This raises a question: given a limited token budget, how should the budget be allocated to maximize accuracy? Existing approaches typically adapt the number of rollouts to each query while keeping the voting rule fixed. However, the voting rule itself shall be adapted as well. In particular, commonly used schemes such as self-consistency and best-of-n can be viewed as special cases of a broader Boltzmann voting family, and the choice of voting rule directly affects the attainable accuracy ceiling of test-time compute. Motivated by this observation, we formulate the rollout count and voting rule as joint, query-dependent decisions and optimize them under a token-budget constraint. We show that no fixed voting rule is uniformly optimal across all token budgets, whereas the adaptive voting rule yields a cost–accuracy Pareto frontier that dominates the frontier associated with any fixed rule. We further derive a closed-form characterization of the large-budget asymptotics of the Pareto frontier, identifying both the factors that determine the accuracy ceiling and those that govern the rate at which this ceiling is approached. Guided by these results, we propose PRICE (**P**riced **R**ollouts and **I**nference-time voting-rule **C**hoice under a token budg**E**t), an algorithm that jointly and adaptively selects the rollout count and voting rule as a function of the token budget. On MATH-500, PRICE consistently outperforms six test-time compute baselines across all evaluated budgets and achieves comparable accuracy using up to 2.8–3.5$\times$ fewer token budgets.

1 INTRODUCTION

Large language models improve their accuracy at inference time when a deployer draws more rollouts and aggregates them with a pre-specified voting rule. Specifically, Wang et al. (2023) introduced *self-consistency* (SC), which commits the modal answer across rollouts. Later work considers a per-rollout score, such as a process-reward-model output (Lightman et al., 2024; Wang et al., 2024) or the model's self-assessed score (Kadavath et al., 2022; Fu et al., 2026; Kang et al., 2025), and uses it to return the top-scored rollout (*best-of-n*, BoN) or to weight the vote (*score-weighted voting*). All these methods empirically achieve substantial accuracy gains over a single rollout (Snell et al., 2024; Brown et al., 2024).

These improvements come at a cost. More rollouts consume more tokens, so a deployer must balance inference accuracy against token usage. In practice, token usage is often constrained by a fixed budget, such as a subscription limit or a company wide token allotment. This budget must cover a stream of queries over a given period. The key question is thus not whether more compute improves accuracy but how to allocate a limited token budget across queries to maximize inference accuracy.

Prior work adapting rollout allocation. To save tokens, recent methods adapt the number of rollouts across queries, based on the intuition that easier queries often require fewer rollouts. For example, they stop generating rollouts once the vote stabilizes (Aggarwal et al., 2023; Li et al., 2024) or the model becomes sufficiently confident (Fu et al., 2026). Other methods allocate more rollouts

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.

based on predicted query difficulty (Damani et al., 2025). However, these methods adapt inference only at the rollout level. They decide how many rollouts each query receives while fixing the voting rule used to aggregate them.

Adaptive voting rule. As the same set of rollouts can produce different answers under different voting rules, it suggests that the voting rule itself should be adapted to each query rather than fixed in advance. We unify commonly used voting rules through the *Boltzmann weighted vote*, parameterized by a voting temperature $\tau \geq 0$ that controls how strongly higher scored rollouts are favored. Choosing a voting rule then reduces to choosing τ. Specifically, SC corresponds to $\tau = 0$, the exponential weighting used by CISC (Taubenfeld et al., 2025) corresponds to an intermediate τ, and BoN is recovered as $\tau \to \infty$.

Seen from this perspective, we find that the voting rule determines which queries can be answered correctly through repeated rollouts and therefore shapes the accuracy ceiling of test-time inference. For a given query, repeated rollouts can recover the correct answer only if that correct answer receives the largest voting weight. Under SC ($\tau = 0$), this requires the correct answer to appear more frequently than any alternative. Increasing τ places more weight on highly scored rollouts, which can recover the correct answer when it is relatively rare but receives high scores. However, the same sharper weighting can fail when the correct answer is frequent but receives lower scores than an incorrect alternative. Thus, the voting rule changes which queries can ultimately be answered correctly from repeated sampling, and different queries may favor different temperatures. By adapting τ across queries, adaptive voting can correctly answer a broader set of queries than any single fixed voting rule, thereby raising the attainable accuracy ceiling of test-time compute.

We thus aim to answer the following question:

*How to adaptively choose **the number of rollouts to draw** and **the voting rule to apply** for each query, so that the accuracy is maximized under **a token budget**?*

Line of work	adaptive rollout count	adaptive voting rule	budget constraint	Pareto frontier
Fixed-count methods: SC, BoN, weighted voting, CISC (Wang et al., 2023; Cobbe et al., 2021; Li et al., 2023; Kang et al., 2025; Taubenfeld et al., 2025)	–	–	–	–
Adaptive-count methods: Adaptive-Consistency, ESC, DeepConf (Aggarwal et al., 2023; Li et al., 2024; Fu et al., 2026)	✓	–	–	–
Budget-aware allocation: PETS, OptPO, Solve-then-Learn (Liu et al., 2026; Wang et al., 2025; Zhai et al., 2026)	✓	–	rollouts	–
PRICE (ours)	✓	✓	tokens	✓

Table 1: Comparison of PRICE against existing works along four dimensions: (i) if the rollout count is adaptive; (ii) if the voting rule is adaptive; (iii) whether the method takes the budget as an explicit constraint and, if so, whether the budget is imposed at the rollout level or at the token level; and (iv) if any Pareto frontier guarantee is established. For a more detailed discussion, see Appendix A.

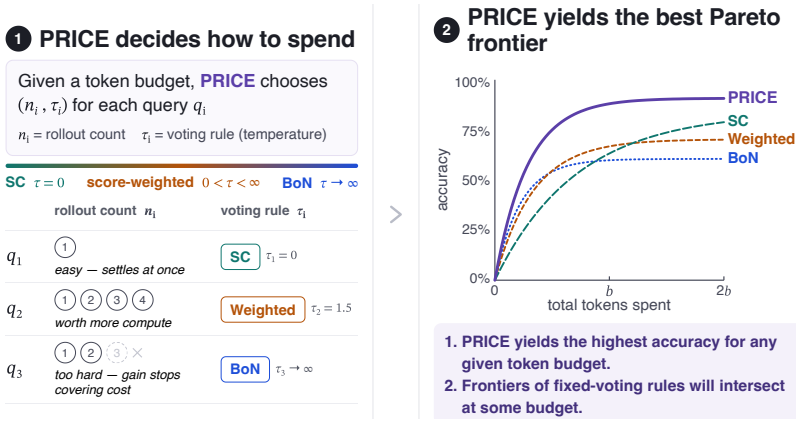


Figure 1: Illustration of the proposed PRICE method.

Table 1 compares existing lines of work with our proposed PRICE method, and Figure 1 illustrates the resulting accuracy–cost tradeoff. Our contributions are threefold.

First, we formulate LLM test-time compute as joint, query-dependent rollout and voting-rule selection under a token budget. We show that the Pareto frontiers of fixed voting rules can cross, so there is no optimal fixed rule across all budgets and queries, while adaptive voting dominates every fixed rule and can attain a higher accuracy ceiling by solving a broader set of queries correctly.

Second, we characterize the large-budget behavior of adaptive voting. We derive its Pareto frontier asymptotics in closed form, identifying both its accuracy ceiling and convergence rate to the ceiling. The gap to the ceiling decays exponentially at a rate determined by the hardest queries whose errors decay most slowly per token. On MATH-500, this gap closes 14–$17\times$ more slowly than the error of a typical query, explaining why additional tokens yield little gain once typical queries are answered correctly.

Third, we develop PRICE, the algorithm that jointly chooses how many rollouts to draw and how to vote for each query under a prespecified token budget. On MATH-500, PRICE outperforms six deployed baselines across all evaluated budget levels and achieves comparable accuracy using up to 2.8–$3.5\times$ fewer token budgets.

2 Why Voting Should Adapt Across Queries

2.1 Rollouts and the Boltzmann weighted vote

A deployer faces queries $q \sim \mathcal{Q}$, where $\mathcal{Q}$ is the distribution of queries, and pays for the tokens to answer these queries. Conditional on a query q, whose correct answer $a^*(q)$ is unknown, the LLM generates an independent and identically distributed (i.i.d.) sequence of rollout triples $\{(\ell_i, \phi_i, a_i)\}_i$, where each triple records the following.

- The *token cost* $\ell_i \in \mathbb{N}$ and the realized length of the rollout are measured in tokens.
- The *score* $\phi_i \in [0, 1]$, a deployer-chosen signal of the rollout's quality, observed once the rollout ends, e.g., the output of a process reward model (He et al., 2024), the generator's own evaluation (Kadavath et al., 2022), or its group confidence (Fu et al., 2026).
- The *answer* $a_i \in \mathcal{A}$, extracted from the rollout, from a finite answer space $\mathcal{A}$.

All per-query quantities above depend on q and we keep q implicit when working at a single query. After n rollouts, self-consistency (SC) returns the modal answer, and best-of-n (BoN) returns the answer of the highest-scoring rollout. Both belong to the family of voting rules below.

Definition 2.1 (Boltzmann weighted vote)**.** *Fix a query q and a temperature $\tau \geq 0$. After n rollouts, the Boltzmann weighted vote assigns to each answer the weight*

$$W_\tau(a, n; q) := \sum_{i=1}^{n} e^{\tau \phi_i} \, \mathbf{1}\{a_i = a\}, \qquad a \in \mathcal{A}, \tag{1}$$

and commits to the answer of the largest weight, $\widehat{a}_\tau(n; q) := \arg\max_{a \in \mathcal{A}} W_\tau(a, n; q)$. Ties are resolved following the rules in Appendix D.1.

At $\tau = 0$, every rollout carries unit weight and the vote is SC. For $0 < \tau < \infty$, the weight $e^{\tau \phi_i}$ increases with the score, and τ amplifies the higher-scored rollouts. As $\tau \to \infty$ the highest-scored rollout outweighs all others and the vote becomes BoN.

2.2 Fixed and adaptive voting policies

A recent line of work adapts the rollout count, using fewer rollouts when the extracted answers stabilize quickly to save compute. However, these methods keep the voting rule fixed, with a fixed τ determined in advance. We formally define the fixed and adaptive voting policies.

Definition 2.2 (Fixed- and adaptive-voting policies)**.** *A policy is a measurable map $\pi : q \mapsto \pi_q = (\tau(q), n(q))$. Given a query q, the policy commits to a temperature $\tau(q)$ and a rollout count $n(q)$. A* **fixed-voting policy** *adapts $n(q)$ to the query but uses one fixed $\tau(q) \equiv \tau \in [0, \tau_{\max}]$. In contrast,*

an **adaptive-voting policy** *adapts both $n(q)$ and $\tau(q)$. We write their per-query action sets as*

$$\Pi(\tau) := \{(\tau, n) : n \in \mathbb{N}\}, \qquad \Pi^{\mathrm{ad}} := \{(\tau, n) : \tau \in [0, \tau_{\max}],\ n \in \mathbb{N}\}.$$

Throughout, our theoretical work focuses on a finite cutoff $\tau_{\max} < \infty$ for clarity, while the experiments include BoN ($\tau = \infty$) as one additional action. Although the policy class of Definition 2.2 is discrete since the rollout count is an integer, we allow a policy to randomize over its actions so that the class becomes continuous for the theoretical analysis.

2.3 Consistency set for solvable queries

The first benefit of adapting the voting rule, or equivalently the temperature, is that adaptive voting can answer a broader set of queries correctly than any fixed rule. To formalize this, we identify, for each voting rule, the queries that it answers correctly in the limit of infinite rollouts.

Definition 2.3 (Consistency sets)**.** *A temperature τ is* consistent *for a query q if the Boltzmann vote at τ eventually returns the correct answer, $\widehat{a}_\tau(n; q) \to a^*(q)$ almost surely as $n \to \infty$. The* consistency set *of τ and the* adaptive consistency set *are*

$$\mathcal{C}(\tau) := \{q :\ \tau \text{ is consistent for } q\}, \qquad \mathcal{C}^{\mathrm{ad}} := \bigcup_{\tau \in [0, \tau_{\max}]} \mathcal{C}(\tau). \tag{2}$$

We call the queries in $\mathcal{C}^{\mathrm{ad}}$ solvable *by some voting rule, equivalently by some temperature.*

Since $\mathcal{C}(\tau) \subseteq \mathcal{C}^{\mathrm{ad}}$ for every τ,

$$\mathbb{P}_{q\sim\mathcal{Q}}[\mathcal{C}^{\mathrm{ad}}] \ \geq \ \sup_{\tau \in [0, \tau_{\max}]} \mathbb{P}_{q\sim\mathcal{Q}}[\mathcal{C}(\tau)], \tag{3}$$

so adaptive voting answers at least as many queries as any fixed rule. Appendix D.1.2 provides further characterization on the consistency sets and the condition when the inequality in (3) is strict.

3 Optimal Joint Allocation under Entropic-Risk Token Budget

3.1 Entropic-risk token budget

An adaptive-voting policy assigns each query q a pair $\pi_q = (\tau, n) \in \Pi^{\mathrm{ad}}$ and earns the population accuracy

$$\mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[r^{\pi_q} \mid q]\big], \quad \text{with} \quad r^{\pi_q} := \mathbf{1}\{\widehat{a}_\tau(n; q) = a^*(q)\}, \tag{4}$$

which is the probability that the returned answer is correct. Since it draws n rollouts, it also incurs the accumulated token cost $\ell^{\pi_q} := \sum_{i=1}^{n} \ell_i(q)$.

For the token budget, a natural first choice is to impose a cap on the mean cost, $\mathbb{E}_{q\sim\mathcal{Q}}[\mathbb{E}[\ell^{\pi_q} \mid q]] \leq b$ (Wan et al., 2026). However, realized costs can vary substantially across queries, since rollouts on hard queries may be several times longer than those on easy ones. A policy that satisfies only a mean-cost constraint can therefore still incur excessive cost on a substantial fraction of queries. To account for this variability, we control cost on the moment-generating function (MGF) scale using a pre-specified risk parameter $\gamma > 0$, and impose the *entropic-risk* constraint at the token-budget level $b > 0$,

$$\tfrac{1}{\gamma} \log \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q]\big] \leq b \quad \Longleftrightarrow \quad \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q]\big] \leq e^{\gamma b}. \tag{5}$$

As $\gamma \to 0_+$, it reduces to the constraint on the mean cost. By Markov's inequality, a policy satisfying (5) obeys, for every $t > 0$, $\mathbb{P}\big[\ell^{\pi_q} > b + t\big] \leq e^{-\gamma t}$. Therefore, the fraction of queries overspending by more than t tokens decays exponentially in t at rate γ. For instance, $\gamma = \log(20)/1024 \approx 2.9 \times 10^{-3}$ keeps the fraction of queries overspending by more than 1024 tokens below 5%.

3.2 Adaptive voting dominates fixed rules, while fixed rules can cross

The best accuracy attainable at budget b by adaptive-voting policies is

$$R^{\mathrm{ad}}(b) := \sup_{\pi_q} \Big\{ \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[r^{\pi_q} \mid q]\big],\ \text{s.t. } (5),\ \pi_q \in \Pi^{\mathrm{ad}} \Big\}. \tag{6}$$

We call a policy that attains the supremum in (6) an ***optimal adaptive policy*** and write

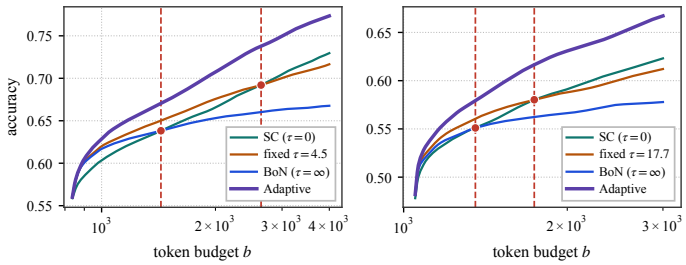
$$\pi^\star(b) \in \arg\max_{\pi_q} \Big\{ \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[r^{\pi_q} \mid q]\big],\ \text{s.t. (5)},\ \pi_q \in \Pi^{\mathrm{ad}} \Big\}, \tag{7}$$

when the maximum exists. Every policy draws at least one rollout, so budgets are taken above the one-rollout threshold $b > b_{\min} := \frac{1}{\gamma}\log \mathbb{E}_{q\sim\mathcal{Q}}[\mathbb{E}[e^{\gamma \ell_1(q)} \mid q]]$. The ***adaptive frontier*** is defined as

$$\mathcal{P}^{\mathrm{ad}} = \text{Upper Envelope}\Big\{ (b, R^{\mathrm{ad}}(b)) : b > b_{\min} \Big\}.$$

Similarly, we denote $R(b;\tau)$ as the best accuracy attainable at budget b by the fixed-voting policies at temperature τ, which substitutes the adaptive class Π^{ad} in (6) with the fixed-temperature class $\Pi(\tau)$. Its corresponding fixed-temperature frontier is denoted as $\mathcal{P}(\tau)$.

Next, we present the central benefit of adaptive voting, through two complementary theorems: (i) no fixed voting rule is universally best over all budgets, while (ii) the adaptive class dominates them all. We start with the first result, showing that any two fixed-τ frontiers can cross.



(a) Qwen2.5-1.5B: fixed rules swap the lead, adaptive on top (b) Llama-3.2-3B: same pattern

Figure 2: Fixed-voting frontiers cross, while the adaptive frontier dominates

Theorem 1 (Fixed-temperature frontiers can cross)**.** *Fix any two temperatures $\tau_1 \neq \tau_2$ in $[0, \tau_{\max}]$. There exist a query distribution $\mathcal{Q}$ and a score ϕ on which the frontiers $\mathcal{P}(\tau_1)$ and $\mathcal{P}(\tau_2)$ cross: there are budgets $b_1 < b_2$ with $R(b_1;\tau_1) > R(b_1;\tau_2)$ and $R(b_2;\tau_1) < R(b_2;\tau_2)$.*

The theorem states that for every pair of temperatures, there is a query population on which their frontiers cross, so no fixed temperature is optimal across all populations and budgets. Next, we show that the adaptive frontier dominates any fixed-temperature frontier.

Theorem 2 (Adaptivity dominates)**.** *For every $\tau \in [0, \tau_{\max}]$ and every $b > b_{\min}$, $R^{\mathrm{ad}}(b) \geq R(b;\tau)$.*

Figure 2 illustrates both results on MATH-500 for two LLM generators, following the protocol in Appendix C.4. In both panels, the best fixed temperature changes at the red markers, showing that no single temperature is optimal across all token budgets. In contrast, the adaptive frontier lies above every fixed-temperature frontier at every budget.

3.3 Solving the token budgeted problem with one dual price

With a dual multiplier $\lambda \geq 0$ on the budget constraint, the Lagrangian of (6) is

$$\mathcal{L}(\pi,\lambda) = \mathbb{E}_{q\sim\mathcal{Q}}\Big[\mathbb{E}[r^{\pi_q} \mid q] - \lambda\, \mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q]\Big] + \lambda e^{\gamma b}.$$

To make the Lagrangian more explicit in terms of the decision variables τ and n, we represent the per-query accuracy as

$$V_\tau(n;q) := \mathbb{E}[r^{\pi_q} \mid q] = \mathbb{P}\big[\widehat{a}_\tau(n;q) = a^*(q) \,\big|\, q\big]. \tag{8}$$

It is the probability of the vote at temperature τ that correctly solves the query q after n rollouts. As the rollout length is i.i.d. given q, the cost of n rollouts is $\mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q] = M(q)^n$ with $M(q) := \mathbb{E}[e^{\gamma \ell_1(q)} \mid q] \in (1, \infty)$ being the single-rollout cost MGF.

Proposition 1 (Strong duality and query-level decomposition)**.** *For $b > b_{\min}$, under the regularity conditions of Appendix D.1, strong duality holds and the problem decouples across queries:*

$$R^{\mathrm{ad}}(b) = \inf_{\lambda\geq 0} \sup_{\pi} \mathcal{L}(\pi,\lambda) = \inf_{\lambda\geq 0} \Big\{ \mathbb{E}_{q\sim\mathcal{Q}}\Big[\sup_{\tau,n}\big\{V_\tau(n;q) - \lambda M(q)^n\big\}\Big] + \lambda e^{\gamma b} \Big\}.$$

Strong duality holds since we show that the optimal value R^{ad} is concave in the MGF-scale budget $e^{\gamma b}$, by applying randomization to the original discrete rollout counts. In the rightmost expression, the supremum is taken *inside* the expectation, because the Lagrangian is separable across queries and the policy acts on each query independently. Then, the inner problem can be solved for each query at the common dual price λ. Therefore, to solve $R^{\mathrm{ad}}(b)$, it suffices to find the optimal per-query policy $\pi_q = (\tau(q), n(q)) \in \Pi^{\mathrm{ad}}$ at each dual price λ, and then take the infimum over λ.

At the supporting price λ_b^* of the outer infimum, which depends on the budget level b, the optimal adaptive policy therefore selects for each query

$$(\tau_b^*(q), n_b^*(q)) \in \arg\max_{\tau,n} \big\{ V_\tau(n;q) - \lambda_b^* M(q)^n \big\}. \tag{9}$$

3.4 What sets the plateau of LLM inference and the speed approaching it

A natural question is how the best accuracy $R^{\mathrm{ad}}(b)$ along the Pareto frontier $\mathcal{P}^{\mathrm{ad}}$ behaves as the budget b grows. This subection derives the closed-form large-budget asymptotics of $R^{\mathrm{ad}}(b)$. Specifically, we characterize its accuracy ceiling as $b \to \infty$, and the rate approaching this ceiling.

The ceiling of accuracy under unlimited budget. The best population accuracy with unlimited rollouts is

$$R^{\infty,\mathrm{ad}} := \mathbb{E}_{q\sim\mathcal{Q}}\Big[\sup_{\tau\in[0,\tau_{\max}],\, n\in\mathbb{N}} V_\tau(n;q)\Big]. \tag{10}$$

For a query $q \in \mathcal{C}^{\mathrm{ad}}$, the supremum equals 1 by Definition 2.3. For a query $q \notin \mathcal{C}^{\mathrm{ad}}$, no temperature achieves perfect accuracy in the infinite-rollout limit, although a suitable temperature and finite rollout count may still answer it correctly with positive probability. The ceiling therefore decomposes into the probability mass of queries that adaptive voting can answer correctly with probability approaching one and the residual accuracy on the remaining queries:

$$R^{\infty,\mathrm{ad}} = \mathbb{P}_{q\sim\mathcal{Q}}[\mathcal{C}^{\mathrm{ad}}] + \mathbb{E}_{q\sim\mathcal{Q}}\Big[\mathbf{1}\{q\notin\mathcal{C}^{\mathrm{ad}}\}\sup_{\tau,n} V_\tau(n;q)\Big]. \tag{11}$$

The first term is the consistency coverage of adaptive voting, which is at least as large as that of any fixed voting rule by (3).

The convergence rate to the accuracy ceiling. Given the ceiling, we ask how fast $R^{\mathrm{ad}}(b)$ approaches it as the budget grows. We show in Appendix D.1.3 that if a temperature τ is consistent for q, each additional rollout reduces the error $1 - V_\tau(n;q)$ exponentially in n. For a solvable query $q \in \mathcal{C}^{\mathrm{ad}}$, we therefore define the per-rollout error-decay rate at a consistent temperature, and its best value over the consistent temperatures.

$$I_\tau(q) := \lim_{n\to\infty} -\tfrac{1}{n}\log\big(1 - V_\tau(n;q)\big) \ \text{ for } q \in \mathcal{C}(\tau), \qquad I^{\mathrm{ad}}(q) := \sup_{\tau:\, q\in\mathcal{C}(\tau)} I_\tau(q).$$

Note that $I^{\mathrm{ad}}(q)$ is large when the error on query q decays quickly, i.e., a certain voting rule rapidly separates the correct answer from the incorrect ones. On the cost side, $\frac{1}{\gamma}\log M(q) = \frac{1}{\gamma}\log\mathbb{E}[e^{\gamma\ell_1(q)}|q]$ is the number of tokens that one rollout consumes in the entropic-risk measure. Their ratio is the query's ***exchange rate***,

$$\frac{I^{\mathrm{ad}}(q)}{\frac{1}{\gamma}\log M(q)} = \frac{\text{per-rollout error decay}}{\text{per-rollout token cost}} = \text{error decay per token}. \tag{12}$$

When the budget is large, the easy queries and unsolvable queries have saturated, and the remaining budget flows to the hardest but still solvable queries. Thus, the rate at which the population approaches its ceiling is determined by the worst exchange rate on the hardest solvable queries.

Theorem 3 (Large-budget law of Pareto frontier)**.** *Assume $\mathbb{P}_{q\sim\mathcal{Q}}[\mathcal{C}^{\mathrm{ad}}] > 0$ and the regularity conditions in Appendix D.4. Then, with the worst-case exchange rate*

$$r^{\mathrm{ad}} := \operatorname*{ess\,inf}_{q\in\mathcal{C}^{\mathrm{ad}}} \frac{I^{\mathrm{ad}}(q)}{\frac{1}{\gamma}\log M(q)} \in (0,\infty),$$

as the token budget $b \to \infty$,

$$R^{\infty,\mathrm{ad}} - R^{\mathrm{ad}}(b) = e^{-r^{\mathrm{ad}}\, b\,(1+o(1))}, \qquad \textit{i.e.,} \qquad \lim_{b\to\infty} -\tfrac{1}{b}\log\big(R^{\infty,\mathrm{ad}} - R^{\mathrm{ad}}(b)\big) = r^{\mathrm{ad}}. \tag{13}$$

Theorem 3 implies that accuracy saturates exponentially in the budget, rising toward the ceiling $R^{\infty,\mathrm{ad}}$ with the gap closing at the rate r^{ad}. Figure 3 in §4.1 validates this rate on MATH-500, where it is 14–17× below the median exchange rate of the queries.

The proof in Appendix D.4 builds on large-deviation theory to characterize asymptotically optimal per-query choices: for solvable queries, the voting rule maximizes the error-decay rate, while the rollout count balances the decreasing error against the increasing priced cost. We then choose the common price to satisfy the population budget and construct a feasible policy, yielding an upper bound on the gap $R^{\infty,\mathrm{ad}} - R^{\mathrm{ad}}(b)$. Finally, we use duality to derive a lower bound on the gap. The upper and lower bounds match in their exponential rates, establishing the claimed asymptotics.

4 Experiments

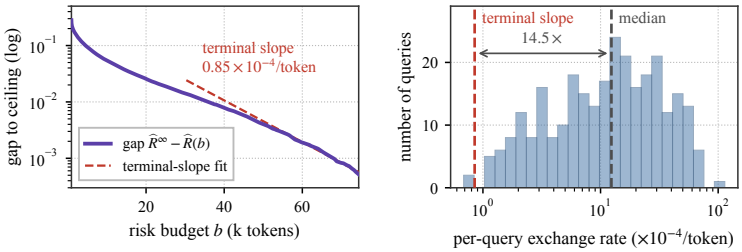
We introduce PRICE (**P**riced **R**ollouts and **I**nference-time voting-rule **C**hoice under a token budg**E**t), a family of algorithms motivated by the optimal policy in §3.3. The optimal adaptive policy in (9) requires two query-specific primitives, the accuracy curves $V_\tau(\cdot; q)$ and the cost MGF $M(q)$. We therefore develop two versions of PRICE that differ in how these primitives are obtained.

PRICE-oracle estimates them from a labeled pool of the query's own rollouts and solves the resulting constrained problem. It provides an oracle benchmark for joint rollout allocation and voting rule selection, but it is not yet ready for deployments as the labeled rollouts needed to estimate the accuracy curves are unavailable for a new query. **PRICE-deployed** instead estimates both primitives from a labeled calibration corpus and updates the estimates as the rollouts of the new query are generated. The detailed algorithms are provided in Appendix B.

We evaluate PRICE on MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024) with two LLM generators, Qwen2.5-1.5B-Instruct and Llama-3.2-3B-Instruct. The vote score ϕ is the generator's self-evaluation (Kadavath et al., 2022), the probability it assigns to the answer being correct, and the vote runs on a grid of 13 temperatures that includes SC ($\tau = 0$) and BoN ($\tau = \infty$). The risk parameter is $\gamma = \log(20)/1024 \approx 2.9 \times 10^{-3}$, which keeps the fraction of queries overspending by more than 1024 tokens below 5%, and a budget b is the realized entropic-risk budget of (5). Appendix C provides the configurations and additional results, including alternative scores.

4.1 PRICE-oracle on MATH-500

The large-budget law against the data. Figure 3 examines how the gap to the ceiling closes as the budget b grows on Qwen2.5-1.5B. Recall that Theorem 3 states that at large budgets, the gap closes exponentially with respect to the budget, at the worst-case exchange rate r^{ad}. The left panel plots the reward gap of PRICE-oracle to its ceiling on a log scale, where it decays almost linearly at large budgets with a terminal slope of 0.85×10^{-4} per risk-adjusted token. The right panel places this slope within the distribution of the per-query exchange rates (12) fitted on the solvable queries. The slope sits at the extreme left edge of the distribution, 14.5× below the median rate. The gap therefore closes at the pace of the slowest solvable queries rather than the typical ones, as Theorem 3 predicts. Llama-3.2-3B replicates the pattern with a terminal slope 17.1× below its median in Appendix C.8, where we also explain how the rates and the slope are fitted from the finite pools.



(a) Gap to the ceiling decays exponentially in b

(b) Terminal slope sits at the slow edge, 14.5× below the median

Figure 3: The large-budget law against the data (PRICE-oracle, Qwen2.5-1.5B).

Adaptive voting against the best fixed voting rule. Table 2 separates the two decisions, adapting the rollout count and adapting the voting rule. We first measure the benefit of adapting the rollout

count with the voting rule held fixed, for SC ($\tau = 0$), BoN ($\tau = \infty$) and the *hindsight-best* fixed temperature. The hindsight-best fixed temperature is the single temperature chosen after seeing the results that maximizes accuracy at each budget, so that no fixed voting rule can do better than this row at that budget. As shown by the green numbers in Table 2, adapting the count alone improves every voting rule at every budget.

We then compare PRICE-oracle, which further adapts the voting rule, against the hindsight-best fixed rule. PRICE-oracle is above it at every budget on both LLM generators, with the gain highlighted in red colors. This is the gain of adaptive voting rather than fixing one in advance, and it comes on top of the gain from adaptive count alone.

	Qwen2.5-1.5B						Llama-3.2-3B					
Method budget $b \rightarrow$	2k	5k	8k	12k	20k	30k	2k	5k	8k	12k	20k	30k
SC ($\tau = 0$)												
fixed count	48.4	55.8	59.6	62.6	65.3	66.6	40.9	47.4	51.7	54.5	57.3	59.0
adaptive count	56.3	61.6	64.0	65.8	67.5	68.4	49.7	54.6	57.0	58.5	60.4	61.5
	(+7.9)	(+5.7)	(+4.4)	(+3.2)	(+2.3)	(+1.7)	(+8.8)	(+7.2)	(+5.4)	(+4.0)	(+3.1)	(+2.5)
BoN ($\tau = \infty$)												
fixed count	49.3	48.8	45.6	41.4	35.7	31.1	40.9	42.2	39.6	36.8	33.0	29.7
adaptive count	53.8	55.1	55.4	55.5	55.6	55.6	46.9	48.0	48.2	48.4	48.5	48.5
	(+4.5)	(+6.3)	(+9.8)	(+14.1)	(+19.8)	(+24.5)	(+6.0)	(+5.7)	(+8.6)	(+11.6)	(+15.6)	(+18.8)
Best fixed τ (hindsight)												
fixed count	49.3	56.9	60.4	63.2	65.6	66.8	40.9	47.4	51.7	54.5	57.3	59.0
adaptive count	56.9	62.2	64.5	66.2	67.8	68.5	49.8	54.6	57.0	58.5	60.4	61.5
	(+7.6)	(+5.3)	(+4.0)	(+3.0)	(+2.2)	(+1.7)	(+8.9)	(+7.2)	(+5.4)	(+4.0)	(+3.1)	(+2.5)
PRICE-oracle	**58.9**	**64.8**	**67.4**	**69.1**	**70.9**	**71.7**	**51.5**	**56.6**	**58.7**	**60.3**	**62.3**	**63.5**
adaptive count & vote	(+2.0)	(+2.6)	(+2.9)	(+2.9)	(+3.1)	(+3.2)	(+1.7)	(+2.0)	(+1.7)	(+1.8)	(+1.9)	(+2.0)

Table 2: PRICE-oracle against fixed voting rules on MATH-500 (accuracy in %, at matched realized budget). Green numbers are the gains of the adaptive count over the fixed count with the same voting rule, and red numbers are the gains of PRICE-oracle over the best fixed τ with adaptive count.

4.2 PRICE-DEPLOYED AGAINST BASELINES

We compare PRICE-deployed against six deployed test-time-compute methods, in two families.

- *Fixed-count* methods commit a pre-fixed rollout count n on every query: fixed-n SC, fixed-n BoN, and CISC (Taubenfeld et al., 2025), a member of the Boltzmann vote family.
- *Adaptive-count* methods stop sampling per query but pin the voting rule: Adaptive-Consistency (Aggarwal et al., 2023), ESC (Li et al., 2024), and DeepConf (Fu et al., 2026).

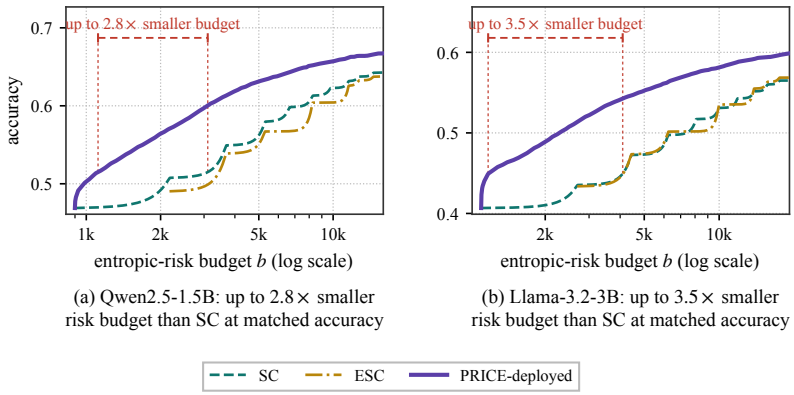
	Qwen2.5-1.5B						Llama-3.2-3B					
Method budget $b \rightarrow$	3.7k	5.3k	6.7k	11.6k	14.5k	32.0k	4.4k	6.2k	8.1k	13.8k	17.4k	40.4k
Fixed-count												
Self-consistency	54.9	58.0	59.8	63.2	64.2	67.1	47.3	49.7	51.7	55.2	56.5	59.9
Best-of-n	49.8	48.1	46.0	41.2	38.9	30.4	42.7	41.3	39.8	35.9	33.9	27.1
CISC	54.9	55.0	59.8	62.3	64.3	66.8	47.3	47.3	51.8	54.3	56.5	59.7
Adaptive-count												
Adaptive-Consistency	46.9	47.3	48.9	48.9	48.9	48.9	40.7	40.7	40.7	40.7	44.4	44.4
ESC	53.9	56.7	56.7	62.6	63.7	65.9	47.4	50.2	50.2	55.5	56.9	59.2
DeepConf (offline)	52.2	58.4	58.4	63.9	63.9	66.9	43.9	50.2	50.2	55.8	55.8	59.5
PRICE-deployed	**61.3**	**63.3**	**64.3**	**66.2**	**66.6**	**67.6**	**54.6**	**56.3**	**57.5**	**59.2**	**59.7**	**61.0**
	(+6.4)	(+4.9)	(+4.5)	(+2.2)	(+2.2)	(+0.5)	(+7.3)	(+6.1)	(+5.7)	(+3.3)	(+2.8)	(+1.0)

Table 3: Accuracy (%) at matched realized budget b on MATH-500. The red numbers are PRICE-deployed minus the best baseline in the column.

Table 3 reports the accuracy attained at matched token budget, and makes two points.

PRICE-deployed wins uniformly. It beats every baseline at every budget on both LLM generators. The baselines themselves are hard to separate, since SC, CISC, ESC and DeepConf track one another to within a few points at matched budgets.

The lead widens as the budget shrinks. PRICE-deployed beats the best baseline by $+6.4$ pp at $b = 3.7$k versus $+0.5$ pp at $b = 32.0$k on Qwen2.5-1.5B, and by $+7.3$ pp at $b = 4.4$k versus $+1.0$ pp at $b = 40.4$k on Llama-3.2-3B.



(a) Qwen2.5-1.5B: up to 2.8× smaller risk budget than SC at matched accuracy

(b) Llama-3.2-3B: up to 3.5× smaller risk budget than SC at matched accuracy

Figure 4: The frontier of PRICE-deployed against SC and ESC. The red dimension lines mark the accuracy at which the budget saving against SC is largest.

Figure 4 traces the frontier of PRICE-deployed against SC and ESC on the same realized cost axis. The frontier of PRICE-deployed lies above both at every budget on both LLM generators. At matched accuracy, PRICE-deployed requires a smaller token budget than both baselines, using up to $2.8\times$ and $3.5\times$ fewer tokens against SC on Qwen2.5-1.5B and Llama-3.2-3B, respectively.

5 Related Work

Adaptive rollouts and answer selection. *Fixed-count* methods draw a prespecified number of rollouts and select an answer by majority vote (Wang et al., 2023), by the highest score (Cobbe et al., 2021; Kang et al., 2025), or by score-weighted voting (Li et al., 2023; Lightman et al., 2024; Taubenfeld et al., 2025). *Adaptive-count* methods stop sampling once the vote or its confidence stabilizes (Aggarwal et al., 2023; Li et al., 2024; Fu et al., 2026; Manvi et al., 2024; Komiyama et al., 2026), but keep the voting rule fixed in advance, whereas PRICE chooses the rollout count and the voting rule jointly for each query.

Budget-aware constrained formulations. Compute-optimal and learned allocators select generation strategies or sample counts under a compute budget (Snell et al., 2024; Damani et al., 2025; Brown et al., 2024; Wu et al., 2025; Han et al., 2025). PETS, OptPO and Solve-then-Learn allocate sample counts under sampling or average-compute constraints (Liu et al., 2026; Wang et al., 2025; Zhai et al., 2026), and CLEAR (Wan et al., 2026) allocates tokens across queries through a global shadow price, with a per-query generation-length limit in a single-rollout regime as its allocation variable. We instead optimize the rollout count and a continuously tunable voting rule jointly under an entropic-risk token constraint in the multiple-rollout regime. Appendix A discusses these works in more detail, other forms of test-time compute, and the connections to constrained optimization.

6 Conclusion

We cast LLM test-time compute as constrained problem over policies that choose, per query, both how many rollouts to draw and how to vote over them, under an entropic-risk token budget. The voting rule decides which queries repeated sampling can solve, and adapting it per query enlarges the solvable set. We show that adaptive voting policy dominates every fixed voting rules in terms of accuracy at matched budgets, while the Pareto frontiers of fixed voting rules can cross. Moreover, we establish the closed-form asymptotic characterization of the adaptive frontier. Then we introduce the algorithm PRICE, which beats six deployed baselines at every evaluated budget and matches the accuracy of baselines with up to 2.8–$3.5\times$ smaller token budgets.

AI USE STATEMENT

We used AI tools in this work as follows. Among tasks requiring disclosure, AI tools assisted with checking and strengthening the rigor of proofs in the Appendix, implementing code for the experiments in Section 4 and the Appendix, and drafting interpretations of experimental results. All such outputs were verified against the underlying data and revised by the authors.

We did not use AI tools to generate synthetic data, develop the theoretical framework, formulate the mathematical claims, or provide the key ideas underlying the proofs. Translation, dataset cleaning or reformatting, and qualitative data analysis were not relevant to this work.

Among tasks for which disclosure is recommended, we used AI tools to assist with code and figure generation and to draft and edit portions of the manuscript for readability. All AI-assisted proofs were checked by the authors, all AI-assisted code was reviewed and its outputs verified, and all AI-assisted text was revised by the authors. We take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of AI.

ETHICS STATEMENT

All experiments use a public mathematical reasoning benchmark, MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), and two open-weight generators, Qwen2.5-1.5B-Instruct and Llama-3.2-3B-Instruct. The work involves no human subjects, no personal or private data, and no release of new datasets or models. No societal consequences that need to be specifically highlighted.

REPRODUCIBILITY STATEMENT

The theoretical results of §2 and §3 are proved in Appendix D. Both versions of PRICE algorithms are given in Appendix B. The configurations of experiments in §4 and additional experimental results are provided in Appendix C. All datasets and models used are publicly available and cited in the paper.

REFERENCES

P. Aggarwal, A. Madaan, Y. Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12375–12396, 2023. doi: 10.18653/v1/2023.emnlp-main.761.

R. R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *Annals of Mathematical Statistics*, 31(4):1015–1027, 1960. doi: 10.1214/aoms/1177705674.

S. Balseiro, H. Lu, and V. Mirrokni. Dual mirror descent for online allocation problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 613–628, 2020. URL https://proceedings.mlr.press/v119/balseiro20a.html.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/.

B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv:2407.21787*, 2024.

A. Charnes and W. W. Cooper. Chance-constrained programming. *Management Science*, 6(1):73–79, 1959. doi: 10.1287/mnsc.6.1.73. URL https://doi.org/10.1287/mnsc.6.1.73.

L. Chen, J. Q. Davis, B. Hanin, P. Bailis, I. Stoica, M. Zaharia, and J. Zou. Are more LLM calls all you need? Towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

M. Damani, I. Shenfeld, A. Peng, A. Bobu, and J. Andreas. Learning how hard to think: Input-adaptive allocation of LM computation. In *International Conference on Learning Representations (ICLR)*, 2025.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2010. doi: 10.1007/978-3-642-03311-7.

H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter, 3rd edition, 2011.

Y. Fu, X. Wang, H. Zhang, Y. Tian, and J. Zhao. Deep think with confidence. In *International Conference on Learning Representations (ICLR)*, 2026.

T. Han, Z. Wang, C. Fang, S. Zhao, S. Ma, and Z. Chen. Token-budget-aware LLM reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025. doi: 10.18653/v1/2025.findings-acl.1274. URL https://aclanthology.org/2025.findings-acl.1274/.

J. He, T. Wei, R. Yan, J. Liu, C. Wang, Y. Gan, S. Tu, C. Y. Liu, L. Zeng, X. Wang, B. Wang, Y. Li, F. Zhang, J. Xu, B. An, Y. Liu, and Y. Zhou. Skywork-o1 open series. Zenodo, version 1.0.0, November 2024. URL https://huggingface.co/Skywork/Skywork-o1-Open-PRM-Qwen-2.5-7B.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

Z. Kang, X. Zhao, and D. Song. Scalable best-of-N selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49:237–252, 1998. doi: 10.1057/palgrave.jors.2600523. URL https://www.statslab.cam.ac.uk/~frank/rate.pdf.

J. Komiyama, D. Oba, and M. Oyamada. Best-of-infinity: Asymptotic performance of test-time LLM ensembling. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3qiCnLf3jf.

P. Kuang, Y. Wang, X. Han, Y. Liu, K. Xu, and H. Wang. Optimal aggregation of LLM and PRM signals for efficient test-time scaling. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=x85kiYqL4y.

G. Lan and Z. Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76:461–498, 2020. doi: 10.1007/s10589-020-00179-x. URL https://doi.org/10.1007/s10589-020-00179-x.

Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou, and W. Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5315–5333, 2023. doi: 10.18653/v1/2023.acl-long.291.

Y. Li, P. Yuan, S. Feng, B. Pan, X. Wang, B. Sun, H. Wang, and K. Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Z. Liu, H. Qu, X. Yin, H. Sun, Y. Han, T. Chen, and Z. Deng. PETS: A principled framework towards optimal trajectory allocation for efficient test-time self-consistency. *arXiv:2602.16745*, 2026.

R. Manvi, A. Singh, and S. Ermon. Adaptive inference-time compute: LLMs can predict if they can do better, even mid-generation. *arXiv:2410.02725*, 2024. URL https://arxiv.org/abs/2410.02725.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, F.-F. Li, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*, 2024.

A. Taubenfeld, T. Sheffer, E. Ofek, A. Feder, A. Goldstein, Z. Gekhman, and G. Yona. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. doi: 10.18653/v1/2025.findings-acl.1030.

X. Wan, S. Zhu, J. Cai, G. Chen, X. Huang, W. Zhou, and M. Sun. The shadow price of reasoning: Economic perspective on optimal budget allocation for LLMs. *arXiv:2606.03092*, 2026.

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Y. Wang, J. Wang, R. Chen, T. Zeng, X.-Y. Wei, and Q. Li. OptPO: Optimal rollout allocation for test-time policy optimization. *arXiv:2512.02882*, 2025.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

P. Whittle. *Risk-Sensitive Optimal Control*. Wiley, 1990.

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

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.nips.cc/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Z. Zhai, B. Li, B. Xiao, M. Li, and X. Wang. Adaptive test-time compute allocation for reasoning LLMs via constrained policy optimization. *arXiv:2604.14853*, 2026.

A Related Work

Adaptive rollouts and answer selection. Methods of test-time compute involving drawing multiple rollouts are classified by whether they adapt the rollout count. *Fixed-count* methods use a prespecified rollout count across queries and then apply a voting rule to select an answer. For example, self-consistency takes the modal answer (Wang et al., 2023), best-of-n returns the highest-scored rollout (Cobbe et al., 2021; Kang et al., 2025), and weighted voting aggregates answers using verifier scores or confidence (Li et al., 2023; Lightman et al., 2024; Taubenfeld et al., 2025). Within our Boltzmann vote family indexed by the temperature τ, SC uses $\tau = 0$, CISC (Taubenfeld et al., 2025) uses a fixed finite τ, and BoN corresponds to $\tau = \infty$. Further work studies non-monotone accuracy as the sample count grows and calibrates aggregation weights using generator and verifier signals (Chen et al., 2024; Kuang et al., 2026). The latter also considers negative weights, which lie outside our Boltzmann vote family.

Adaptive-count methods instead adjust the rollout count for each query, stopping once the vote or its confidence stabilizes, as in Adaptive-Consistency, ESC, and DeepConf (Aggarwal et al., 2023; Li et al., 2024; Fu et al., 2026). Related approaches predict whether additional generation is worthwhile or use answer agreement to approximate the infinite-sample majority vote (Manvi et al., 2024; Komiyama et al., 2026). All these methods use a prespecified aggregation mechanism, whereas PRICE jointly chooses the rollout count and voting rule for each query.

Other LLM test-time compute. Apart from the multiple rollout regime this work focuses on, LLM test-time compute can improve accuracy by extending an individual reasoning trajectory or searching over intermediate reasoning states. Chain-of-thought generates intermediate steps before answering, while Tree of Thoughts explores alternative reasoning paths (Wei et al., 2022; Yao et al., 2023). More recently, s1 uses budget forcing to control how long a model reasons (Muennighoff et al., 2025). These approaches complement the methods that generate multiple rollouts and then select an answer from them.

Budget-aware constrained formulations. Another line of work studies the tradeoff between inference cost and accuracy. Repeated-sampling studies characterize how coverage and selection performance scale with inference compute (Brown et al., 2024). Compute-optimal (Snell et al., 2024) and learned allocators (Damani et al., 2025) select generation strategies or sample counts under a compute budget . Related studies examine compute-optimal combinations of model size and inference strategy (Wu et al., 2025), or control reasoning length through token budgets (Han et al., 2025). PETS allocates trajectories to maximize agreement with the infinite-budget majority vote under a rollout-based budget (Liu et al., 2026). OptPO uses Bayesian sequential testing to stop rollouts and conducts test-time policy optimization (Wang et al., 2025). Solve-then-Learn (Zhai et al., 2026) allocates rollout counts under an average-compute constraint via Lagrangian relaxation. Our focus is the joint optimization of rollout count and a continuously tunable voting rule under an entropic-risk token constraint. More recently, CLEAR (Wan et al., 2026) allocates tokens across queries via a global shadow price over a parametric per-query utility model. Its core allocation variable is a per-query generation-length limit in a single-rollout regime. In contrast, our work focuses on the multiple-rollout regime, where we decide how many rollouts to allocate to each query and how to aggregate the resulting answers across rollouts.

Connections to constrained optimization. Deciding how to spend a token budget across queries to maximize accuracy is a resource-allocation problem. A common formulation constrains expected cost (Lan & Zhou, 2020), which controls the average token cost but still allows individual queries to overspend tokens severely. Chance constraints instead limit the probability that costs exceed a given threshold (Charnes & Cooper, 1959). Our entropic-risk budget considers the constraint on the exponential scale, which bounds the probability of large budget overspending (Whittle, 1990; Föllmer & Schied, 2011). To solve the budget constrained problem, we adopt Lagrangian relaxation, which assigns a common shadow price to token usage across queries (Boyd & Vandenberghe, 2004; Kelly et al., 1998). Each query then independently decides its rollout count and voting rule by trading off expected accuracy against this priced token cost. When queries arrive sequentially, we adopt the standard online allocation method to adjust the shadow price through exponential updates (Balseiro et al., 2020).

B ALGORITHMIC DESIGN: PRICE, ORACLE AND DEPLOYED

We now introduce PRICE (**P**riced **R**ollouts and **I**nference-time voting-rule **C**hoice under a token budg**E**t), a family of algorithms motivated by the optimal policy in §3.3. The optimal adaptive policy in (9) requires two query-specific primitives, the accuracy curves $V_\tau(\cdot; q)$ and the cost MGF $M(q)$. We therefore develop two versions of PRICE that differ in how these primitives are obtained. PRICE-oracle estimates them from a labeled pool of the query's own rollouts and solves the resulting constrained problem. It provides an oracle benchmark for joint rollout allocation and voting rule selection, but it is not yet ready for deployments as the labeled rollouts needed to estimate the accuracy curves are unavailable for a new query. PRICE-deployed instead estimates both primitives from a labeled calibration corpus and updates the estimates as the rollouts of the new query are generated. Both versions take a token budget b as input and convert it into a shadow price λ that governs how many rollouts each query receives. We evaluate both algorithms on MATH-500 in §4.

Throughout this section, $\gamma > 0$ is the risk parameter in (5), $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$ is a grid of K voting temperatures, and $\Lambda \subset (0, \infty)$ is a finite grid of candidate dual prices.

B.1 PRICE-ORACLE

In the oracle setup, the two primitives of every query are computed from the query's own rollouts. Suppose we have Q queries $q_1, \ldots, q_Q$, each with a known correct answer $a^*(q)$ and a pool of m rollouts $\{(\ell_i, \phi_i, a_i)\}_{i \le m}$, which record the token cost, the score, and the answer of each rollout. Recall that the per-query accuracy curve $V_\tau(n; q)$ is the probability that the vote returns the correct answer at temperature τ over n i.i.d. rollouts, so we estimate it by resampling the pool. Specifically, we draw S uniformly random orderings $\sigma_1, \ldots, \sigma_S$ of the pool, and for each ordering σ_s we run the vote at temperature τ on its first n rollouts, whose winner we denote by $\widehat{a}_\tau(n; \sigma_s)$.

We estimate the per-query accuracy by the fraction of orderings on which the vote is correct,

$$\widehat{V}_\tau(n; q) := \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}\big\{\widehat{a}_\tau(n; \sigma_s) = a^*(q)\big\}, \tag{14}$$

for every $\tau \in \mathcal{T}$ and $n \le m$. The cost MGF is estimated by its empirical mean over the pool,

$$\widehat{M}(q) := \frac{1}{m} \sum_{i=1}^{m} e^{\gamma \ell_i}. \tag{15}$$

At a fixed price $\lambda \in \Lambda$, PRICE-oracle substitutes these estimates into the per-query problem (50) and selects

$$\big(\widehat{\tau}_\lambda(q), \widehat{n}_\lambda(q)\big) \in \arg\max_{\tau \in \mathcal{T},\, n \le m} \big\{\widehat{V}_\tau(n; q) - \lambda\, \widehat{M}(q)^{\,n}\big\}. \tag{16}$$

For a prespecified budget level b, Proposition 1 shows that there is a corresponding supporting price λ_b^*. Next, we describe how to find its empirical counterpart $\widehat{\lambda}_b$ on the grid Λ. Note that at each trial price λ, (16) selects a count $\widehat{n}_\lambda(q_i)$ for each query q_i. For each ordering σ_s of the query's pool, we sum the token costs of the first $\widehat{n}_\lambda(q_i)$ rollouts and denote this total token cost by $\widehat{c}_\lambda(q_i; \sigma_s)$. Aggregating over the Q queries and the S orderings as in (5), we define the realized entropic-risk token cost $\widehat{C}(\lambda)$ and select the smallest grid price that meets the budget level b,

$$\widehat{C}(\lambda) := \frac{1}{\gamma} \log\left(\frac{1}{QS} \sum_{i=1}^{Q} \sum_{s=1}^{S} e^{\gamma \widehat{c}_\lambda(q_i; \sigma_s)}\right), \qquad \widehat{\lambda}_b := \min\big\{\lambda \in \Lambda : \widehat{C}(\lambda) \le b\big\}. \tag{17}$$

Here, we consider that the price grid Λ is wide enough that its largest price meets the budget.

We use bisection search to find $\widehat{\lambda}_b$, because $\widehat{C}(\lambda)$ is non-increasing in λ. Specifically, fix a query q, and let (τ, n) and (τ', n') be its selected pairs at two prices $\lambda < \lambda'$. By (16), it holds that

$$\begin{aligned} \widehat{V}_\tau(n; q) - \lambda\, \widehat{M}(q)^n &\ge \widehat{V}_{\tau'}(n'; q) - \lambda\, \widehat{M}(q)^{n'}, \\ \widehat{V}_{\tau'}(n'; q) - \lambda' \widehat{M}(q)^{n'} &\ge \widehat{V}_\tau(n; q) - \lambda' \widehat{M}(q)^n. \end{aligned}$$

Adding these two inequalities cancels the accuracy terms and yields

$$(\lambda' - \lambda)\big(\widehat{M}(q)^{n'} - \widehat{M}(q)^{n}\big) \le 0.$$

Since rollout lengths are positive, we have $\widehat{M}(q) > 1$, and hence $n' \le n$ whenever $\lambda < \lambda'$. Consequently, the selected rollout count $\widehat{n}_\lambda(q_i)$ is non-increasing in λ for each q_i. On a fixed ordering, a shorter prefix spends no more tokens, so every $\widehat{c}_\lambda(q_i; \sigma_s)$, and therefore $\widehat{C}(\lambda)$, is non-increasing in λ.

After $\widehat{\lambda}_b$ is selected, PRICE-oracle commits the pair in (16) at this price and returns the corresponding vote, as summarized in Algorithm 1.

Algorithm 1 PRICE-oracle

Input: queries $q_1, \ldots, q_Q$ with their rollout pools and correct answers; budget b; risk γ; grids $\mathcal{T}$ and Λ

▷ *Phase 1: estimation of the primitives and bisection search of $\widehat{\lambda}_b$*
for each query q **do**
  Estimate $\widehat{V}_\tau(n; q)$ and $\widehat{M}(q)$ using (14) and (15)
end for
Bisection search for $\widehat{\lambda}_b$ using (17)
▷ *Phase 2: commit*
for each query q **do**
  Commit the pair in (16) at $\lambda = \widehat{\lambda}_b$ and return the corresponding vote
end for

Given a specified budget level b, Algorithm 1 yields an estimate of the point on the Pareto frontier $\mathcal{P}^{\mathrm{ad}}$ associated with b. And the entire frontier is traced by varying the budget level b. By the Lagrangian form in Proposition 1, each budget b corresponds to a supporting price λ_b^*, so sweeping the budget is equivalent to sweeping the price λ. Therefore, to display the entire Pareto frontier in later §4, we directly sweep $\lambda \in \Lambda$ for Algorithm 1.

We emphasize that PRICE-oracle uses each query's own rollouts to estimate its accuracy curves and cost MGF, which is not implementable for a new query. However, it can be interpreted as follows. When the primitives are given on the queries' own labeled rollouts, it tells us how much gain the optimal adaptive policy can offer over the baseline methods; see §4.1.

B.2 PRICE-DEPLOYED

In the deployed setting, we have neither a rollout pool nor the correct answer of a new query. We thus predict its accuracy curves and cost MGF from previously observed queries. Specifically, we construct a labeled calibration corpus and compute both primitives for each calibration query by (14) and (15). We split the corpus into fitting queries, on which we fit two predictors, and held-out queries, which are reserved for the price search below. The first predicts the accuracy curves across temperatures and rollout counts using the query embedding and label-free statistics from the rollouts generated so far, including answer agreement, score distributions, and rollout lengths. The second predicts the cost MGF from the query embedding and updates this prediction using the observed rollout lengths. As new rollouts arrive, both predictions are refreshed, allowing the estimates to improve online. With a slight abuse of notation, we keep writing $\widehat{V}_\tau(n; q)$ and $\widehat{M}(q)$ for the current predictions. Appendix B.3 details the calibration corpus and fitting procedure.

Another major difference from PRICE-oracle is in how the rollout count is chosen. Instead of committing the count that maximizes the net value in (16), PRICE-deployed uses a sequential stopping rule. After each rollout n, it evaluates the predicted accuracy at the best temperature, $\widehat{V}^\star(n'; q) = \max_{\tau \in \mathcal{T}} \widehat{V}_\tau(n'; q)$, at the current count and the next few counts $n' > n$, and records the accumulated token cost $L_n = \sum_{i \le n} \ell_i$. At a price λ, it stops when the largest predicted average accuracy gain over the next w rollouts no longer exceeds the priced marginal cost of another rollout,

$$\max_{1 \le j \le w} \frac{\widehat{V}^\star(n+j; q) - \widehat{V}^\star(n; q)}{j} \;\le\; \lambda\, e^{\gamma L_n}\big(\widehat{M}(q) - 1\big). \tag{18}$$

Algorithm 2 PRICE-deployed

Input: labeled calibration corpus with pooled rollouts, split into fitting and held-out queries; budget b; risk γ; grids $\mathcal{T}$ and Λ; window w; step sizes η_t; query stream $q_1, q_2, \ldots$

▷ *Phase 1: calibration and bisection search of* $\widehat{\lambda}_b$
Fit the accuracy and cost predictors on the fitting queries, using targets from (14) and (15)
Bisection search for $\widehat{\lambda}_b \in \Lambda$ using (17) on the held-out calibration queries
Initialize $\lambda_1 \leftarrow \widehat{\lambda}_b$
▷ *Phase 2: sequential rollouts and online price updates*
for $t = 1, 2, \ldots$ **do**
  Set $q \leftarrow q_t$, $\lambda \leftarrow \lambda_t$, and $L_0 \leftarrow 0$
  for $n = 1, 2, \ldots$ **do**
    Draw rollout (ℓ_n, ϕ_n, a_n); set $L_n \leftarrow L_{n-1} + \ell_n$
    Obtain the current predictions $\widehat{V}_\tau(k; q)$, for $\tau \in \mathcal{T}$ and $n \le k \le n + w$, and $\widehat{M}(q)$
    Break if (18) holds
  end for
  ▷ *Commit*
  Select $\widehat{\tau}(q) \in \arg\max_{\tau \in \mathcal{T}} \widehat{V}_\tau(n; q)$
  Output the vote over the n observed rollouts at $\widehat{\tau}(q)$
  ▷ *Price update*
  Set $L^{(t)} \leftarrow L_n$ and update λ_{t+1} using (19)
end for

On the right-hand side, one more rollout is predicted to raise the MGF cost from $e^{\gamma L_n}$ to $e^{\gamma L_n}\widehat{M}(q)$. The look-ahead window w prevents premature stopping when the next rollout alone offers little gain but several additional rollouts may offer more. We use $w = 4$ in our experiments. After stopping, PRICE-deployed aggregates the n observed rollouts using the estimated best temperature $\widehat{\tau}(q) \in \arg\max_{\tau \in \mathcal{T}} \widehat{V}_\tau(n; q)$.

We adopt the sequential stopping rule rather than the net value maximization in (16) for three reasons. First, the accuracy curve estimates improve as more rollouts are observed, so deciding after each rollout uses a better estimate than deciding once at the start. Second, L_n records the tokens already consumed, allowing the rule to stop a query whose rollouts have already run long. Third, the deployed adaptive-count baselines also use sequential stopping, so the comparison in §4.2 places all methods under the same decision protocol. This sequential heuristic differs from the adaptive policies analyzed in §3, which commit to the rollout count and temperature before generating any rollouts. Therefore, the theoretical guarantees for the committed policies do not directly apply to PRICE-deployed. We evaluate its performance empirically in §4.2, and leave a formal analysis of its sequential behavior to future work.

Given a specified budget level b, it remains to find the estimate of its corresponding supporting price λ_b^*. We initialize the price with $\widehat{\lambda}_b$, found by the same bisection search as in Algorithm 1, now run on the held-out calibration queries. We then update the price across the online query stream. Specifically, let q_t be the t-th arriving query and λ_t its price, initialized at $\lambda_1 = \widehat{\lambda}_b$. This price stays fixed throughout the query's rollouts. After the query finishes with total token cost $L^{(t)}$, we update the price using a positive step size η_t via stochastic gradient ascent,

$$\lambda_{t+1} = \lambda_t \exp\Big\{\eta_t \Big[e^{\gamma(L^{(t)} - b)} - 1\Big]\Big\}. \tag{19}$$

The term in square brackets is the realized MGF cost divided by its target $e^{\gamma b}$, minus one. Thus overspending raises the price for subsequent queries, while underspending lowers it. Algorithm 2 combines the calibration, sequential stopping, and online price update.

As with PRICE-oracle, Algorithm 2 gives the adaptive policy at one specified token budget level b. To display the frontier across budget levels, it suffices to sweep each price $\lambda \in \Lambda$, and we present the resulting frontier in §4.2. Appendix C.10 further examines how closely a price calibrated in advance on held-out queries meets the requested budget for the new incoming queries.

B.3 CALIBRATION OF PRICE-DEPLOYED

The reward-curve predictor outputs the full temperature-by-count array $\widehat{V}_\tau(m; q)$ for the 13 temperatures and counts $m \leq 64$. Its inputs are a 32-dimensional projection of the normalized query embedding and label-free statistics of the rollouts drawn so far, namely how much the answers agree, the mean, spread and extremes of the raw self-evaluation scores, the rollout lengths, summaries of the answer favored at each temperature and the current count. The predictor is a ridge regression whose penalty is selected on the tune queries by the error of the predicted curves, and its targets are the reward curves of the calibration queries computed as in Algorithm 1. Predictions are clipped to $[0, 1]$. The array is recomputed at counts $1, 2, 4, 8, 16, 32, 48$ and 64 and reused in between, while the stopping test of Algorithm 2 runs at every count.

The cost predictor is a ridge regression of $\log \widehat{M}(q)$ on the same embedding features. After each rollout it is blended with the empirical MGF of the lengths observed so far, with a weight on the prior that decays as more rollouts arrive. The window of the stopping rule, $w = 4$, was selected over $w = 1$ by the accuracy of the resulting frontier on the tune queries. The main tables read the price of each column budget retrospectively on the test sweep, and Appendix C.10 applies prices chosen in advance on the audit queries. No test label enters any of these choices.

C EXPERIMENT DETAILS

We distinguish three experimental settings. The deployed comparison evaluates a controller fitted on separate calibration queries. The split-pool oracle estimates each query's primitives on one half of its rollout pool and evaluates the resulting actions on the other half. The in-sample diagnostics estimate and evaluate reward curves on the same pool to illustrate the theoretical mechanisms. These settings use different score fits and cost readouts, specified below.

C.1 DATA, SCORES AND REPLAY

Rollout pools and grading. Every experiment replays a fixed pool of 128 rollouts per MATH-500 query, generated once per generator at sampling temperature 0.8, top-p 1.0 and at most 2048 new tokens per rollout. Queries whose reference answer coincides with the abstention cluster are excluded. A voted answer is graded correct when it belongs to the verified set of correct answers of the query, so equivalent correct answers are not forced into a single string. Abstentions cast no vote, and a vote with no valid answer is counted as incorrect.

Calibration corpus. PRICE-deployed learns from 7,425 MATH-train queries with 64 rollouts each, split into 5,197 fit, 1,115 tune and 1,113 audit queries. Repeated question texts stay in the same split, and no question text overlaps MATH-500. Fit queries determine the score CDF, the embedding projection and the predictors, tune queries select hyperparameters, and audit queries determine prospective prices after every other choice is frozen.

Self-evaluation score. For each question and extracted candidate answer, the generator is asked whether the answer is correct, and the raw score is the probability it assigns to the True token variants after normalizing over the True and False variants (Kadavath et al., 2022). The vote uses the rank transform

$$\phi(p) = \frac{\sum_{x \in \mathcal{P}_{\text{fit}}} \mathbf{1}\{x \leq p\}}{|\mathcal{P}_{\text{fit}}| + 1},$$

fit once on the rollouts of the fit queries and applied unchanged everywhere else. In the split-pool oracle of §4.1, the transform is instead fit on the pool half A of each split direction and applied to both halves, so B scores never enter the fit. Llama's deployed scores were recomputed with a pinned implementation of the verification prompt, whereas its oracle experiment uses an earlier score cache, so the oracle and deployed Llama results are not compared with each other.

Voting grids and ties. The inverse-temperature grids are

$$\mathcal{T}_{\text{Qwen}} = \{0, 0.1853, 0.3705, 0.7411, 1.4821, 2.5937, 3.7053, 5.558, 8.1517, 11.857, 15.4141, 21.5279, \infty\},$$
$$\mathcal{T}_{\text{Llama}} = \{0, 0.2345, 0.469, 0.938, 1.876, 3.283, 4.69, 7.035, 10.318, 15.008, 19.5104, 27.2489, \infty\},$$

coefficients in $e^{\tau\phi}$. Equal vote totals at a finite temperature are broken in favor of the answer holding the highest single score, and equal scores by a fixed enumeration of the answers. The split-pool oracle breaks ties by the enumeration alone. External baselines keep their own conventions.

Replay and budget readouts. Each test query is replayed along 64 fixed orderings of its pool. A reported budget is the entropic risk of the realized token sums, $b = \frac{1}{\gamma} \log \mathbb{E}[e^{\gamma L}]$, averaged over queries, orderings and any randomization between two policies, and accuracy is averaged the same way. Every method is swept, and its sweep is read at a column budget by mixing the two adjacent policies on the MGF scale $e^{\gamma b}$, so that the mixture spends the budget exactly. Fixed-count methods are read at the count that spends the budget, and the other methods at the best feasible point of their sweep, which can leave budget unused once accuracy saturates. Token counts cover generation only and exclude scoring, embeddings, prediction and calibration.

C.2 Baselines and budget matching

All six baselines are replayed on the same pools, orderings and horizon with their original voting and stopping conventions. SC and BoN use the same self-evaluation replay as PRICE. CISC weights each rollout by $e^{p/T}$ with the raw self-evaluation probability p, and its divisor temperature T is selected on the tune queries from 80 geometrically spaced values by the average accuracy at five and ten rollouts, giving $T = 1.12$ for Qwen and $T = 0.89$ for Llama. SC sweeps the counts $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 19, 22, 27, 32, 39, 48, 56, 64\}$, BoN sweeps $\{1, 2, 4, 8, 16, 32, 48, 64\}$ and CISC sweeps $\{1, 2, 3, 5, 7, 10, 15, 20, 30\}$. Adaptive-Consistency uses its beta-stability stopping statistic with thresholds $\{0.5, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.99, 0.995, 0.999\}$ at a cap of 64 rollouts. ESC stops at the first unanimous window and sweeps the window and the cap, with windows 2 to 10 and caps 2 to 64. DeepConf keeps its lowest-group, bottom-ten-percent and tail confidence statistics, sweeps the retained fractions 0.1 and 0.9 and the counts $\{2, 4, 8, 16, 24, 32, 48, 64\}$, and weights the retained answers by confidence.

An operating point of any method records its population accuracy and its realized MGF cost $C = \mathbb{E}[e^{\gamma L}]$. Reading a sweep at a budget b with $C_0 \le e^{\gamma b} \le C_1$ for two adjacent points mixes them with weight

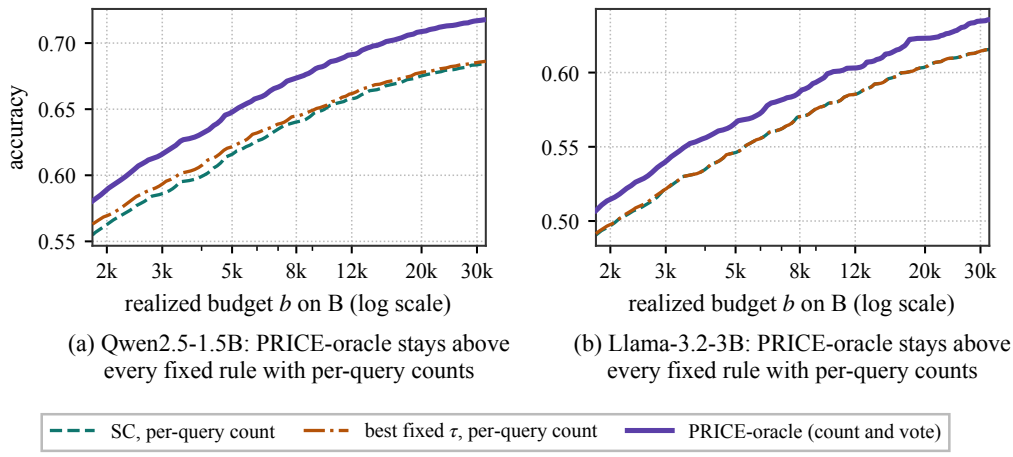
$$\theta = \frac{e^{\gamma b} - C_0}{C_1 - C_0}, \qquad R_b = (1 - \theta) R_0 + \theta R_1,$$

where the choice between the two policies is made independently of the query's rollouts. The six column budgets of Table 3 are the realized budgets of SC and BoN at $n = 3, 4, 5, 8, 10$ and 22 rollouts, so their entries are exact operating points, and BoN's row follows its decline with n rather than its best count.

C.3 Split-pool oracle

The oracle experiment of §4.1 splits each 128-rollout pool into two disjoint halves A and B, in four random partitions evaluated in both directions. Each half supplies 256 orderings, on which the reward curves at counts $1, \ldots, 64$ and all 13 temperatures and the cost MGF are computed as in Algorithm 1. The A primitives determine the actions at $\lambda = 0$ and at 16,385 geometrically spaced prices from 10^{-180} to 10^{6}, separately for the joint policy and for each fixed temperature with its own per-query counts. The chosen actions are executed on B, whose labels and token sums give the accuracy and the realized budget of every price. The hindsight-best fixed temperature is read on B at each column budget. The fixed-count SC and BoN rows use one common count for all queries, averaged over the eight split directions, and are read at the column budget by mixing the two adjacent counts.

Figure 5 shows the resulting frontiers.



(a) Qwen2.5-1.5B: PRICE-oracle stays above every fixed rule with per-query counts

(b) Llama-3.2-3B: PRICE-oracle stays above every fixed rule with per-query counts

Figure 5: Split-pool frontiers of PRICE-oracle with the self-evaluation score ((a) Qwen2.5-1.5B, (b) Llama-3.2-3B). PRICE-oracle, the best global temperature and SC all choose per-query counts from the A primitives and are evaluated with the actual B token sums. The best global temperature is selected at each budget after seeing the results, so the orange curve is an envelope over fixed rules rather than a single policy.

C.4 In-sample frontiers of Figure 2

Figure 2 uses the in-sample protocol of Algorithm 1 on each query's full pool, with the self-evaluation score rank-normalized on that pool. The reward curves at every temperature and count are estimated from random orderings of the pool, the cost of n rollouts is $\widehat{M}(q)^n$ with $\widehat{M}(q)$ the mean of $e^{\gamma \ell}$ over the pool, and every frontier is the exact solution of the priced problem on these estimates, swept over the price. The panels show the Level-3 queries of MATH-500, the difficulty band on which the fixed rules separate most. The intermediate fixed temperature is one interior grid value, the same in both panels. A red dot marks a budget at which the lead between two fixed rules swaps. Appendix C.5 repeats the figure on the full population with two further scores.

C.5 Full-population crossing frontiers

Figure 2 in the main text shows the Level-3 queries of MATH-500 under the self-evaluation score, the band where the frontiers separate most. Figure 6 repeats it on the full population and the full budget range, adding the Likelihood and Self-certainty scores. The three Qwen2.5-1.5B panels each show a lead swap among the fixed rules (red dots), and where the swap hugs the low-budget edge the inset plots SC's accuracy gap to each competitor, whose sign flip through zero is the crossing. On the full Llama-3.2-3B population, BoN leads SC only at the smallest feasible budgets and by about 0.2 points, so the bottom row carries no markers, and the Level-3 band of Figure 2 is where the swap is visible on both generators. In every panel the adaptive frontier stays above all three fixed rules.

Temperature grid and γ. The temperature grid of this in-sample protocol is expressed in units of a per-generator, per-score spread s, so that the pinned intermediate rule $\tau = 0.4/s$ sits at the same relative position in every panel. The spreads s and per-generator values γ^* are estimated on a held-out calibration split of 100 queries, and all frontiers are then replayed on the full 500 queries. The per-generator values are $\gamma^* = 1.5 \times 10^{-3}$ for Llama-3.1-8B and 2.9×10^{-3} for Llama-3.2-3B and Qwen2.5-1.5B, and the shared headline value $\gamma = 2.9262 \times 10^{-3}$ is their median. At each price, the replay solves the per-query problem (16) on the finite temperature and count grid by enumeration and commits the resulting pair, as in Algorithm 1. Plotted frontiers are the upper concave hull of 100 sampled population policies in $B = e^{\gamma b}$.

C.6 Adaptive voting across scores

The main text uses the self-evaluation score. To see how the choice of score moves the picture, we repeat the in-sample oracle comparison of Appendix C.4 with five commonly deployed per-rollout

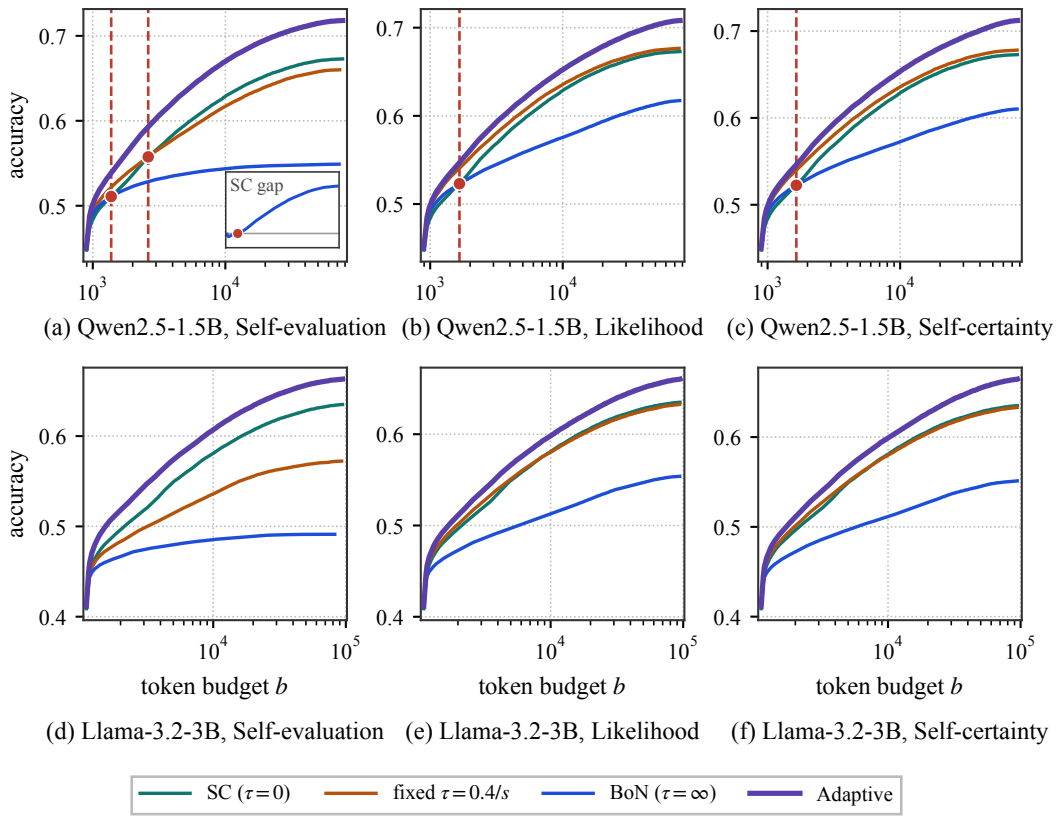


Figure 6: Figure 2 on the full MATH-500 population, with the Likelihood and Self-certainty scores added; panels (a)–(c) Qwen2.5-1.5B, (d)–(f) Llama-3.2-3B. Red dots mark the budgets at which the lead among the fixed rules swaps, and the inset in panel (a) shows the pairwise gap of SC over each competitor crossing zero. On the full Llama-3.2-3B population BoN leads SC only marginally at the smallest budgets, so panels (d)–(f) carry no markers.

scores on the same rollouts, adding a third generator, Llama-3.1-8B. Self-evaluation (Kadavath et al., 2022) is the probability the generator assigns to its answer being correct when asked, as in the main text. Self-certainty (Kang et al., 2025) measures how far the generator's next-token distributions along the rollout are from uniform, averaged over the rollout. Likelihood is the length-normalized log-probability of the rollout under the generator. DeepConf (Fu et al., 2026) is the lowest group confidence, the minimum over sliding windows of the mean token confidence. The Skywork process reward model (PRM) (He et al., 2024) is a separate model trained to judge reasoning steps, and we use its score for the final step. The first four are self-assessments computed from the generator's own outputs, whereas the PRM queries a second model. Each score is rank-normalized by its empirical CDF over the pool before entering the vote. For every score column we run SC, BoN and the full 13-temperature grid, and report the column's hindsight-best fixed temperature, the single temperature that maximizes accuracy separately at every budget. Skywork PRM is included so that the full score-quality axis is visible, but it stands apart, since its per-rollout scoring cost is a forward pass of a separate model and lies outside the token accounting (Table 4, dagger), which is why the main-text comparison is verifier-free.

In Table 4, adaptive voting is the best rule in every cell. Over all 45 (generator, score, budget) cells it beats the SC anchor and the hindsight-best fixed temperature. Fixed score-based rules fall below SC exactly on the weaker scores, where the value of adapting the rule is largest, and the gap contracts as the score becomes more informative.

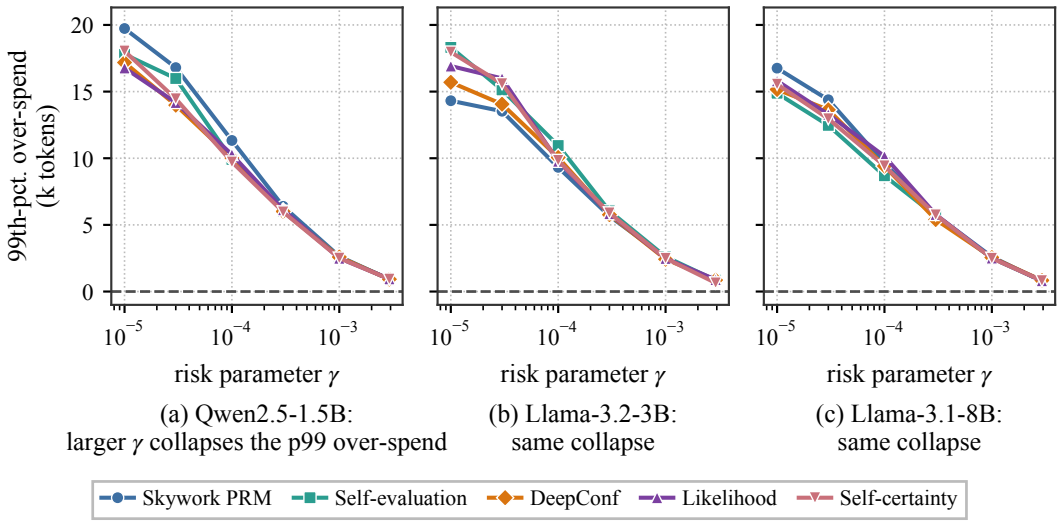
C.7 The realized cost tail under different risk parameters

Figure 7 shows how the risk parameter shapes the realized cost tail of PRICE-oracle under the in-sample protocol of Appendix C.4. Sweeping $\gamma \in \{10^{-5}, 3\times10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3},$

Table 4: **Score-grid robustness readout** (MATH-500; accuracy % at budgets $b \in \{2, 5, 30\}$k; subscripts = difference from the SC anchor). "Best fixed τ" is the hindsight oracle chosen per (column, budget). †PRM queries a separate verifier whose scoring cost is outside the token budget.

	Qwen2.5-1.5B			Llama-3.2-3B			Llama-3.1-8B		
Method	b=2k	b=5k	b=30k	b=2k	b=5k	b=30k	b=2k	b=5k	b=30k
SC (τ=0)	53.7	59.6	66.3	49.6	54.9	61.8	47.8	54.0	62.3
Self-evaluation									
BoN (τ=∞)	52.3$_{-1.4}$	53.8$_{-5.8}$	54.8$_{-11.5}$	46.7$_{-3.0}$	48.0$_{-6.9}$	49.1$_{-12.8}$	45.4$_{-2.4}$	48.1$_{-5.9}$	50.3$_{-12.0}$
best fixed τ (hindsight)	55.5$_{+1.8}$	60.6$_{+1.0}$	66.5$_{+0.3}$	50.1$_{+0.5}$	54.9	61.8	48.7$_{+0.9}$	54.8$_{+0.9}$	62.5$_{+0.2}$
PRICE-oracle	**57.2**$_{+3.5}$	**63.5**$_{+3.9}$	**70.6**$_{+4.3}$	**51.8**$_{+2.2}$	**57.4**$_{+2.5}$	**64.6**$_{+2.8}$	**50.2**$_{+2.4}$	**57.2**$_{+3.3}$	**66.3**$_{+4.0}$
Self-certainty									
BoN (τ=∞)	53.0$_{-0.7}$	55.6$_{-4.0}$	59.8$_{-6.5}$	47.2$_{-2.4}$	49.7$_{-5.2}$	53.7$_{-8.1}$	46.5$_{-1.3}$	49.9$_{-4.0}$	54.4$_{-7.9}$
best fixed τ (hindsight)	55.2$_{+1.6}$	60.5$_{+0.9}$	67.0$_{+0.8}$	50.2$_{+0.5}$	55.2$_{+0.2}$	61.8	48.8$_{+1.0}$	54.5$_{+0.5}$	62.3
PRICE-oracle	**56.1**$_{+2.4}$	**61.9**$_{+2.4}$	**69.3**$_{+3.1}$	**51.1**$_{+1.5}$	**56.7**$_{+1.8}$	**64.1**$_{+2.3}$	**49.7**$_{+1.9}$	**56.0**$_{+2.0}$	**64.4**$_{+2.1}$
Likelihood									
BoN (τ=∞)	53.0$_{-0.7}$	55.8$_{-3.7}$	60.3$_{-5.9}$	47.3$_{-2.4}$	49.7$_{-5.2}$	53.9$_{-7.9}$	46.4$_{-1.4}$	49.8$_{-4.1}$	54.4$_{-7.9}$
best fixed τ (hindsight)	55.3$_{+1.6}$	60.5$_{+0.9}$	66.9$_{+0.6}$	50.2$_{+0.5}$	55.2$_{+0.3}$	61.8	48.8$_{+1.0}$	54.5$_{+0.6}$	62.3
PRICE-oracle	**56.1**$_{+2.4}$	**61.9**$_{+2.3}$	**69.1**$_{+2.9}$	**51.1**$_{+1.5}$	**56.6**$_{+1.7}$	**63.8**$_{+2.0}$	**49.7**$_{+1.9}$	**55.9**$_{+1.9}$	**64.2**$_{+1.9}$
DeepConf									
BoN (τ=∞)	53.0$_{-0.7}$	55.8$_{-3.8}$	60.4$_{-5.9}$	48.0$_{-1.7}$	50.8$_{-4.1}$	55.5$_{-6.3}$	47.1$_{-0.7}$	50.8$_{-3.2}$	56.4$_{-5.9}$
best fixed τ (hindsight)	55.2$_{+1.5}$	60.4$_{+0.8}$	66.6$_{+0.4}$	50.4$_{+0.8}$	55.4$_{+0.5}$	61.8	49.0$_{+1.2}$	54.6$_{+0.7}$	62.3
PRICE-oracle	**56.0**$_{+2.4}$	**61.8**$_{+2.2}$	**68.9**$_{+2.7}$	**51.3**$_{+1.7}$	**56.8**$_{+1.9}$	**63.8**$_{+2.0}$	**49.8**$_{+2.1}$	**55.9**$_{+1.9}$	**64.3**$_{+2.1}$
Skywork PRM†									
BoN (τ=∞)	60.8$_{+7.1}$	67.3$_{+7.7}$	74.3$_{+8.0}$	55.4$_{+5.7}$	61.6$_{+6.6}$	68.9$_{+7.0}$	53.0$_{+5.2}$	59.6$_{+5.7}$	67.1$_{+4.8}$
best fixed τ (hindsight)	60.8$_{+7.1}$	67.5$_{+7.9}$	74.8$_{+8.5}$	55.4$_{+5.7}$	61.6$_{+6.7}$	69.1$_{+7.2}$	53.0$_{+5.2}$	59.8$_{+5.8}$	67.9$_{+5.6}$
PRICE-oracle	**61.1**$_{+7.4}$	**68.3**$_{+8.7}$	**76.7**$_{+10.5}$	**55.8**$_{+6.2}$	**62.8**$_{+7.8}$	**72.0**$_{+10.2}$	**53.5**$_{+5.8}$	**61.1**$_{+7.1}$	**70.6**$_{+8.3}$

$2.9\times10^{-3}\}$ at matched work ($\bar{n} \approx 8$ at every γ, using at each γ the policy whose mean count is closest to 8), it reports the 99th-percentile over-spend: the smallest t such that 99% of queries cost at most $b + t$ tokens. **Raising γ collapses this over-spend from 14–20k tokens to under 1k on every generator and score, while the mean cost stays essentially flat** (4,647 $\rightarrow$ 4,258 tokens on Qwen–PRM). At the largest swept γ, the value used in the main text, fewer than 1% of replayed costs exceed their risk budget by 1024 tokens. Accuracy is not held fixed across γ in this sweep, and the over-spend is measured relative to b_γ, which itself grows with γ.



(a) Qwen2.5-1.5B: larger γ collapses the p99 over-spend

(b) Llama-3.2-3B: same collapse

(c) Llama-3.1-8B: same collapse

Figure 7: **Empirical overspend relative to the risk budget** ((a) Qwen2.5-1.5B, (b) Llama-3.2-3B, (c) Llama-3.1-8B; shared y range): p99 over-spend beyond b vs. γ at matched work ($\bar{n} \approx 8$ per γ), five scores. 99% of queries spend at most b plus the value shown; the dashed gray line marks zero over-spend, and the collapse is uniform across scores and generators.

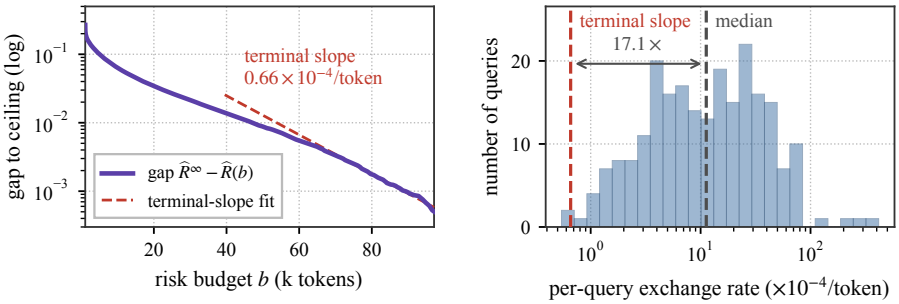
C.8 THE LARGE-BUDGET LAW ON LLAMA-3.2-3B AND HORIZON SENSITIVITY

Population gap. For the rate diagnostic of Figure 3 we rebuild the reward curves of every query with the deployed self-evaluation scores from 512 orderings of its pool, truncated to $n \leq 64$, and solve the priced problem exactly over all counts and temperatures at 400 log-prices. The budget axis is $\frac{1}{\gamma} \log \mathbb{E}_q[\widehat{M}(q)^{n_q}]$ for the chosen counts n_q, with $\widehat{M}(q)$ the mean of $e^{\gamma \ell}$ over all 128 lengths of the query's pool, as in Algorithm 1. The population gap is the distance to the finite-horizon ceiling $\mathbb{E}_q \max_{n \leq N, \tau} \widehat{V}_\tau(n; q)$, retained where it exceeds 5×10^{-4}, and the terminal slope is fitted to the log gap over the upper quarter of the retained budget range. The fits use 73 points at 56k–74k tokens on Qwen and 94 points at 74k–97k on Llama, with $R^2 = 0.993$ on both.

Per-query exchange rates. A query enters the solvable set when some temperature reaches $\widehat{V}_\tau(N; q) \geq 0.999$. At each such temperature we fit $\log(1 - \widehat{V}_\tau(n; q))$ against n over the counts with positive empirical error, requiring at least three of them, take the steepest fitted decay, and divide it by $\log \widehat{M}(q)/\gamma$ to obtain the exchange rate (12). On Llama-3.2-3B the terminal slope is 0.66×10^{-4} per risk-adjusted token against a median exchange rate of 11.26×10^{-4}, a $17.1\times$ separation at the 0.4th percentile of the fitted rates (Figure 8). Table 5 repeats the fits at horizons 32 and 48, where the slope stays at the slow edge of the query rates.

Generator	N	proxy	fitted	slope	slowest rate	slope percentile
Qwen2.5-1.5B	32	271	247	2.10	1.53	2.02
	48	296	272	1.21	1.08	0.37
	64	309	285	0.85	0.81	0.70
Llama-3.2-3B	32	215	179	1.66	0.64	1.12
	48	248	212	1.06	0.64	1.89
	64	265	229	0.66	0.64	0.44

Table 5: Horizon sensitivity of the rate diagnostic, with consistency threshold 0.999. Slope and slowest fitted rate are in units of 10^{-4} per risk-adjusted token. The final column is the percentage of fitted query rates at or below the population terminal slope.



(a) Gap to the ceiling decays exponentially in b

(b) Terminal slope sits at the slow edge, $17.1\times$ below the median

Figure 8: The large-budget law against the data on Llama-3.2-3B, with the protocol of Figure 3. The terminal slope is 0.66×10^{-4} per risk-adjusted token, $17.1\times$ below the median of the fitted exchange rates.

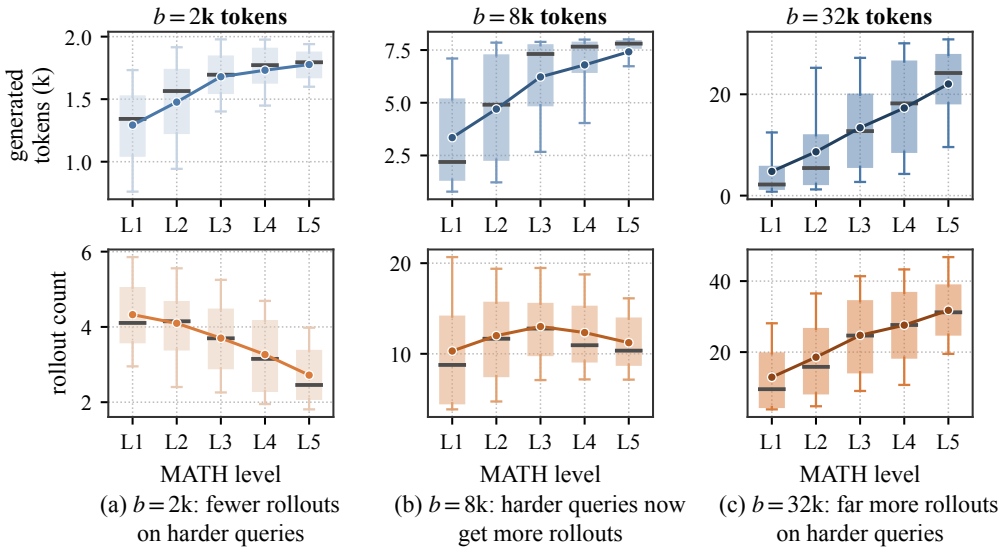
C.9 WHERE THE TOKEN BUDGET GOES

Having seen in §4.2 that PRICE-deployed beats the baselines, we now open the box and ask how it does so. MATH-500 is stratified by difficulty into levels 1–5, which lets us investigate how different difficulty levels consume the budget. We fix a small, a medium, and a large budget ($b = 2$k, 8k, and 32k tokens) and read Figure 9 row by row.

The top row shows the generated tokens per level. At every budget, the token spend rises with difficulty, but the rising speed changes with the budget. When the budget is scarce ($b = 2$k), Levels 3 to 5 use almost the same amount of tokens. When the budget is ample ($b = 32$k), Level 5 receives

4.6× tokens of Level 1 because the easy queries have already saturated and the additional budget flows to the harder ones.

The bottom row shows the rollout counts that PRICE picks. At the small budget, the count decreases with the level. This does not contradict the top row, since harder queries typically generate longer rollouts, so fewer rollouts still cost more tokens. When tokens are scarce ($b = 2$k), PRICE protects the rollout counts of the easy queries and trims the rollouts of the hard ones, whose draws cost more. As the budget grows to $b = 32$k, the hardest level receives about three times the rollouts of the easiest because, with sufficient budget, the harder the query, the more rollouts it needs. Llama-3.2-3B replicates the pattern below.



(a) $b = 2$k: fewer rollouts on harder queries (b) $b = 8$k: harder queries now get more rollouts (c) $b = 32$k: far more rollouts on harder queries

Figure 9: Allocation by MATH level of PRICE-deployed (Qwen2.5-1.5B; boxes = IQR with p10–p90 whiskers, dark-gray median, mean dot-line).

Figure 10 repeats Figure 9 on Llama-3.2-3B with the same budgets 2k, 8k and 32k, read from the retrospective frontier of PRICE-deployed. Each observation is a query's mean generated tokens or rollout count over its 64 orderings, and the boxes summarize the queries of one MATH level. The difficulty level is never an input of the controller. At the three budgets the population mean generated tokens are 1,651, 6,181 and 15,168 on Qwen and 1,559, 5,664 and 15,673 on Llama. On Llama the Level 1 and Level 5 mean tokens at 32k are 4,353 and 25,299, a 5.8× ratio, and as on Qwen the hardest level receives fewer rollouts than the easiest at 2k but more at 8k and 32k.

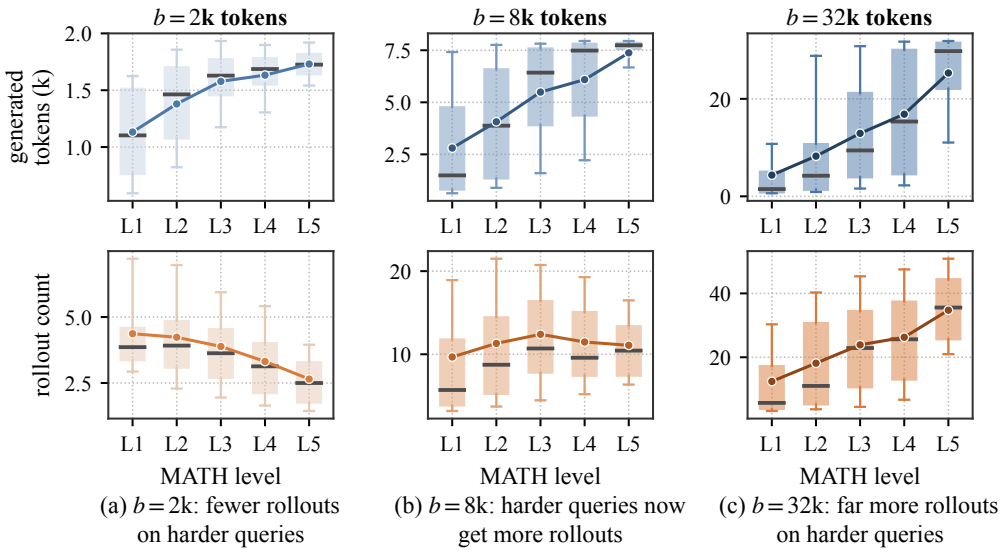
C.10 Prices chosen in advance

The main tables read the price of each budget retrospectively on the test sweep, which treats PRICE-deployed like the baselines but leaves open whether a deployer can meet a budget fixed in advance. To test this, we select the two adjacent prices and their mixture weight on the 1,113 audit queries of the calibration corpus, with the controller frozen, and apply them unchanged to MATH-500. Table 6 reports the requested and the achieved risk budgets together with the accuracy. The achieved budget exceeds the request by 0.1–0.9% on Qwen and by 0.5–3.8% on Llama, and the accuracy is within 0.2 points of the retrospective readout, so a price calibrated on independent queries transfers with a small overspend.

D Proofs

D.1 Preliminaries: weighted voting as competing random walks

The appendices prove every result of the main text, and they open with this preliminary section, which develops once the analytic backbone all later proofs share: the Boltzmann weighted vote



(a) $b = 2$k: fewer rollouts on harder queries (b) $b = 8$k: harder queries now get more rollouts (c) $b = 32$k: far more rollouts on harder queries

Figure 10: Allocation by MATH level on Llama-3.2-3B (PRICE-deployed with the self-evaluation score), with the same budgets and summaries as Figure 9. Token spending increases with difficulty at every budget, while the rollout count contrast between Level 1 and Level 5 reverses between 2k and 8k.

	Qwen2.5-1.5B						Llama-3.2-3B					
budget $b \to$	3.7k	5.3k	6.7k	11.6k	14.5k	32.0k	4.4k	6.2k	8.1k	13.8k	17.4k	40.4k
Requested budget (tokens)	3,680	5,291	6,658	11,553	14,516	31,983	4,426	6,235	8,072	13,784	17,426	40,399
Achieved budget (tokens)	3,713	5,305	6,663	11,574	14,545	32,040	4,593	6,380	8,251	13,999	17,613	40,600
overspend (%)	+0.9	+0.3	+0.1	+0.2	+0.2	+0.2	+3.8	+2.3	+2.2	+1.6	+1.1	+0.5
Accuracy (%), price chosen in advance	61.4	63.3	64.3	66.2	66.6	67.6	54.8	56.5	57.6	59.2	59.7	60.9
Accuracy (%), price read retrospectively	61.3	63.3	64.3	66.2	66.6	67.6	54.6	56.3	57.5	59.2	59.7	61.0

Table 6: Prices chosen in advance. For each budget of Table 3, the price of PRICE-deployed is selected on the audit queries of the calibration corpus and applied unchanged to MATH-500. The achieved entropic-risk budget overshoots the request by at most 0.9% on Qwen2.5-1.5B and 3.8% on Llama-3.2-3B, and the accuracy matches the retrospective readout of Table 3.

reduces to a family of *competing random walks*, one per wrong answer, and each question the paper asks about the vote becomes a question about those walks at a progressively finer resolution. §D.1.1 records the decomposition, the equivalence between winning the vote and keeping every walk positive, and an elementary error floor that needs no asymptotics. §D.1.2 settles the law-of-large-numbers regime, where drifts decide the race, and proves the consistency results of §2. §D.1.3 carries out the large-n error analysis: on a consistent query the error is a lower large deviation of positive-drift walks, with exact exponential rate given by Cramér's theorem and constant-prefactor refinement by the Bahadur–Rao theorem. §D.1.4 makes those estimates uniform in the temperature, the form the temperature-optimization proofs consume. §D.1.5 converts the rates into the fact about the reward curve $V_\tau(\cdot; q)$ quoted repeatedly downstream: its constant error floor at inconsistent temperatures. The remaining subsections are organized by main-text section and consume this toolkit; Table 7 records where each main-text result is proved.

Result	**Proved in**
Theorem D.1 (consistency criterion, stated in Appendix D.1.2)	Appendix D.1.2
Theorems 1 and 2	Appendix D.2
Proposition 1 (strong duality)	Appendix D.3
Lemma D.9 (attainment of the per-query maximum)	Appendix D.3
Proposition D.1 (per-query large-budget optimum, stated in the appendix)	Appendix D.4.2
Proposition D.2 (asymptotically optimal policy, stated in Appendix D.4.4)	Appendix D.4.4
Theorem 3 (asymptotic frontier)	Appendix D.4

Table 7: Where each result is proved.

Formal policy convention. Definition 2.2 is formalized as follows. A randomized committed policy is a measurable probability kernel $q \mapsto \mu_q$ on Π^{ad} (or on $\Pi(\tau)$ for a fixed-voting policy). Conditional on q, the action $(\tau, n) \sim \mu_q$ is drawn independently of the rollouts; reward and cost average over this auxiliary draw as well. A point-mass kernel is called *pure* and is identified with its pair $\pi_q = (\tau(q), n(q))$. In constraints such as $\pi_q \in \Pi$, the notation means $\mathrm{supp}(\mu_q) \subseteq \Pi$. Randomization convexifies the population budget problem, while Lemma D.8 shows that each per-query Lagrangian optimizer can be chosen pure.

Two standing conventions are in force throughout the appendices. First, ties at the weighted-vote argmax of Definition 2.1 are resolved as follows: at $\tau > 0$ by the fixed deterministic rule (a null event under the regularity below, Lemma D.9), and at $\tau = 0$, where count ties have positive probability, first by the total score $\sum_{i \le n} \phi_i \mathbf{1}\{a_i = a\}$ of the tied answers, then by the deterministic rule. The $\tau = 0$ clause is the $\tau \downarrow 0$ limit of the vote ($e^{\tau\phi} = 1 + \tau\phi + O(\tau^2)$, so a small temperature compares counts first and total scores second); it is what makes $\tau \mapsto V_\tau(n; q)$ continuous at 0 and hence the maximum over τ in (50) attained (Lemma D.9). It changes $V_0(n; q)$ only on count-tie events at finite n, which are exponentially rare on any query where $\tau = 0$ is consistent, so the consistency and leading exponential-rate conclusions are unchanged. Finite-budget reward values and frontiers can depend on the tie convention. Second, the following mild regularity condition holds; it is a modeling condition on the underlying score law. Empirical-CDF scores take finitely many values and need not satisfy it; the finite-grid empirical search attains its maximum without a continuity argument. The only exceptions are the two explicit counterexample constructions (the instance of Example D.1 and the two-query population built for Theorem 1), which use discrete scores and invoke only estimates that do not require the condition.

We also impose the standard measurability needed for population optimization: $q \mapsto M(q)$ and $(q, \tau) \mapsto V_\tau(n; q)$ are measurable for every n, and the pointwise suprema below admit measurable ϵ-maximizing selectors. Under these conditions the displayed expectations are well defined and the interchange of pointwise optimization and integration in Proposition 1 is valid. This condition is automatic, for example, for a standard Borel query space with a measurable rollout kernel and the Carathéodory integrands supplied by Assumption 1.

Assumption 1 (Score regularity)**.** *For every q and every a with $p(a \mid q) > 0$, the class-conditional score law $p(\phi \in \cdot \mid a, q)$ is atomless.*

D.1.1 The pairwise decomposition and the win event

Fix a query q with correct answer $a^* = a^*(q)$, and collect the wrong answers the LLM actually produces,

$$\mathcal{A}_q := \{a \in \mathcal{A} : \ a \neq a^*, \ p(a \mid q) > 0\},$$

the *support competitors*. Whether the vote returns a^* is a pairwise comparison against each of them, and each comparison is a random walk: for $a \in \mathcal{A}_q$ and $\tau \in [0, \tau_{\max}]$, define the per-rollout **lead increment** and the **lead walk**:

$$\begin{aligned} X_i^{(a)}(\tau) &:= e^{\tau\phi_i}\big(\mathbf{1}\{a_i = a^*\} - \mathbf{1}\{a_i = a\}\big), \\ S_n^{(a)}(\tau) &:= \sum_{i \le n} X_i^{(a)}(\tau) = W_\tau(a^*, n; q) - W_\tau(a, n; q), \end{aligned} \tag{20}$$

the accumulated Boltzmann weight by which a^* leads a after n rollouts. Conditional on q the increments are i.i.d. and bounded, $|X_i^{(a)}(\tau)| \le e^{\tau_{\max}}$. For clarity, we suppress the dependence of $S_n^{(a)}$ on τ and q, without ambiguity. When $|\mathcal{A}| = 2$ there is a single walk and the picture is a plain race: the vote returns the correct answer exactly when its lead is positive. With several competitors the same statement holds walk by walk.

Lemma D.1 (Win event)**.** *Fix q, $\tau \in [0, \tau_{\max}]$, and $n \ge 1$. Then*

$$\Big\{\min_{a \in \mathcal{A}_q} S_n^{(a)} > 0\Big\} \subseteq \{\widehat{a}_\tau(n; q) = a^*\}, \qquad \Big\{\min_{a \in \mathcal{A}_q} S_n^{(a)} < 0\Big\} \subseteq \{\widehat{a}_\tau(n; q) \neq a^*\},$$

and the error event is sandwiched between the strict and non-strict lead losses:

$$\bigcup_{a \in \mathcal{A}_q} \{S_n^{(a)} < 0\} \ \subseteq \ \{\widehat{a}_\tau(n; q) \neq a^*\} \ \subseteq \ \bigcup_{a \in \mathcal{A}_q} \{S_n^{(a)} \le 0\}. \tag{21}$$

The proof of this lemma is elementary and omitted; it is a direct consequence of the definitions (20).

Remark 1 (Ties and leading-rate bounds). *The two inclusions in* (21) *differ only on the tie events* $\{S_n^{(a)} = 0\}$*, which the fixed rule resolves. Throughout the appendices we use* (21) *one-sidedly, upper-bounding the error by the non-strict union and lower-bounding it by the strict one, so every stated bound is valid regardless of the tie rule. Moreover, ties are quantitatively negligible: for* $\tau > 0$*, Assumption 1 makes a tie at equal* positive *weights a null event, so* $\mathbb{P}(S_n^{(a)} = 0) = \big(1 - p(a^* \mid q) - p(a \mid q)\big)^n$*, the probability that no rollout hits* a^* *or* a*.*

Next, we record an error floor for Boltzmann weighted voting that requires no drift information at all. If all n rollouts return the same support competitor $a \neq a^*$, every lead $\{X_i^{(a)}(\tau)\}_{i=1}^n$ is non-positive and the vote errs by Lemma D.1, so whenever $p(a^* \mid q) < 1$,

$$1 - V_\tau(n; q) = \mathbb{P}(\widehat{a}_\tau(n; q) \neq a^* \mid q) \geq \mathbb{P}(a_1 = \cdots = a_n = a \mid q) = p(a \mid q)^n \geq p_{\min}^n, \tag{22}$$

for every τ and every n, where

$$p_{\min} := \min\{p(a' \mid q) : p(a' \mid q) > 0\};$$

we call (22) the *unanimity floor*.

D.1.2 The long-run regime: consistency

This section states and proves the consistency criterion deferred from §2.3.

Consistency is determined by the expected voting weight. Each rollout contributes weight $e^{\tau\phi}$ to its generated answer, so the expected weight that answer a receives per rollout is

$$\omega_\tau(a; q) := \mathbb{E}\big[\tfrac{1}{n} W_\tau(a, n; q)\big] = p(a \mid q)\, \mathbb{E}\big[e^{\tau\phi} \;\big|\; a, q\big], \tag{23}$$

which is the answer's generation probability multiplied by its mean exponentiated score. Next, we characterize the consistent sets $\mathcal{C}(\tau)$ and $\mathcal{C}^{\mathrm{ad}}$ through the expected voting weight.

Definition D.1 (Identifiability margin). *For a query* q *and a temperature* $\tau \in [0, \tau_{\max}]$*, the* identifiability margin *is*

$$\delta_\tau^*(q) := \omega_\tau\big(a^*(q); q\big) - \max_{a \neq a^*(q)} \omega_\tau(a; q).$$

It measures how much larger the correct answer's expected voting weight is than that of its strongest incorrect competitor, so $\delta_\tau^*(q) > 0$ means that the correct answer is identified in the infinite-rollout limit under temperature τ. At $\tau = 0$, it is the probability gap $p(a^*(q) \mid q) - \max_{a \neq a^*(q)} p(a \mid q)$ between the correct answer and incorrect ones.

On the law-of-large-numbers scale the race is decided by drifts. Taking expectations in (20) and recalling the expected per-rollout weight $\omega_\tau(a; q)$ of Definition D.1,

$$\mathbb{E}\big[X_1^{(a)}(\tau) \;\big|\; q\big] = \omega_\tau(a^*; q) - \omega_\tau(a; q) \geq \omega_\tau(a^*; q) - \max_{a' \neq a^*} \omega_\tau(a'; q) = \delta_\tau^*(q), \qquad a \in \mathcal{A}_q, \tag{24}$$

with equality for any competitor attaining the maximum in Definition D.1. The identifiability margin $\delta_\tau^*(q)$ is thus the worst-case drift of the lead walks: it is positive exactly when every lead drifts upward, for all $a \in \mathcal{A}_q$. Theorem D.1 is the law-of-large-numbers consequence of this picture.

Theorem D.1 (Consistency criterion). *Fix* $\tau \in [0, \tau_{\max}]$*. For every query* q*,* $\widehat{a}_\tau(n; q) \to a^*(q)$ *almost surely as* $n \to \infty$ *if and only if* $\delta_\tau^*(q) > 0$*. Consequently,*

$$\mathcal{C}(\tau) = \{q : \delta_\tau^*(q) > 0\}, \qquad \mathcal{C}^{\mathrm{ad}} = \{q : \exists \tau \in [0, \tau_{\max}],\, \delta_\tau^*(q) > 0\}.$$

Proof of Theorem D.1. By the strong law of large numbers, simultaneously for the finitely many $a \in \mathcal{A}_q$,

$$\tfrac{1}{n} S_n^{(a)} \xrightarrow{a.s.} \mathbb{E}\big[X_1^{(a)}(\tau) \;\big|\; q\big].$$

Sufficiency. If $\delta_\tau^*(q) > 0$, then by (24) every limit is at least $\delta_\tau^*(q)$, so almost surely $\min_{a \in \mathcal{A}_q} S_n^{(a)} \geq n\, \delta_\tau^*(q)/2 > 0$ for all large n, and Lemma D.1 gives $\widehat{a}_\tau(n; q) = a^*$ for all large n, that is $\widehat{a}_\tau(n; q) \to a^*(q)$ a.s.

Necessity. Pick a competitor $a^\flat$ attaining the maximum in Definition D.1, so that $\frac{1}{n}S_n^{(a^\flat)} \to \delta_\tau^*(q)$ a.s. by (24). If $\delta_\tau^*(q) < 0$ then $S_n^{(a^\flat)} \to -\infty$ a.s., so by Lemma D.1 $\widehat{a}_\tau(n;q) \neq a^*$ for all large n (no uniqueness among competitors is needed). If $\delta_\tau^*(q) = 0$ then $\omega_\tau(a^*;q) = \omega_\tau(a^\flat;q) > 0$, so $p(a^* \mid q)$ and $p(a^\flat \mid q)$ are both positive and the increment $X_1^{(a^\flat)}(\tau)$ takes a positive and a negative value with positive probability: $S_n^{(a^\flat)}$ is a non-degenerate centred random walk. By Chung–Fuchs recurrence $\liminf_n S_n^{(a^\flat)} = -\infty$ a.s., so $S_n^{(a^\flat)} < 0$ infinitely often, and on those n Lemma D.1 gives $\widehat{a}_\tau(n;q) \neq a^*$; hence $\widehat{a}_\tau(n;q) \not\to a^*$ a.s. ■

Theorem D.1 characterizes exactly which queries can be answered correctly through repeated rollouts. For a fixed temperature, increasing the number of rollouts only makes the empirical voting weights concentrate around their expectations. Thus, repeated rollouts can recover the correct answer if and only if it has the largest expected voting weight, or equivalently, if $\delta_\tau^*(q) > 0$.

Example D.1 (Strict coverage gain on a heterogeneous population)**.** *Let $\mathcal{Q}$ put mass $\frac{1}{2}$ on each of two queries q_A and q_B. On q_A the correct answer is a minority among the rollouts but scores high, and on q_B the correct answer is a majority but scores low. We construct an instance in which $\delta_\tau^*(q_A) > 0$ if and only if $\tau > \tau_c$ and $\delta_\tau^*(q_B) > 0$ if and only if $\tau < \tau_c$, for a threshold $\tau_c \in (0, \tau_{\max})$. Every fixed temperature therefore solves at most one of the two queries, $\sup_\tau \mathbb{P}_q[\mathcal{C}(\tau)] = \frac{1}{2}$, while the adaptive class solves both, $\mathbb{P}_q[\mathcal{C}^{\mathrm{ad}}] = 1$. The detailed specification is deferred to Appendix D.1.2.*

Details of Example D.1. Set $\tau_{\max} = 1$ and let $\mathcal{Q}$ put mass $\frac{1}{2}$ on each of two queries q_A, q_B. On both, the LLM produces only the correct answer a^* and one wrong answer a, and the score is deterministic given the answer, $\phi \in \{0.1, 0.9\}$.

- On q_A the correct answer is a minority with the high score, $p(a^* \mid q_A) = 0.4$ with $\phi = 0.9$ and $p(a \mid q_A) = 0.6$ with $\phi = 0.1$, so $\delta_\tau^*(q_A) = 0.4\, e^{0.9\tau} - 0.6\, e^{0.1\tau} > 0$ if and only if $\tau > \tau_c := \log(3/2)/0.8 \approx 0.51$.
- On q_B the roles are reversed, $p(a^* \mid q_B) = 0.6$ with $\phi = 0.1$ and $p(a \mid q_B) = 0.4$ with $\phi = 0.9$, so $\delta_\tau^*(q_B) = 0.6\, e^{0.1\tau} - 0.4\, e^{0.9\tau} > 0$ if and only if $\tau < \tau_c$.

By Theorem D.1, $\mathcal{C}(\tau) = \{q_A\}$ for $\tau > \tau_c$, $\mathcal{C}(\tau) = \{q_B\}$ for $\tau < \tau_c$, and $\mathcal{C}(\tau_c) = \emptyset$. Hence $\sup_\tau \mathbb{P}_q[\mathcal{C}(\tau)] = \frac{1}{2}$, while $\mathcal{C}^{\mathrm{ad}}$ contains both queries and $\mathbb{P}_q[\mathcal{C}^{\mathrm{ad}}] = 1$.

The example reveals why adaptive voting helps. When the correct answer is rare but well scored, a larger τ amplifies its score advantage. When it is frequent but poorly scored, a smaller τ relies more on frequency. Because these cases favor different temperatures, no single fixed rule performs consistently across both.

D.1.3 Large-n error decay: Cramér and Bahadur–Rao rates

On a consistent query every lead drifts upward, so an error at large n is a *lower large deviation* of positive-drift walks, an exponentially rare event. This subsection fixes q with $0 < p(a^* \mid q) < 1$ (if $p(a^* \mid q) = 1$ then $V_\tau(\cdot;q) \equiv 1$ and there is nothing to prove) and develops the exact rate. For every $a \in \mathcal{A}_q$, it satisfies $p(a \mid q) > 0$, so the increment $X_1^{(a)}(\tau) = e^{\tau\phi}\mathbf{1}\{a_i = a^*\} - e^{\tau\phi}\mathbf{1}\{a_i = a\}$ takes a positive value ($e^{\tau\phi} \geq 1$ on $\{a_i = a^*\}$) and a negative value ($-e^{\tau\phi} \leq -1$ on $\{a_i = a\}$), each with positive probability. Its log-moment-generating function

$$\Lambda_a(\theta;\tau) := \log \mathbb{E}\big[e^{\theta X_1^{(a)}(\tau)} \mid q\big]$$

is everywhere finite (the increments are bounded) and has three standard properties the proofs use.

- First, it is smooth and convex, with derivatives the tilted mean and variance: writing $\mathbb{E}^\theta$ for expectation under the tilted law $d\mathbb{P}^\theta/d\mathbb{P} \propto e^{\theta X_1^{(a)}(\tau)}$,

$$\Lambda_a'(\theta;\tau) = \mathbb{E}^\theta\big[X_1^{(a)}(\tau) \mid q\big], \qquad \Lambda_a''(\theta;\tau) = \mathrm{Var}^\theta\big(X_1^{(a)}(\tau) \mid q\big) > 0, \tag{25}$$

  the strict positivity because the increment takes both signs with positive probability, so it is non-constant under every equivalent tilt; hence $\Lambda_a(\cdot;\tau)$ is strictly convex and Λ_a' strictly increasing.

- Second, its slope at the origin is the drift (24):

$$\Lambda_a'(0;\tau) = \mathbb{E}[X_1^{(a)}(\tau) \mid q].$$

- Third, since $X_1^{(a)} \le -1$ on $\{a_i = a\}$, every $\theta \le 0$ has $\mathbb{E}[e^{\theta X_1^{(a)}}] \ge p(a \mid q)\, e^{|\theta|}$, so

$$\Lambda_a(\theta;\tau) \;\ge\; \log p(a \mid q) + |\theta| \qquad \text{for every } \theta \le 0, \tag{26}$$

  and in particular $\Lambda_a(\theta;\tau) \to +\infty$ as $\theta \to -\infty$.

An error is a lower deviation of the lead, so the relevant tilts are non-positive, and the **per-competitor rate** is

$$I_a(\tau) := -\inf_{\theta \le 0} \Lambda_a(\theta;\tau) \;\in\; [0,\infty), \tag{27}$$

non-negative because the infimum is at most $\Lambda_a(0;\tau) = 0$. The next lemma records when the rate is positive, where the minimizing tilt sits, and that both vary continuously in the temperature.

Lemma D.2 (Rate–drift dichotomy, tilt cap, and continuity)**.** *Fix q with $0 < p(a^* \mid q) < 1$ and $a \in \mathcal{A}_q$.*

(i) *If $\mathbb{E}[X_1^{(a)}(\tau) \mid q] > 0$, then $\Lambda_a(\cdot;\tau)$ attains its minimum over $\mathbb{R}$ at a unique $\theta_a(\tau) \in (-\infty, 0)$, characterized by $\Lambda_a'(\theta_a(\tau);\tau) = 0$; the rate is $I_a(\tau) = -\Lambda_a(\theta_a(\tau);\tau) > 0$, and the tilt is capped: $|\theta_a(\tau)| \le -\log p(a \mid q) \le \theta_+ := -\log p_{\min}$.*

(ii) *If $\mathbb{E}[X_1^{(a)}(\tau) \mid q] \le 0$, then $I_a(\tau) = 0$.*

(iii) *$\tau \mapsto I_a(\tau)$ is continuous on $[0, \tau_{\max}]$.*

In particular, $I_a(\tau) > 0$ if and only if the lead over a drifts strictly upward.

Proof of Lemma D.2. ***Part (i).*** The slope $\Lambda_a'(\cdot;\tau)$ is continuous and strictly increasing with $\Lambda_a'(0;\tau) = \mathbb{E}[X_1^{(a)}(\tau) \mid q] > 0$, and as $\theta \to -\infty$ the tilted mean tends to $\operatorname{ess\,inf} X_1^{(a)} \le -1 < 0$. Rising from a negative limit to a positive value at the origin, it has a unique zero $\theta_a(\tau) \in (-\infty, 0)$, the unique minimizer of $\Lambda_a(\cdot;\tau)$ over $\mathbb{R}$ by convexity; since the minimizer is negative, restricting the infimum in (27) to $\theta \le 0$ loses nothing. As $\theta_a(\tau)$ is a strict minimizer and $\Lambda_a(0;\tau) = 0$, the minimum value is strictly negative, so $I_a(\tau) > 0$. For the cap, evaluate (26) at $\theta_a(\tau)$: $\log p(a \mid q) + |\theta_a(\tau)| \le \Lambda_a(\theta_a(\tau);\tau) \le \Lambda_a(0;\tau) = 0$.

Part (ii). By convexity,

$$\Lambda_a(\theta;\tau) \ge \Lambda_a(0;\tau) + \theta\, \Lambda_a'(0;\tau) = \theta\, \mathbb{E}[X_1^{(a)}(\tau) \mid q] \ge 0$$

for every $\theta \le 0$ when the drift is non-positive, so the infimum in (27) is attained at $\theta = 0$ and $I_a(\tau) = 0$.

Part (iii). For bounded ϕ the map $(\theta, \tau) \mapsto \Lambda_a(\theta;\tau)$ is jointly continuous. By (26), any $\theta < \log p(a \mid q)$ has $\Lambda_a(\theta;\tau) > 0 \ge \inf_{\theta' \le 0} \Lambda_a(\theta';\tau)$, so for every τ

$$I_a(\tau) = -\min_{\theta \in [\log p(a|q),\, 0]} \Lambda_a(\theta;\tau),$$

a minimum of a jointly continuous function over a fixed compact interval, hence continuous in τ. ■

The slowest competitor sets the rate of the vote. Define

$$I_\tau(q) := \min_{a \in \mathcal{A}_q} I_a(\tau), \qquad a^\circ \in \arg\min_{a \in \mathcal{A}_q} I_a(\tau), \tag{28}$$

fixing any one minimizer as a° if several tie (for a perfect query, $p(a^* \mid q) = 1$, the competitor set is empty and we use the convention $I_\tau(q) := \infty$). By Lemma D.2 and (24), $I_\tau(q) > 0$ if and only if $\delta_\tau^*(q) > 0$, and $\tau \mapsto I_\tau(q)$ is continuous on $[0, \tau_{\max}]$. When $\delta_\tau^*(q) > 0$, write $\theta^\circ := \theta_{a^\circ}(\tau) < 0$ for the slowest tilt and

$$\sigma_\circ^2 := \Lambda_{a^\circ}''(\theta^\circ;\tau) = \operatorname{Var}^{\theta^\circ}\big(X_1^{(a^\circ)}(\tau) \mid q\big) > 0 \tag{29}$$

for the tilted variance, positive by (25).

Two bounds now read the error (21) off the walks, one in each direction, and both are quoted throughout the appendices. In one direction, for any $\theta \le 0$, $\{S_n^{(a)} \le 0\} \subseteq \{e^{\theta S_n^{(a)}} \ge 1\}$, so Markov's inequality and independence give $\mathbb{P}(S_n^{(a)} \le 0) \le \mathbb{E}[e^{\theta X_1^{(a)}}]^n = e^{n\Lambda_a(\theta;\tau)}$; minimizing over $\theta \le 0$ and applying the non-strict side of (21) with a union bound yields the Chernoff bound, valid for every τ and every n:

$$\mathbb{P}\big(S_n^{(a)} \le 0\big) \;\le\; e^{-nI_a(\tau)} \qquad \text{and} \qquad 1 - V_\tau(n;q) \;\le\; \mathbb{P}\left(\bigcup_{a\in\mathcal{A}_q} \{S_n^{(a)} \le 0\}\right) \;\le\; |\mathcal{A}|\, e^{-nI_\tau(q)}. \tag{30}$$

In the other direction, an exponential change of measure localizes the slowest walk just below the origin.

Lemma D.3 (Tilted lower bound). *Fix q with $0 < p(a^* \mid q) < 1$ and $\tau \in [0, \tau_{\max}]$ with $\delta_\tau^*(q) > 0$. With $a^\circ, \theta^\circ, \sigma_\circ^2$ as in (28)–(29) and $\mathbb{P}^{\theta^\circ}$ the law tilting each increment $X_1^{(a^\circ)}(\tau)$ of the slowest walk by $e^{\theta^\circ X_1^{(a^\circ)}(\tau)}$, for every $n \ge 1$*

$$1 - V_\tau(n;q) \;\ge\; \mathbb{P}\big(S_n^{(a^\circ)} < 0\big) \;\ge\; e^{-nI_\tau(q)}\, e^{-|\theta^\circ|\sqrt{n}}\, \mathbb{P}^{\theta^\circ}\big(-\sqrt{n} \le S_n^{(a^\circ)} < 0\big), \tag{31}$$

and under $\mathbb{P}^{\theta^\circ}$ the increments are i.i.d. with mean 0 and variance $\sigma_\circ^2$.

Proof of Lemma D.3. The first inequality is the strict side of (21). Under the tilted law the increments remain i.i.d., with mean $\Lambda'_{a^\circ}(\theta^\circ;\tau) = 0$ (the defining property of the tilt) and variance $\sigma_\circ^2$, by (25). Undoing the tilt via $d\mathbb{P}/d\mathbb{P}^{\theta^\circ} = e^{-\theta^\circ S_n^{(a^\circ)} + n\Lambda_{a^\circ}(\theta^\circ;\tau)}$ and keeping only the window $\{-\sqrt{n} \le S_n^{(a^\circ)} < 0\}$, on which $e^{-\theta^\circ S_n^{(a^\circ)}} = e^{|\theta^\circ| |S_n^{(a^\circ)}|} \ge e^{-|\theta^\circ|\sqrt{n}}$,

$$\begin{aligned}\mathbb{P}\big(S_n^{(a^\circ)} < 0\big) \;&\ge\; \mathbb{E}^{\theta^\circ}\Big[\mathbf{1}\big\{-\sqrt{n} \le S_n^{(a^\circ)} < 0\big\}\, e^{-\theta^\circ S_n^{(a^\circ)} + n\Lambda_{a^\circ}(\theta^\circ;\tau)}\Big] \\ &\ge\; e^{-nI_\tau(q)}\, e^{-|\theta^\circ|\sqrt{n}}\, \mathbb{P}^{\theta^\circ}\big(-\sqrt{n} \le S_n^{(a^\circ)} < 0\big),\end{aligned}$$

using $\Lambda_{a^\circ}(\theta^\circ;\tau) = -I_{a^\circ}(\tau) = -I_\tau(q)$. ∎

The two bounds already sandwich the error's exponential order; the crude rate below needs nothing further. The sharp, constant-prefactor form needs the slowest walk to be non-degenerate, which we isolate now.

Assumption 2 (Unique, non-lattice slowest competitor). *At the pair (q,τ) the slowest increment $X_1^{(a^\circ)}(\tau)$ is non-lattice, and the minimum rate (28) is attained at a single competitor a° (no other competitor shares the slowest rate $I_\tau(q)$).*

Lemma D.4 (Error rate of the vote). *Fix q with $p(a^* \mid q) < 1$ and $\tau \in [0, \tau_{\max}]$ with $\delta_\tau^*(q) > 0$.*

(i) Crude rate. $-\frac{1}{n}\log(1 - V_\tau(n;q)) \to I_\tau(q)$, *equivalently* $1 - V_\tau(n;q) = e^{-nI_\tau(q)(1+o(1))}$.

(ii) Sharp rate. *If in addition Assumption 2 holds, then*

$$1 - V_\tau(n;q) = \frac{c^\circ}{\sqrt{n}}\, e^{-nI_\tau(q)}\,(1 + o(1)), \qquad c^\circ := \frac{1}{|\theta^\circ|\,\sigma_\circ\sqrt{2\pi}} > 0. \tag{32}$$

Proof of Lemma D.4. ***Part (i): crude rate.*** The Chernoff bound (30) gives $-\frac{1}{n}\log(1 - V_\tau(n;q)) \ge I_\tau(q) - \frac{1}{n}\log|\mathcal{A}|$. For the reverse direction, the tilted walk of Lemma D.3 is a mean-zero i.i.d. sum with variance $\sigma_\circ^2$, so the central limit theorem applied to $-S_n^{(a^\circ)}/(\sigma_\circ\sqrt{n})$ gives

$$\mathbb{P}^{\theta^\circ}\big(-\sqrt{n} \le S_n^{(a^\circ)} < 0\big) = \mathbb{P}^{\theta^\circ}\big(0 < -S_n^{(a^\circ)} \le \sqrt{n}\big) \;\longrightarrow\; \Phi(1/\sigma_\circ) - \Phi(0) \;>\; 0,$$

so (31) yields $-\frac{1}{n}\log(1 - V_\tau(n;q)) \le I_\tau(q) + |\theta^\circ|\, n^{-1/2} + o(1)$, sandwiching the rate at $I_\tau(q)$. (This limit is Cramér's theorem for the lower tail (Dembo & Zeitouni, 2010); we keep the self-contained tilting argument because §D.1.4 reuses its constants.)

Part (ii): sharp rate. The walk $-S_n^{(a^\circ)}$ has i.i.d. non-lattice increments with strictly negative mean, log-MGF $\theta \mapsto \Lambda_{a^\circ}(-\theta;\tau)$ minimized at $|\theta^\circ| > 0$ with minimum value $-I_\tau(q)$ and second derivative $\sigma_\circ^2$ there, so the Bahadur–Rao theorem (Bahadur & Ranga Rao, 1960; Dembo & Zeitouni, 2010) gives

$$\mathbb{P}\big(S_n^{(a^\circ)} \le 0\big) = \mathbb{P}\big(-S_n^{(a^\circ)} \ge 0\big) = \frac{c^\circ}{\sqrt{n}}\, e^{-nI_\tau(q)}\,(1+o(1)), \qquad c^\circ = \frac{1}{|\theta^\circ|\,\sigma_\circ\sqrt{2\pi}}. \tag{33}$$

For the upper bound on the error, the non-strict side of (21) and the union bound isolating the slowest walk give

$$1 - V_\tau(n;q) \;\le\; \mathbb{P}\big(S_n^{(a^\circ)} \le 0\big) + \sum_{a\in\mathcal{A}_q,\, a\neq a^\circ} \mathbb{P}\big(S_n^{(a)} \le 0\big),$$

and by the uniqueness clause of Assumption 2 the second-slowest rate $I_2 := \min_{a\in\mathcal{A}_q,\, a\neq a^\circ} I_a(\tau)$ strictly exceeds $I_\tau(q)$ (with $I_2 = \infty$ if a° is the only competitor), so (30) bounds the sum by $|\mathcal{A}|\, e^{-nI_2} = o(n^{-1/2}e^{-nI_\tau(q)})$, negligible against (33). For the lower bound, the strict side of (21) gives $1 - V_\tau(n;q) \ge \mathbb{P}(S_n^{(a^\circ)} < 0) = \mathbb{P}(S_n^{(a^\circ)} \le 0) - \mathbb{P}(S_n^{(a^\circ)} = 0)$, and the tie probability is negligible as well: Assumption 2 rules out the lattice case $\tau = 0$, and for $\tau > 0$ Remark 1 gives $\mathbb{P}(S_n^{(a^\circ)} = 0) = (1 - p(a^* \mid q) - p(a^\circ \mid q))^n$ (zero if $p(a^* \mid q) + p(a^\circ \mid q) = 1$), while $\mathbb{E}[e^{\theta X_1^{(a^\circ)}}] \ge (1 - p(a^* \mid q) - p(a^\circ \mid q)) + p(a^\circ \mid q)e^{|\theta|}$ for $\theta \le 0$ shows $I_\tau(q) = -\Lambda_{a^\circ}(\theta^\circ;\tau) < -\log(1 - p(a^* \mid q) - p(a^\circ \mid q))$, so the tie term is $o(n^{-1/2}e^{-nI_\tau(q)})$. The error thus matches $\mathbb{P}(S_n^{(a^\circ)} \le 0)$ to first order on both sides, giving (32). ■

D.1.4 Uniform bounds in the temperature

Optimizing the temperature requires the rate estimates of §D.1.3 to hold uniformly in τ, and the obstacle is the constants: the tilt θ° and the tilted variance $\sigma_\circ^2$ vary with τ, and near the consistency boundary the rate $I_\tau(q)$ vanishes. The resolution is that the two τ-dependent factors in (31) admit bounds whose constants depend only on $(p_{\min}, \tau_{\max})$: the tilt is capped by $\theta_+ = -\log p_{\min}$ (Lemma D.2(i)), and the window probability is floored as follows.

Lemma D.5 (Uniform window floor). *Fix q with $0 < p(a^* \mid q) < 1$. There are $\kappa_1 > 0$ and $n_3 \in \mathbb{N}$, depending only on $(p_{\min}, \tau_{\max})$, such that for every $\tau \in [0, \tau_{\max}]$ with $\delta_\tau^*(q) > 0$ and every $n \ge n_3$,*

$$\mathbb{P}^{\theta^\circ}\big(-\sqrt{n} \le S_n^{(a^\circ)} < 0\big) \;\ge\; \kappa_1.$$

Proof of Lemma D.5. Under $\mathbb{P}^{\theta^\circ}$ the walk has mean zero and increments bounded by $e^{\tau_{\max}}$. Since $e^{\theta^\circ y}$ is non-increasing in y (as $\theta^\circ < 0$), the tilt only raises the mass of the lower values, so $\mathbb{P}^{\theta^\circ}(X_1^{(a^\circ)} \le -1) \ge \mathbb{P}(X_1^{(a^\circ)} \le -1) = p(a^\circ \mid q) \ge p_{\min}$. The tilted variance and third absolute moment are therefore pinned to τ-free ranges,

$$\sigma_\circ^2 = \mathbb{E}^{\theta^\circ}\big[(X_1^{(a^\circ)})^2 \mid q\big] \in [p_{\min},\, e^{2\tau_{\max}}], \qquad \mathbb{E}^{\theta^\circ}\big[\big|X_1^{(a^\circ)}\big|^3 \mid q\big] \le e^{3\tau_{\max}}. \tag{34}$$

Writing the window in terms of the negated walk, $\mathbb{P}^{\theta^\circ}(-\sqrt{n} \le S_n^{(a^\circ)} < 0) = \mathbb{P}^{\theta^\circ}(-S_n^{(a^\circ)} \le \sqrt{n}) - \mathbb{P}^{\theta^\circ}(-S_n^{(a^\circ)} \le 0)$, and applying the Berry–Esseen theorem (universal constant C_0) to the standardized sum $-S_n^{(a^\circ)}/(\sigma_\circ\sqrt{n})$ at the two endpoints $1/\sigma_\circ$ and 0, with $1/\sigma_\circ \ge e^{-\tau_{\max}}$ and (34),

$$\mathbb{P}^{\theta^\circ}\big(-\sqrt{n} \le S_n^{(a^\circ)} < 0\big) \;\ge\; \big(\Phi(e^{-\tau_{\max}}) - \Phi(0)\big) - \frac{2C_0\, e^{3\tau_{\max}}}{p_{\min}^{3/2}\sqrt{n}} \;\ge\; \kappa_1 > 0$$

for all $n \ge n_3$, with κ_1 and n_3 depending only on $(p_{\min}, \tau_{\max})$. ■

Corollary D.1 (Margin-free error floor). *Fix q with $0 < p(a^* \mid q) < 1$, and let $\theta_+ = -\log p_{\min}$, $c_2 := -\log \kappa_1$, and n_3 be as above. Then for every $\tau \in [0, \tau_{\max}]$ with $\delta_\tau^*(q) > 0$ and every $n \ge n_3$,*

$$1 - V_\tau(n;q) \;\ge\; e^{-nI_\tau(q) - \theta_+\sqrt{n} - c_2}. \tag{35}$$

Proof of Corollary D.1. Combine the tilted lower bound (31) with the tilt cap $|\theta^\circ| \le \theta_+$ of Lemma D.2(i) and the window floor of Lemma D.5. ■

The constants in (35) are free of the consistency margin: the floor holds right down to the consistency boundary, which is exactly the form the temperature choice in Appendix D.4 requires. Combining it with the Chernoff bound (30) gives a two-sided rate estimate, uniform over any closed sub-range of consistent temperatures.

Lemma D.6 (Uniform rate on a closed sub-range). *Fix q with $p(a^* \mid q) < 1$ and a margin $\delta_0 > 0$. There are a threshold $n_1(\delta_0) < \infty$ and a deterministic $\varepsilon_n(\delta_0) \downarrow 0$, both independent of $\tau \in \Theta_{\delta_0} := \{\tau : \delta^*_\tau(q) \geq \delta_0\}$, with*

$$\left| -\tfrac{1}{n}\log\big(1 - V_\tau(n;q)\big) - I_\tau(q) \right| \leq \varepsilon_n(\delta_0) \qquad \text{for all } \tau \in \Theta_{\delta_0},\ n \geq n_1(\delta_0).$$

Proof of Lemma D.6. Every $\tau \in \Theta_{\delta_0}$ is consistent, so both one-sided bounds of this section apply, and their corrections are already free of τ: the Chernoff bound (30) gives $-\frac{1}{n}\log(1 - V_\tau(n;q)) \geq I_\tau(q) - \frac{1}{n}\log|\mathcal{A}|$, while the margin-free floor (35) gives $-\frac{1}{n}\log(1 - V_\tau(n;q)) \leq I_\tau(q) + \theta_+ n^{-1/2} + c_2/n$ for $n \geq n_3$. Take $n_1(\delta_0) := n_3$ and $\varepsilon_n(\delta_0) := \theta_+ n^{-1/2} + \frac{1}{n}\big(\log|\mathcal{A}| + c_2\big)$, which happen not to depend on δ_0 at all. ■

D.1.5 The error floor at inconsistent temperatures

The rates above yield the fact about the reward curve $V_\tau(\cdot;q)$ that the budgeted analysis consumes repeatedly, a floor: at an inconsistent temperature the error never vanishes, however many rollouts are drawn. Beyond its role in the temperature choice (Appendix D.4), this floor is what caps the ceiling of an inconsistent temperature in the frontier-crossing construction (Appendix D.2).

Lemma D.7 (Error floor at inconsistent temperatures). *Fix q with $p(a^* \mid q) < 1$. There is a constant $\kappa = \kappa(q) > 0$, depending only on $(p_{\min}, \tau_{\max})$, such that*

$$1 - V_\tau(n;q) \ \geq\ \kappa \qquad \text{for all } n \in \mathbb{N} \text{ and all } \tau \in [0, \tau_{\max}] \text{ with } \delta^*_\tau(q) \leq 0. \tag{36}$$

In particular, $\sup_n V_\tau(n;q) \leq 1 - \kappa < 1$ at every inconsistent temperature.

Proof of Lemma D.7. The reason is intuitive: if τ is inconsistent, some support competitor does not trail a^* in expected weight, so the vote keeps a constant chance of returning it however many rollouts are drawn. Concretely, $\delta^*_\tau(q) \leq 0$ yields a competitor $a \in \mathcal{A}_q$ whose lead walk has non-positive drift $\mu := \mathbb{E}[X_1^{(a)}(\tau)] \leq 0$ by (24), bounded increments $|X_1^{(a)}| \leq e^{\tau_{\max}}$, and second moment $\mathbb{E}[(X_1^{(a)})^2] \geq p_{\min}$ (the squared increment is $e^{2\tau\phi} \geq 1$ on $\{a_i = a\}$, an event of probability at least $p_{\min}$). Write $\sigma^2 := \mathrm{Var}(X_1^{(a)})$.

For small n, the unanimity floor (22) gives $1 - V_\tau(n;q) \geq p_{\min}^n$. For large n we bound the strict event $\{S_n^{(a)} < 0\}$, on which the vote cannot return a^* by Lemma D.1. Two regimes cover every $\mu \leq 0$.

- *Case 1: $\mu \leq -c_1$, where $c_1 := \sqrt{p_{\min}/2}$.* The mean $n\mu$ alone carries the walk below the origin: the event $\{|S_n^{(a)} - n\mu| < n|\mu|\}$ is contained in $\{S_n^{(a)} < 0\}$, so Chebyshev's inequality, together with $\sigma^2 \leq \mathbb{E}[(X_1^{(a)})^2] \leq e^{2\tau_{\max}}$ and $\mu^2 \geq c_1^2 = p_{\min}/2$, gives

$$\mathbb{P}\big(S_n^{(a)} < 0\big) \ \geq\ \mathbb{P}\big(|S_n^{(a)} - n\mu| < n|\mu|\big) \ \geq\ 1 - \frac{\sigma^2}{n\mu^2} \ \geq\ 1 - \frac{2e^{2\tau_{\max}}}{n\,p_{\min}} \ \geq\ \tfrac{1}{4} \tag{37}$$

  for $n \geq n_2 := \lceil 8e^{2\tau_{\max}}/(3p_{\min})\rceil$.

- *Case 2: $-c_1 < \mu \leq 0$.* Here the variance carries a floor, $\sigma^2 = \mathbb{E}[(X_1^{(a)})^2] - \mu^2 \geq p_{\min} - p_{\min}/2 = p_{\min}/2$. Standardize the walk as $Z_n := (S_n^{(a)} - n\mu)/(\sigma\sqrt{n})$ and threshold at $-\sigma\sqrt{n}$, strictly below the origin so as to avoid any atom of $S_n^{(a)}$ there, the non-positive drift only helping:

$$\mathbb{P}\big(S_n^{(a)} < 0\big) \ \geq\ \mathbb{P}\big(S_n^{(a)} < -\sigma\sqrt{n}\big) \ =\ \mathbb{P}\Big(Z_n < -1 - \tfrac{\sqrt{n}\,\mu}{\sigma}\Big) \ \geq\ \mathbb{P}(Z_n < -1), \tag{38}$$

the last step because $\mu \le 0$. The increments are i.i.d. with third absolute moment $\rho := \mathbb{E}|X_1^{(a)} - \mu|^3 \le (2e^{\tau_{\max}})^3$, so the Berry–Esseen theorem (universal constant C_0), applied to the standardized sum of the negated increments at the continuity point 1, with the variance floor $\sigma^2 \ge p_{\min}/2$, gives

$$\mathbb{P}(Z_n < -1) \;=\; \mathbb{P}(-Z_n > 1) \;\ge\; \big(1-\Phi(1)\big) - \frac{C_0\rho}{\sigma^3\sqrt{n}} \;\ge\; \big(1-\Phi(1)\big) - \frac{8C_0\, e^{3\tau_{\max}}}{(p_{\min}/2)^{3/2}\sqrt{n}} \;\ge\; \tfrac{1}{8} \tag{39}$$

for all $n \ge n_2$, enlarging n_2 if needed, since $1 - \Phi(1) > 0.15$ and the correction vanishes.

Both regimes give $\mathbb{P}(S_n^{(a)} < 0) \ge \frac{1}{8}$ for $n \ge n_2$, with n_2 depending only on $(p_{\min}, \tau_{\max})$; combined with the unanimity floor (22) for $n < n_2$, the choice $\kappa := \min(p_{\min}^{n_2}, \frac{1}{8})$ proves (36). ■

D.2 Proofs for §3.2: adaptivity dominates fixed temperatures

Throughout the proof appendices we abbreviate the population reward (4) and the population MGF cost $\mathbb{E}_{q\sim\mathcal{Q}}[\mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q]]$ of §3.1 of a policy $\pi = (\pi_q)_q$ as

$$R(\pi) \;:=\; \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[r^{\pi_q} \mid q]\big], \qquad C(\pi) \;:=\; \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q]\big],$$

so that the budget constraint reads $C(\pi) \le B$ with $B = e^{\gamma b}$. To keep the two budget scales distinct, throughout the proof appendix define

$$\begin{aligned}\overline{R}^{\mathrm{ad}}(B) &:= \sup\{R(\pi) : C(\pi) \le B,\ \pi_q \in \Pi^{\mathrm{ad}}\} \\ &= R^{\mathrm{ad}}(\gamma^{-1}\log B), \\ \overline{R}(B;\tau) &:= R(\gamma^{-1}\log B;\tau).\end{aligned}$$

This section proves two complementary facts. The adaptive frontier dominates every fixed-temperature frontier at every budget (Theorem 2), while among fixed temperatures there is no dominance at all: any two cross on a suitable population (Theorem 1).

Proof of Theorem 2. Fix τ and a feasible budget $B \ge \mathbb{E}_{q\sim\mathcal{Q}}\big[\mathbb{E}[e^{\gamma \ell_1(q)} \mid q]\big]$. Every population policy π with $\pi_q \in \Pi(\tau)$ for all q also satisfies $\pi_q \in \Pi^{\mathrm{ad}}$, because $\Pi(\tau) \subseteq \Pi^{\mathrm{ad}}$ by definition. Hence the feasible set defining $\overline{R}(B;\tau)$,

$$\{\pi :\ C(\pi) \le B, \pi_q \in \Pi(\tau)\} \subseteq \big\{\pi :\ C(\pi) \le B, \pi_q \in \Pi^{\mathrm{ad}}\big\},$$

is a subset of the one defining $\overline{R}^{\mathrm{ad}}(B)$. Therefore, $\overline{R}^{\mathrm{ad}}(B) \ge \overline{R}(B;\tau)$, for every $\tau \in [0, \tau_{\max}]$. ■

Proof of Theorem 1. ***Idea.*** Relabel the temperatures so that $\tau_1 < \tau_2$. We construct two query types. Both temperatures are consistent on q_A, but the error decays faster at τ_1. Only τ_2 is consistent on q_B. Giving q_A most of the population mass makes τ_1 preferable at an intermediate budget; at sufficiently large budgets, the additional coverage of q_B makes τ_2 preferable. We prove that

- At a large budget, the rescue tips the population toward τ_2;
- At a smaller, finite budget, the advantage on q_A outweighs the loss on q_B.

Construction. We set $\tau_c := \frac{1}{2}(\tau_1 + \tau_2)$, so $\tau_1 < \tau_c < \tau_2$. Let the query distribution $\mathcal{Q}$ place mass $1-\varepsilon$ on a bulk query q_A and mass ε on a rescue query q_B, with $\varepsilon \in (0,1)$ to be fixed later. There are two answers $\mathcal{A} = \{a^*, a\}$, and every rollout has the same deterministic length ℓ, so the per-rollout cost MGF is a common constant $M := e^{\gamma\ell} > 1$ for both queries; the uniform policy drawing n rollouts on every query has cost $B = M^n$. The score is two-valued, $\phi \in \{0,1\}$. Specifically, we construct the following instance.

- *Bulk* q_A (mass $1-\varepsilon$): it satisfies $p(a^* \mid q_A) = 0.7$ and $p(a \mid q_A) = 0.3$, with $\mathbb{P}(\phi = 1 \mid a^*, q_A) = 0.3$ and $\mathbb{P}(\phi = 1 \mid a, q_A) = 0.7$. Then

$$\begin{aligned}\delta^*_\tau(q_A) &= p(a^* \mid q_A)\,[\mathbb{P}(\phi = 0 \mid a^*, q_A) + \mathbb{P}(\phi = 1 \mid a^*, q_A)e^\tau] \\ &\quad - p(a \mid q_A)\,[\mathbb{P}(\phi = 0 \mid a, q_A) + \mathbb{P}(\phi = 1 \mid a, q_A)e^\tau] \\ &= 0.7\,(0.7 + 0.3e^\tau) - 0.3\,(0.3 + 0.7e^\tau) = 0.40 > 0 \quad \text{for every } \tau,\end{aligned}$$

  so by Theorem D.1, the bulk q_A is consistent at every temperature, with the ceiling

$$\sup_{n\in\mathbb{N}} V_{\tau_1}(n; q_A) = \sup_{n\in\mathbb{N}} V_{\tau_2}(n; q_A) = 1,$$

  where $V_\tau(\cdot; q)$ is the per-query reward curve defined in (8).

- *Rescue* q_B (mass ε): with deterministic per-class scores $\phi = 1$ on a^* and $\phi = 0$ on a, and $p_B := p(a^* \mid q_B) = \frac{1}{1+e^{\tau_c}} < \frac{1}{2}$. Then

$$\delta^*_\tau(q_B) = p_B\, e^\tau - (1 - p_B) > 0 \quad \iff \quad e^\tau > \frac{1-p_B}{p_B} = e^{\tau_c} \quad \iff \quad \tau > \tau_c,$$

  so the rescue q_B is consistent at τ_2 but not at τ_1 (recall that $\tau_c = \frac{1}{2}(\tau_1 + \tau_2) > 0$ since $\tau_2 > \tau_1 \geq 0$), with ceilings

$$s_B := \sup_{n\in\mathbb{N}} V_{\tau_1}(n; q_B) \leq 1 - \kappa_B < 1, \qquad \sup_{n\in\mathbb{N}} V_{\tau_2}(n; q_B) = 1,$$

  the first by the error floor at inconsistent temperatures (Lemma D.7, with a constant $\kappa_B > 0$), the second by Theorem D.1.

Two endpoints. We first study the two endpoints of budget and compare the population reward on τ_1, τ_2.

- *One rollout ($B = M$).* The budget $B = M$ forces exactly one rollout on every query: any policy draws $T_q \geq 1$ rollouts at cost $M^{T_q} \geq M$, so the constraint $\mathbb{E}_q[M^{T_q}] \leq M$ can hold only if $T_q \equiv 1$. A single rollout commits its own answer $a_1 \in \{a^*, a\}$ irrespective of the temperature, so for both $\tau \in \{\tau_1, \tau_2\}$

$$\overline{R}(M; \tau) = \mathbb{E}_{q\sim\mathcal{Q}}\big[V_\tau(1; q)\big] = \mathbb{E}_{q\sim\mathcal{Q}}\big[p(a^* \mid q)\big] = (1-\varepsilon)\,0.7 + \varepsilon\, p_B,$$

  so $\overline{R}(M; \tau_1) = \overline{R}(M; \tau_2)$.

- *Large budget.* At a large budget, both types are consistent at τ_2, so $\overline{R}(B; \tau_2) \to 1$ as $B \to \infty$, whereas at τ_1 the inconsistent rescue caps the frontier: whatever (possibly random) count a feasible fixed-τ_1 policy spends on the rescue, its accuracy there is at most s_B, so

$$\overline{R}(B; \tau_1) \;\leq\; (1-\varepsilon)\cdot 1 + \varepsilon\, s_B \;=\; 1 - \varepsilon(1 - s_B) \;<\; 1.$$

  Choosing any B_2 large enough that $\overline{R}(B_2; \tau_2) > 1 - \varepsilon(1 - s_B)$ gives $\overline{R}(B_2; \tau_1) < \overline{R}(B_2; \tau_2)$.

Therefore, the hotter temperature τ_2 leads at B_2. It remains to find a budget $B_1 < B_2$ at which the cooler temperature τ_1 leads, such that $\overline{R}(B_1; \tau_1) > \overline{R}(B_1; \tau_2)$. We devote the four steps below to it.

Step 1: reduce the comparison to the bulk q_A**.** Fix a budget $B = M^n$, for $n \in \mathbb{N}$. We lower-bound the τ_1 reward and upper-bound the τ_2 reward, and find that the comparison depends only on the bulk q_A.

For τ_1, rolling every query exactly n times has a population cost M^n, hence it is affordable; dropping the nonnegative rescue contribution,

$$\overline{R}(M^n; \tau_1) \;\geq\; (1-\varepsilon)\, V_{\tau_1}(n; q_A) + \varepsilon\, V_{\tau_1}(n; q_B) \;\geq\; (1-\varepsilon)\, V_{\tau_1}(n; q_A). \tag{40}$$

For τ_2, we need an upper bound valid for every feasible fixed-τ_2 policy, including randomized ones. Fix a feasible policy π, and let T be its randomized committed count on q_A; randomization being

independent of the rollouts, the bulk accuracy is the mixture $\mathbb{E}_T[V_{\tau_2}(T; q_A)]$ and the bulk cost is $\mathbb{E}_T[M^T]$. Bounding the rescue accuracy by 1,

$$R(\pi) \le (1-\varepsilon)\mathbb{E}_T[V_{\tau_2}(T; q_A)] + \varepsilon. \tag{41}$$

Due to the budget constraint, the token cost for the bulk q_A must be smaller than M^n:

$$(1-\varepsilon)\mathbb{E}_T[M^T] \le M^n, \quad \Longrightarrow \quad \mathbb{E}_T[M^T] \le \frac{M^n}{1-\varepsilon}.$$

Choose the length ℓ large enough that $(1-\varepsilon)M = (1-\varepsilon)e^{\gamma\ell} > 1$. Then running the bulk past n rollouts is rare:

$$\mathbb{P}(T \ge n+1) = \mathbb{P}(M^T \ge M^{n+1}) \le \frac{\mathbb{E}[M^T]}{M^{n+1}} \le \frac{M^n/(1-\varepsilon)}{M^{n+1}} = \frac{1}{(1-\varepsilon)M}.$$

Now split the bulk accuracy according to whether the committed count exceeds n. On $\{T \le n\}$, the vote uses at most n rollouts, so $V_{\tau_2}(T; q_A) \le \max_{1\le t\le n} V_{\tau_2}(t; q_A)$; on $\{T \ge n+1\}$, we bound the accuracy by 1, an event of probability at most $\frac{1}{(1-\varepsilon)M}$. Hence

$$\mathbb{E}_T[V_{\tau_2}(T; q_A)] \le \max_{1\le t\le n} V_{\tau_2}(t; q_A) + \frac{1}{(1-\varepsilon)M}.$$

Substituting into (41) and taking the supremum over feasible π, we obtain

$$\overline{R}(M^n; \tau_2) \le (1-\varepsilon) \max_{1\le t\le n} V_{\tau_2}(t; q_A) + \varepsilon + \frac{1}{M}.$$

Subtracting this from (40), the cooler temperature τ_1 leads at budget M^n once the bulk advantage clears the rescue's ε and $1/M$:

$$\boxed{(1-\varepsilon)\Big[V_{\tau_1}(n; q_A) - \max_{1\le t\le n} V_{\tau_2}(t; q_A)\Big] \;>\; \varepsilon + \tfrac{1}{M}} \quad \Longrightarrow \quad \overline{R}(M^n; \tau_1) > \overline{R}(M^n; \tau_2). \tag{42}$$

From here on, we aim to show the boxed inequality, where only the bulk q_A appears; the rest of the proof secures it.

Step 2: the bulk vote is a biased walk, and the temperature only adds noise. On the bulk q_A there is a single support competitor a, so the pairwise decomposition of §D.1.1 reduces to one lead walk (20),

$$S_t(\tau) := W_\tau(a^*, t; q_A) - W_\tau(a, t; q_A) = \sum_{i=1}^{t} X_i(\tau),$$
$$\text{where} \quad X_i(\tau) := e^{\tau\phi_i}\big(\mathbf{1}\{a_i = a^*\} - \mathbf{1}\{a_i = a\}\big),$$

and the vote returns the correct a^* when the lead is positive and errs when it is negative (Lemma D.1). The increments $X_i(\tau)$ are i.i.d.; under q_A, classifying a rollout by its answer and its score gives the increment's four-point law,

$$X_i(\tau) = \begin{cases} +1 & \text{with prob } 0.7\cdot 0.7 = 0.49 \quad (a_i = a^*,\ \phi_i = 0), \\ +e^\tau & \text{with prob } 0.7\cdot 0.3 = 0.21 \quad (a_i = a^*,\ \phi_i = 1), \\ -1 & \text{with prob } 0.3\cdot 0.3 = 0.09 \quad (a_i = a,\ \phi_i = 0), \\ -e^\tau & \text{with prob } 0.3\cdot 0.7 = 0.21 \quad (a_i = a,\ \phi_i = 1). \end{cases}$$

Reading off the first two moments,

$$\mathbb{E}[X_i(\tau) \mid q_A] = 0.40 \ \ (\text{every } \tau), \qquad \mathrm{Var}\big(X_i(\tau) \mid q_A\big) = 0.42\,(1 + e^{2\tau}).$$

The drift toward a^* is the temperature-free margin $\delta^*_\tau(q_A) = 0.40$, while the variance grows with τ. The larger variance suggests slower error decay at τ_2, but variance alone does not prove a comparison of error probabilities. Step 3 establishes the required comparison of exponential rates directly from the moment-generating functions.

Step 3: large-deviation bounds for the bulk error. The bulk error after t rollouts is $\psi_\tau(t) := 1 - V_\tau(t; q_A)$. Because $V_\tau = 1 - \psi_\tau$, the bulk advantage in the box (42) is

$$V_{\tau_1}(n; q_A) - \max_{1\le t\le n} V_{\tau_2}(t; q_A) = \min_{1\le t\le n} \psi_{\tau_2}(t) - \psi_{\tau_1}(n), \tag{43}$$

so we want an upper bound on $\psi_{\tau_1}(n)$ and a lower bound on $\min_{1\le t\le n} \psi_{\tau_2}(t)$. Both come from the large-deviation bounds in §D.1.3, instantiated at the bulk: the walk $S_t(\tau)$ sums i.i.d. bounded increments of *positive* mean 0.40, so a non-positive lead is a lower large deviation, governed by the log-MGF $\Lambda_X(\theta; \tau) := \log \mathbb{E}[e^{\theta X_1(\tau)} \mid q_A]$ and the rate (27) of the single competitor,

$$I(\tau) \;:=\; I_\tau(q_A) \;=\; -\inf_{\theta\le 0} \Lambda_X(\theta; \tau) \;>\; 0, \tag{44}$$

positive at every τ by the rate–drift dichotomy (Lemma D.2, the drift being $0.40 > 0$), which also places the minimizing tilt of $\Lambda_X(\cdot; \tau)$ over $\mathbb{R}$ at a unique $\theta^*(\tau) < 0$.

(i) Upper bound for $\psi_{\tau_1}(n)$. With a single competitor, the Chernoff bound (30) reads

$$\psi_{\tau_1}(n) \;\le\; \mathbb{P}\big(S_n(\tau_1) \le 0 \mid q_A\big) \;\le\; e^{-nI(\tau_1)}. \tag{45}$$

(ii) The hotter rate is smaller. Grouping the four-point law by score, with $e^{\theta e^\tau} + e^{-\theta e^\tau} = 2\cosh(\theta e^\tau)$,

$$\begin{aligned}\mathbb{E}\big[e^{\theta X_1(\tau)} \;\big|\; q_A\big] &= 0.49e^{\theta} + 0.21e^{\theta e^\tau} + 0.09e^{-\theta} + 0.21e^{-\theta e^\tau} \\ &= \underbrace{0.49\, e^{\theta} + 0.09\, e^{-\theta}}_{\phi=0:\ \tau\text{-free}} + \underbrace{0.42\, \cosh(\theta e^\tau)}_{\phi=1}.\end{aligned}$$

The temperature enters only through the term $\cosh(\theta e^\tau)$, which for each fixed $\theta \neq 0$ strictly increases in τ ($|\theta| e^\tau$ is strictly increasing and cosh is even and increasing in $|\cdot|$). Hence $\Lambda_X(\theta; \tau_1) < \Lambda_X(\theta; \tau_2)$ for every $\theta \neq 0$. Evaluating at the nonzero minimizer $\theta^*(\tau_2)$ of the hotter log-MGF,

$$\inf_\theta \Lambda_X(\theta; \tau_1) \le \Lambda_X(\theta^*(\tau_2); \tau_1) < \Lambda_X(\theta^*(\tau_2); \tau_2) = \inf_\theta \Lambda_X(\theta; \tau_2),$$

that is $I(\tau_1) > I(\tau_2) > 0$. Fix any rate ρ strictly between them, $I(\tau_2) < \rho < I(\tau_1)$.

(iii) Lower bound for $\min_{1\le t\le n} \psi_{\tau_2}(t)$. By the crude rate of Lemma D.4(i), applied to the consistent pair (q_A, τ_2),

$$\tfrac{1}{t} \log \psi_{\tau_2}(t) \;\longrightarrow\; -I(\tau_2) > -\rho,$$

so there is an integer threshold $t_0 \ge 2$ with

$$\psi_{\tau_2}(t) \ge e^{-t\rho} \ge e^{-n\rho} \qquad (t_0 \le t \le n),$$

where the second inequality holds because $\rho > 0$ and $t \le n$ make $e^{-t\rho}$ decreasing in t. For the finitely many small counts $1 \le t \le t_0$, the unanimity floor (22) already keeps the error positive: $\psi_{\tau_2}(t) \ge p(a \mid q_A)^t = 0.3^t > 0$. Let

$$p_0 := \min_{1\le t<t_0} \psi_{\tau_2}(t) > 0.$$

Because $e^{-n\rho} \to 0$ while p_0 is a fixed positive constant, for all n large enough such that $e^{-n\rho} \le p_0$, then

$$\psi_{\tau_2}(t) \ge p_0 \ge e^{-n\rho} \qquad \forall 1 \le t < t_0.$$

Combining the two ranges, we obtain that for all n large enough that $n \ge t_0$ and $e^{-n\rho} \le p_0$,

$$\min_{1\le t\le n} \psi_{\tau_2}(t) \;\ge\; e^{-n\rho}. \tag{46}$$

Step 4: assemble. By (43), (45), and (46), the bulk advantage obeys

$$V_{\tau_1}(n; q_A) - \max_{1\le t\le n} V_{\tau_2}(t; q_A) = \min_{1\le t\le n} \psi_{\tau_2}(t) - \psi_{\tau_1}(n) \;\ge\; e^{-n\rho} - e^{-n\, I(\tau_1)}.$$

Since $\rho < I(\tau_1)$ the first term dominates: once n is large enough that $e^{-n(I(\tau_1)-\rho)} \le \frac{1}{2}$, the advantage is at least $\frac{1}{2} e^{-n\rho}$.

Let n° be any integer exceeding all three large-count thresholds at once: $n^\circ \geq t_0$ and $e^{-n^\circ \rho} \leq p_0$ in (46), and the threshold making $e^{-n^\circ (I(\tau_1)-\rho)} \leq \frac{1}{2}$. For this n°, the bulk advantage at n° is at least

$$V_{\tau_1}(n^\circ; q_A) - \max_{1 \leq t \leq n^\circ} V_{\tau_2}(t; q_A) \geq \tfrac{1}{2} e^{-n^\circ \rho} =: c > 0.$$

The box (42) then asks only for $(1-\varepsilon)c > \varepsilon + 1/M$.

Choose $\varepsilon \leq \min\{\frac{c}{4}, \frac{1}{2}\}$ and then the length ℓ (hence $M = e^{\gamma \ell}$) large enough that $\frac{1}{M} < \frac{c}{4}$ and $(1-\varepsilon)M > 1$; then $(1-\varepsilon)c \geq \frac{1}{2}c$ while $\varepsilon + \frac{1}{M} < \frac{c}{4} + \frac{c}{4} = \frac{c}{2}$, so the box holds. For this (ε, ℓ), the budget $B_1 := M^{n^\circ}$ satisfies $\overline{R}(B_1; \tau_1) > \overline{R}(B_1; \tau_2)$.

Finally, $B_1 = M^{n^\circ}$ is finite, so enlarging the large-budget B_2 above, if necessary, keeps $B_1 < B_2$. We have produced $B_1 < B_2$ with $\overline{R}(B_1; \tau_1) > \overline{R}(B_1; \tau_2)$ and $\overline{R}(B_2; \tau_1) < \overline{R}(B_2; \tau_2)$, so the frontiers $\mathcal{P}(\tau_1)$ and $\mathcal{P}(\tau_2)$ cross after the monotone change of variable $b = \gamma^{-1} \log B$. As $\tau_1 \neq \tau_2$ were arbitrary, no fixed temperature dominates another across all query populations and budgets.

■

D.3 Proofs for §3.3: duality, the per-query inner problem

This section solves the budgeted problem over the adaptive class. The plan is classical: dualize the single population budget so the problem decouples into independent per-query problems (Proposition 1 and Lemma D.8), then observe that each per-query problem separates as in (50) and that its maximum is attained (Lemma D.9).

D.3.1 Strong duality and decoupling

Proof of Proposition 1. Randomization makes the optimal reward concave in the MGF budget B. A supporting line to this value function supplies the dual price. We then justify optimizing separately for each query.

We first make the randomization explicit, since both later steps rest on it. A randomized per-query policy is a probability measure μ_q on $\Pi^{\mathrm{ad}} = [0, \tau_{\max}] \times \mathbb{N}$; an action is drawn from this measure independently of the rollouts. The per-query reward and cost are the μ_q-averages of the pure values,

$$\mathbb{E}[r^{\pi_q} \mid q] = \int_{\Pi^{\mathrm{ad}}} V_\tau(n; q)\, d\mu_q(\tau, n), \qquad \mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q] = \int_{\Pi^{\mathrm{ad}}} M(q)^n\, d\mu_q(\tau, n), \tag{47}$$

so that $R(\pi) = \mathbb{E}_q \int V_\tau(n; q)\, d\mu_q$ and $C(\pi) = \mathbb{E}_q \int M(q)^n\, d\mu_q$. In the dual problem we restrict to finite-cost kernels, which include every policy feasible at a finite budget. Both reward and cost are *linear* in the kernel: for kernels μ_q^0, μ_q^1 and $t \in [0,1]$, the mixed kernel $\mu_q^t := t\mu_q^1 + (1-t)\mu_q^0$ has

$$\int V_\tau(n; q)\, d\mu_q^t = t \int V_\tau(n; q)\, d\mu_q^1 + (1-t) \int V_\tau(n; q)\, d\mu_q^0,$$

and likewise for the cost integrand $M(q)^n$. Hence R and C are affine along kernel mixtures, the fact used twice below.

Step 1 (Concavity of $\overline{R}^{\mathrm{ad}}$). The value $\overline{R}^{\mathrm{ad}}$ is non-decreasing, since enlarging B enlarges the feasible set. For concavity fix budgets B_0, B_1 and $\epsilon > 0$, and pick policies π_0, π_1 with $C(\pi_j) \leq B_j$ and $R(\pi_j) \geq \overline{R}^{\mathrm{ad}}(B_j) - \epsilon$. For $t \in [0,1]$, the policy that runs π_1 with probability t and π_0 otherwise, independently per query, is admissible because the adaptive class allows exactly this per-query randomization. By the affineness just shown, it has cost $tC(\pi_1) + (1-t)C(\pi_0) \leq tB_1 + (1-t)B_0$ and reward $tR(\pi_1) + (1-t)R(\pi_0)$, so

$$\overline{R}^{\mathrm{ad}}\big(tB_1 + (1-t)B_0\big) \;\geq\; tR(\pi_1) + (1-t)R(\pi_0) \;\geq\; t\overline{R}^{\mathrm{ad}}(B_1) + (1-t)\overline{R}^{\mathrm{ad}}(B_0) - \epsilon.$$

Letting $\epsilon \to 0$ gives concavity.

Step 2 (Supporting line is the dual). The map $u \mapsto \overline{R}^{\mathrm{ad}}(B + u)$ is concave, finite (as $\overline{R}^{\mathrm{ad}} \leq 1$), and, since $B > \mathbb{E}_q[M(q)]$ is above the one-rollout cost, has $u = 0$ in the interior of its domain.

A supporting line at $u = 0$ provides a slope $\lambda^* \geq 0$ with $\overline{R}^{\mathrm{ad}}(B+u) - \lambda^* u \leq \overline{R}^{\mathrm{ad}}(B)$ for all admissible u. Reading this at an arbitrary finite-cost policy π with $u = C(\pi) - B$ gives $R(\pi) - \lambda^*(C(\pi) - B) \leq \overline{R}^{\mathrm{ad}}(B)$ for every such π, that is $\sup_\pi \mathcal{L}(\pi, \lambda^*) \leq \overline{R}^{\mathrm{ad}}(B)$. The reverse inequality is weak duality, since $\mathcal{L}(\pi, \lambda) \geq R(\pi)$ whenever $C(\pi) \leq B$. The two meet:

$$\overline{R}^{\mathrm{ad}}(B) = \inf_{\lambda \geq 0} \sup_{\pi} \mathcal{L}(\pi, \lambda) = \inf_{\lambda > 0} \sup_{\pi} \mathcal{L}(\pi, \lambda).$$

The last equality follows by taking $\lambda \downarrow 0$ if the supporting slope is zero; the dual objective at positive λ is at most its $\lambda = 0$ value plus λB.

Step 3 (Decoupling). We write the per-policy Lagrangian

$$L_q(\pi_q, \lambda) \;:=\; \mathbb{E}[r^{\pi_q} \mid q] - \lambda\, \mathbb{E}[e^{\gamma \ell^{\pi_q}} \mid q], \tag{48}$$

so that $\mathcal{L}(\pi, \lambda) = \mathbb{E}_{q \sim \mathcal{Q}}[L_q(\pi_q, \lambda)] + \lambda B$ has no term coupling distinct queries, and the population policy acts on each query independently, so the supremum passes inside the expectation,

$$\sup_{\pi} \mathcal{L}(\pi, \lambda) = \mathbb{E}_q\Big[\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda)\Big] + \lambda B.$$

For $\lambda > 0$, the measurable ϵ-maximizing selectors can be chosen with integrable cost. Indeed, comparison with the one-rollout action gives

$$V_\tau(n; q) - \lambda M(q)^n \geq -\lambda M(q) - \epsilon, \qquad M(q)^n \leq M(q) + (1+\epsilon)/\lambda.$$

Thus the selector has finite population cost, and its integrated value is within ϵ of the pointwise supremum. Letting $\epsilon \downarrow 0$ proves the interchange. At $\lambda = 0$, take a measurable ϵ-maximizing action and replace it by a one-rollout action wherever its cost exceeds K. The resulting policy has cost at most $K + M(q)$, and its reward converges to that of the selector as $K \to \infty$ by dominated convergence. This proves the interchange at zero price as well. Lemma D.8 evaluates the pure-action objective, completing the proof.

■

The decoupling leaves one object to evaluate, the per-query inner supremum. The next lemma shows it is a plain joint supremum over pure pairs (τ, n): randomization, having only convexified the population problem, adds nothing query by query.

Lemma D.8 (Per-query inner problem as a joint supremum)**.** *For every $\lambda > 0$ and every q,*

$$\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = \sup_{\tau \in [0, \tau_{\max}],\, n \in \mathbb{N}} \big\{ V_\tau(n; q) - \lambda\, M(q)^n \big\}. \tag{49}$$

Proof of Lemma D.8. ***Pure pairs.*** For a pure pair $\pi_q = (\tau, n)$, the reward term is $\mathbb{E}[r^{\pi_q} \mid q] = V_\tau(n; q)$ by (8), and the cost term is $\mathbb{E}[e^{\gamma \sum_{i \leq n} \ell_i} \mid q] = M(q)^n$ because the lengths are i.i.d. conditional on q. Hence $L_q((\tau, n), \lambda) = V_\tau(n; q) - \lambda M(q)^n$.

Randomization adds nothing. A randomized π_q draws (τ, n) from a measure μ_q independent of the rollouts, so by (47) its Lagrangian value is the μ_q-mixture of the pure values, $L_q(\pi_q, \lambda) = \int \big[V_\tau(n; q) - \lambda M(q)^n\big] d\mu_q(\tau, n)$. A mixture never exceeds the supremum of the quantity it mixes, and the pure pairs are themselves admissible, so the two suprema coincide:

$$\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = \sup_{\tau \in [0, \tau_{\max}],\, n \in \mathbb{N}} \big\{ V_\tau(n; q) - \lambda M(q)^n \big\}.$$ ■

D.3.2 Recovering a constrained policy from a supporting price

The pure maximizers of a priced problem need not individually meet a requested budget. Let $B = e^{\gamma b}$, let $\lambda^* \geq 0$ be a supporting price from the strong-duality proof, and write

$$g_q(\lambda^*) = \sup_{\tau, n} \{ V_\tau(n; q) - \lambda^* M(q)^n \}.$$

Conditional on primal attainment, a finite-cost kernel μ^* is constrained-optimal if and only if it is supported on the priced maximizers for almost every query, is feasible, and satisfies complementary slackness:

$$C(\mu^*) \le B, \qquad \lambda^*\big(C(\mu^*) - B\big) = 0.$$

Indeed, the duality gap for any feasible μ is the sum of two nonnegative terms,

$$R^{\mathrm{ad}}(b) - R(\mu) = \mathbb{E}_q\left[g_q(\lambda^*) - \int \big(V_\tau(n;q) - \lambda^* M(q)^n\big)\, d\mu_q\right] + \lambda^*\big(B - C(\mu)\big).$$

It vanishes exactly under these conditions. At $\lambda^* = 0$ only reward is priced, and a slack budget is allowed; pointwise attainment at zero price is not asserted without primal attainment.

For two finite-cost policies supported on maximizers at the same price, with costs $C_- < C_+$ and $C_- \le B \le C_+$, choose the higher-cost policy independently before sampling with probability

$$\theta = \frac{B - C_-}{C_+ - C_-}.$$

The resulting policy has cost B and remains on the same supporting face, so it is constrained-optimal. Mixing arbitrary policies from a sampled price sweep also interpolates their costs and rewards, but does not establish this exact optimality condition.

Why mixing can be necessary. Consider one binary query with correct-answer probability 0.6, unit rollout length, and $\gamma = \log 2$, so $M = 2$. Correct scores are uniform on $[0.75, 1]$ and wrong scores on $[0, 0.25]$. For any finite positive temperature, one rollout has accuracy 0.6 and two have accuracy 0.84, since any correct rollout wins the two-rollout vote. At $\lambda^* = 0.12$, both priced values equal 0.36; every $n \ge 3$ has value at most $1 - 0.12(8) = 0.04$. At MGF budget $B = 3$, the one-rollout choice is feasible but suboptimal and the two-rollout choice is infeasible. Mixing them equally yields cost 3 and reward 0.72, exactly the dual bound $0.36 + 0.12(3)$. Thus, choosing a pure priced maximizer alone does not recover the constrained optimum.

D.3.3 Attainment of the per-query maximum

For $\lambda > 0$, we solve the inner problem $\sup_{\tau,n}\{V_\tau(n;q) - \lambda M(q)^n\}$ of §3.3. Because the cost $M(q)^n$ depends on the count alone,

$$\sup_{\tau,n}\big\{V_\tau(n;q) - \lambda M(q)^n\big\} = \max_{n\in\mathbb{N}}\big\{V^\star(n;q) - \lambda M(q)^n\big\}, \qquad V^\star(n;q) := \max_{\tau\in[0,\tau_{\max}]} V_\tau(n;q), \tag{50}$$

and the following lemma shows that both maxima are attained, so that the per-query optimal policy (9) is well defined. The decomposition in (50) implies that the optimal count is chosen against the best vote $\tau \in [0, \tau_{\max}]$ for each query. It is the one place where the $\tau = 0$ tie-breaking convention of Definition 2.1 is used.

Lemma D.9 (Attainment of the per-query maximum)**.** *Fix q and $\lambda > 0$, and let Assumption 1 hold.*

(i) (Ties are null.) *For every $n \ge 1$ and $\tau > 0$, two distinct answers attain $\max_a W_\tau(a, n; q)$ with probability zero; at $\tau = 0$, two distinct answers attain the maximal count with equal total scores with probability zero.*

(ii) (Continuity in the temperature.) *For every $n \ge 1$, $\tau \mapsto V_\tau(n;q)$ is continuous on $[0, \tau_{\max}]$; hence $V^\star(n;q) = \max_\tau V_\tau(n;q)$ is attained.*

(iii) (Finite search.) *Every maximizer of $n \mapsto V^\star(n;q) - \lambda M(q)^n$ satisfies $n \le N(\lambda, q) := \max\{n \in \mathbb{N} : M(q)^n \le M(q) + 1/\lambda\}$, and the maximum is attained.*

Proof of Lemma D.9. Write $N_a := \sum_{i\le n} \mathbf{1}\{a_i = a\}$ and $S_a := \sum_{i\le n} \phi_i \mathbf{1}\{a_i = a\}$ for the count and the total score of answer a.

(i). Condition on the answer pattern $(a_1, \dots, a_n)$; since the rollout triples are i.i.d., the scores are then independent with $\phi_i \sim p(\cdot \mid a_i, q)$, each atomless. Fix $a \ne a'$ and let $A := \{i : a_i = a\}$, $A' := \{i : a_i = a'\}$. If $A = A' = \emptyset$ both weights vanish while $\max_a W_\tau(a, n; q) > 0$; if exactly

one is empty the weights differ. If both are nonempty, pick $i_0 \in A$ and condition further on $(\phi_j)_{j\neq i_0}$: the tie $\sum_{i\in A} e^{\tau\phi_i} = \sum_{i\in A'} e^{\tau\phi_i}$ pins $e^{\tau\phi_{i_0}}$, hence ϕ_{i_0} (as $\tau > 0$), to a single value, an event of probability zero by atomlessness. At $\tau = 0$, a tie at the maximal count forces $A, A' \neq \emptyset$, and $S_a = S_{a'}$ pins ϕ_{i_0} to a single value likewise. Sum over the finitely many pairs and patterns.

(ii). Fix n. We show $\widehat{a}_\tau(n;q) \to \widehat{a}_{\tau_0}(n;q)$ almost surely as $\tau \to \tau_0$ within $[0, \tau_{\max}]$; dominated convergence of the bounded indicators $\mathbf{1}\{\widehat{a}_\tau(n;q) = a^*\}$ then gives continuity, and a continuous function on the compact interval $[0, \tau_{\max}]$ attains its maximum. For $\tau_0 > 0$: by (i) the maximizer $\widehat{a} := \widehat{a}_{\tau_0}(n;q)$ is almost surely unique, the finitely many strict inequalities $W_{\tau_0}(\widehat{a}, n; q) > W_{\tau_0}(a, n; q)$ persist on a neighbourhood of τ_0 by continuity of $\tau \mapsto W_\tau(\cdot, n; q)$, and there $\widehat{a}_\tau(n;q) = \widehat{a}$. For $\tau_0 = 0$: by (i) the lexicographic maximizer $\widehat{a}$ of (N_a, S_a), which is $\widehat{a}_0(n;q)$ under Definition 2.1, is almost surely unique. Since $\phi_i \in [0,1]$, $W_\tau(a, n; q) = N_a + \tau S_a + r_\tau(a)$ with $|r_\tau(a)| \leq 2n\tau^2$ for $\tau \leq 1$. For $a \neq \widehat{a}$, either $N_a \leq N_{\widehat{a}} - 1$, and then $W_\tau(\widehat{a}, n; q) - W_\tau(a, n; q) \geq 1 - n\tau - 4n\tau^2 > 0$ for all small τ, or $N_a = N_{\widehat{a}}$ and $S_a < S_{\widehat{a}}$, and then the difference is at least $\tau(S_{\widehat{a}} - S_a) - 4n\tau^2 > 0$ for all small τ. Hence $\widehat{a}_\tau(n;q) = \widehat{a}_0(n;q)$ for all sufficiently small $\tau > 0$. This is where the $\tau = 0$ convention enters: an arbitrary count tie-break can make $V_0(n;q)$ differ from its right-hand limit, so compactness alone would no longer establish attainment.

(iii). For every n, $V^\star(n;q) - \lambda M(q)^n \leq 1 - \lambda M(q)^n$, while at $n = 1$ the objective is at least $-\lambda M(q)$. If $M(q)^n > M(q) + 1/\lambda$ then $1 - \lambda M(q)^n < -\lambda M(q)$, so such n are strictly dominated by $n = 1$. The supremum is therefore a maximum over the finite set $\{1, \ldots, N(\lambda, q)\}$, and $N(\lambda, q) \geq 1$. ■

The unlimited-budget ceiling. The expression in (10) is indeed the limit of the budgeted value. It is an upper bound for every policy. Conversely, choose a measurable ϵ-maximizing action for $\sup_{\tau,n} V_\tau(n;q)$ and replace it by one rollout wherever its MGF cost exceeds K. The modified policy has finite population cost at most $K + \mathbb{E}_q[M(q)]$. As $K \to \infty$, its reward converges by dominated convergence to within ϵ of $R^{\infty,\mathrm{ad}}$. Letting $\epsilon \downarrow 0$ proves $R^{\mathrm{ad}}(b) \to R^{\infty,\mathrm{ad}}$. This argument requires neither finite-count attainment of every per-query ceiling nor the stronger rate assumptions below.

D.4 Proofs for §3.4: the asymptotic adaptive frontier

We prove the large-budget law in three steps: solve the per-query problem at a fixed temperature, optimize the temperature, and aggregate over queries. The ceiling is $R^{\infty,\mathrm{ad}} = \mathbb{E}_q[\sup_{\tau,n} V_\tau(n;q)]$. By strong duality,

$$\overline{R}^{\mathrm{ad}}(B) = \inf_{\lambda>0} \Big\{ \mathbb{E}_q\big[\sup_{\tau,n}(V_\tau(n;q) - \lambda M(q)^n)\big] + \lambda B \Big\}. \tag{51}$$

The per-query estimates concern $\lambda \downarrow 0$. Turning them into a population statement requires additional assumptions, stated explicitly in Assumption 3; in particular, pointwise asymptotics do not by themselves justify averaging over queries.

D.4.1 Per-query large-budget optimum at a fixed temperature

We start at the innermost layer of (51): the per-query inner problem $\sup_n\{V_\tau(n;q) - \lambda M(q)^n\}$ at a single query q and a single temperature τ, in the vanishing-price limit; the temperature itself is optimized in §D.4.2. Throughout this section, write the log inverse price and the fixed-temperature price exponent

$$\zeta := \log(1/\lambda) \longrightarrow \infty, \qquad \alpha_\tau(q) := \frac{I_\tau(q)}{I_\tau(q) + \log M(q)} \in (0,1) \quad \text{for } \delta^*_\tau(q) > 0,$$

where $I_\tau(q)$ is the slowest Cramér rate for the query q defined in (28). For non-perfect queries we extend $\alpha_\tau(q)$ to all temperatures by the same ratio, so it is zero at inconsistent temperatures; for perfect queries we set $\alpha_\tau(q) = 1$.

Fix a temperature τ with $\delta^*_\tau(q) > 0$. Writing the reward as one minus its gap, the inner objective at this τ is $V_\tau(n;q) - \lambda M(q)^n = 1 - h_\tau(n)$, where

$$h_\tau(n) := \big(1 - V_\tau(n;q)\big) + \lambda M(q)^n \tag{52}$$

is the sum of the reward gap and the cost; maximizing the net value over n is the same as minimizing h_τ. On the exponential scale each term is governed by a single rate. The reward gap falls at rate $I_\tau(q)$ by the recalled Cramér rate in Lemma D.4, while the cost rises at rate $\log M(q)$, since $\lambda M(q)^n = e^{-\zeta + n \log M(q)}$. Hence

$$h_\tau(n) = \underbrace{e^{-nI_\tau(q)(1+o(1))}}_{\text{falls, rate } I_\tau(q)} + \underbrace{e^{-\zeta + n \log M(q)}}_{\text{rises, rate } \log M(q)}. \tag{53}$$

At leading exponential order, the optimum balances the decreasing error term against the increasing cost term. This argument does not require the finite-count reward curve to be monotone. Equating the two exponents suggests

$$n^* I_\tau(q) = \zeta - n^* \log M(q) \quad \Longleftrightarrow \quad n^* = \frac{\zeta}{I_\tau(q) + \log M(q)},$$

and the common exponent there is $n^* I_\tau(q) = \alpha_\tau(q)\zeta$. So each term of $h_\tau(n^*)$ is about $e^{-\alpha_\tau(q)\zeta}$. The next lemma proves this balance and the resulting count asymptotics.

Lemma D.10 (Fixed-temperature inner value). *Fix q with $p(a^* \mid q) < 1$ and τ with $\delta^*_\tau(q) > 0$. Let $n_\tau(\lambda; q)$ be any integer minimizer of h_τ. As $\lambda \to 0_+$, it satisfies*

$$n_\tau(\lambda; q) = \frac{\zeta}{I_\tau(q) + \log M(q)} \, (1 + o(1)), \tag{54}$$

and the resulting inner value is

$$\sup_{n \in \mathbb{N}} \left[V_\tau(n; q) - \lambda M(q)^n \right] = 1 - e^{-\alpha_\tau(q)\, \zeta (1 + o(1))}. \tag{55}$$

Proof of Lemma D.10. Abbreviate $I := I_\tau(q)$ and $\alpha := \alpha_\tau(q) = I/(I + \log M(q)) \in (0, 1)$. By Lemma D.4, there is a deterministic sequence $\delta_n \to 0$ with $1 - V_\tau(n; q) = e^{-n(I+\delta_n)}$. With the exact cost, the two terms of (52) are

$$1 - V_\tau(n; q) = e^{-n(I+\delta_n)}, \qquad \lambda M(q)^n = e^{-\zeta + n \log M(q)}, \tag{56}$$

so $h_\tau(n) = e^{-n(I+\delta_n)} + e^{-\zeta + n \log M(q)}$. Since the net value is $1 - h_\tau$, the value claim (55) is equivalent to

$$\min_{n \in \mathbb{N}} h_\tau(n) = e^{-\alpha\zeta(1+o(1))}, \tag{57}$$

which we prove by two matching inequalities.

Step 1 (upper bound). Evaluate h_τ at the balance count rounded upward,

$$n^* := \left\lceil \frac{\zeta}{I + \log M(q)} \right\rceil = \frac{\zeta}{I + \log M(q)} + O(1). \tag{58}$$

As $\lambda \to 0_+$ we have $\zeta \to \infty$, hence $n^* \to \infty$ and $\delta_{n^*} \to 0$. The two exponents at n^* in (56) are then both $\alpha\zeta(1 + o(1))$: for the reward term,

$$n^*(I + \delta_{n^*}) = \frac{\zeta I}{I + \log M(q)} + o(\zeta) = \alpha\zeta(1 + o(1)),$$

and for the cost term, using $1 - \frac{\log M(q)}{I + \log M(q)} = \alpha$,

$$\zeta - n^* \log M(q) = \zeta \Big(1 - \tfrac{\log M(q)}{I + \log M(q)} \Big) + O(1) = \alpha\zeta(1 + o(1)).$$

Substituting into (56),

$$\min_n h_\tau(n) \le h_\tau(n^*) = e^{-\alpha\zeta(1+o(1))} + e^{-\alpha\zeta(1+o(1))} = e^{-\alpha\zeta(1+o(1))},$$

the factor of two absorbed into the $(1 + o(1))$ in the exponent.

Step 2 (lower bound). We show no count pushes h_τ below $e^{-\alpha\zeta(1+o(1))}$, i.e, $\min_n h_\tau(n) \ge e^{-\alpha\zeta(1+o(1))}$. Fix a level $s > 0$ and suppose some n achieves $h_\tau(n) = e^{-s\zeta}$. Since $h_\tau(n)$ is a sum of two nonnegative terms, *each* term is at most $e^{-s\zeta}$, which by (56) reads

$$\text{(reward)} \quad e^{-n(I+\delta_n)} \le e^{-s\zeta} \implies n(I + \delta_n) \ge s\zeta, \tag{59}$$

$$\text{(cost)} \quad e^{-\zeta + n \log M(q)} \le e^{-s\zeta} \implies n \log M(q) \le (1 - s)\zeta. \tag{60}$$

Then (60) forces $1 - s > 0$. Dividing (59) by (60) cancels n:

$$\frac{I + \delta_n}{\log M(q)} \geq \frac{s\zeta}{(1-s)\zeta} = \frac{s}{1-s}.$$

Cross-multiplying (both sides positive) and gathering the s terms,

$$\begin{aligned}(I + \delta_n)(1 - s) \geq s \log M(q) &\iff s\big(I + \log M(q) + \delta_n\big) \leq I + \delta_n \\ &\iff s \leq \frac{I + \delta_n}{I + \log M(q) + \delta_n}.\end{aligned}$$

It remains to read the right-hand side as $\alpha + o(1)$, uniformly over the counts that can attain a level. A count below any fixed N cannot: the unanimity floor (22) keeps $h_\tau(n) \geq p_{\min}^N$ for $n \leq N$, so such counts only attain levels $s = O(1/\zeta)$. For $n \geq N$, the map $\delta \mapsto (I + \delta)/(I + \log M(q) + \delta)$ is increasing, so $s \leq (I + \bar{\delta}_N)/(I + \log M(q) + \bar{\delta}_N)$ with $\bar{\delta}_N := \sup_{m \geq N} \delta_m \to 0$ as $N \to \infty$. Sending $\zeta \to \infty$ and then $N \to \infty$, every level attained by some n satisfies $s \leq \alpha + o(1)$, such that $h_\tau(n) \geq e^{-(\alpha + o(1))\zeta}$ for every n. Consequently, $\min_n h_\tau(n) \geq e^{-\alpha\zeta(1+o(1))}$.

Conclusion. Steps 1 and 2 give (57), hence (55) as $V_\tau(n; q) - \lambda M(q)^n = 1 - h_\tau$. To show the expression of the balance count $n_\tau(\lambda; q)$ in (54), let $n_\tau(\lambda; q)$ minimize h_τ. By (57),

$$h_\tau(n_\tau(\lambda; q)) = e^{-\alpha\zeta(1+o(1))},$$

therefore

$$s := -\frac{1}{\zeta} \log h_\tau(n_\tau(\lambda; q)) = \alpha + o(1).$$

In particular $h_\tau(n_\tau(\lambda; q)) \to 0$, while the unanimity floor (22) bounds h_τ away from zero on any bounded range of counts, so $n_\tau(\lambda; q) \to \infty$ and $\delta_{n_\tau(\lambda;q)} \to 0$. Applying (59)–(60) at this level pins

$$n_\tau(\lambda; q) \cdot (I + o(1)) \geq (\alpha + o(1))\zeta, \qquad n_\tau(\lambda; q) \cdot \log M(q) \leq (1 - \alpha + o(1))\zeta.$$

Since $\alpha = I/(I + \log M(q))$, the above inequalities imply $n_\tau(\lambda; q) = \zeta/(I + \log M(q))\,(1 + o(1))$. ■

D.4.2 Choosing the temperature

Fix $q \in \mathcal{C}^{\mathrm{ad}}$ with $0 < p(a^* \mid q) < 1$. With the fixed-temperature value of §D.4.1 in hand, we now let the temperature vary, the second layer of (51). By Lemma D.10 the inner value at temperature τ is $1 - e^{-\alpha_\tau(q)\zeta(1+o(1))}$. Its leading exponential behavior improves as the exponent $\alpha_\tau(q) = \frac{I_\tau(q)}{I_\tau(q) + \log M(q)}$ increases. This exponent is increasing in $I_\tau(q)$, so a temperature maximizing $I_\tau(q)$ is a natural candidate for leading-order optimality. Lemma D.11 below justifies the maxima in the notation

$$\begin{aligned}\tau^\star(q) &\in \arg\max_{\tau:\, \delta_\tau^*(q) > 0} I_\tau(q), \quad \text{a selected rate maximizer,} \\ I^{\mathrm{ad}}(q) &:= \max_{\tau:\, \delta_\tau^*(q) > 0} I_\tau(q), \quad \text{with the largest price exponent} \\ \alpha^{\mathrm{ad}}(q) &:= \frac{I^{\mathrm{ad}}(q)}{I^{\mathrm{ad}}(q) + \log M(q)}.\end{aligned} \tag{61}$$

The difficulty in the proof is that the supremum over τ and the large-ζ limit must commute, which needs the per-τ error term in Lemma D.10 controlled uniformly in τ. The uniform bounds of §D.1.4 supply exactly that control, and this subsection deploys them in two results, proved in order.

- First, the rate-maximizer $\tau^\star(q)$ exists and sits at a strictly positive consistency margin: the temperatures that matter form a compact band away from the boundary (Lemma D.11), on which the two-sided uniform rate of Lemma D.6 applies.
- Second, the margin-free floor of Corollary D.1 excludes every other temperature, including those near the consistency boundary the band does not reach, which shows that a rate-maximizing temperature is asymptotically sufficient (Proposition D.1).

We start with the first result.

Lemma D.11 (The rate-maximizer stays away from the consistency boundary). *For every $q \in \mathcal{C}^{\mathrm{ad}}$ there is a margin $\delta_0(q) > 0$ such that the maximizer $\tau^\star(q)$ lies in the closed band $\Theta_{\delta_0(q)} := \{\tau : \delta^*_\tau(q) \geq \delta_0(q)\}$, and the restriction does not lower the supremum: $\alpha^{\mathrm{ad}}(q) = \sup_{\tau \in \Theta_{\delta_0(q)}} \alpha_\tau(q)$.*

Proof of Lemma D.11. Since $q \in \mathcal{C}^{\mathrm{ad}}$, some temperature is consistent, so the best adaptive rate and its exponent are strictly positive: $I^{\mathrm{ad}}(q) > 0$ and $\alpha^{\mathrm{ad}}(q) = I^{\mathrm{ad}}(q)/(I^{\mathrm{ad}}(q) + \log M(q)) > 0$. If $p(a^* \mid q) = 1$ the claim is immediate: every temperature is consistent with margin $\delta^*_\tau(q) = \omega_\tau(a^*; q) \geq 1$, continuous and positive on the compact $[0, \tau_{\max}]$, so $\delta_0(q) := \min_\tau \delta^*_\tau(q) > 0$ gives $\Theta_{\delta_0(q)} = [0, \tau_{\max}]$, and under the convention $I_\tau(q) = \infty$ of (28) every exponent equals $\alpha_\tau(q) = 1 = \alpha^{\mathrm{ad}}(q)$. So assume $p(a^* \mid q) < 1$; then Lemma D.2 applies to every competitor.

Step 1 (positive rate iff positive margin). By the rate–drift dichotomy of Lemma D.2 and the drift identity (24), the vote rate $I_\tau(q) = \min_{a \in \mathcal{A}_q} I_a(\tau)$ of (28) is positive if and only if every lead walk drifts strictly upward, that is if and only if $\delta^*_\tau(q) > 0$. Since $M(q) > 1$, the exponent $\alpha_\tau(q) = \frac{I_\tau(q)}{I_\tau(q) + \log M(q)}$ is positive exactly when the rate is, so

$$\alpha_\tau(q) > 0 \iff I_\tau(q) > 0 \iff \delta^*_\tau(q) > 0. \tag{62}$$

Step 2 (continuity in the temperature). By Lemma D.2(iii), each competitor rate $\tau \mapsto I_a(\tau)$ is continuous on $[0, \tau_{\max}]$; hence the vote rate $I_\tau(q) = \min_{a \in \mathcal{A}_q} I_a(\tau)$ and the exponent $\alpha_\tau(q)$ are continuous on $[0, \tau_{\max}]$.

Step 3 (a compact band inside the consistent set). Consider the superlevel set of the exponent at half its maximum,

$$S := \left\{\tau \in [0, \tau_{\max}] : \alpha_\tau(q) \geq \tfrac{1}{2}\alpha^{\mathrm{ad}}(q)\right\}.$$

By Step 2 it is closed, hence compact. It is non-empty, since some consistent temperature has exponent within $\frac{1}{2}\alpha^{\mathrm{ad}}(q)$ of the positive supremum. And it lies strictly inside the consistent set: for $\tau \in S$ the exponent $\alpha_\tau(q) \geq \frac{1}{2}\alpha^{\mathrm{ad}}(q) > 0$, so the margin is positive there by (62). The margin being continuous and positive on the compact S, it attains a positive minimum, and we set $\delta_0(q) := \min_{\tau \in S} \delta^*_\tau(q) > 0$, so that $S \subseteq \Theta_{\delta_0(q)}$.

Step 4 (the maximizer lies in the band). Off S the exponent is $\alpha_\tau(q) < \frac{1}{2}\alpha^{\mathrm{ad}}(q)$, strictly below the supremum, so the supremum over all consistent temperatures equals the supremum over S. Being continuous on the compact S, the exponent attains its maximum there at some $\tau^\star(q) \in S \subseteq \Theta_{\delta_0(q)}$ with $\alpha_{\tau^\star(q)}(q) = \alpha^{\mathrm{ad}}(q)$. Restricting to $\Theta_{\delta_0(q)}$ therefore moves neither the maximizer nor the supremum, $\alpha^{\mathrm{ad}}(q) = \sup_{\tau \in \Theta_{\delta_0(q)}} \alpha_\tau(q)$. ■

Lemma D.11 delivers a compact band $\Theta_{\delta_0(q)}$ that contains a rate-maximizer, and on any such band the two-sided uniform rate estimate of Lemma D.6 (§D.1.4) applies: the pointwise convergence of Lemma D.4 holds at a single deterministic error bound across the whole band, the form that Assumption 3 builds on in §D.4.3. In particular the maximizer exists, and applying Lemma D.10 at it delivers the achievability half of the per-query optimum below. It remains to rule out every other temperature, including those near the consistency boundary that the band does not reach; doing so shows that the single temperature $\tau^\star(q)$ attains the per-query optimum to leading exponential order (asymptotic sufficiency, not uniqueness or exact finite-price optimality).

Proposition D.1 (Explicit per-query policy as the price vanishes). *Fix $q \in \mathcal{C}^{\mathrm{ad}}$ with $p(a^* \mid q) < 1$. As $\lambda \to 0_+$,*

$$\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = 1 - e^{-\alpha^{\mathrm{ad}}(q)\,\zeta(1+o(1))}, \tag{63}$$

realized to exponential precision by the single temperature $\tau^\star(q)$ of (61) and the rollout count

$$n_\lambda(q) := n_{\tau^\star(q)}(\lambda; q) = \frac{\zeta}{I^{\mathrm{ad}}(q) + \log M(q)}\,(1 + o(1)). \tag{64}$$

Remark 2. *On a perfect query, $p(a^* \mid q) = 1$, where $I^{\mathrm{ad}}(q) = \infty$, $\alpha^{\mathrm{ad}}(q) = 1$, and (64) is undefined, the conclusion (63) holds directly: $V_\tau(n; q) \equiv 1$, so the supremum equals $1 - \lambda M(q) = 1 - e^{-\zeta(1+O(1/\zeta))}$, realized at a single rollout, and we set $n_\lambda(q) := 1$.*

Proof of Proposition D.1. By (49) and the decomposition $V_\tau(n;q) - \lambda M(q)^n = 1 - h_\tau(n)$ of (52), the per-query supremum is

$$\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = 1 - \inf_{\tau \in [0,\tau_{\max}]} \min_{n \in \mathbb{N}} h_\tau(n). \tag{65}$$

So to achieve (63), it suffices to show

$$\inf_{\tau \in [0,\tau_{\max}]} \min_{n \in \mathbb{N}} h_\tau(n) = e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}.$$

By Lemma D.11, there exists a rate-maximizer $\tau^\star(q) \in \arg\max_{\tau:\delta^*_\tau(q)>0} I_\tau(q)$ that attains the best exponent $\alpha_{\tau^\star(q)}(q) = \alpha^{\mathrm{ad}}(q)$. Using Lemma D.10, we establish that

$$\min_n h_{\tau^\star(q)}(n) = e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}$$

with the minimizer count in (64). Therefore, we have shown one direction that

$$\inf_{\tau \in [0,\tau_{\max}]} \min_{n \in \mathbb{N}} h_\tau(n) \le e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}.$$

The rest of the proof is to show the reverse inequality.

The reverse inequality requires a bound uniform in τ: the optimizer can change with λ, so pointwise convergence at each fixed temperature is insufficient. We first exclude inconsistent temperatures, then apply the margin-free error bound to all remaining temperatures.

Step 1 (inconsistent temperatures stay far from the ceiling). By the error floor at inconsistent temperatures (Lemma D.7), there is a constant $\kappa = \kappa(q) > 0$, depending only on $(p_{\min}, \tau_{\max})$, such that $1 - V_\tau(n;q) \ge \kappa$ for all $n \in \mathbb{N}$ and all inconsistent τ (those with $\delta^*_\tau(q) \le 0$), which is (36).

On one hand, (36) and $h_\tau(n) = (1 - V_\tau(n;q)) + \lambda M(q)^n \ge 1 - V_\tau(n;q)$ force every inconsistent τ to keep

$$\min_n h_\tau(n) \;\ge\; \min_n \big(1 - V_\tau(n;q)\big) \;\ge\; \kappa > 0. \tag{66}$$

On the other hand, $q \in \mathcal{C}^{\mathrm{ad}}$ supplies the consistent rate-maximizer $\tau^\star(q)$, at which Lemma D.10 gives

$$\min_n h_{\tau^\star(q)}(n) \;=\; e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))} \;\xrightarrow{\lambda \to 0_+}\; 0, \tag{67}$$

equivalently $\sup_n [V_{\tau^\star(q)}(n;q) - \lambda M(q)^n] = 1 - e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))} \to 1$. Hence for all ζ large enough that $e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))} < \kappa$, the infimum is unaffected by discarding inconsistent temperatures:

$$\inf_{\tau \in [0,\tau_{\max}]} \min_n h_\tau(n) = \inf_{\tau:\, \delta^*_\tau(q)>0} \min_n h_\tau(n). \tag{68}$$

Step 2 (a reward-gap floor, uniform over all consistent temperatures). From here on we work only with the consistent temperatures $\{\tau : \delta^*_\tau(q) > 0\}$ left by Step 1. On this set the vote rate $I_\tau(q)$ is positive but may be arbitrarily small near the consistency boundary, so any estimate whose constants degrade with the margin is useless there. This is exactly what the margin-free floor of Corollary D.1 was built for: with $\theta_+ = -\log p_{\min}$ and constants $c_2 < \infty$, $n_3 \in \mathbb{N}$ depending only on $(p_{\min}, \tau_{\max})$,

$$1 - V_\tau(n;q) \;\ge\; e^{-nI_\tau(q) - \theta_+\sqrt{n} - c_2}$$

for every consistent τ and every $n \ge n_3$, which is (35).

Step 3 (bound every achievable exponent). It remains to bound, uniformly over the consistent temperatures left by Step 1, how small $h_\tau(n)$ can be. We measure each value on the log inverse price scale: for a consistent τ and a count n, let

$$s \;:=\; -\frac{1}{\zeta} \log h_\tau(n) \tag{69}$$

be the exponential rate $h_\tau(n) = e^{-s\zeta}$ it attains, so that bounding $\min_n h_\tau(n)$ from below is the same as bounding this rate s from above. Every attained level obeys $s < 1$, since $h_\tau(n) \ge \lambda M(q)^n >$

$\lambda = e^{-\zeta}$ (as $M(q) > 1$); we may also take $s \geq 0$, since a negative level makes the bound (72) below trivial.

If $n < n_3$, the unanimity floor (22) gives $h_\tau(n) \geq p_{\min}^{n_3}$, so $s = O(\zeta^{-1})$. If $n \geq n_3$, then $h_\tau(n) = e^{-s\zeta}$ forces each of its two nonnegative terms below $e^{-s\zeta}$. The cost term gives

$$\lambda M(q)^n = e^{-\zeta + n \log M(q)} \leq e^{-s\zeta} \implies n \log M(q) \leq (1-s)\zeta, \quad \text{hence} \quad \sqrt{n} \leq \sqrt{\zeta / \log M(q)}, \tag{70}$$

while the reward term, bounded below by the margin-free floor (35), gives

$$e^{-nI_\tau(q) - \theta_+\sqrt{n} - c_2} \leq 1 - V_\tau(n;q) \leq e^{-s\zeta} \implies nI_\tau(q) \geq s\zeta - \theta_+\sqrt{n} - c_2 \geq s\zeta - C'\sqrt{\zeta}, \tag{71}$$

where $C' := \theta_+ / \sqrt{\log M(q)} + c_2$ for $\zeta \geq 1$, the bound on $\sqrt{n}$ from (70) absorbing the $\theta_+\sqrt{n}$ term. Dividing the reward bound (71) by the cost bound (70) cancels the unknown count n (the bound is trivial if $s\zeta \leq C'\sqrt{\zeta}$),

$$\frac{I_\tau(q)}{\log M(q)} = \frac{nI_\tau(q)}{n \log M(q)} \geq \frac{s\zeta - C'\sqrt{\zeta}}{(1-s)\zeta} = \frac{s - C'\zeta^{-1/2}}{1-s},$$

and cross-multiplying (both sides positive) then solving for s yields

$$s \leq \frac{I_\tau(q) + C'\zeta^{-1/2} \log M(q)}{I_\tau(q) + \log M(q)} \leq \alpha_\tau(q) + C'\zeta^{-1/2}, \tag{72}$$

with C' independent of τ and n.

The exponent increases with the rate, and $I^{\mathrm{ad}}(q) = \max_{\tau:\, \delta_\tau^*(q) > 0} I_\tau(q)$ is the largest rate, so $\alpha_\tau(q) \leq \alpha^{\mathrm{ad}}(q)$ for every consistent τ; together with the small-n case this gives $s \leq \alpha^{\mathrm{ad}}(q) + C'\zeta^{-1/2}$ in both cases, uniformly in consistent τ. Equivalently $\min_n h_\tau(n) \geq e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}$ for every consistent τ, so with (68),

$$\inf_{\tau \in [0, \tau_{\max}]} \min_n h_\tau(n) \geq e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}.$$

Together with the achievability shown at the start, this proves $\inf_\tau \min_n h_\tau(n) = e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}$, hence (63) via (65), realized by the temperature $\tau^\star(q)$ and the count (64). ■

D.4.3 Aggregating over the population

For a measurable query set E, write $\mathbb{E}_{q \in E}[f(q)] := \mathbb{E}_q[\mathbf{1}\{q \in E\} f(q)]$; this is an unnormalized restricted expectation. Essential infima over E are taken with respect to the restricted query distribution. The previous two subsections give each query's optimal value; the outer layers of (51) remain, the population average $\mathbb{E}_q$ and the price-to-budget trade. The single tool is the Laplace principle, that an exponential average is dominated by the extreme value of its exponent; we state it (Lemma D.12), record the regularity it needs (Assumption 3), and then eliminate λ through a primal–dual sandwich (Theorem D.2).

Lemma D.12 (Laplace principle)**.** *Let ν be a probability measure and X a bounded measurable function with essential supremum $\bar{X}$ and essential infimum $\underline{X}$. As $A \to \infty$,*

$$\tfrac{1}{A} \log \int e^{AX}\, d\nu \to \bar{X}, \qquad \tfrac{1}{A} \log \int e^{-AX}\, d\nu \to -\underline{X}.$$

Proof of Lemma D.12. Since $X \leq \bar{X}$ a.s., $\int e^{AX} d\nu \leq e^{A\bar{X}}$, the upper bound. For the lower bound, $E_\epsilon := \{X > \bar{X} - \epsilon\}$ has $\nu(E_\epsilon) > 0$ by definition of $\bar{X}$, so $\int e^{AX} d\nu \geq \nu(E_\epsilon) e^{A(\bar{X} - \epsilon)}$ and $\frac{1}{A} \log \int e^{AX} d\nu \geq \bar{X} - \epsilon + \frac{1}{A} \log \nu(E_\epsilon)$; send $A \to \infty$, then $\epsilon \downarrow 0$. The second claim is the first applied to $-X$. ■

Proposition D.1 describes what happens for each query as the price $\lambda = e^{-\zeta}$ tends to zero. To obtain a population frontier, we need these estimates to remain valid after averaging over queries. The conditions below explain when this step is justified. We use the same temperature $\tau^\star(q)$ and integer count $n_\lambda(q)$ as in Proposition D.1; no new policy is introduced here.

All query-dependent policy choices are measurable, following the convention in Appendix D.1. Recall that $\alpha^{\mathrm{ad}}(q) = I^{\mathrm{ad}}(q)/(I^{\mathrm{ad}}(q) + \log M(q))$, with $\alpha^{\mathrm{ad}}(q) = 1$ on perfect queries, and write

$$\alpha^{\mathrm{ad}}_{\min} := \operatorname*{ess\,inf}_{q \in \mathcal{C}^{\mathrm{ad}}} \alpha^{\mathrm{ad}}(q). \tag{73}$$

Assumption 3 (Laplace regularity)**.** (L1) Bounded per-rollout cost. *There exist* $1 < m_- \le m_+ < \infty$ *such that* $m_- \le M(q) \le m_+$ *for* $\mathcal{Q}$*-a.e.* q*.*

(L2) Positive minimum error-decay rate. $\operatorname{ess\,inf}_{q \in \mathcal{C}^{\mathrm{ad}}} I^{\mathrm{ad}}(q) > 0$.

(L3) Finite-cost attainment on unsolvable queries. *For* $\mathcal{Q}$*-a.e.* $q \notin \mathcal{C}^{\mathrm{ad}}$*, a pair* $(\widehat{\tau}_q, \widehat{n}_q)$ *with finite count attains* $\sup_{\tau,n} V_\tau(n; q)$*, and*

$$K^- := \mathbb{E}_q\Big[M(q)^{\widehat{n}_q} \mathbf{1}\{q \notin \mathcal{C}^{\mathrm{ad}}\}\Big] < \infty.$$

(L4) Uniform per-query asymptotics. *There is a deterministic* $\epsilon_\zeta \to 0$*, independent of* q*, such that for all sufficiently large* ζ*, with* $\lambda = e^{-\zeta}$ *and for* $\mathcal{Q}$*-a.e.* $q \in \mathcal{C}^{\mathrm{ad}}$*:*

(a) *The optimal priced value satisfies*

$$\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = 1 - e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+u_q(\zeta))}, \qquad |u_q(\zeta)| \le \epsilon_\zeta.$$

(b) *For a non-perfect query, the temperature* $\tau^\star(q)$ *and count* $n_\lambda(q)$ *of Proposition D.1 satisfy the separate error and cost bounds*

$$\left|\frac{1}{\zeta}\log\big(1 - V_{\tau^\star(q)}(n_\lambda(q); q)\big) + \alpha^{\mathrm{ad}}(q)\right| \le \epsilon_\zeta,$$
$$\left|\frac{n_\lambda(q)\log M(q)}{\zeta} - (1 - \alpha^{\mathrm{ad}}(q))\right| \le \epsilon_\zeta.$$

Next, we explain each condition in order. First, (L1) bounds the cost MGF uniformly over queries. A natural way to satisfy it is to cap the number of tokens in each rollout: if $1 \le \ell_1 \le \ell_{\max}$ for a common token cap $\ell_{\max}$, then $e^\gamma \le M(q) \le e^{\gamma \ell_{\max}}$ for every query. The lower bound simply says that every rollout consumes tokens. Condition (L2) says that even the slowest solvable query has a positive error-decay rate. This is automatic for a finite query population: each non-perfect query in $\mathcal{C}^{\mathrm{ad}}$ has $I^{\mathrm{ad}}(q) > 0$, and the minimum of finitely many positive rates is positive. For an infinite population, positive rates can approach zero, so a uniform lower bound must be required separately.

For (L3), consider a query outside $\mathcal{C}^{\mathrm{ad}}$, which cannot reach accuracy one by increasing its rollout count. This condition asks that its best attainable accuracy instead be reached at some finite count $\widehat{n}_q$. These queries can then be assigned their best actions at a fixed aggregate cost; only the solvable queries determine how quickly the population approaches its ceiling as the budget grows.

The finite-count requirement holds, for example, if $\delta^*_\tau(q) < 0$ for *every* $\tau \in [0, \tau_{\max}]$. At every temperature, an incorrect answer then has a larger expected voting weight than the correct answer. Since the margin is continuous on this compact interval, it is bounded away from zero on the negative side. The bounded lead increments therefore give $\sup_\tau V_\tau(n; q) \to 0$ as $n \to \infty$ (a variance bound on the sample mean suffices, uniformly in τ). If the correct answer has positive sampling probability, one rollout already has positive accuracy, so the maximum must occur among finitely many counts. Continuity in τ, from Lemma D.9, then gives a maximizing temperature at one of those counts. If the correct answer is never sampled, the ceiling is zero and is already attained at one rollout. Thus strict negativity at all temperatures is a sufficient condition for finite-count attainment; merely having $\delta^*_\tau(q) \le 0$ is not the same condition. For a finite query population, the aggregate cost of these finite-count actions is automatically finite. For an infinite population, (L3) also requires this cost to be integrable.

Finally, (L4) asks that the approximation errors in the preceding per-query asymptotic formulas vanish uniformly over queries, so that we can use these formulas inside the population averages. Without uniform control, queries whose approximations converge increasingly slowly could change the aggregate rate.

Part (a) controls the optimal priced value. Its deficit from one includes both the voting error and the priced cost: for a fixed action (τ, n),

$$1 - \big[V_\tau(n;q) - \lambda M(q)^n\big] = \big[1 - V_\tau(n;q)\big] + \lambda M(q)^n.$$

This value estimate yields the dual bound on population accuracy. To construct a policy that meets a given budget, we also need its actual error and its cost separately. This is the role of part (b): under the policy of Proposition D.1, the error decays with exponent $\alpha^{\mathrm{ad}}(q)$, while the MGF cost grows with exponent $1-\alpha^{\mathrm{ad}}(q)$, both on the scale $\zeta = \log(1/\lambda)$. The cost estimate tells us how small the price can be while meeting the budget; the error estimate then tells us the accuracy achieved at that price. The two exponents follow from the count balance $n_\lambda(q) = \zeta/(I^{\mathrm{ad}}(q) + \log M(q))(1+o(1))$ and the error rate $1 - V_{\tau^\star(q)}(n;q) = e^{-nI^{\mathrm{ad}}(q)(1+o(1))}$.

For a finite query population, both parts of (L4) follow from the per-query results: take the largest approximation error over the finitely many queries. Each error tends to zero, so their maximum does too. Thus (L4)(b) does not add a new per-query asymptotic claim; it requires uniformity of the error and cost estimates already derived. Perfect queries need only one rollout, have zero voting error, and contribute bounded cost under (L1). Their priced value is exactly $1-\lambda M(q)$, so they satisfy part (a) directly and are excluded from the log-error expression in part (b).

Together, (L1) and (L2) ensure $\alpha^{\mathrm{ad}}_{\min} > 0$. The additional condition $\alpha^{\mathrm{ad}}_{\min} < 1$ in the theorem below ensures that non-perfect solvable queries remain; their error decay determines the frontier rate.

We can now state and prove the MGF-scale form of the asymptotic frontier.

Theorem D.2 (Asymptotic adaptive Pareto frontier, MGF scale)**.** *Assume $\mathbb{P}_q[\mathcal{C}^{\mathrm{ad}}] > 0$, Assumption 3, and $\alpha^{\mathrm{ad}}_{\min} < 1$. Then, as $B \to \infty$,*

$$R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) \;=\; B^{-\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min}) + o(1)}. \tag{74}$$

Remark 3. *Using $\overline{R}^{\mathrm{ad}}(e^{\gamma b}) = R^{\mathrm{ad}}(b)$ gives $R^{\infty,\mathrm{ad}} - R^{\mathrm{ad}}(b) = e^{-r^{\mathrm{ad}} b(1+o(1))}$ with $r^{\mathrm{ad}} := \gamma\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min})$. This is the worst-case exchange rate: with risk-adjusted rollout length $\ell_\gamma(q) := \gamma^{-1}\log M(q)$,*

$$\gamma\,\frac{\alpha^{\mathrm{ad}}(q)}{1-\alpha^{\mathrm{ad}}(q)} = \gamma\,\frac{I^{\mathrm{ad}}(q)}{\log M(q)} = \frac{I^{\mathrm{ad}}(q)}{\ell_\gamma(q)},$$

and since $x \mapsto x/(1-x)$ is increasing, taking essential infima over $q \in \mathcal{C}^{\mathrm{ad}}$ gives $r^{\mathrm{ad}} = \operatorname{ess\,inf}_{q\in\mathcal{C}^{\mathrm{ad}}} I^{\mathrm{ad}}(q)/\ell_\gamma(q)$. The endpoint $\alpha^{\mathrm{ad}}_{\min} = 1$ is degenerate, forcing $\mathcal{Q}$-a.e. query in $\mathcal{C}^{\mathrm{ad}}$ to be perfect, where a single rollout already attains the ceiling; the hypothesis $\alpha^{\mathrm{ad}}_{\min} < 1$ excludes it.

Proof of Theorem 3 (main text). Theorem D.2 gives $R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) = B^{-\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min})+o(1)}$ as $B \to \infty$. Substituting the token-budget relation $B = e^{\gamma b}$ and the identification $r^{\mathrm{ad}} = \gamma\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min}) = \operatorname{ess\,inf}_{q\in\mathcal{C}^{\mathrm{ad}}} I^{\mathrm{ad}}(q)/\ell_\gamma(q)$ of the preceding remark yields $R^{\infty,\mathrm{ad}} - R^{\mathrm{ad}}(b) = e^{-r^{\mathrm{ad}} b(1+o(1))}$ as $b \to \infty$, which is (13). ■

Proof of Theorem D.2. Throughout, the reward gap and the cost are aggregated by the Laplace principle over the *non-perfect consistent queries* $\{q \in \mathcal{C}^{\mathrm{ad}} : p(a^* \mid q) < 1\}$, on which the per-query exponents are genuine; perfect queries ($p(a^* \mid q) = 1$) and inconsistent queries ($q \notin \mathcal{C}^{\mathrm{ad}}$) are carried separately and shown not to affect the rate.

The λ-policy and its per-query asymptotics. For $\lambda = e^{-\zeta}$, let the λ-policy π_λ commit the per-query pair

$$\pi_\lambda(q) := \begin{cases} (\tau^\star(q),\, n_\lambda(q)), & q \in \mathcal{C}^{\mathrm{ad}},\ p(a^* \mid q) < 1 \quad \text{(Proposition D.1)}, \\ (0,\, 1), & q \in \mathcal{C}^{\mathrm{ad}},\ p(a^* \mid q) = 1, \\ (\widehat{\tau}_q,\, \widehat{n}_q), & q \notin \mathcal{C}^{\mathrm{ad}} \quad \text{(clause (L3))}. \end{cases} \tag{75}$$

The first case uses the policy of Proposition D.1, and the third uses the finite-count action in (L3). In particular, $n_\lambda(q)$ is the integer count defined in that proposition, not a rounded version of its leading-order approximation. Multiplying the balance count (64) by $I^{\mathrm{ad}}(q)$ and by $\log M(q)$ separates the two exponents,

$$n_\lambda(q)\, I^{\mathrm{ad}}(q) = \alpha^{\mathrm{ad}}(q)\zeta(1+o(1)), \qquad n_\lambda(q)\log M(q) = (1-\alpha^{\mathrm{ad}}(q))\zeta(1+o(1)),$$

so that, by the rate Lemma D.4 in the q-uniform form of (L4)(b), every non-perfect consistent query obeys

$$1 - V_{\tau^\star(q)}(n_\lambda(q); q) = e^{-\alpha^{\mathrm{ad}}(q)\zeta + \eta_q(\zeta)\zeta}, \qquad M(q)^{n_\lambda(q)} = e^{(1-\alpha^{\mathrm{ad}}(q))\zeta + \eta'_q(\zeta)\zeta}, \tag{76}$$

with $|\eta_q(\zeta)|, |\eta'_q(\zeta)| \le \epsilon_\zeta \to 0$ uniformly over $\{q \in \mathcal{C}^{\mathrm{ad}} : p(a^* \mid q) < 1\}$; a perfect query contributes zero reward gap and cost $M(q) \le m_+ = e^{o(\zeta)}$.

This uniformity is what licenses every Laplace step below. For any bounded exponent $X_q \in [0,1]$ and any error $|\eta_q(\zeta)| \le \epsilon_\zeta$, the sandwich

$$e^{X_q\zeta - \epsilon_\zeta\zeta} \;\le\; e^{X_q\zeta + \eta_q(\zeta)\zeta} \;\le\; e^{X_q\zeta + \epsilon_\zeta\zeta}$$

shifts $\frac{1}{\zeta}\log$ of any average by at most $\epsilon_\zeta \to 0$, so Lemma D.12 applies as if the exponents were exact; the additive form of the error is all the argument uses.

Step 1 (cost and reward of the λ-policy). Split the cost over $\mathcal{C}^{\mathrm{ad}}$ and its complement. The complement contributes at most the constant K^- by (L3). On $\mathcal{C}^{\mathrm{ad}}$ the exponent $1 - \alpha^{\mathrm{ad}}(q) \in [0,1]$ is bounded, so Lemma D.12 applied to the conditional law $\mathcal{Q}(\cdot \mid \mathcal{C}^{\mathrm{ad}})$ with $X = 1 - \alpha^{\mathrm{ad}}$ gives

$$\frac{1}{\zeta}\log \mathbb{E}_{q\in\mathcal{C}^{\mathrm{ad}}}\big[M(q)^{n_\lambda(q)}\big] \;\longrightarrow\; \operatorname*{ess\,sup}_{q\in\mathcal{C}^{\mathrm{ad}}}\big(1-\alpha^{\mathrm{ad}}(q)\big) = 1 - \alpha^{\mathrm{ad}}_{\min} > 0,$$

the restriction and the conditioning differing only by the constant mass factor $\mathbb{P}_q[\mathcal{C}^{\mathrm{ad}}] > 0$, immaterial on this scale. The $\mathcal{C}^{\mathrm{ad}}$ term grows exponentially while the complement adds a constant, so the total cost is

$$C(\pi_\lambda) = e^{(1-\alpha^{\mathrm{ad}}_{\min})\zeta + o(\zeta)}. \tag{77}$$

On the reward side, π_λ attains the ceiling exactly on $q \notin \mathcal{C}^{\mathrm{ad}}$ and on perfect queries, so the shortfall is supported on the non-perfect consistent queries, where the summand is $e^{-\alpha^{\mathrm{ad}}(q)\zeta + \eta_q(\zeta)\zeta}$ by (76). Applying Lemma D.12 to $-X$ with $X = \alpha^{\mathrm{ad}}$, whose essential infimum over that set is $\alpha^{\mathrm{ad}}_{\min}$ by (73), gives

$$R^{\infty,\mathrm{ad}} - R(\pi_\lambda) = \mathbb{E}_{q\in\mathcal{C}^{\mathrm{ad}}}\big[1 - V_{\tau^\star(q)}(n_\lambda(q); q)\big] = e^{-\alpha^{\mathrm{ad}}_{\min}\zeta(1+o(1))}. \tag{78}$$

Step 2 (primal direction: the gap is at most the power law). Fix $\epsilon \in (0,1)$ and run the λ-policy at

$$\zeta_B := \frac{(1-\epsilon)\log B}{1 - \alpha^{\mathrm{ad}}_{\min}}.$$

By the cost (77), $C(\pi_\lambda) = B^{(1-\epsilon)(1+o(1))} \le B$ for all large B, so π_λ is feasible and $\overline{R}^{\mathrm{ad}}(B) \ge R(\pi_\lambda)$. The gap (78) at ζ_B is then

$$R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) \;\le\; e^{-\alpha^{\mathrm{ad}}_{\min}\zeta_B(1+o(1))} \;=\; B^{-(1-\epsilon)\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min})\,(1+o(1))}.$$

Since $\epsilon \in (0,1)$ was arbitrary,

$$\limsup_{B\to\infty} \frac{\log(R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B))}{\log B} \;\le\; -\frac{\alpha^{\mathrm{ad}}_{\min}}{1 - \alpha^{\mathrm{ad}}_{\min}}. \tag{79}$$

Step 3 (dual direction: the gap is at least the power law). The matching lower bound comes from weak duality, which we record directly. For any feasible π (so $C(\pi) \le B$) and any $\lambda \ge 0$, the per-query Lagrangian L_q of (48) averages to $\mathbb{E}_q[L_q(\pi_q, \lambda)] = R(\pi) - \lambda C(\pi)$, hence

$$R(\pi) \;=\; \mathbb{E}_q[L_q(\pi_q,\lambda)] + \lambda C(\pi) \;\le\; \mathbb{E}_q\Big[\sup_{\pi_q\in\Pi^{\mathrm{ad}}} L_q(\pi_q,\lambda)\Big] + \lambda B, \tag{80}$$

the last step using $\lambda \ge 0$ and $C(\pi) \le B$. The right-hand side of (80) does not depend on π, so taking the supremum over all feasible π on the left, and recalling $\overline{R}^{\mathrm{ad}}(B) = \sup\{R(\pi) : C(\pi) \le B\}$, gives for every $\lambda \ge 0$ the dual bound

$$\overline{R}^{\mathrm{ad}}(B) \;\le\; \mathbb{E}_q\Big[\sup_{\pi_q\in\Pi^{\mathrm{ad}}} L_q(\pi_q,\lambda)\Big] + \lambda B. \tag{81}$$

We identify the per-query optimum in (81) at $\lambda = e^{-\zeta}$, region by region. Writing $L_q(\pi_q, \lambda) = V_\tau(n; q) - \lambda M(q)^n$ for the pair (τ, n) that defines π_q, the supremum is $\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = \sup_{n,\tau} \big(V_\tau(n; q) - \lambda M(q)^n\big)$. On the non-perfect consistent set, clause (L4)(a) evaluates it to $1 - e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+u_q(\zeta))}$ with $|u_q(\zeta)| \le \epsilon_\zeta$. On a perfect query a single rollout already reaches the ceiling, so the supremum is attained at $n = 1$ and equals $1 - \lambda M(q)$; using $\alpha^{\mathrm{ad}}(q) = 1$ and $0 < \log M(q) \le \log m_+$,

$$1 - \sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) = \lambda M(q) = e^{-\zeta(1 - \log M(q)/\zeta)} = e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))},$$

the same leading exponent as the non-perfect case at $\alpha^{\mathrm{ad}}(q) = 1$. On $q \notin \mathcal{C}^{\mathrm{ad}}$ the priced cost $\lambda M(q)^n \ge 0$ makes $\sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda) \le \sup_{n,\tau} V_\tau(n; q)$.

Because $R^{\infty,\mathrm{ad}} = \mathbb{E}_q[\sup_{n,\tau} V_\tau(n; q)]$ and the leading term of (81) are the same $\mathcal{Q}$-average over q, subtracting (81) from $R^{\infty,\mathrm{ad}}$ combines them into one expectation,

$$R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) \;\ge\; \mathbb{E}_q\Big[\sup_{n,\tau} V_\tau(n; q) - \sup_{\pi_q \in \Pi^{\mathrm{ad}}} L_q(\pi_q, \lambda)\Big] - \lambda B, \tag{82}$$

whose integrand the three evaluations above make nonnegative everywhere: on the non-perfect consistent set $\sup_{n,\tau} V_\tau = 1$, so it equals $e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+u_q(\zeta))}$; on a perfect query it equals $\lambda M(q) = e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}$; and on $q \notin \mathcal{C}^{\mathrm{ad}}$ it is nonnegative. Discarding the nonnegative inconsistent part $q \notin \mathcal{C}^{\mathrm{ad}}$ preserves the inequality and leaves the average over $\mathcal{C}^{\mathrm{ad}}$, on which the integrand is $e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}$ with relative error at most $\max(\epsilon_\zeta, \log m_+/\zeta) \to 0$ uniformly in q. Lemma D.12 then evaluates this average exactly as it did the primal gap (78) in Step 1: applied to $-X$ with $X = \alpha^{\mathrm{ad}}$, whose essential infimum over $\mathcal{C}^{\mathrm{ad}}$ is $\alpha^{\mathrm{ad}}_{\min}$ by (73), it gives

$$R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) \;\ge\; \mathbb{E}_{q \in \mathcal{C}^{\mathrm{ad}}}\big[e^{-\alpha^{\mathrm{ad}}(q)\zeta(1+o(1))}\big] - \lambda B \;=\; e^{-\alpha^{\mathrm{ad}}_{\min}\zeta(1+o(1))} - e^{-\zeta}B. \tag{83}$$

Choose

$$\zeta := \frac{(1+\epsilon)\log B}{1 - \alpha^{\mathrm{ad}}_{\min}}, \qquad \text{so that} \qquad e^{-\zeta}B = B^{-(\alpha^{\mathrm{ad}}_{\min}+\epsilon)/(1-\alpha^{\mathrm{ad}}_{\min})},$$

while the first term of (83) is $B^{-(1+\epsilon)\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min})\,(1+o(1))}$. Since $\alpha^{\mathrm{ad}}_{\min} < 1$ gives $(1+\epsilon)\alpha^{\mathrm{ad}}_{\min} < \alpha^{\mathrm{ad}}_{\min} + \epsilon$, the first term dominates, so for all large B

$$R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B) \;\ge\; \tfrac{1}{2}\, B^{-(1+\epsilon)\,\alpha^{\mathrm{ad}}_{\min}/(1-\alpha^{\mathrm{ad}}_{\min})\,(1+o(1))}.$$

Letting $\epsilon \downarrow 0$,

$$\liminf_{B\to\infty} \frac{\log(R^{\infty,\mathrm{ad}} - \overline{R}^{\mathrm{ad}}(B))}{\log B} \;\ge\; -\frac{\alpha^{\mathrm{ad}}_{\min}}{1 - \alpha^{\mathrm{ad}}_{\min}}. \tag{84}$$

The matching bounds (79) and (84) are (74). ■

D.4.4 An asymptotically optimal policy at large budgets

Theorem 3 gives the large-budget asymptotics of the adaptive frontier, and we now give an explicit policy that attains it. At finite budgets, the optimal policy (9) is only implicitly defined. As $b \to \infty$, however, the explicit policy below is budget-feasible and closes the gap to the ceiling at the optimal rate of Theorem 3, so it is optimal to leading exponential order.

Proposition D.2 (An asymptotically optimal adaptive policy)**.** *Under the assumptions of Theorem 3, let $\pi^\dagger_b$ be the adaptive policy that, for each query, commits the following pair.*

- *For a solvable query $q \in \mathcal{C}^{\mathrm{ad}}$ with $p(a^* \mid q) < 1$,*

$$\tau^\dagger(q) \in \arg\max_{\tau:\, q \in \mathcal{C}(\tau)} I_\tau(q), \qquad n^\dagger_b(q) = \frac{(\gamma + r^{\mathrm{ad}})\, b}{I^{\mathrm{ad}}(q) + \log M(q)}\,(1 + o(1)). \tag{85}$$

- *For a perfect query $q \in \mathcal{C}^{\mathrm{ad}}$ with $p(a^* \mid q) = 1$, any temperature and $n^\dagger_b(q) = 1$.*

- *For an unsolvable query $q \notin \mathcal{C}^{\mathrm{ad}}$, the finite-count action that attains its ceiling $\sup_{\tau,n} V_\tau(n;q)$, which does not depend on the budget.*

Then, as $b \to \infty$, $\pi_b^\dagger$ is budget-feasible and attains the law of Theorem 3,

$$R^{\infty,\mathrm{ad}} - R(\pi_b^\dagger) = e^{-\,r^{\mathrm{ad}}\, b\,(1+o(1))}. \tag{86}$$

Note that the maximizer of (9) at a finite budget need not coincide with $(\tau^\dagger, n_b^\dagger)$ defined in the above proposition. For the solvable queries, the decisions can separate at large budgets. The temperature $\tau^\dagger(q)$ depends only on the query, specifically the one at which its error decays fastest, which is independent of the budget. In contrast, the rollout count $n_b^\dagger(q)$ grows in proportion to the budget b, with the rate depending on the query's error decay rate $I^{\mathrm{ad}}(q)$ and token cost $\log M(q)$.

Justification of Proposition D.2. We justify the budget-indexed allocation (85) and the gap law (86) of Proposition D.2, whose policy $\pi_b^\dagger$ is the λ_b-policy defined in (75). Set $\kappa := 1 - \alpha_{\min}^{\mathrm{ad}} > 0$ and $t_b := \gamma b/\kappa$. Equation (77) gives $\log C(\pi_{e^{-\zeta}}) = \kappa\zeta + o(\zeta)$. Consequently, there is a deterministic sequence $\epsilon_b \to 0$ such that the choice

$$\zeta_b = (1-\epsilon_b)t_b, \qquad \lambda_b = e^{-\zeta_b}$$

is feasible for every sufficiently large b. To see this without assuming cost monotonicity, let

$$d_b := \sup_{\zeta \ge t_b/2} \left| \frac{\log C(\pi_{e^{-\zeta}})}{\kappa\zeta} - 1 \right| \longrightarrow 0$$

and take $\epsilon_b = 2d_b + t_b^{-1/2}$ for large b. Then $\epsilon_b < 1/2$, $\zeta_b \ge t_b/2$, and

$$\log C(\pi_{\lambda_b}) \le \kappa\zeta_b(1+d_b) \le \kappa t_b = \gamma b.$$

The subleading reduction in ζ_b ensures feasibility while preserving its leading order. Since $\gamma/\kappa = \gamma + r^{\mathrm{ad}}$, substituting $\zeta_b = (\gamma + r^{\mathrm{ad}})b(1+o(1))$ into (64) gives (85). With the perfect and residual queries handled as in (75), equation (78) yields

$$R^{\infty,\mathrm{ad}} - R(\pi_{\lambda_b}) = e^{-\alpha_{\min}^{\mathrm{ad}}\zeta_b(1+o(1))} = e^{-r^{\mathrm{ad}}b(1+o(1))}.$$

Thus this family of feasible policies attains the same leading exponential rate as the optimal frontier. Here $n_{\lambda_b}(q)$ is the integer minimizer at the selected rate-maximizing temperature, as defined in Lemma D.10. Thus no separate rounding rule is needed. The leading count formula alone does not certify finite-budget feasibility; the price choice above specifies the required subleading adjustment.