THINK MORE OR READ MORE? A TOKEN-BUDGET THEORY OF REASONING WITH SHARED CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning systems increasingly improve answers by incorporating additional context into the model's input, from multi-agent debate, where agents read their peers' responses, to self-refinement, where a model revisits its own drafts. In these systems, communication consumes context as one agent's output becomes another's input. We study the regime where token reading and generation draw on one fixed token window, so every token spent reading is a token that cannot be spent reasoning. We develop a **token-budget theory** that formalizes this trade-off, modeling accuracy as a function of the reasoning budget remaining after reading and aggregating votes with a Condorcet jury model that permits correlated errors. For multi-agent debate, the theory gives a closed-form prediction for the optimal number of agents. This quantity is estimated from single-agent measurements, allowing the optimal group size to be determined without running debate, and the same accounting extends to a heterogeneous set of different models. We empirically verify our theory and show that, at the predicted group size, our method outperforms strong baselines under the tightest calibration spend, where debate-calibrated rules can afford only one or two questions. Self-refinement obeys the same accounting, with its own draft in place of a peer message. The theory further predicts that communication pays only when the gain in the accuracy ceiling exceeds the reasoning room spent on reading, with this condition tightening as the window shrinks. Across five open models and four reasoning tasks, self-consistency matches or beats deployed debate at 93% of 980 matched window-and-group-size settings, with the gap being largest under tight windows and nearly disappearing at a loose one, consistent with the theory. Under a shared window, how much room a model has left to *think matters more than what it reads*.

1 INTRODUCTION

Large language models (LLMs) are increasingly used in reasoning pipelines that reuse model-generated text as input to subsequent computation. Multi-agent debate exchanges answers among agents (Du et al., 2024; Liang et al., 2024), while self-refinement feeds a model its own previous drafts (Madaan et al., 2023). Such reuse can improve reasoning by providing additional information, but under a fixed token budget, it also raises a fundamental question: not *whether* added feedback helps, but *how much* it is worth in terms of the reasoning budget it consumes.

Majority-voting analyses predict that adding agents amplifies a correct majority until correlated errors cap the gain (Ladha, 1992; Liu et al., 2026). However, on multi-hop reasoning tasks, a single long chain-of-thought (Wei et al., 2022) often matches or outperforms a multi-agent system under an equal thinking budget (Tran & Kiela, 2026). Debate also does not consistently outperform self-consistency on an accuracy-per-token-cost frontier (Smit et al., 2024; Bertalanic & Fortuna, 2026), echoing the broader trade-off between token expenditure and accuracy in sampling and search (Snell et al., 2025; Brown et al., 2024). The token budget deserves separate attention because it corresponds directly to the cost incurred in practice. At current API prices, inference can range from cents to tens of dollars per million tokens, depending on the model and provider (OpenAI, 2026; Google, 2026; Anthropic, 2026). For multi-agent debate, the cost grows further with the number of agents n (group size): adding agents increases not only the amount of reasoning generated, but also the amount of peer text that must be read and processed. Existing work attacks this cost either by compressing agent communication or by fixing the reasoning budget, but treats communication and reasoning

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

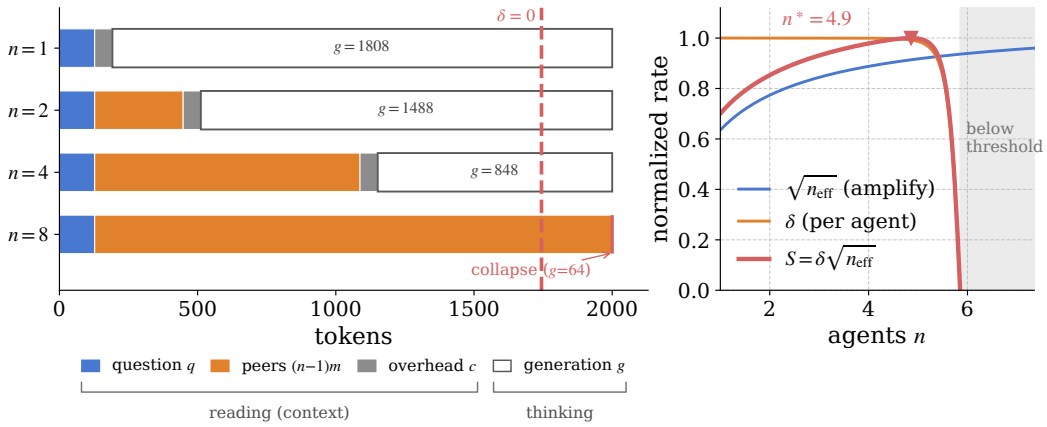


Figure 1: Overview of our method (Qwen2.5-7B, GSM8K). One window W pays for both reading (question plus peer messages) and thinking (generation g), and each extra agent adds m tokens of reading. The per-agent edge δ is the gap between the probability of the correct answer and that of the strongest wrong answer: larger δ means a stronger correct-answer margin, while $\delta \leq 0$ means that a wrong answer is at least as likely. **Left**: The dashed line marks the smallest thinking room at which $\delta > 0$; a group that reads past it (here n=8) leaves too few tokens for reasoning and its accuracy collapses. **Right**: The effective sample size n_{eff} measures the independent-vote equivalent of n correlated agents, while δ falls as thinking room shrinks. Their product S, the synergy index, therefore captures the trade-off between per-agent accuracy and voting amplification and peaks at $n^* = 4.9$. The shaded band marks $\delta \leq 0$, where votes amplify a wrong majority and extra agents cannot help.

as separate costs, despite both consuming the same token budget (§2). Once the token budget is viewed this way, a trade-off appears: with n small, more of the budget remains for thinking, while with n large, most of it is spent on reading peer messages, leaving little room to reason. The question is then how to estimate the best group size n^* for multi-agent debate under this trade-off. Existing approaches measure n by sweeping or adaptively searching the design space, for instance, successive halving (Jamieson & Talwalkar, 2016). Both require debate runs to identify the operating point, without explaining *why* one group size is optimal or *when* adding more agents stops improving accuracy.

We therefore introduce a **token-budget theory** that answers both questions in closed form. Consider homogeneous debate in which each agent, working under a window of W tokens, must first pay for what it reads, namely the question and its peers' messages, together H tokens, plus a small fixed overhead c. The remainder, $g = W - H - c$, is the room it has left to reason. Instead of assigning each agent a fixed chance of voting correctly, we introduce its *ability*, which is the accuracy it reaches when its reasoning is capped at g tokens. Ability grows with thinking room, too little room can prevent an agent from reaching an answer, while beyond a sufficient room, additional tokens yield little further gain. We let it grow with g (Eq. 3) via the fittable thinking-gain function ϕ, whose functional form and parameters can be calibrated from held-out calibration data. We then combine the resulting votes with a Condorcet jury model that allows correlated errors, developing the per-answer margin δ (Eq. 6), the group accuracy $A(n)$, the effective sample size n_{eff} and the synergy index $S(n) = \delta\sqrt{n_{\text{eff}}}$ (Eq. 10) in Section 3. Figure 1 provides an overview of our method.

Our model yields a closed-form prediction of an **interior optimal group size** n^* that *moves with the budget*. It is read off single-agent measurements: we measure the ability curve with a cap sweep and estimate one error-correlation parameter at the *initial round*, then predict group size without running the debate. With a sufficiently large budget, our predictions approach the classical correlation-limited ensemble, where the main benefit comes from aggregating independent evidence. The practical advantage of our approach is therefore greatest in the tight-budget regime, where choosing the group size correctly matters most and our method achieves the best accuracy-budget trade-off. At the predicted size, it incurs only a 1.4% accuracy gap from a per-cell oracle, using between 26 and 79 calibration questions depending on the cell, where the debate-calibrated rules

use only one or two at matched spend. The accounting extends to a heterogeneous model set, where the rule chooses which models to include as well as how many.

Our framework also clarifies the cost of communication relative to self-consistency (Wang et al., 2023). Self-consistency leaves the full window available for reasoning, whereas debate spends part of that budget processing peer responses, so communication must provide sufficient informational benefit to justify this cost. In our experiments, self-consistency matches or beats deployed debate at 93% of the 980 matched operating points, while a paired test finds debate significantly ahead at only 0.5%. The gap is largest under tight budgets and nearly disappears at a loose budget, where communication consumes a smaller share of the available reasoning budget. Self-refinement, which rereads only its own last draft without voting, pays the same fixed reading cost every round with none of the amplification that voting provides, so its accuracy collapses faster than debate's as rounds accumulate.

Our contributions are as follows. **First**, we introduce a **token-budget theory** that predicts the **optimal debate size** n^* in closed form and extends to heterogeneous model sets, outperforming strong baselines under the tightest calibration spend without running a debate. **Second**, the theory reveals when **debate's voting benefits outweigh its communication cost** relative to self-consistency under the same token budget. **Third**, ablations show that **reading cost** drives collapse and the interior optimum, with robustness tests covering longer reasoning and practical token pricing.

2 Related Work

Multi-agent debate in LLMs. LLM instances critique and revise each other's answers to improve reasoning relative to a single pass (Du et al., 2024; Liang et al., 2024), and voting over more agents can improve accuracy in practice (Wang et al., 2023; Li et al., 2024a). Each additional agent, however, also introduces peer messages that other agents must process, increasing the communication burden of the group. Existing work addresses this burden by reducing communication through sparse speaker selection, group partitioning, or retaining only conflicting messages (Li et al., 2024b; Liu et al., 2024b; Zeng et al., 2025; Nguyen et al., 2026). A complementary line studies whether debate is worthwhile under a matched sampling or thinking budget, with several works finding that self-correction or a single long reasoning chain can match or outperform debate (Huang et al., 2024; Tran & Kiela, 2026; Bertalanic & Fortuna, 2026). A third line promotes diversity through heterogeneous model sets, arguing that homogeneous evaluation overvalues debate (Zhang et al., 2025), attributing vanilla debate's weakness to lost candidate diversity (Zhu et al., 2026), and showing that mixed panels revise less harmfully than homogeneous ones (Nilayam et al., 2026). Beyond debate, this idea appears in mixtures of agents (Wang et al., 2025), model routing (Ong et al., 2025), and cost-aware cascades (Chen et al., 2023).

Estimating the operating point. Existing approaches search the design space, e.g., via successive halving (Jamieson & Talwalkar, 2016), or learn the operating point from data, but both require debate runs for calibration. Chen et al. (2024) make voting accuracy non-monotone in call count through query-difficulty diversity, but their optimum is refit separately at each budget, whereas ours transfers from a single-agent cap sweep. Liu et al. (2026) give the closest theoretical treatment, deriving an amplification-collapse threshold in closed form from a power-law ability curve. Their ability curve is calibrated in simulation and remains concave, so it cannot capture the near-zero accuracy below the minimum length needed for a solution, the regime where our collapse occurs. Their accounting also holds the per-agent generation allowance fixed, so adding an agent costs no reasoning room, making their collapse a consequence of correlation alone rather than reading. We recover their setting as the $m = 0$ case (§ 4.3).

Token Budgeting in LLMs. A separate line studies how to allocate a fixed token budget across samples or reasoning length for a single agent, including token budgets for chain-of-thought, budget forcing at decoding time, budget-aligned tree search, and the non-monotone effect of reasoning length on task success (Han et al., 2025; Muennighoff et al., 2025; Miyamoto et al., 2026; Qi, 2026). Controlled comparisons at matched sampling or search budgets find that voting, rather than talking, carries the gain (Choi et al., 2025; Kaesberg et al., 2025; Wang et al., 2024a), while Wunderlich et al. (2026) report the opposite ordering under per-agent rather than shared contexts. These methods budget within reasoning or sampling, leaving peer text outside the account.

3 METHOD

We formalize the token-budget problem behind debate, define per-agent ability as a function of the room left to think, and turn it into a group-accuracy prediction. We call the resulting predictor **TB-MAD**, for token-budget multi-agent debate: it reads a single-agent probe sweep and returns the group size, with no debate run.

3.1 PROBLEM FORMULATION

Consider n LLM agents that debate a question from a task Q over several rounds under one token budget W, the *window*. Reading and thinking draw on the same W, where tokens spent reading peer messages are tokens no longer available for the agent's own chain of thought, and vice versa. Each peer message costs about m tokens on average to read once. Under last-round visibility, which we use throughout, agent i retains the question (q tokens), a fixed overhead (c), and $n-1$ peer messages, so its retained history and remaining room to think are

$$H \;=\; q + (n-1)\,m, \qquad\qquad g \;=\; W - H - c. \tag{1}$$

Retaining the full transcript would increase H at every n and can only decrease g, so the penalty we measure is a lower bound. Self-refinement uses a different accounting. The group size is one, so there are no peer messages, but the agent retains its own last draft. Its history is $H = q + m$ and its room is $g = W - q - m - c$. In contrast, Eq. 1 with $n = 1$ describes a debate agent with no peer messages, corresponding to self-consistency.

Adding an agent raises H and lowers g by m tokens, which cuts both ways. When W is generous, g stays past the point where extra tokens stop buying accuracy and a larger group is close to free. When W is tight, the agent pushes g below the room a chain needs to reach an answer, and the lost generation budget outweighs the extra vote. We call this *collapse*: under a shared budget, more agents can reduce accuracy. Given W, we want the group size that maximizes task accuracy,

$$\max_{n}\; A_Q(n, W) \quad \text{s.t.} \quad g(n) = W - H(n) - c \geq 0, \tag{2}$$

where $A_Q(n, W)$ is the debate group's accuracy at size n and window W. Since a debate agent's vote depends on how much of the window it got to reason in, solving Eq. 2 needs a model of per-agent ability as a function of g, which we give next.

3.2 PER-AGENT ABILITY AND THE THINKING GAIN

Rather than assuming fixed voting accuracy, we let that chance depend on what the window leaves an agent to think. The *per-agent ability* $\beta_{\text{eff}}(g)$ is the accuracy of a single agent whose generation is capped at g tokens, measured directly with a cap sweep and no debate. We drop the Q subscript below, since every quantity is on the same task.

$$\beta_{\text{eff}}(g) \;=\; p_{\max} \quad \underbrace{\phi(g)}_{\text{thinking gain}}, \qquad \phi(g) \;=\; \beta_{\text{eff}}(g)/p_{\max} \;\in\; [0, 1], \tag{3}$$

where $p_{\max} = \lim_{g\to\infty} \beta_{\text{eff}}(g)$ is the ceiling ability with unlimited room, so the *thinking gain* ϕ is the fraction of its own ceiling an agent reaches at cap g.

Assumption 1 (Admissible gain)**.** *$\phi : [0, \infty) \to [0, 1]$ is continuously differentiable, strictly increasing ($\phi' > 0$) on $(0, \infty)$, and saturating ($\phi(\infty)=1$), and satisfies $\phi(0)=0$.*

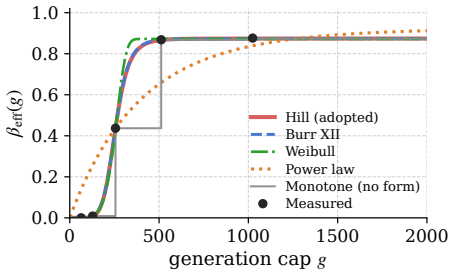
More room cannot hurt, since a longer allowance can always reproduce a shorter chain, and it eventually stops helping. This is consistent with the diminishing return in test-time-compute scaling literature (Snell et al., 2025).

Choice of ϕ. The cap sweeps are *sigmoidal*, near zero below the length a solution needs, rising sharply around it, then saturating. Held-out error selects the sigmoid *family* over the concave power law prior work adopts (Liu et al., 2026). Isotonic regression extrapolates competitively out of range but is far worse under leave-one-out. Within the sigmoid family, the HILL form has the lowest total error (Appendix Table 9). We adopt it,

$$\beta_{\text{eff}}(g) \;=\; p_{\max}\,\frac{g^{\kappa}}{g^{\kappa} + g_0^{\kappa}}, \tag{4}$$

Table 1: Candidate forms for the thinking gain ϕ, adopted form in bold. *Right:* all five fitted to one measured cap sweep (Qwen2.5-7B/GSM8K).

Form	$\phi(g)$
Hill (Hill, 1910)	$\frac{g^\kappa}{g^\kappa + g_0^\kappa}$
Burr XII (Burr, 1942)	$1 - \left(1 + \left(\frac{g}{g_0}\right)^\kappa\right)^{-\lambda}$
Weibull (Weibull, 1951)	$1 - e^{-\left(\frac{g}{g_0}\right)^\kappa}$
Power law (Pareto, 1964)	$1 - (1 + \frac{g}{g_0})^{-\kappa}$
Monotone (Robertson et al., 1988)	isotonic step



with scale g_0 and shape $\kappa > 0$, fitted on a separate calibration set. Under this form, $\phi(g)$ could be interpreted as the probability that a solution fits within a budget of g tokens, making g_0 a median solution length (Appendix B.1).

3.3 Group Accuracy and the Synergy Index

To turn per-agent ability into group accuracy, we use a correlation-aware Condorcet model (de Condorcet, 1785; Boland, 1989; Ladha, 1992) that captures how additional agents help until correlated errors diminish the gain. The following three simplifications make it tractable. We read β_{eff} as a per-question probability where Eq. 4 defines a task average (Chen et al., 2024), and treat the agents as *homogeneous* so all share $\beta_{\text{eff}}(g(n))$. Moreover, we assume their errors are *equicorrelated*, with every pair sharing a ρ independent of n, estimated at round 0 before any agent reads a peer and so measurable without running a debate (Appendix B.1). ρ plays the role that diversity measures play in the ensemble literature (Krogh & Vedelsby, 1994; Kuncheva & Whitaker, 2003), capturing a similar trade-off between individual accuracy and error independence.

For a given question, let the correct answer's margin over the strongest wrong answer be

$$\delta(n) = \beta_{\text{eff}}\big(g(n)\big) - q_{\max}(n), \qquad q_{\max} = (1 - \beta_{\text{eff}})\, c_w, \tag{5}$$

where $q_{\max}$ is the highest-probability wrong answer and c_w is the *error concentration*, the probability of producing the strongest wrong answer conditional on being wrong. On a K-way task, c_w is bounded between $1/(K-1)$ and 1, with the lower bound attained when errors are spread uniformly over the wrong options and the upper bound when every error picks the same wrong option. We hold c_w fixed across n and g, so all budget dependence enters through β_{eff}. Thus

$$\delta = (1 + c_w)\beta_{\text{eff}} - c_w, \tag{6}$$

which is positive exactly when $\beta_{\text{eff}} > \tau \equiv c_w/(1 + c_w)$. Across answers, c_w is analogous to ρ across agents: ρ captures dependence in whether agents are correct, while c_w captures dependence in which wrong answer they choose when they are incorrect. Both are measured at round 0 and held fixed and debiased for small samples (Appendix D.4).

Score agent j by $X_j = +1$ when it answers correctly, $X_j = -1$ when it gives the strongest wrong answer, and $X_j = 0$ otherwise, so the group prefers the correct answer exactly when $\bar{X}_n > 0$, with $\mathbb{E}[X_j] = \delta$ by homogeneity. Write $\sigma^2 = \text{Var}(X_j)$ for this per-agent variance, common to all j by homogeneity (Appendix D.5 gives its closed form in β_{eff} and c_w). Equicorrelation with $\rho \in (0, 1)$ gives $\text{Cov}(X_i, X_j) = \rho\sigma^2$ and hence

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2\big(1 + \rho(n-1)\big)}{n} = \frac{\sigma^2}{n_{\text{eff}}(n)}, \tag{7}$$

$$n_{\text{eff}}(n) = \frac{n}{1 + \rho(n-1)} \xrightarrow[n\to\infty]{} \frac{1}{\rho}. \tag{8}$$

Thus n correlated agents provide the effective sample size n_{eff}, which is bounded by $1/\rho$ and yields diminishing returns (Kish, 1965). Approximating $\bar{X}_n$ by a normal with the same moments,

$$A(n) = \Pr[\bar{X}_n > 0] \simeq \Phi\left(\frac{\delta(n)\sqrt{n_{\text{eff}}(n)}}{\sigma}\right), \tag{9}$$

where Φ is the standard normal CDF and $\delta(n)$ carries $g(n)$ from Eq. 1. Eq. 9 is an *aggregation head*: given a per-agent ability, it maps that ability to the accuracy of the resulting vote. For debate, feeding it the pre-revision ability $\beta_{\text{eff}}(g(n))$ gives the initial-round vote accuracy $A(n)$. After revision, the same head applies to the agents' updated ability, so revision changes the ability entering the head instead of the aggregation mechanism itself. We therefore use the pre-revision head to determine the optimal operating point n^*, while treating revision as a separate accuracy lift whose effect is evaluated on the deployed debate.

Because Φ is monotone and $\sigma > 0$, we can maximize the signed *synergy index*

$$\arg\max_n A(n) = \arg\max_n S(n), \qquad S(n) = \delta(n)\sqrt{n_{\text{eff}}(n)}. \tag{10}$$

$\delta(n)$ is the per-agent edge the budget supplies; $\sqrt{n_{\text{eff}}(n)}$ the amplification voting supplies. Keeping the sign of δ matters for the objective's behaviour just outside $\mathcal{F}$. Votes amplify a negative edge as readily as a positive one, so S falls once δ turns negative and the rule returns $n = 1$, where an unsigned measure would keep climbing. Since n^* is read off $S(n)$ rather than accuracy itself, it does not depend on Φ being Gaussian: any strictly increasing map from margin to accuracy yields the maximizer. As n grows the amplification rises while $\delta(n)$ falls, because each agent costs m tokens of room and so lowers β_{eff}, producing the interior optimum. Eq. 10 treats σ as a constant fit once per cell; once the vote score is fixed, σ is instead a function of n through $\beta_{\text{eff}}(n)$, and substituting that derived $\sigma(n)$ moves $\hat{n}^*$ at only 7 of 140 swept operating points (Appendix D.5). We summarize the procedure in Appendix Alg. 1.

Self-refinement as the degenerate case. Self-refinement reads its own previous draft, corresponding to $n{=}1$ while still paying the reading cost. Hence $g_{\text{ref}}(W) = W - q - m - c$, reducing δ through the generation-floor effect, while $\sqrt{n_{\text{eff}}} = 1$ provides no amplification. Repeating rounds therefore does not improve S unless revision increases the ability of subsequent generations.

3.4 Theory: the collapse regime

We now analyze what happens when ability is tied to the window, using a single model in which reading consumes reasoning budget and yields the following results.

The interior best size. Throughout, n is continuous, $g(n) = W - q - c - (n-1)m$, and we restrict to the feasible region $\mathcal{F} = \{n : \delta(n) \geq 0\}$. The constraint is set by δ, not g: $\delta(n)$ reaches zero strictly before $g(n)$ does, so $\mathcal{F} = n : \delta(n) \geq 0 \subset n : g(n) \geq 0$, while $\phi(g) \to 0$ continuously as $g \to 0$.

Proposition 1 (Single-peaked accuracy)**.** *Let ϕ satisfy Assumption 1, with $m > 0$, $p_{\max} \in (0, 1]$, $\rho \in [0, 1)$, $\delta(1) > 0$, and suppose some larger group helps ($S(n) > S(1)$ for some $n \in \text{int}(\mathcal{F})$). Then δ is strictly decreasing on $\mathcal{F} = [1, n_0]$, where $n_0 > 1$ is the unique group size at which the average agent's edge over the strongest distractor vanishes ($\delta(n_0){=}0$), and $S(n)^2$ attains an interior maximum $n^* \in (1, n_0)$.*

Proof. See Appendix A.1. □

Every form in Table 1 has $\phi(0){=}0$, so $\delta < 0$ once the reasoning budget is exhausted at $n_{\text{exh}} = 1 + (W - q - c)/m$ (distinct from the deployment cap $n_{\max}$ used elsewhere); hence n_0 exists and $\mathcal{F}$ lies strictly inside the domain of ϕ. The mechanism is a race between two rates, one shrinking like $1/n^2$ as an extra agent *adds* amplification, the other growing without bound as it *costs* per-agent edge with $\delta \to 0^+$ near n_0. Continuity on a compact interval and a vanishing edge at n_0 give existence; uniqueness follows from log-concavity of the shifted edge δ(Appendix A.1). The $1/n^2$ rate limits how fast $n^*(W)$ can grow. Each extra agent costs m tokens of room while its amplification gain shrinks, keeping n^* below the feasibility cap $n_{\max}$ (Appendix Cor. 2). As $W \to \infty$, thinking room ceases to bind and the classical correlation ceiling returns, so an interior optimum arises only when the window is scarce (Appendix A.2).

Reading cannot beat voting at matched aggregation. Self-consistency draws n chains from one agent without exchanging messages, so each reads only the question ($g_{\text{SC}}{=}W - q - c$), while a

debate agent also retains $n-1$ peer messages and keeps only $g(n) \leq g_{\mathrm{SC}}$. Thus, debate's "synergy" is a *voting* effect unless peer text raises $\Delta p_{\max}$, which we measure in §4.2.

Proposition 2 (Self-consistency dominates the initial-round vote)**.** *Let $A^0_{\mathrm{deb}}(n, W)$ be initial-round debate vote accuracy of Eq. 9, before any revision. Under Assumption 1, at matched aggregation (same n, ρ, σ), $A_{\mathrm{SC}}(n, W) \geq A^0_{\mathrm{deb}}(n, W)$ for every $n \geq 1$ and W, with equality at n=1. If peer text raises the ceiling by $\Delta p_{\max}$, then debate can exceed self-consistency at (n, W) only when*

$$\Delta p_{\max}\, \phi\big(g(n)\big) \;>\; p_{\max}\big(\phi(g_{\mathrm{SC}}) - \phi(g(n))\big), \tag{11}$$

and the dominance result above is the special case $\Delta p_{\max} \leq 0$.

Proof. See Appendix A.3. □

Self-Refinement. For self-refinement (n=1 case), Prop. 2 becomes $A_{\mathrm{ref}}(W) \leq p_{\max}\phi(g_{\mathrm{SC}})$, a single untouched agent is the ceiling unless re-reading raises $p_{\max}$. §4.2 measures it.

4 Experiments

4.1 Setup

Models, Data And Metrics. We run five open instruction models across four families from 3–12B: **Llama-3.2-3B** (Llama Team, AI @ Meta, 2024), **Qwen2.5-7B** (Qwen Team, 2025), **Falcon3-7B** (Technology Innovation Institute, 2024), **Gemma-2-9B** (Gemma Team, 2024) and **Gemma-3-12B** (Gemma Team, 2025). We use four reasoning benchmarks, two difficulty levels in math and two non-math domains: **GSM8K** (Cobbe et al., 2021), **MATH500** (Lightman et al., 2024), ten-way **MMLU-Pro** (Wang et al., 2024b) and five-way **CSQA** (Talmor et al., 2019) (Appendix Table 3). For all tasks, we report accuracy in percentage.

Baselines. We compare our closed-form prediction against eight alternatives grouped by calibration cost. *Debate-calibrated*, using outcomes from actual debate runs that our method does not require: **Empirical Argmax** selects the best accuracy among tested (n, W) configurations, **Successive Halving** (Jamieson & Talwalkar, 2016) spends the same budget adaptively, and a **Quadratic Surface** in (n, W). *No calibration:* **naive** n=1 and **naive** n=$n_{\max}$. *Probe-calibrated, as ours is:* **Learned** β_{eff} replaces the functional form with gradient-boosted trees; **Star Exponent** (Liu et al., 2026) substitutes the power-law ability curve into our aggregator and budget model, which isolates the effect of the functional form; and **Difficulty Mixture** (Chen et al., 2024) estimates per-question accuracy from probe sweeps and simulates voting, with the per-call allowance adjusted under our window accounting as n varies. More details are in Appendix B.2.

Implementation Details. We report *regret* as the percentage-point accuracy gap between the selected group size and an oracle that chooses the highest-accuracy configuration for each model–dataset pair. We sweep the grid of Table 4, extending to $W \in \{16k, 32k\}$ and n up to 32 to test the large-W limit. For all baselines, the first 40% of test samples are used for calibration and the remaining 60% for evaluation, with calibration costs accounted for as described in Appendix B.3. All runs use vLLM (Kwon et al., 2023) at temperature 1.0 across three seeds, pooling samples across seeds before fitting rather than fitting each seed separately and averaging the estimates. This choice has negligible effect on the results (Appendix B.1). More details are in Appendix B.

4.2 Main Results

Fitting And Predicting Performance. $\beta_{\mathrm{eff}}(g)$ is measured by capping one agent's generation at g, then fitting $(p_{\max}, g_0, \kappa)$, with the HILL form chosen on held-out error from five candidate forms (Appendix Table 9). Using this fit, TB-MAD achieves the lowest regret, averaging 1.40% and staying below 2% across all models and eight comparisons at equal calibration spends (Table 2). The difference is largely driven by calibration cost: debate-calibrated rules afford far fewer questions, since debate expands input and output tokens by up to $49\times$ and $11\times$ (Appendix Table 11). Among probe-based methods, Star Exponent uses a power law that has higher held-out RMSE than the selected HILL form (Appendix Table 9), while Difficulty Mixture estimates a per-question difficulty

Table 2: Held-out regret against the per-cell oracle (percentage points) when each method chooses n at an equal *token* spend. **Bold** is the lowest regret in a column and underline the second lowest, the **Avg.** column included. Per-cell values for all calibration spend in Appendix Tables 24–25.

Method	1.0M budget					2.0M budget					Avg.
	L-3B	Q-7B	F-7B	G2-9B	G3-12B	L-3B	Q-7B	F-7B	G2-9B	G3-12B	
Debate-calibrated											
Successive Halving	7.13	12.05	9.18	5.63	11.39	5.52	4.93	6.39	3.38	8.84	7.44
Empirical Argmax	4.88	3.17	3.37	3.19	1.41	4.48	2.38	2.77	3.05	1.62	3.03
Quadratic Surface	4.20	2.46	2.43	2.98	2.27	4.25	2.24	2.95	2.07	**0.91**	2.68
No calibration											
Naive $n{=}n_{\max}$	17.86	35.46	27.28	16.74	36.70	17.86	35.46	27.28	16.74	36.70	26.81
Naive $n{=}1$	7.88	3.22	3.87	4.32	**1.03**	7.88	3.22	3.87	4.32	1.03	4.07
Probe-calibrated											
Difficulty Mixture	5.81	1.32	2.11	3.54	1.30	5.37	1.43	2.16	3.86	1.38	2.83
Star Exponent	3.82	1.16	1.47	2.06	1.22	4.04	1.18	1.36	1.97	1.25	1.95
Learned β_{eff}	2.72	**1.14**	1.07	**1.27**	2.20	2.39	1.05	1.05	**1.10**	1.84	1.58
TB-MAD (Ours)	**2.29**	1.29	**0.93**	1.40	1.90	**1.88**	**0.83**	**0.84**	1.15	1.47	**1.40**

that is more data-hungry; Learned β_{eff} is the closest but trails while sacrificing interpretability. Actual n^* predictions at 1M spend are in Appendix Figure 7. Broken out by window, TB-MAD is within 1% of the best rule at six of the seven windows, trailing by 1.16% at $W{=}2500$ (Appendix Table 20). Its performance is therefore not driven by a single region of the sweep.

TB-MAD Scaling Behavior. TB-MAD provides the best regret-cost trade-off, with regret decreasing as the calibration set grows (Figure 2). Its plain regret is already among the lowest at 10 calibration questions and stays roughly flat beyond 60, since one probe-fitted ability curve gains little from more data. Debate-calibrated baselines start higher and keep dropping, closing part of that gap while their cost climbs (left panel). At equal spend, TB-MAD achieves the lowest regret and saturates once the maximum number of calibration questions is reached, whereas debate-based baselines keep improving as more tokens are spent, and Successive Halving overtakes TB-MAD at the largest budget.

Robustness to Heterogeneous Settings. We extend the evaluation to a heterogeneous setting with K unique models, each appearing with the same model count, and isolate the allocation problem (Appendix C). The probe-calibrated methods are again the strongest, and Learned β_{eff} edges ahead of TB-MAD. Our rule trades a small amount of regret for a closed form and an explicit model-selection gate (Appendix Tables 12 and 13).

Generalization To Self-Refinement. Revision inherits the same generation-floor effect as debate, making self-refinement similarly token-inefficient: tight windows can reduce accuracy before any refinement benefit emerges. Useful refinement also requires both headroom and self-correction ability, so gains remain modest relative to the added token cost (Appendix D.1).

When More Agents Do Not Pay. Reading peers costs generation, which resolves two questions often studied separately. First, communication does not pay as majority voting with *no* communication matches or beats debate at 93% of the 980 matched (n, W) operating points with $n \geq 2$ (Appendix Table 18). A paired bootstrap on the same cells is more favourable still. Self-consistency is significantly better in 57% of cases, debate in just 0.5%, and the remaining 42.5% are indistinguishable. The benefit of voting over debate has also been reported at matched sampling budgets (Smit et al., 2024; Choi et al., 2025). More importantly, Eq. 11 says when the ordering should reverse, and we test that prediction across seven windows rather than at a single operating point. At $W{=}1200$, the margin is $+42.0\%$, or $+13.7\%$ restricting to cells where g is not clipped to the floor. It falls below 1 point by $W{=}6\text{k}$ and is indistinguishable from zero through 32k on Qwen2.5-7B/GSM8K (Appendix Table 16), in line with Appendix Cor. 3's correlation ceiling. The measured debate accuracy includes one revision round, so the observed margin also includes a revision lift not covered

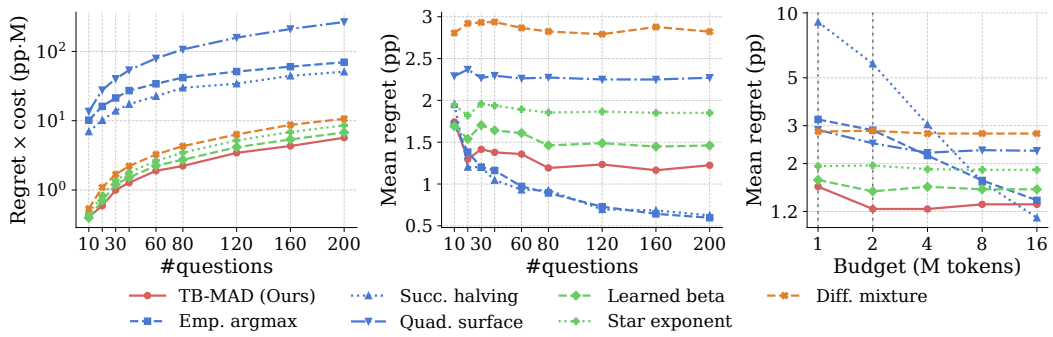


Figure 2: Two ways to fix a calibration budget, equal questions (left and middle) and equal spend (right, dotted verticals at the two budgets of Table 2), and two readings of the equal-questions sweep: cost-adjusted (left) and plain regret (middle).

by Prop. 2, estimated at 0.0–6.7% (Table 14, Appendix D.1). Second, a group is not always worth forming: a single agent matches or beats the debate group at 47% of the 980 cells (Appendix Table 18, single≥group column), with $\delta > 0$ determining when grouping pays. Single-agent probes recover the sign of δ correctly in 81% of cases, compared with 56% for always grouping. The gap $A(n, W) - A(1, W)$ is interior in n and collapses under tight windows across all five models (Appendix Figure 8), a generation-floor effect since raising $g_{\min}$ recovers the worst cell (Appendix Figure 10). Directly measuring $\Delta p_{\max}$ per agent at fixed g, gives a small, sign-mixed estimate across four model-task cells (Appendix Table 15), so we find no reliable ceiling gain from reading a peer, consistent with debate's apparent gains coming from the vote and not from $\Delta p_{\max}$. Further details are in Appendix D.2.

4.3 Ablations

The reading cost m. m is the token cost of reading one peer message (Eq. 1), determining how quickly added agents reduce the group's shared thinking room. Both the collapse and interior optimum arise from $m > 0$. Setting $m{=}0$ holds g fixed regardless of group size, as in debate where each agent has its own window. The theory then predicts the largest feasible group at every window, with no interior optimum and no collapse (Appendix Figure 9). More details are in Appendix D.6.

The price of a token. We ablate the output-to-input price ratio $r{=}c_{\text{out}}/c_{\text{in}}$ to reflect practical cost models, where reading and generating tokens can have different prices. Making input more expensive reduces their performance accordingly and TB-MAD keeps the lowest regret at every ratio tried on these spends (Appendix Table 19), while cheaper input buys a larger cost-optimal group at the same spend (Appendix Figure 6).

Is the collapse a long-context artifact? A longer transcript could hurt through position (Liu et al., 2024a; Hsieh et al., 2024) or length alone (Levy et al., 2024), but these effects do not explain the collapse. Moving the question to the end of the transcript, after the peer messages, at fixed H and g changes accuracy by less than 1% on average (Appendix Table 23). Question position moves the collapse cell without lifting it out of collapse.

5 Conclusion

We develop a token-budget theory of when additional context is worth its cost by placing reading and thinking in a shared window. The theory predicts optimal group size and identifies a critical window below which a single agent can outperform debate. Empirically, our method achieves the lowest regret under tight budgets, while self-consistency often matches or outperforms debate at lower cost. The theory offers a simple guide for *when* to read more, think more, or add agents. Limitations and Future work are discussed in Appendix E.

AI USE STATEMENT

In this work, we used generative AI tools solely to support writing, including proofing, language polishing, and improving readability. We did not use them to develop research ideas, conduct experiments, or generate the substantive content of the work. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

TB-MAD lowers the token cost of deploying multi-agent debate by choosing a group size before any debate is run. Its predictions are calibrated per model and task, so a size chosen outside the regime the probe sweep covers can lose accuracy with no signal that it has. We use only public benchmarks and open-weight models, with no human subjects or personal data.

REPRODUCIBILITY STATEMENT

All model equations, proofs and implementation details are in Appendices B and A.1–A.3. We will release the code, the banked debate transcripts and the table and figure generation scripts openly for research purposes upon publication.

REFERENCES

Anthropic. Pricing, 2026. URL https://platform.claude.com/docs/en/about-claude/pricing.

Blaz Bertalanic and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. *arXiv preprint arXiv:2605.00914*, 2026.

Philip J. Boland. Majority systems and the Condorcet jury theorem. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 38(3):181–189, 1989.

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

Irving W Burr. Cumulative frequency functions. *The Annals of Mathematical Statistics*, 1942.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more LLM calls all you need? towards the scaling properties of compound AI systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2403.02419.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Nicolas de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 21–27 Jul 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google. Gemini developer api pricing, 2026. URL https://ai.google.dev/gemini-api/docs/pricing.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025.

Archibald Vivian Hill. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *The Journal of Physiology*, 1910.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *Conference on Language Modeling (COLM)*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *AISTATS*, 2016.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025. arXiv:2502.19130.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1994.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 2003.

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

Krishna K. Ladha. The condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3), 1992.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.14848.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research (TMLR)*, 2024a. arXiv:2402.05120.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.427. URL https://aclanthology.org/2024.findings-emnlp.427/.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050; defines the MATH500 subset.

Bang Liu, Linglong Kong, and Jian Pei. Phase transition for budgeted multi-agent synergy. *arXiv preprint arXiv:2601.17311*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12, 2024a.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*, 2024b.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sora Miyamoto, Daisuke Oba, and Naoaki Okazaki. Aligning tree-search policies with fixed token budgets in test-time scaling of LLMs. *arXiv preprint arXiv:2602.09574*, 2026.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20286–20332, 2025.

Manh Nguyen, Anh Nguyen, Dung Nguyen, Svetha Venkatesh, and Hung Le. Hear both sides: Efficient multi-agent debate via diversity-aware message retention. *arXiv preprint arXiv:2603.20640*, 2026.

Prashanti Nilayam, Kiran Kumar Ramanna, Prashil Tumbade, and Sankalp Nayak. Heterogeneous llm debate under adversarial peers: Honest gains, replacement costs, and resilience. *arXiv preprint arXiv:2606.19826*, 2026.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 34433–34448, 2025.

OpenAI. Pricing, 2026. URL `https://developers.openai.com/api/docs/pricing`.

Vilfredo Pareto. *Cours d'économie politique*. Librairie Droz, 1964.

Xuan Qi. Brief is better: Non-monotonic chain-of-thought budget effects in function-calling language agents. *arXiv preprint arXiv:2604.02155*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Tim Robertson, Richard Dykstra, and FT Wright. Order restricted statistical inference. *Wiley*, 1988.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.17371.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations (ICLR)*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*, 2019.

Technology Innovation Institute. The Falcon 3 family of open models. https://huggingface.co/blog/falcon3, 2024. Technology Innovation Institute technical report.

Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets. *arXiv preprint arXiv:2604.02460*, 2026.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a. arXiv:2402.18272.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.11171.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024b. arXiv:2406.01574.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2201.11903.

Waloddi Weibull. A statistical distribution function of wide applicability. *Journal of applied mechanics*, 1951.

Florian Valentin Wunderlich, Lars Benedikt Kaesberg, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Multi-agent reasoning improves compute efficiency: Pareto-optimal test-time scaling. *arXiv preprint arXiv:2605.01566*, 2026.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, Xitai Jin, Tianying Tiana Chen, Jing Li, and Xiaohua Xu. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of NAACL*, 2025. arXiv:2502.04790.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, et al. Stop overvaluing multi-agent debate: We must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. *arXiv preprint arXiv:2601.19921*, 2026.

APPENDIX

Each appendix, the main-text section it supports, and its subsections, one per line with its page.

A Proofs *(supports §3.4)* 14
A.1 Single-peakedness 14
A.2 The large-window limit 16
A.3 Self-consistency weakly dominates debate 16

B Implementation Details *(supports §3.2, §4.1)* 17
B.1 Implementing TB-MAD 17
B.2 Implementing baselines 18

C Generalization To Heterogeneous Settings *(supports §4.2)* 20
C.1 Extended method 21
C.2 Updated theory analysis 22
C.3 Results 23

D Additional Results *(supports §4.2, §4.3)* 24
D.1 Self-refine, full results 24
D.2 Voting against talking, and the single agent 24
D.3 Repricing at API rates 26
D.4 Is the error concentration c_w constant? 26
D.5 Does σ depend on n? 27
D.7 Generation-floor recovery 28

E Limitations And Future Work *(supports §5)* 28

A PROOFS

A.1 SINGLE-PEAKEDNESS (PROP. 1)

Treat n as continuous, so

$$g(n) = W - q - c - (n-1)m, \qquad \delta(n) = (1 + c_w)\, p_{\max}\, \phi\big(g(n)\big) - c_w.$$

Since ϕ is defined on $[0, \infty)$, we work on $[1, n_{\text{exh}}]$ with $n_{\text{exh}} = 1 + (W - q - c)/m$, the (real-valued, not the deployment cap $n_{\max}$ used elsewhere) group size at which the reasoning budget is exhausted ($g(n_{\text{exh}}) = 0$).

Step 1: δ is strictly decreasing. g is strictly decreasing in n ($g'(n) = -m < 0$) and ϕ is strictly increasing (Assumption 1), so $\phi(g(n))$ is strictly decreasing and hence $\delta(n)$ is also strictly decreasing. Every form in Table 1 has $\phi(0) = 0$, so $\delta(n_{\text{exh}}) = -c_w < 0$. With $\delta(1) > 0$, δ continuous and strictly decreasing, there is a unique root $n_0 \in (1, n_{\text{exh}})$, and $\mathcal{F} = [1, n_0]$ lies strictly inside the domain of ϕ. (This step uses only monotonicity of ϕ, not the differentiability also assumed in Assumption 1.)

Step 2: the log-derivative decomposition. Write $M(n) = S(n)^2 = n_{\text{eff}}(n)\delta(n)^2$ for the squared synergy index of §3.3. Since $\phi \in C^1$ and g is affine, $\delta \in C^1$; on $\text{int}(\mathcal{F})$ we have $\delta > 0$, so $\log M(n) = \log n_{\text{eff}}(n) + 2 \log \delta(n)$ is differentiable and

$$\frac{\mathrm{d}}{\mathrm{d}n} \log M = \underbrace{\frac{n'_{\text{eff}}}{n_{\text{eff}}}}_{\alpha(n)} - \underbrace{\left(-2\frac{\delta'}{\delta}\right)}_{\mathcal{C}(n)}.$$

With $D(n) = 1 + \rho(n-1)$ and $n_{\text{eff}} = n/D$,

$$\frac{n'_{\text{eff}}}{n_{\text{eff}}} = \frac{\mathrm{d}}{\mathrm{d}n}\big(\log n - \log D\big) = \frac{1}{n} - \frac{\rho}{D}$$
$$= \frac{D - \rho n}{nD} = \frac{1-\rho}{n\big(1+\rho(n-1)\big)},$$

is the amplification term $\alpha(n)$. It is strictly positive for $\rho \in [0,1)$, strictly decreasing in n because the denominator nD is increasing. Moreover, its decay rate depends on ρ, with $\alpha(n) = \Theta(n^{-2})$ as $n \to \infty$ for $\rho \in (0,1)$ and $\alpha(n) = 1/n$ at $\rho = 0$. For the cost term, $\delta'(n) = -(1+c_w)p_{\max} m\, \phi'\big(g(n)\big)$, and ϕ' is continuous and strictly positive on the compact interval $[g(n_0), g(1)]$, so $|\delta'|$ is bounded below by some $\varepsilon > 0$ on $\mathcal{F}$. Since $\delta > 0$ and $\delta' < 0$ on the interior, $\mathcal{C}(n) > 0$ there, and as $n \to n_0^-$ we have $\delta \to 0^+$ with $|\delta'| \geq \varepsilon$, so $\mathcal{C}(n) \to +\infty$.

Step 3: existence of an interior maximizer. $M = S^2$ is continuous on the compact interval $\mathcal{F} = [1, n_0]$, so it attains a maximum there. At the right endpoint $\delta(n_0) = 0$ gives $M(n_0) = 0 < M(1)$, and by hypothesis $M(\tilde{n}) > M(1)$ for some $\tilde{n} \in \operatorname{int}(\mathcal{F})$. The maximum therefore exceeds both endpoint values and is attained at some $n^* \in (1, n_0)$. At any such interior maximizer $\log M$ is differentiable (Step 2), so $\alpha(n^*) = \mathcal{C}(n^*)$.

Step 4: uniqueness under log-concavity. $\mathcal{C}(n) = -2(\log\delta)'(n)$ is non-decreasing iff $(\log\delta)'$ is non-increasing iff δ is log-concave on $\mathcal{F}$. If $\mathcal{C}$ is non-decreasing then $\alpha - \mathcal{C}$ is strictly decreasing (strictly decreasing α minus non-decreasing $\mathcal{C}$), hence has exactly one zero. Therefore $\log M$ is strictly quasi-concave on $\operatorname{int}(\mathcal{F})$. The interior maximizer n^* of Step 3 is a stationary point; by uniqueness, it is the only one, so n^* is the unique interior maximum.

Lemma 1 (Uniqueness of the peak)**.** *If δ is log-concave on $\mathcal{F}$, the stationary point of $\log M$ is unique, so n^* is the only interior maximum.*

The ability $\beta(n) = p_{\max}\phi(g)$ is log-concave in g, and g is affine in n, so β is log-concave in n. Since $\log\beta = \log p_{\max} + \log\phi$ and for the adopted HILL gain $\phi(g) = g^\kappa/(g^\kappa + g_0^\kappa)$,

$$(\log\phi)'(g) \;=\; \frac{\kappa\, g_0^\kappa}{g\big(g^\kappa + g_0^\kappa\big)}, \tag{12}$$

which is decreasing in g, so $\log\phi$ is concave in g and hence in n because g is affine in n. The same holds for the power law, where $(\log\phi)' = \kappa/\big(g_0 u(u^\kappa - 1)\big)$ with $u = 1 + g/g_0$ is decreasing in g. For the bounded exponential, $\log(1 - e^{-u})$ has second derivative $-1/\big(4\sinh^2(u/2)\big) < 0$, and $u = g/g_0$ is affine, so log-concavity in g follows. The WEIBULL case needs the chain rule, since $u = (g/g_0)^\kappa$ is convex for the fitted $\kappa > 1$ (Table 5).

The shift also preserves log-concavity, on the whole of $\mathcal{F}$. Recall $\delta(n) = (1+c_w)\beta(n) - c_w$ from the definitions above, and drop the argument n. Since $c_w > 0$ is a constant, $\delta' = (1+c_w)\beta'$ and $\delta'' = (1+c_w)\beta''$ share β'''s sign.

- If $\beta''(n) \leq 0$: on $\operatorname{int}(\mathcal{F})$ we have $\delta(n) > 0$ (Step 1), so $\delta\delta'' \leq 0 \leq (\delta')^2$, giving $\delta\delta'' - (\delta')^2 \leq 0$ directly; log-concavity of β is not needed in this case.
- If $\beta''(n) > 0$: substituting the derivatives above,

$$\delta\delta'' - (\delta')^2 = (1+c_w)\Big[(1+c_w)\big(\beta\beta'' - (\beta')^2\big) - c_w\beta''\Big] \tag{13}$$
$$\leq (1+c_w)\big[0 - c_w\beta''\big] \leq 0, \tag{14}$$

  using $\beta\beta'' - (\beta')^2 \leq 0$ (log-concavity of β, shown above) and $c_w\beta'' \geq 0$.

The two cases exhaust every $n \in \operatorname{int}(\mathcal{F})$, so $\delta\delta'' - (\delta')^2 \leq 0$ throughout, i.e. δ is log-concave on all of $\mathcal{F}$. Lemma 1 then applies unconditionally: $\mathcal{C}$ is non-decreasing on $\mathcal{F}$, so $\alpha - \mathcal{C}$ has exactly one zero and n^* is unique. The only inputs are log-concavity of β (established above for every candidate in Table 1) and $c_w > 0$.

Corollary 1 (n^* decreases in ρ)**.** *Under Prop. 1 and Lemma 1, n^* is strictly decreasing in ρ.*

Proof. n^* solves $\alpha(n) = \mathcal{C}(n)$, and $\mathcal{C}(n) = -2\delta'(n)/\delta(n)$ does not depend on ρ. From Step 2,

$$\frac{\partial \alpha}{\partial \rho}(n) = \frac{\partial}{\partial \rho} \frac{1-\rho}{n(1+\rho(n-1))} = \frac{-1}{\big(1+\rho(n-1)\big)^2} \;<\; 0$$

for every $n > 1$, so α falls pointwise as ρ rises while $\mathcal{C}$ is fixed. $\alpha - \mathcal{C}$ is strictly decreasing in n at the crossing (Step 4), so lowering the α shifts the crossing left and n^* strictly decreases in ρ. □

More correlated errors buy less from an extra vote at every group size, so the optimum sits at a smaller group.

A.2 The large-window limit

Write $\alpha(n) = n'_{\text{eff}}/n_{\text{eff}} = (1-\rho)/\big(n(1+\rho(n-1))\big)$ for the amplification term of A.1 Step 2. Since $\delta'(n) = -(1+c_w)p_{\max} m\, \phi'\big(g(n)\big)$, the cost term depends on n and W only through the room g,

$$C(n) \;=\; -2\frac{\delta'(n)}{\delta(n)} \;=\; \widehat{C}\big(g(n,W)\big), \qquad \widehat{C}(g) \;=\; \frac{2m(1+c_w)p_{\max}\,\phi'(g)}{(1+c_w)p_{\max}\phi(g) - c_w}.$$

Corollary 2 (Slow growth)**.** *Assume the conditions of Prop. 1 and Lemma 1, so that C is non-decreasing in n on F, equivalently $\widehat{C}' \le 0$. Then on any window range where the maximizer stays interior, $n^*(W)$ is differentiable, non-decreasing in W, and*

$$0 \;\le\; \frac{dn^*}{dW} \;=\; \frac{1}{m + \alpha'(n^*)/\widehat{C}'(g)} \;\le\; \frac{1}{m} \;=\; \frac{dn_{\max}}{dW}.$$

The optimal size therefore never grows faster than the feasibility cap, and the gap between them widens as the window loosens.

Proof. The interior maximizer solves $F(n,W) = \alpha(n) - \widehat{C}\big(g(n,W)\big) = 0$ with $g(n,W) = W - q - c - (n-1)m$, so $\partial g/\partial W = 1$ and $\partial g/\partial n = -m$. Hence $\partial F/\partial W = -\widehat{C}'(g)$ and $\partial F/\partial n = \alpha'(n) + m\widehat{C}'(g)$. By A.1 Step 2, α is strictly decreasing, so $\alpha' < 0$; by hypothesis $\widehat{C}' \le 0$, so $\partial F/\partial n < 0$ and the implicit function theorem applies. Then

$$\frac{dn^*}{dW} = -\frac{\partial F/\partial W}{\partial F/\partial n} = \frac{\widehat{C}'(g)}{\alpha'(n^*) + m\widehat{C}'(g)}.$$

Numerator and denominator are both non-positive, giving $dn^*/dW \ge 0$. Dividing through by $\widehat{C}'$ gives the stated form, and $\alpha'/\widehat{C}' \ge 0$ gives the upper bound $1/m$. Since $n_{\max}(W) = 1 + (W - q - c)/m$, that bound is exactly $dn_{\max}/dW$. □

Corollary 3 (The large-window limit)**.** *As $W \to \infty$, thinking room stops binding, $\delta(n) \to (1 + c_w)p_{\max} - c_w$ is constant in n, and $S(n)^2$ becomes monotone increasing. For $\rho > 0$ it saturates at $\big((1+c_w)p_{\max} - c_w\big)^2/\rho$ as $n_{\text{eff}} \to 1/\rho$, the classical correlation ceiling. At $\rho = 0$ the errors are independent, $n_{\text{eff}} = n$ is unbounded, and no ceiling exists. An interior best size therefore exists only while the window is scarce, and once it is not, the classical picture returns.*

Proof. $g(n) \to \infty$ forces $\phi\big(g(n)\big) \to 1$, so $\delta' \to 0$ and $\widehat{C} \to 0$. With $\alpha > 0$ everywhere, $\log M$ has no interior stationary point and M is monotone increasing on F, saturating at the stated ceiling as $n_{\text{eff}} \to 1/\rho$ when $\rho > 0$. □

A.3 Self-consistency weakly dominates debate (Prop. 2)

$g_{\text{SC}} = W - q - c \ge g(n)$ since $(n-1)m \ge 0$; ϕ increasing gives $\phi(g_{\text{SC}}) \ge \phi(g(n))$; multiplying by $p_{\max}$ gives the ability bound $\beta_{\text{eff}}(g_{\text{SC}}) \ge \beta_{\text{eff}}(g(n))$, with equality iff $n{=}1$ (since $(n-1)m = 0$ only at $n{=}1$). The margin $\delta = (1+c_w)\beta_{\text{eff}} - c_w$ is strictly increasing in β_{eff} for $c_w \ge 0$, and Φ is increasing, so the head of Eq. 9 is increasing in β_{eff} at fixed (n, ρ, σ, c_w). The ability bound therefore carries over to $A_{\text{SC}}(n,W) \ge A^0_{\text{deb}}(n,W)$. □

Note that $\Delta p_{\max}$ is a gain in the *ceiling* from conditioning on peer content and not in generation length, so the left side of Eq. 11 must be bought outright; we report it in Appendix D.2.

Table 3: Benchmark details. All benchmarks use 500 samples, including the first 40% used for calibration and the remaining 60% for evaluation.

Benchmark	Task	Sample size	HuggingFace path
GSM8K	Grade-school math (open)	500	`openai/gsm8k`
MATH500	Competition math (open)	500	`HuggingFaceH4/MATH-500`
MMLU-Pro	Multi-domain QA (10-way)	500	`TIGER-Lab/MMLU-Pro`
CSQA	Commonsense QA (5-way)	500	`tau/commonsense_qa`

Table 4: Operating grid. Generation room is the leftover $g = \mathrm{clip}(W - H - c, g_{\min}, g_{\mathrm{cap}})$, so a larger group at fixed W thinks less. The probe sweep runs to 4096. Deployment clips g to $g_{\mathrm{cap}} = 2048$, and the out-of-range test in Table 9 fits on caps up to 2048 and predicts at 4096. Table 17's long-CoT comparison raises g_{cap} to 8192 to give the single agent every advantage.

Parameter	Value
Window W (tokens)	$\{1.2, 1.6, 2.0, 2.5, 3.2, 4.5, 6.5\}$k
Agents n	$\{1, \ldots, 8\}$
Question length q (mean, tokens)	GSM8K 128, MATH500 146, MMLU-Pro 258, CSQA 105
Per-message length m	320
Overhead c	64
Revision rounds R	1 (so R+1=2 generation passes)
Generation room g	$\mathrm{clip}(W - H - c,\ g_{\min},\ g_{\mathrm{cap}})$
$g_{\min},\ g_{\mathrm{cap}}$	64, 2048
Probe caps g	$\{64, 128, 256, 512, 1024, 2048, 4096\}$
Seeds	$\{42, 43, 44\}$ (pooled)

B Implementation Details

We provide the configuration details for all hyperparameters in Table 4. All experiments are conducted in a single H100 (80G).

B.1 Implementing TB-MAD

Fitting β_{eff}. The three parameters $(p_{\max}, g_0, \kappa)$ are fitted to the single-agent *thinking* sweep by bound-constrained nonlinear least squares. Let (g_i, A_i, N_i) denote the per-level mean accuracies A_i at reasoning budget g_i, computed from N_i samples. We minimize the sample-weighted sum of squared residuals:

$$(\hat{p}_{\max}, \hat{g}_0, \hat{\kappa}) = \arg\min_{p_{\max}, g_0, \kappa} \sum_i N_i \big(A_i - \beta_{\text{eff}}(g_i)\big)^2, \tag{15}$$

where $\beta_{\text{eff}}(g) = p_{\max} g^\kappa / (g^\kappa + g_0^\kappa)$ (Eq. 4), with $p_{\max} \in (0, 1]$ and $g_0, \kappa > 0$. Table 5 reports the fits with three goodness-of-fit measures across all settings. R^2 alone is uninformative, reaching $R^2 \geq 0.993$ in every cell; we therefore use RMSE to select between candidate forms. We report a 95% CI on $\hat{p}_{\max}$ (Table 5) from a percentile bootstrap: resample the sweep's per-level records with replacement, recompute pooled accuracy at each level, and refit, over 500 refits per cell. Several sweep points sit numerically above the point estimate $\hat{p}_{\max}$, but none exceed the CI's upper bound in any of the twenty cells, so every such crossing is within sampling noise and no cell needs refitting.

Pooling the seeds. Every β_{eff} is fitted on the pools from three different seeds (42, 43, 44), because ϕ is nonlinear in $(p_{\max}, g_0, \kappa)$: averaging three per-seed parameter vectors gives a curve that is not the average of the three fitted curves. Refitting Eq. 4 separately for each seed changes n^* by only 0.02 agents on average, while the regret of each seed-specific n^* relative to the per-cell oracle is identical to two decimal places in 18 of 20 cells. The question set is fixed across seeds, so this is generation-sampling noise only.

Does g_0 correspond to a real length scale? On Qwen2.5-7B and Falcon3-7B, the fitted Hill g_0 lands within 4% of the median completion length on GSM8K and within 15% on MMLU-Pro (Table 6).

Table 5: Single-agent Hill fits on the cap sweep, all four tasks. $p_{\max}$, its 95% bootstrap CI (resampling sweep records, 500 refits), RMSE and LOO are on the accuracy scale, reported as percentages; LOO is the held-out RMSE, leaving out one swept cap at a time.

Model	Task	$p_{\max}$	95% CI	g_0	κ	RMSE	LOO	R^2
Qwen2.5-7B	GSM8K	87.4	[86.8, 88.0]	256	6.9	0.17	1.89	1.000
	MATH500	67.6	[66.5, 68.9]	433	3.7	0.21	0.27	1.000
	MMLU-Pro	49.9	[48.8, 51.1]	437	3.7	0.37	1.88	1.000
	CSQA	80.5	[79.7, 81.4]	246	9.8	0.09	8.32	1.000
Falcon3-7B	GSM8K	81.0	[80.4, 81.7]	244	4.7	0.37	2.43	1.000
	MATH500	61.0	[59.7, 62.4]	475	3.5	0.49	1.68	1.000
	MMLU-Pro	47.7	[46.6, 49.0]	414	2.8	0.65	1.17	0.999
	CSQA	74.9	[74.0, 75.8]	168	3.0	1.76	4.03	0.996
Llama-3.2-3B	GSM8K	69.0	[68.2, 69.8]	171	5.5	0.15	3.48	1.000
	MATH500	34.4	[33.4, 35.6]	222	3.5	0.14	0.34	1.000
	MMLU-Pro	32.8	[31.9, 33.8]	325	3.4	0.80	2.10	0.997
	CSQA	67.8	[67.0, 68.7]	201	6.2	0.17	4.53	1.000
Gemma-2-9B	GSM8K	81.3	[80.5, 82.1]	144	5.0	0.36	3.77	1.000
	MATH500	40.2	[39.1, 41.3]	226	3.7	0.07	0.19	1.000
	MMLU-Pro	50.4	[49.2, 51.5]	274	3.3	1.77	5.43	0.992
	CSQA	77.9	[77.0, 78.7]	137	5.4	1.66	13.15	0.996
Gemma-3-12B	GSM8K	92.9	[92.2, 93.4]	184	4.2	0.35	1.65	1.000
	MATH500	75.3	[74.0, 77.1]	459	2.4	0.41	0.87	1.000
	MMLU-Pro	60.4	[59.0, 61.9]	478	2.9	1.02	2.55	0.998
	CSQA	80.5	[79.6, 81.4]	176	8.6	0.74	12.41	1.000

Table 6: Fitted Hill midpoint $\hat{g}_0$ against the measured median completion length, per model and task. Median lengths are read off the largest generation cap (4096 tokens), where under 1% of samples are truncated. *ratio* is $\hat{g}_0$ over the median.

	GSM8K			MMLU-Pro		
Model	$\hat{g}_0$	median len.	ratio	$\hat{g}_0$	median len.	ratio
Qwen2.5-7B	256	264	0.97	437	490	0.89
Falcon3-7B	244	254	0.96	414	484	0.85
Mean ratio		0.96			0.87	

Estimating $\hat{\rho}$, and its invariance From the initial round's answers, we compute pairwise agreement of *errors* across agents (same wrong answer) and map it to the equicorrelation ρ in Eq. 7. Eq. 7 assumes that ρ does not vary with either n or W. We verify this empirically (Table 8): (1) *Constant in n:* re-estimating $\hat{\rho}$ at each group size changes it by at most $\mathrm{range}_n \leq 0.05$, so ρ does not grow with the group; and (2) *Constant in W:* re-estimating $\hat{\rho}$ at each window changes it by at most $\mathrm{range}_W \leq 0.05$ across all twenty model-task cells. Error correlation appears to be a property of the model and task, rather than of the budget on these cells.

B.2 Implementing Baselines

The five algorithms are collected in Alg. 1–5; $\mathcal{D}$ is the calibration sample and $g(n, W)$ the room the window leaves.

Calibrated on measured debate cells:

- **Empirical argmax** measures every (n, W) cell on $\mathcal{D}$ and selects the size with the highest $\hat{A}(n, W)$ (Alg. 2).

Table 7: Seed-to-seed spread of the ability fit, of the resulting n^*, and of the regret that spread costs, as standard deviations over three per-seed refits. $\hat{p}_{\max}$ and regret are in percentage points, $\hat{g}_0$ in tokens, $\hat{\kappa}$ and n^* dimensionless. Each per-seed fit uses that seed's full thinking sweep and its regret is scored against the per-cell oracle on the full bank. **Mean** regret is scored from a fit refit on a single held-out seed at a time, not the pooled (all-seed) fit the rest of the paper uses, so it is not directly comparable to the pooled-fit regret in Table 21; the near-zero **s.d.** shows that which seed is dropped does not matter, not that single-seed refitting matches the pooled result.

		s.d. over the three per-seed fits			s.d. of n^*	regret (pp)	
Model	**Task**	$\hat{p}_{\max}$	$\hat{g}_0$	$\hat{\kappa}$	**mean**	**mean**	**s.d.**
Qwen2.5-7B	GSM8K	0.02	0.1	0.30	0.026	2.01	0.00
	MATH500	0.31	1.1	0.05	0.020	0.78	0.00
	MMLU-Pro	0.22	3.7	0.05	0.023	1.09	0.00
	CSQA	0.07	1.7	1.66	0.064	0.70	0.00
Falcon3-7B	GSM8K	0.22	0.7	0.07	0.024	0.20	0.00
	MATH500	0.09	2.7	0.01	0.016	1.31	0.00
	MMLU-Pro	0.56	5.4	0.04	0.004	1.20	0.00
	CSQA	0.15	0.4	0.08	0.033	1.11	0.00
Llama3.2-3B	GSM8K	0.22	1.1	0.14	0.016	0.11	0.00
	MATH500	0.19	1.8	0.14	0.064	2.08	0.75
	MMLU-Pro	0.24	0.7	0.10	0.000	4.40	0.00
	CSQA	0.10	0.9	0.09	0.008	0.52	0.00
Gemma2-9B	GSM8K	0.05	0.7	0.07	0.008	0.20	0.00
	MATH500	0.12	2.6	0.06	0.044	0.81	0.00
	MMLU-Pro	0.23	2.0	0.02	0.016	0.80	0.00
	CSQA	0.13	0.2	0.17	0.016	0.51	0.00
Gemma3-12B	GSM8K	0.12	0.6	0.04	0.020	0.30	0.01
	MATH500	0.04	2.3	0.02	0.008	0.33	0.00
	MMLU-Pro	0.17	0.7	0.03	0.004	1.73	0.00
	CSQA	0.03	1.1	0.15	0.004	6.34	0.00
Mean		0.16	1.5	0.16	0.021	1.33	–

- **Successive halving** (Jamieson & Talwalkar, 2016) evaluates all sizes on a small sample, keeps the better half, and doubles the sample until one size remains or the budget is exhausted (Alg. 3).
- **Quadratic surface** fits a least-squares quadratic in $\{1, n, W, n^2, W^2, nW\}$ to the measured cells and selects its best predicted size per window (Alg. 2).

Calibrated on single-agent probes only, as ours is:

- **Learned** β_{eff} uses our aggregator with a gradient-boosted-tree estimate $\hat{\beta}_{\text{gbt}}$ fitted to the $|\mathcal{G}|$=7 probe caps, with no parametric shape (Alg. 4).
- **Star exponent** (Liu et al., 2026) uses our aggregator, $\hat{\rho}$, and budget model, replacing ϕ with $p_{\max}\big(1-(1+g/g_0)^{-\kappa}\big)$, the POW form of Table 1 (Alg. 4).
- **Difficulty mixture** (Chen et al., 2024) estimates per-question accuracy $p_i(g)$ from the probe sweep and simulates votes. The original fixes the per-call allowance, so p_i cannot vary with n; we instead evaluate p_i at the remaining room $g(n, W)$, interpolated in $\log g$, and add our correlation $\hat{\rho}$ and tie penalty c_w (Alg. 5).

No calibration: **naive** n=1 and **naive** n=$n_{\max}$ always use one agent or the largest group; neither reads data, so neither has an algorithm.

Table 8: Round-0 error correlation $\hat{\rho}$ re-estimated at each window. range$_W$ and range$_n$ are $\max \hat{\rho} - \min \hat{\rho}$ over the swept windows and over the group sizes.

		$\hat{\rho}$ at W ($\times 10^3$ tokens)								
Model	Task	1.2	1.6	2.0	2.5	3.2	4.5	6.5	range$_W$	range$_n$
Qwen2.5-7B	GSM8K	0.36	0.35	0.36	0.35	0.34	0.35	0.35	0.01	0.02
	MATH500	0.71	0.69	0.68	0.68	0.68	0.68	0.68	0.04	0.01
	MMLU-Pro	0.60	0.57	0.57	0.56	0.56	0.56	0.56	0.04	0.02
	CSQA	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.01	0.03
Falcon3-7B	GSM8K	0.41	0.41	0.40	0.40	0.41	0.41	0.40	0.01	0.01
	MATH500	0.68	0.65	0.64	0.64	0.65	0.64	0.64	0.05	0.01
	MMLU-Pro	0.59	0.56	0.57	0.56	0.56	0.56	0.55	0.04	0.01
	CSQA	0.51	0.50	0.50	0.50	0.50	0.50	0.50	0.01	0.02
Llama3.2-3B	GSM8K	0.41	0.41	0.40	0.41	0.42	0.42	0.42	0.01	0.02
	MATH500	0.52	0.52	0.52	0.53	0.52	0.52	0.52	0.01	0.02
	MMLU-Pro	0.44	0.43	0.43	0.43	0.42	0.43	0.44	0.02	0.00
	CSQA	0.42	0.42	0.42	0.42	0.42	0.42	0.43	0.01	0.02
Gemma2-9B	GSM8K	0.57	0.58	0.58	0.58	0.57	0.57	0.57	0.01	0.01
	MATH500	0.64	0.64	0.65	0.64	0.64	0.64	0.65	0.01	0.02
	MMLU-Pro	0.66	0.66	0.66	0.66	0.67	0.66	0.66	0.01	0.01
	CSQA	0.57	0.59	0.58	0.58	0.57	0.57	0.58	0.02	0.02
Gemma3-12B	GSM8K	0.65	0.65	0.66	0.66	0.65	0.66	0.65	0.01	0.04
	MATH500	0.77	0.80	0.78	0.77	0.77	0.78	0.77	0.02	0.02
	MMLU-Pro	0.65	0.64	0.64	0.63	0.63	0.63	0.63	0.02	0.01
	CSQA	0.55	0.54	0.55	0.55	0.55	0.55	0.54	0.01	0.04

Table 9: Fitting performance of thinking gain ϕ, over five models and four datasets, per cell in Table 5, Appendix B.3. RMSE is on the accuracy scale in percent; leave-one-out holds out one swept cap at a time and out-of-range predicts at 4096 from caps up to 2048. Adopted form in bold.

Selection criterion	HILL	WEIB	POW	BURR	iso
Leave-one-out RMSE	**3.61**	5.21	11.74	4.49	10.02
Out-of-range RMSE	0.38	0.75	6.85	0.33	**0.31**
Total	**3.99**	5.96	18.59	4.82	10.33

Ours: **TB-MAD** fits $(p_{\max}, g_0, \kappa)$ on the probe sweep and evaluates $p_{\max}\phi(g; g_0, \kappa)$ at the remaining room for each size (Alg. 1).

B.3 Calibration Cost

TB-MAD, Star exponent, and learned β_{eff} estimate ρ and c_w from data rather than treating them as free constants. Each calibration question requires at least three independent round-0 answers, since no peer history is available yet and three samples are the minimum needed for the correlation estimate. We charge each the token cost of measuring them, $0.1\times$ a single probe question (Table 11, so all baselines are matched to the same calibration spend throughout (Figure 2, right panel; Appendix Table 19). Table 10 additionally reports the wall-clock cost of calibration, comparing our single-agent probe with the full (n, W) grid used by the Empirical Argmax debate baseline.

C Generalization to heterogeneous settings

We extend our method to heterogeneous model settings, where agents have different abilities β_{eff}. To isolate the effect of model heterogeneity, we consider a *uniform allocation* setting in which the K models are assigned the same number of agents, while every agent keeps the same room $g(n)$

Table 10: Calibration wall-clock in minutes, 200 questions. Probe is our single-agent sweep; debate (empirical argmax) is the full (n, W) grid. *ratio* is debate over probe per dataset, then averaged over the four.

	GSM8K		MATH500		CSQA		MMLU-Pro		
Model	probe	debate	probe	debate	probe	debate	probe	debate	ratio
Llama3.2-3B	6	21	11	34	6	38	11	71	5×
Qwen2.5-7B	8	25	13	43	7	55	13	93	5×
Falcon3-7B	8	31	17	54	6	57	14	105	6×
Gemma2-9B	7	29	13	62	6	68	13	147	8×
Gemma3-12B	12	62	34	146	10	171	31	336	9×

Table 11: Input and output tokens each calibration rule spends, as a multiple of the single-agent probe sweep, summed over the twenty cells at the full 200-question pool. The probe (except Difficulty Mixture) multiple is $1\times$ by definition; debate re-reads every peer message, so its input multiple is the largest.

Calibration rule	**Input**	**Output**
Empirical Argmax	$49\times$	$11\times$
Quadratic Surface	$49\times$	$11\times$
Successive Halving	$33\times$	$9\times$
Difficulty Mixture	$0.9\times$	$0.8\times$
Round-0 probe (ρ, c_w)	$0.1\times$	$0.1\times$
Probe sweep (TB-MAD, learned β_{eff}, Star Exponent)	$1\times$	$1\times$

and each model i has fixed parameters $(p_{\max,i}, g_{0,i}, \kappa_i)$. The goal is the optimal number of agents across the K models under a total budget of n agents, with n/K agents assigned to each model.

C.1 Extended method

Per-model edges and candidate subsets. Under uniform allocation, every agent receives

$$g(n) = \text{clip}\big(W - q - c - (n-1)m,\ g_{\min},\ g_{\text{cap}}\big), \tag{16}$$

so heterogeneity enters through the model-specific edge

$$\mathbb{E}[X_j] = \delta_j(n) = (1 + c_w)\, p_{\max,j}\, \phi_j\big(g(n)\big) - c_w. \tag{17}$$

A negative edge indicates a model that systematically points toward the wrong answer, so mixing it with positive-edge models can dilute the group edge. We therefore allow the n agents to draw either from all K models or from any positive-edge singleton when the fitted edges have mixed signs:

$$\bar{\delta}_S(n) = \frac{1}{|S|} \sum_{i \in S} \delta_i(n), \tag{18}$$

with candidate set

$$\mathcal{S}(n) = \begin{cases} \{\{1, \dots, K\}\} \cup \{\{i\} : \delta_i(n) > 0\}, & \min_i \delta_i(n) \le 0 < \max_i \delta_i(n), \\ \{\{1, \dots, K\}\}, & \text{otherwise.} \end{cases} \tag{19}$$

Thus $|S| = 1$ is simply a homogeneous group from model i, whereas the full set remains available at every n.

Correlation becomes block structure. Let $\hat{\rho}_w$ and $\hat{\rho}_b$ denote round-0 error correlations for agents backed by the same and different models, respectively. For a subset S with equal blocks of $n/|S|$ agents,

$$n_{\text{eff}}(n, |S|) = \frac{n}{1 + \hat{\rho}_w(n/|S| - 1) + \hat{\rho}_b(n - n/|S|)}. \tag{20}$$

This reduces to the homogeneous $n/(1+\hat{\rho}_w(n-1))$ when $|S|=1$ and, as $n\to\infty$, approaches $|S|/(\hat{\rho}_w+(|S|-1)\hat{\rho}_b)$. Hence mixing raises the asymptotic ceiling whenever $\hat{\rho}_b<\hat{\rho}_w$, i.e., cross-model errors are less correlated.

Accuracy and operating point. With common variance σ^2, $\mathrm{Var}(\bar{X}_n)=\sigma^2/n_{\mathrm{eff}}(n,|S|)$, giving

$$A(n,S)=\Phi\Big(\bar{\delta}_S(n)\sqrt{n_{\mathrm{eff}}(n,|S|)}/\sigma\Big). \tag{21}$$

Since Φ is monotone, maximizing accuracy is equivalent to

$$\mathcal{V}(n,S)=\mathrm{sign}\big(\bar{\delta}_S(n)\big)\,n_{\mathrm{eff}}(n,|S|)\,\bar{\delta}_S(n)^2. \tag{22}$$

We preserve the sign of the margin to avoid amplifying a negative-edge subset, consistent with the treatment of $S(n)$ in § 3.3. The edges are obtained from single-agent probes, while only $\hat{\rho}_w,\hat{\rho}_b$ require round-0 measurements (independent answers before any exchange, not the full multi-round debate); calibrating all K models therefore costs roughly K times a set-blind probe rule. As in the homogeneous setting (§4.2), we charge this round-0 probe cost against the calibration budget of every rule that reads $\hat{\rho}_w,\hat{\rho}_b$ from the debate bank rather than treating them as free constants, summed once per distinct model in the model set; Tables 12 and 13 already reflect it.

The selection gate. Algorithm 6 evaluates $\mathcal{V}(n,S)$ over the swept grid and $\mathcal{S}(n)$, then selects

$$(\hat{n}^*,\hat{S})\in\arg\max_{n\in\{K,2K,\ldots,n_{\max}\},\ S\in\mathcal{S}(n)}\mathcal{V}(n,S). \tag{23}$$

The gate admits singletons only when fitted edges disagree in sign; otherwise it reduces to the homogeneous rule. Allowing singletons at every n would simply favor the strongest model whenever $\bar{\delta}$ dominates, with little corresponding change in n_{eff}, which explains the weaker performance of *TB-MAD-hetero w/o gate* in Tables 12 and 13.

C.2 Updated theory analysis

Proposition 3 (Heterogeneous aggregation head)**.** *Under uniform allocation into $|S|$ equal blocks, a common σ, within/between round-0 correlations (ρ_w,ρ_b), and the ±1 scoring of §3.3, the group accuracy is*

$$A(n)=\Phi\Big(\bar{\delta}_S(n)\,\sqrt{n_{\mathrm{eff}}(n,|S|)}/\sigma\Big),$$

with n_{eff} as in Eq. 20. Whenever $\bar{\delta}_S(n)>0$, $\arg\max_n A(n)=\arg\max_n n_{\mathrm{eff}}(n,|S|)\,\bar{\delta}_S(n)^2$, and every step of §3.4 that uses δ only through $A(n)$ holds verbatim with $\delta\to\bar{\delta}_S$.

Proof sketch. The covariance of $(X_1,\ldots,X_n)$ has constant diagonal σ^2, constant within-block off-diagonal $\rho_w\sigma^2$, and constant cross-block off-diagonal $\rho_b\sigma^2$. Summing gives $\mathrm{Var}(\bar{X}_n)=n^{-2}\big[n\sigma^2+(n/|S|-1)n\rho_w\sigma^2+(n-n/|S|)n\rho_b\sigma^2\big]=\sigma^2/n_{\mathrm{eff}}(n,|S|)$. The normal approximation of $\bar{X}_n$ with mean $\bar{\delta}_S$ and this variance gives $A(n)$; monotonicity of Φ and $\sigma>0$ give the $\arg\max$ identity. □

We do not have experiments isolating the two results below, so we only note whether each still holds and under which conditions.

Prop. 2 (self-consistency dominates the initial-round vote). Still holds on a model set when no member's ceiling is raised by peer text, $\Delta p_{\max,i}\le 0$ for all i: $g_{\mathrm{SC}}\ge g(n)$ and each ϕ_i increasing make $\bar{\delta}$ larger under self-consistency, so the dominance carries over with no new assumption. It can fail once some member is weaker than a peer that sends it information its own model lacks, $\Delta p_{\max,i}>0$. The crossing condition is then

$$\sum_i \Delta p_{\max,i}\,\phi_i\big(g(n)\big)\ >\ \sum_i p_{\max,i}\big(\phi_i(g_{\mathrm{SC}})-\phi_i(g(n))\big), \tag{24}$$

the homogeneous Eq. 11 summed over members.

Table 12: Held-out regret against the per-cell oracle (percentage points) when each method chooses n on the heterogeneous model-set bank at a matched *token* budget, the model-set analogue of Table 2 with the same marks. Per-cell values in Tables 31–32.

Method	1.0M budget				2.0M budget				Avg.
	Q+G3	Q+L	F+G2	G3+L	Q+G3	Q+L	F+G2	G3+L	
Debate-calibrated									
Successive Halving	8.96	10.88	7.97	7.47	4.75	6.87	4.82	5.50	7.15
Quadratic Surface	2.61	7.01	2.68	2.63	4.60	6.62	2.79	4.05	4.13
Empirical Argmax	2.93	7.09	3.30	3.07	2.96	5.18	3.64	3.19	3.92
No calibration									
Naive $n{=}n_{\max}$	30.82	27.81	17.81	22.85	30.82	27.81	17.81	22.85	24.82
Naive $n{=}2$	3.66	6.89	3.70	3.40	3.66	6.89	3.70	3.40	4.41
Probe-calibrated									
Difficulty Mixture	2.45	3.77	2.60	2.19	2.81	4.34	2.67	2.09	2.86
Star Exponent	1.92	3.66	1.95	1.60	1.31	3.30	1.96	1.50	2.15
Learned β_{eff}	1.56	2.68	**1.66**	**1.30**	1.13	2.35	**1.58**	1.31	1.70
TB-MAD-hetero w/o gate	1.33	**0.82**	2.06	2.41	1.48	**1.12**	1.76	1.37	1.54
TB-MAD-hetero (Ours)	**0.97**	1.87	2.05	1.34	**0.81**	2.07	1.95	**0.85**	**1.49**

The interior optimum. The result extends under the stated restrictions. Existence (Step 3 of Appendix A.1) requires only that $\bar{\delta}_S$ be continuous, strictly decreasing, and cross zero once, all of which continue to hold here. In particular, even when $g_{0,i}$ and κ_i vary across members, $\bar{\phi} = |S|^{-1}\sum_{i\in S}\phi_i$ remains strictly increasing as a positive average of strictly increasing functions, so the feasibility boundary $\bar{\delta}_S{=}0$ has a unique crossing in g. Local uniqueness at this crossing also carries over without assuming log-concavity of $\bar{\phi}$. Whenever $\bar{\delta}_S \to 0^+$ with $\bar{\delta}''_S \le 0$, the boundary argument in Appendix A.1 applies unchanged, since it relies only on positivity of $\bar{\delta}_S$ and the sign of its second derivative. Away from the boundary, however, log-concavity of $\bar{\phi}$ need not hold for a heterogeneous average and is not assumed here.

C.3 Results

We evaluate on the model sets below, spanning members of similar and of differing capability, on GSM8K, MATH500, MMLU-Pro and CSQA as in the homogeneous setting. Agent i speaks as member $i \bmod K$, so a set forms K equal blocks only at n divisible by K, which fixes each grid.

- $K{=}2$, $n \in \{2,4,6,8\}$: Qwen2.5-7B+Gemma3-12B, Qwen2.5-7B+Llama3.2-3B, Falcon3-7B+Gemma2-9B, Gemma3-12B+Llama3.2-3B.
- $K{=}3$, $n \in \{3,6\}$: Qwen2.5-7B+Gemma3-12B+Llama3.2-3B, Qwen2.5-7B+Falcon3-7B+Gemma2-9B.
- $K{=}4$, $n \in \{4,8\}$: Qwen2.5-7B+Falcon3-7B+Gemma2-9B+Gemma3-12B, Qwen2.5-7B+Falcon3-7B+Gemma3-12B+Llama3.2-3B.

Per-member $(p_{\max,i}, g_{0,i}, \kappa_i)$ come from the single-agent sweeps and $\hat{\rho}_w, \hat{\rho}_b$ from round-0 answers, exactly as on the homogeneous cells.

Tables 12 and 13 report held-out regret against the per-cell oracle when each method picks (n, S), against the eight baselines of Table 2. Probe-calibrated methods are the strongest at every K. At $K{=}2$ our method is second best, just behind learned β_{eff}. For $K > 2$ the star exponent matches it: as K grows the group ability averages K member curves, so the curve shape that separates the two fits at $K{=}2$ washes out. Removing the gate is worse on average at both K ($1.49 \to 1.54$ on the pairs, $1.66 \to 2.75$ on the larger sets), which confirms the design choice.

Table 13: Held-out regret (percentage points) on the balanced-K sets, two K=3 sets on grid $\{3, 6\}$ and two K=4 on $\{4, 8\}$, at the budgets of Table 2 with the same marks. Each naive column is a fixed n, blank where a set's grid omits it; a row spanning one K gets no average. Per-cell values in Tables 29–30.

Method	1.0M budget				2.0M budget				Avg.
	Q+G3+L	Q+F+G2	Q+F+G2+G3	Q+F+G3+L	Q+G3+L	Q+F+G2	Q+F+G2+G3	Q+F+G3+L	
Debate-calibrated									
Successive Halving	2.97	3.09	4.48	3.66	2.41	1.68	2.77	2.84	2.99
Quadratic Surface	2.58	2.41	1.81	1.64	1.90	1.81	1.91	1.71	1.97
Empirical Argmax	2.14	1.88	2.39	2.16	1.75	1.79	2.02	1.73	1.98
No calibration									
Naive n=8	–	–	20.72	19.46	–	–	20.72	19.46	–
Naive n=6	14.65	14.05	–	–	14.65	14.05	–	–	–
Naive n=3	2.88	2.10	–	–	2.88	2.10	–	–	–
Naive n=4	–	–	2.04	1.88	–	–	2.04	1.88	–
Probe-calibrated									
Difficulty Mixture	2.60	1.86	1.88	1.44	2.36	1.79	1.75	1.54	1.90
Star Exponent	1.64	1.60	1.77	1.35	1.64	1.60	1.70	1.34	1.58
Learned β_{eff}	**1.23**	**1.49**	**1.61**	**1.26**	1.22	**1.34**	**1.54**	**1.21**	**1.36**
TB-MAD-hetero w/o gate	3.77	2.33	3.53	3.01	1.91	2.47	2.88	2.10	2.75
TB-MAD-hetero (Ours)	1.26	1.50	2.72	1.90	**0.99**	1.54	2.05	1.34	1.66

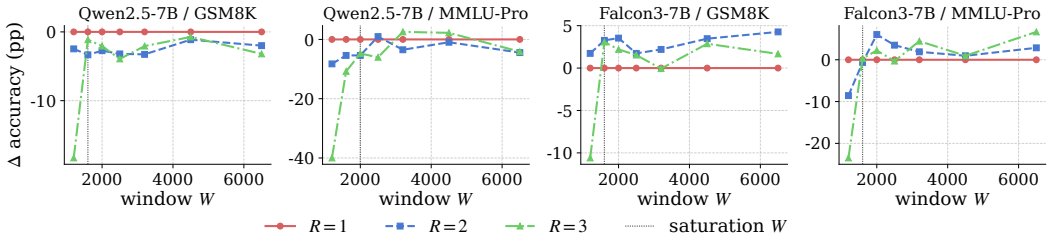


Figure 3: Self-refinement accuracy against the first draft (Δ=$A_R - A_1$, in points) as the window grows, $R \in \{1, 2, 3\}$; R=1 is the base generation without refinement. The dotted vertical line marks the saturation window: from there on, both R=2 and R=3 stay within 7% of the first draft.

D Additional Results

D.1 Self-refinement, full results

Figure 3 shows the same generation-floor pattern across all cells: at the smallest window, increasing from R=1 to R=3 reduces accuracy by 1.3–44.3 points, with the loss recovering as W grows. More refinement steps require a larger window before they become beneficial.

The weakest model (Llama3.2-3B) gains almost nothing ($\Delta \approx 0$), consistent with limited self-correction ability without external feedback (Huang et al., 2024), while ceiling models have little headroom (Qwen/Gemma-3 on GSM8K). A genuine interior R^* appears mainly for mid-capability models with both headroom and self-correction ability (Falcon3, Gemma-2, and most models on MATH500), but the resulting gains remain small relative to the additional token cost (Table 14).

D.2 Voting against talking, and the single agent

Two secondary consequences follow from the same comparison: self-consistency weakly dominates debate (Prop. 2), and forming a group is not always worthwhile ($\delta \leq 0$).

Where the margin is largest, and how firmly. Prop. 2 implies that the self-consistency margin is set by the per-vote ability consumed by peer reading. It is therefore largest when the generation window binds and shrinks as the window loosens, eventually vanishing once both methods reach their best size at sufficiently large W (Figures 4–5). The large gap at the tightest window is partly a clipping effect: many $(n, W$=1200$)$ cells collapse to the same clipped generation room, whereas restricting to unclipped cells reduces the margin from +42.0 to +13.7%. A question-clustered paired

Table 14: Self-refine over $R \in \{1, 2, 3\}$ rounds across settings. A_1: first-draft accuracy; gain*: oracle (grid-best) refine gain over A_1; R^*: the round most often best across the swept windows; drop: A_1 minus the R=3 accuracy at the smallest window.

	GSM8K				MATH500				MMLU-Pro				CSQA			
Model	A_1	gain*	R^*	drop	A_1	gain*	R^*	drop	A_1	gain*	R^*	drop	A_1	gain*	R^*	drop
Qwen2.5-7B	87.6	0.0	1	18.3	64.7	1.3	1	44.3	47.0	2.5	1	40.0	81.4	1.1	1	4.7
Falcon3-7B	81.3	4.3	2	10.6	55.1	3.5	1	31.3	48.3	**6.7**	3	23.5	74.1	0.7	2	1.3
Gemma-2-9B	81.7	2.0	2	1.6	38.2	3.3	2	3.8	50.5	3.2	1	13.7	77.1	0.0	1	1.3
Gemma-3-12B	92.7	0.9	1	7.5	65.7	2.3	1	27.1	54.3	2.9	2	33.3	80.5	0.2	1	2.3
Llama-3.2-3B	69.9	0.0	1	28.8	33.7	0.0	1	15.3	27.3	0.0	1	20.3	66.7	0.0	1	14.1

Table 15: Per-agent round-1 accuracy (%) with no peer read (**No peer**, n=1) against with one peer read (**With peer**, n=2), fixed g=2048 for both, averaged over 3 seeds. $\widehat{\Delta p_{\max}}$ (%) is the difference.

Model	**Task**	**No peer** (round 1)	**With peer** (round 1)	$\widehat{\Delta p_{\max}}$
Qwen2.5-7B	GSM8K	86.7	84.7	-1.97
	MMLU-Pro	49.8	46.0	-3.78
Falcon3-7B	GSM8K	80.6	82.9	+2.33
	MMLU-Pro	46.1	50.2	+4.11

bootstrap with 2000 resamples confirms the cell-level effect; applying the same test per window gives the SC sig. W column in Table 18.

Where the closed form gets the sign wrong. CSQA is the one task where the closed form recovers the δ sign poorly, most severely on Gemma3-12B (predicted 24%, the lowest cell in Table 18). The per-cell fit there has $k \approx 8.6$, an unusually steep Hill curve, fit to only seven calibration points; its leave-one-out RMSE is over an order of magnitude above its in-sample RMSE, the signature of overfitting rather than of a genuine near-step response. Combined with c_w estimates that are the least stable of any task (Table 21), the fitted ϕ crosses the break-even β almost as soon as the window opens, so the closed form predicts the group winning at every swept size while the group in fact never does. All five models show the same direction of error on CSQA, so this is a property of the task's fit stability, not an isolated bug in one cell.

Is $\Delta p_{\max}$ actually positive? Prop. 2's dominance result is the special case $\Delta p_{\max} \leq 0$. We test directly whether peer text raises the per-agent ceiling, instead of measuring it through majority vote. On Qwen2.5-7B and Falcon3-7B across GSM8K and MMLU-Pro, we fix g=2048 and compare a single agent's round-1 accuracy without peer access (n=1) with the round-1 accuracy of two agents after each reads the other's round-0 answer (n=2). This isolates the effect of peer text from voting. $\widehat{\Delta p_{\max}}$ is small and sign-mixed, at 2–4% in magnitude, with the sign determined by model rather than task: positive for Falcon3-7B and negative for Qwen2.5-7B on both tasks (Table 15). Thus, peer reading does not reliably raise the ceiling in these cells, suggesting that debate gains arise primarily from voting rather than from $\Delta p_{\max}$.

Does the margin actually vanish at large W? On Qwen2.5-7B/GSM8K, extending the same comparison past the main grid to $W \in \{6\text{k}, \ldots, 32\text{k}\}$ confirms it does: each method read at its own best group size, the margin is already under 1pp by W=6k and stays within seed noise through 32k (Table 16), consistent with Cor. 3's correlation ceiling. We report one representative case since reaching the ceiling requires double-digit n and is not practical in realistic settings.

Debate against single-agent long-CoT. Contrary to the equal-budget finding of Tran & Kiela (2026), on our matched generated-token budgets, we find the debate group still leads a single agent whose room is grown to g_{cap}=8192 (W extends to 16k). Its chain saturates well below the cap, at about 275 tokens on GSM8K and 533–540 on MMLU-Pro. At matched generated-token budgets,

Table 16: Qwen2.5-7B/GSM8K, self-consistency against debate at large W, each method read at its own best group size $\hat{n}$ over the extended grid $n \in \{1, 2, 4, 8, 13, 21, 32\}$ (columns 2–5). **margin** is debate minus self-consistency ($\pm$ seed s.d. over 3 seeds), and **matched** n is the same gap when self-consistency is instead read at debate's $\hat{n}$. Every margin lies inside its own seed s.d., so the two methods are indistinguishable from W=6k onward.

W	$\hat{n}_{\mathrm{deb}}$	debate	$\hat{n}_{\mathrm{SC}}$	self-cons.	margin	matched n
6k	13	93.67	13	94.07	-0.40 ± 1.11	-0.40
9k	13	93.80	13	94.07	-0.27 ± 0.46	-0.27
12k	13	94.00	21	94.07	-0.07 ± 0.81	$+0.00$
16k	21	94.33	13	93.67	$+0.67 \pm 1.22$	$+0.73$
24k	13	94.07	21	94.00	$+0.07 \pm 1.47$	$+0.13$
32k	21	93.87	21	93.67	$+0.20 \pm 0.53$	$+0.20$

Table 17: Single-agent long-CoT (n=1, room grown to g_{cap}=8192) against the debate group ($n \geq 2$), matched at the tokens debate spends at its own best cell. Gen. tokens instead gives long-CoT's own best window, where more thinking stops improving its accuracy.

Model	Task	Long-CoT A	gen. tokens	Debate A	Debate − long-CoT
Qwen2.5-7B	GSM8K	88.9	276	93.7	$+\mathbf{4.9}$
	MMLU-Pro	49.5	533	49.8	$+\mathbf{0.2}$
Falcon3-7B	GSM8K	81.5	274	91.4	$+\mathbf{9.9}$
	MMLU-Pro	46.5	540	55.3	$+\mathbf{8.9}$

the group leads on every cell (Table 17). Self-consistency recovers the same advantage without message exchange, making the comparison one of breadth versus depth rather than communication.

D.3 Repricing at API rates

In practice, input tokens are often cheaper than output tokens. We define the cost of an operating point (n, W) as $c_{\mathrm{in}} \cdot (\text{input tokens}) + c_{\mathrm{out}} \cdot (\text{output tokens})$, normalize $c_{\mathrm{out}} = 1$, and consider $r \in 1, 2, 2.5, 4$. Figure 6 shows the accuracy-vs-cost Pareto frontier and the cost-optimal group size $n^*(B)$ (instead of $n^*(W)$), where B is the average token per-question cost swept across operating points. The interior optimum persists under realistic pricing, with cheaper input shifting n^* upward at matched budget. On Qwen2.5-7B/MMLU-Pro, however, the optimum remains pinned at $n = 1$: all $n \geq 2$ points collapse far below it due to the g-floor effect in §D.2, regardless of the price ratio.

The same pricing affects the coverage of a fixed calibration budget. Table 19 shows that discounted input costs slightly improve both coverage and performance for debate-calibrated methods, while our method retains the lowest regret at all budgets (1, 2, and 4M) across every price ratio.

D.4 Is the error concentration c_w constant?

On a K-way decision, c_w is bounded by $\frac{1}{K-1} \leq c_w \leq 1$, while the observed values lie well within these bounds. Uniform errors would give c_w=$\frac{1}{9}$ on MMLU-Pro, compared with measured values of 0.42–0.67 (the stable, per-n estimate; Table 21). §3.3 holds c_w fixed in n and g so that all budget dependence in $q_{\max}$ runs through β_{eff}; we test both directions on round-0 answers (Table 21). In n, $\hat{c}_w$ is flat. Eighteen of twenty cells drift by at most 0.05 over the swept group sizes, and the largest drift is 0.08 on Qwen2.5-7B/GSM8K. In g it rises: at g=64 little parses so the surviving wrong mass is thin and spread, at g=4096 complete answers agree on the same wrong value. A single c_w therefore overstates concentration at the bottom of the window range, where the collapse happens. Substituting the per-n estimate $c_w(n)$ for the fixed cell-level value moves n^* at 26 of 140 operating points and cuts mean held-out regret from 1.16pp to 0.80pp, most sharply on Llama3.2-3B/MMLU-Pro ($4.40 \to 0.65$) and MATH500 ($2.71 \to 0.79$pp).

Table 18: Accuracy (%) over the swept windows with $n \geq 2$. Barred columns average over the grid, unbarred read both methods at debate's size $n^*(W)$. $\delta \leq 0$ **share** compares each cell's debate accuracy against the single-agent accuracy pooled over all windows; **single≥group** is the stricter, per-window version of the same comparison, the observed share of the 49 cells with $A(1, W) \geq A(n, W)$ at matching W; **predicted** is the fraction of the $\delta \leq 0$-share signs the closed form recovers. **SC sig.** W is how many of the 7 swept windows have SC significantly ahead (paired bootstrap) at a majority of the 7 group sizes.

		theory			averaged over the grid		at the oracle $n^*(W)$		group vs. single agent			SC vs. debate	
Model	**Task**	$p_{\max}$	$\bar{\delta}$	$\bar{A}_1$	$\bar{A}_{\text{SC}}$	$\bar{A}_{\text{deb}}$	A_{SC}	A_{deb}	$\delta \leq 0$ **share**	**single≥group**	**predicted**	**SC≥deb**	**SC sig.** W
Llama3.2-3B	GSM8K	69.0	+0.23	67.7	**80.3**	69.1	**82.0**	80.7	16/49	16/49	80	49/49	4/7
	MATH500	34.4	-0.05	34.1	**42.0**	32.9	**42.1**	41.5	21/49	22/49	86	36/49	4/7
	MMLU-Pro	32.8	-0.23	31.8	**37.1**	27.9	**37.0**	35.7	23/49	22/49	47	42/49	4/7
	CSQA	67.8	+0.11	65.1	**69.9**	60.1	**71.7**	70.2	16/49	15/49	78	49/49	3/7
Qwen2.5-7B	GSM8K	87.4	+0.37	84.7	**92.2**	73.5	**92.5**	92.3	15/49	15/49	96	48/49	2/7
	MATH500	67.6	+0.11	67.2	**71.6**	48.7	**70.4**	68.3	23/49	24/49	86	46/49	4/7
	MMLU-Pro	49.9	-0.15	47.9	**53.2**	32.4	**52.6**	46.0	36/49	40/49	69	49/49	7/7
	CSQA	80.5	+0.23	80.6	**82.2**	65.7	**82.4**	82.0	23/49	21/49	80	49/49	3/7
Falcon3-7B	GSM8K	81.0	+0.28	83.4	**91.4**	75.6	**91.6**	90.8	13/49	13/49	100	45/49	3/7
	MATH500	61.0	+0.06	60.0	**65.9**	45.6	**65.3**	62.4	24/49	28/49	86	48/49	7/7
	MMLU-Pro	47.7	-0.15	48.6	**53.4**	38.7	**52.4**	50.8	24/49	25/49	96	48/49	5/7
	CSQA	74.9	+0.21	73.8	**76.9**	66.5	**77.7**	75.6	21/49	28/49	73	49/49	7/7
Gemma2-9B	GSM8K	81.3	+0.32	81.1	**87.5**	80.9	**88.3**	86.6	9/49	9/49	94	45/49	7/7
	MATH500	40.2	-0.11	42.2	**48.7**	41.2	48.8	**48.8**	14/49	14/49	98	32/49	3/7
	MMLU-Pro	50.4	-0.14	49.0	**51.1**	40.0	**50.8**	50.6	23/49	23/49	88	40/49	3/7
	CSQA	77.9	+0.26	73.8	**76.2**	71.2	**77.2**	76.3	15/49	17/49	71	48/49	2/7
Gemma3-12B	GSM8K	92.9	+0.45	92.3	**93.9**	78.4	**93.9**	93.8	16/49	16/49	92	48/49	2/7
	MATH500	75.3	+0.15	71.6	**73.8**	52.4	**73.0**	70.7	26/49	33/49	80	48/49	4/7
	MMLU-Pro	60.4	-0.10	57.7	**60.6**	40.2	**60.2**	56.6	26/49	32/49	90	47/49	4/7
	CSQA	80.5	+0.26	78.2	**78.3**	58.6	**78.0**	76.4	49/49	49/49	24	49/49	7/7
All		–	–	64.5	**69.3**	55.0	**69.4**	67.8	**433/980**	**462/980**	**81**	**915/980** (93)	**85/140**

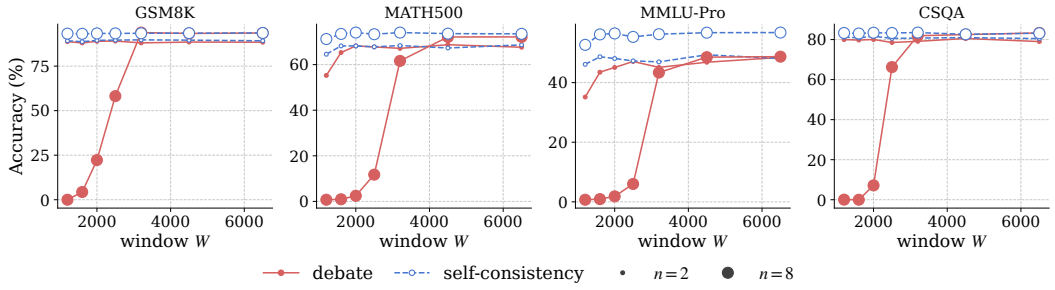


Figure 4: Debate (red, solid) against self-consistency (blue, dashed) as the window grows, one panel per task (Qwen2.5-7B). Marker size is the group size $n \in \{2, 8\}$. The gap closes as the window grows on every panel (Prop. 2).

D.5 Does σ depend on n?

§3.3 treats σ as a single constant per cell, fitted once and reused across all n. However, σ is determined by the three-valued vote score $X_j \in \{+1, 0, -1\}$ once $\Pr[X_j{=}{+}1] = \beta_{\text{eff}}$ and $\Pr[X_j{=}{-}1] = c_w(1{-}\beta_{\text{eff}})$ are fixed. Its variance is therefore $\sigma^2(n) = c_w{+}(1{-}c_w)\beta_{\text{eff}}(n){-}\delta(n)^2$, which varies with n through $\beta_{\text{eff}}(n)$. We test whether replacing the fitted constant with this derived, n-dependent value changes the prediction. Table 22 reruns Algorithm 1 under both choices on the same 5 models $\times$ 4 tasks $\times$ 7 windows $=$ 140 operating points as Table 21, and reports every disagreement in $\hat{n}^*$. The two choices disagree at only 7 of 140 points, always by one agent. Six shift downward, where the derived $\sigma(n)$ favors a smaller group, while one, Llama3.2-3B/MATH500, shifts upward. Unlike the $c_w(n)$ substitution above, which is uniformly conservative, the derived $\sigma(n)$ is not one-sided. We therefore retain the simpler fitted constant in the main text, while noting that it changes only $7/140$ predictions.

D.6 Sensitivity of the prediction to m.

Figure 9 sweeps $m \in \{0, 100, 320, 700\}$ for the closed form against empirical argmax and the true held-out best (oracle), all matched to a 1M-token calibration budget and scored held out. At $m{=}0$

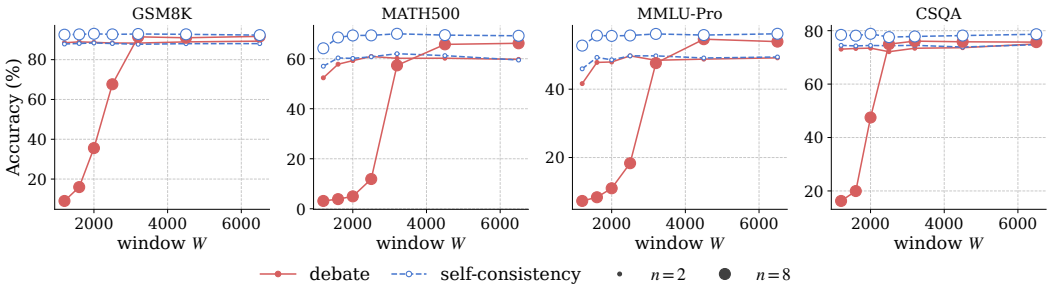


Figure 5: Same as Figure 4, on Falcon3-7B: the gap closes with the window on every task there too.

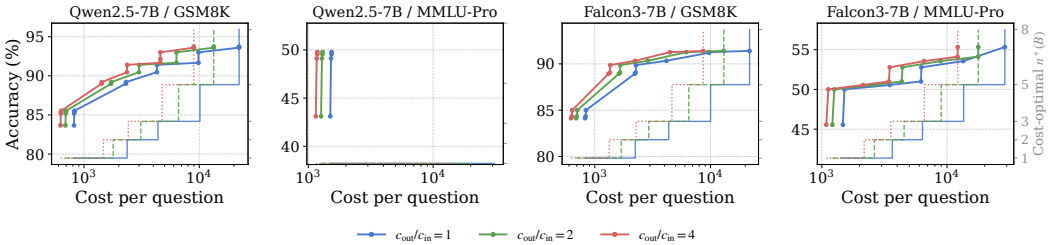


Figure 6: Accuracy against cost per question. Lines are the Pareto frontier (left axis) and steps are the cost-optimal group $n^*(B)$ (right axis), under symmetric pricing and under cheaper input at API rates.

the prediction is pinned at $n_{\max}$ on every window, as the theory requires with no per-agent reading cost; n^* then falls with m at fixed W, with the paper's m=320 the closest of the three swept values to the oracle. Argmax, collapse to K=1 question at this budget, stays flat and low throughout.

D.7 Generation-floor recovery.

Figure 10 sweeps the generation floor $g_{\min}$ at the worst cell (n=8, W=1200), well below the fitted scale g_0. Accuracy recovers as $g_{\min}$ crosses g_0, tracking $\beta_{\text{eff}}(g_{\min})$ with no free parameter. GSM8K ($g_0 \approx 244$–256) is near ceiling by $g_{\min}$=512, while the harder MMLU-Pro ($g_0 \approx 414$–437) continues climbing past 1024. This suggests the collapse reflects a task-dependent reasoning requirement captured by g_0, rather than a fragile artifact of the $g_{\min}$=64 floor reported in §4.2. The floor gives the sharpest manifestation of the effect, not its only one.

E Limitations and Future Work

First, the aggregation model is a tractable approximation rather than an exact account of group behavior. It relies on equicorrelation, a fixed error concentration, and an average-based decision threshold, all of which are empirically evaluated rather than assumed to hold exactly. Moreover, although we investigate heterogeneous model sets, allocation remains uniform across agent counts; adaptive allocation is an interesting direction for future work.

Second, the analysis focuses on self-generated computation, including debate and self-refinement, and does not cover settings where additional context is retrieved or supplied externally, such as retrieval-augmented generation.

Third, all five models have at most 12B parameters and short fitted solution lengths g_0, ranging from 137 to 478 tokens. Models with thousand-token reasoning chains would have much larger g_0, shifting the collapse regime to larger windows and potentially beyond deployed context lengths. We expect the qualitative shape to transfer, but not the specific operating points.

Table 19: Held-out regret (percentage points) at four input/output price ratios and every swept budget. **Bold** is the column minimum, underline indicates an entry fixed at its large-budget value as the whole question is already used for calibration. First group: $c_{\text{out}}/c_{\text{in}}$=1, i.e. Table 2.

Method	$c_{\text{out}}/c_{\text{in}} = 1.0$					$c_{\text{out}}/c_{\text{in}} = 2.0$					$c_{\text{out}}/c_{\text{in}} = 2.5$					$c_{\text{out}}/c_{\text{in}} = 4.0$				
Budget (M)	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
Empirical Argmax	3.20	2.86	2.18	1.67	1.35	3.11	2.51	1.83	1.42	1.08	2.91	2.45	1.89	1.45	1.10	2.87	1.88	1.57	1.25	0.97
Successive Halving	9.08	5.81	3.04	1.65	**1.12**	6.54	4.56	1.83	**1.26**	**1.00**	6.31	4.45	1.73	**1.24**	**0.92**	5.55	2.01	1.39	**1.10**	**0.95**
Quadratic Surface	2.87	2.48	2.25	2.31	2.29	2.55	2.34	2.22	2.25	2.31	2.35	2.32	2.18	2.27	2.36	2.50	2.26	2.45	2.32	2.27
Learned β_{eff}	1.68	1.49	1.56	1.52	1.52	1.65	1.78	1.50	1.52	1.52	1.63	1.80	1.50	1.52	1.52	1.76	1.70	1.52	1.52	1.52
Star Exponent	1.95	1.96	1.88	1.87	1.87	2.05	1.93	1.82	1.87	1.87	1.95	2.08	1.89	1.87	1.87	1.93	2.03	1.87	1.87	1.87
Difficulty Mixture	2.82	2.84	2.76	2.76	2.76	2.69	2.83	2.76	2.76	2.76	2.73	2.82	2.76	2.76	2.76	3.05	2.78	2.76	2.76	2.76
TB-MAD (Ours)	**1.56**	**1.23**	**1.23**	**1.29**	1.29	**1.43**	**1.50**	**1.23**	1.29	1.29	**1.29**	**1.42**	**1.24**	1.29	1.29	**1.50**	**1.40**	**1.26**	1.29	1.29

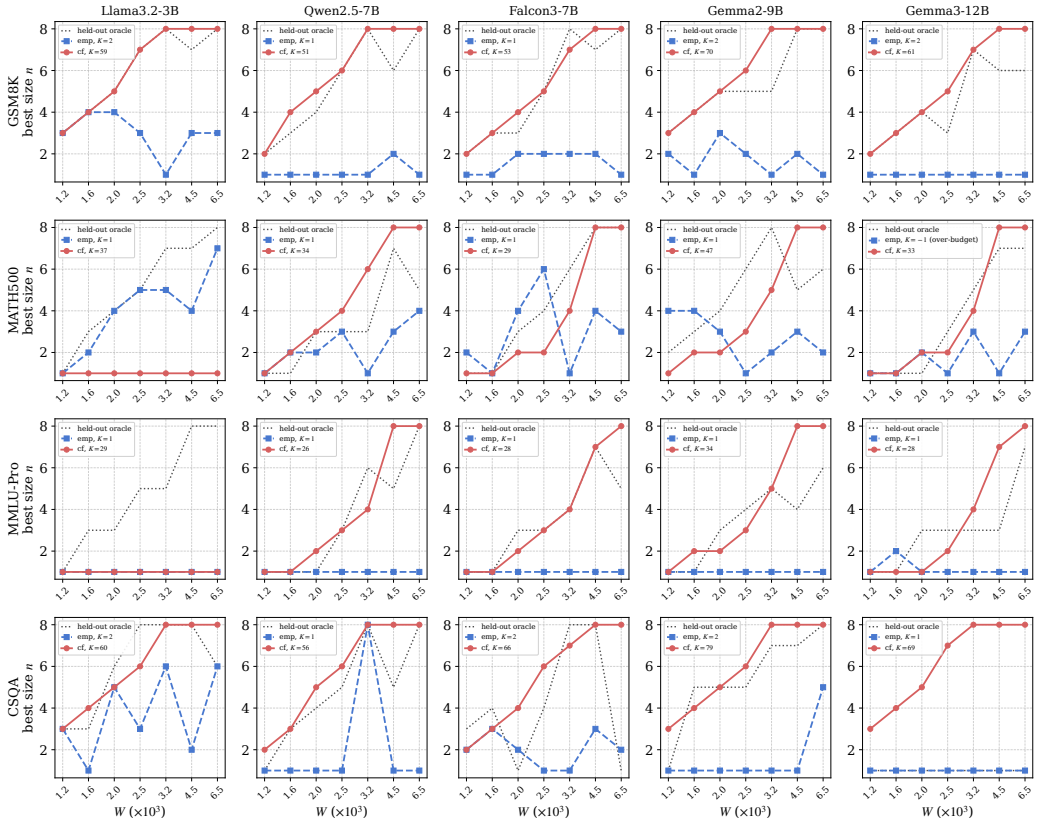


Figure 7: Predicted $n^*(W)$ (red) against empirical $\hat{n}^*(W)$ (green) at 1M calibration tokens.

Fourth, additional revision rounds increase visibility and may shift the trade-off if their revision benefit does not offset the added context cost; Appendix D.1 explores R up to 3 for self-refinement.

Fifth, the whole account assumes one window pays for reading and generation. Under per-agent contexts, reading does not consume generation room, m = 0 in our accounting, and the interior optimum disappears (§ 4.3).

Table 20: Held-out regret (percentage points) by calibration window W, each method's own chosen n averaged over all five models and four tasks. Complements Table 2, which aggregates over W instead. **Bold** is the lowest regret in a row.

	Debate-calibrated			*No calibration*		*Probe-calibrated*			
W	**Halv.**	**Emp.**	**Surf.**	$n_{\max}$	n=1	**Mix**	**Star**	**Learn**	**TB-MAD**
1200	1.03	**0.82**	2.38	58.82	1.56	1.65	2.16	1.91	1.12
1600	1.16	0.98	2.44	54.23	2.58	1.34	2.27	1.97	**0.96**
2000	**1.09**	1.17	2.80	42.72	3.81	1.89	2.59	1.34	1.49
2500	**0.59**	1.04	3.18	25.57	4.16	3.20	2.22	1.30	1.75
3200	**0.62**	1.12	3.32	2.93	5.36	3.93	1.59	0.72	0.87
4500	0.99	1.56	2.49	**0.84**	5.69	3.76	1.17	1.16	1.18
6500	1.27	1.04	0.66	**0.58**	5.96	4.69	1.02	1.02	1.02

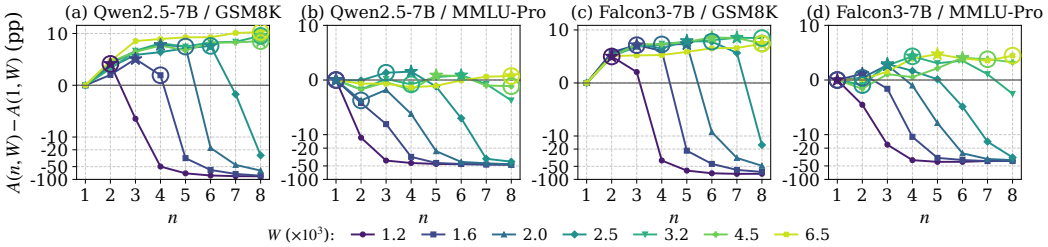


Figure 8: Synergy gap $A(n, W) - A(1, W)$ against n under the shared window W of the standard debate protocol. Stars mark the empirical best size $\hat{n}^*(W)$, measured directly on the grid; open circles mark $n^*(W)$ predicted from single-agent probes alone, with no debate data.

Table 21: Range of c_w over grid of n and g. The first three columns are the same pooled estimate at 1, 8 and 48 samples per question. The last two columns are held-out regret against the per-cell oracle when n^* is chosen with the fixed cell-level c_w against the per-n estimate $c_w(n)$.

		$\hat{c}_w$ by samples per question			$\hat{c}_w$ range		n^*	regret (pp)	
Model	**Task**	1	8	48	**over** g	**over** n	**moved**	c_w **fixed**	$c_w(n)$
Llama3.2-3B	GSM8K	0.00	0.00	0.14	0.00–0.34	0.29–0.33	0/7	0.28	0.28
	MATH500	0.41	0.27	0.20	0.24–0.28	0.24–0.28	4/7	2.71	0.79
	MMLU-Pro	0.03	0.02	0.21	0.02–0.44	0.42–0.45	6/7	4.40	0.65
	CSQA	0.00	0.00	0.32	0.00–0.77	0.72–0.77	0/7	0.52	0.52
Qwen2.5-7B	GSM8K	0.05	0.05	0.15	0.05–0.49	0.35–0.43	0/7	0.50	0.50
	MATH500	0.38	0.31	0.30	0.30–0.54	0.51–0.56	2/7	0.78	0.79
	MMLU-Pro	0.08	0.08	0.25	0.07–0.63	0.58–0.61	2/7	1.09	1.50
	CSQA	0.00	0.00	0.24	0.00–0.90	0.88–0.90	0/7	0.71	0.71
Falcon3-7B	GSM8K	0.05	0.04	0.11	0.04–0.31	0.35–0.40	0/7	0.20	0.20
	MATH500	0.37	0.29	0.28	0.29–0.45	0.44–0.46	1/7	1.31	1.42
	MMLU-Pro	0.10	0.08	0.24	0.07–0.57	0.54–0.58	1/7	1.12	0.57
	CSQA	0.02	0.02	0.34	0.02–0.85	0.81–0.83	1/7	1.11	1.23
Gemma2-9B	GSM8K	0.00	0.00	0.15	0.00–0.43	0.40–0.42	1/7	0.20	0.20
	MATH500	0.14	0.12	0.20	0.11–0.31	0.30–0.35	3/7	0.81	0.12
	MMLU-Pro	0.02	0.02	0.30	0.02–0.65	0.63–0.67	3/7	0.80	0.41
	CSQA	0.01	0.01	0.37	0.01–0.87	0.82–0.84	0/7	0.51	0.51
Gemma3-12B	GSM8K	0.02	0.02	0.14	0.02–0.78	0.73–0.77	0/7	0.30	0.30
	MATH500	0.47	0.37	0.35	0.36–0.67	0.52–0.57	1/7	0.33	0.22
	MMLU-Pro	0.06	0.05	0.23	0.05–0.72	0.58–0.64	1/7	1.70	1.31
	CSQA	0.00	0.00	0.30	0.00–0.93	0.87–0.91	0/7	3.80	3.80
Mean		0.11	0.09	0.24			26/140	1.16	0.80

Table 22: Where the arg max of Algorithm 1 moves when the fitted constant σ is replaced by the derived $\sigma(n)$, per model-task cell across the 7 swept W (columns, in tokens). **Moved** counts the windows where $\hat{n}^*$ differs; each numbered column gives the signed shift at that W ($\hat{n}^*_{\text{derived}} - \hat{n}^*_{\text{const}}$), 0 where the two choices of σ agree.

Model	**Task**	**Moved**	$\hat{n}^*_{\text{derived}} - \hat{n}^*_{\text{const}}$ **by** W						
			1200	**1600**	**2000**	**2500**	**3200**	**4500**	**6500**
Llama3.2-3B	GSM8K	0/7	0	0	0	0	0	0	0
	MATH500	1/7	0	0	0	+1	0	0	0
	MMLU-Pro	0/7	0	0	0	0	0	0	0
	CSQA	0/7	0	0	0	0	0	0	0
Qwen2.5-7B	GSM8K	1/7	0	-1	0	0	0	0	0
	MATH500	0/7	0	0	0	0	0	0	0
	MMLU-Pro	0/7	0	0	0	0	0	0	0
	CSQA	0/7	0	0	0	0	0	0	0
Falcon3-7B	GSM8K	1/7	0	0	0	0	-1	0	0
	MATH500	1/7	0	-1	0	0	0	0	0
	MMLU-Pro	0/7	0	0	0	0	0	0	0
	CSQA	0/7	0	0	0	0	0	0	0
Gemma2-9B	GSM8K	0/7	0	0	0	0	0	0	0
	MATH500	0/7	0	0	0	0	0	0	0
	MMLU-Pro	0/7	0	0	0	0	0	0	0
	CSQA	0/7	0	0	0	0	0	0	0
Gemma3-12B	GSM8K	2/7	0	0	0	-1	-1	0	0
	MATH500	1/7	0	0	-1	0	0	0	0
	MMLU-Pro	0/7	0	0	0	0	0	0	0
	CSQA	0/7	0	0	0	0	0	0	0

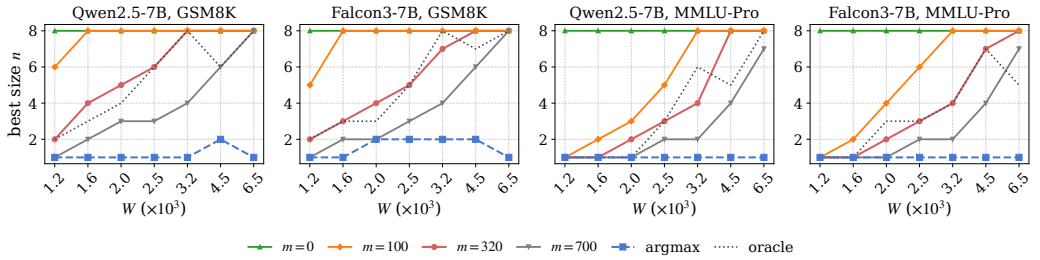


Figure 9: Predicted $n^*(W)$ from the closed form against empirical argmax and the true held-out best (oracle), all matched to a 1M-token calibration budget.

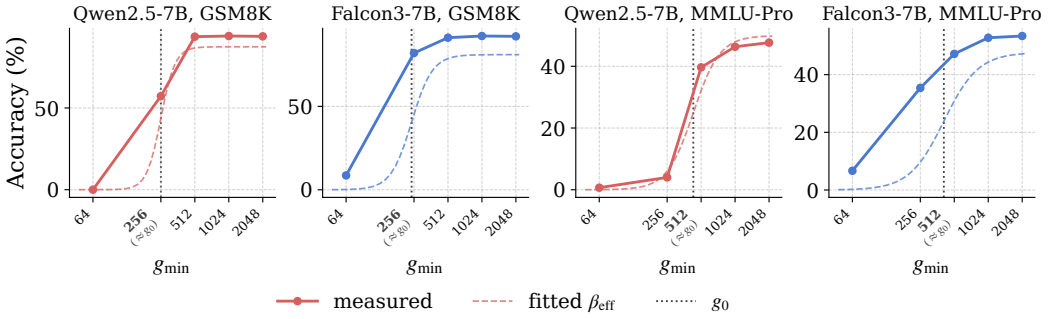


Figure 10: Accuracy against the generation floor $g_{\min}$ at the worst cell (n=8, W=1200), Qwen2.5-7B and Falcon3-7B on GSM8K and MMLU-Pro.

Table 23: Effect of question position: question first (qf) versus buried after the peer messages (qb). Δ=$A_{\mathrm{qf}} - A_{\mathrm{qb}}$ (%).

		grid mean			collapse cell (n=8, W=1200)		
Model	Task	A_{qf}	A_{qb}	Δ	A_{qf}	A_{qb}	Δ
Falcon3-7B	GSM8K	79.2	77.9	+1.24	9.0	0.2	+8.80
	MMLU-Pro	42.1	41.6	+0.48	6.9	1.3	+5.67
Qwen2.5-7B	GSM8K	78.0	78.2	-0.24	0.0	0.1	-0.07
	MMLU-Pro	37.4	36.9	+0.50	0.7	0.5	+0.20
Average		59.2	58.7	+0.49	4.2	0.5	+3.65

Algorithm 1 Predicting group accuracy and the best size at a deployment window

Require: $(p_{\max}, g_0, \kappa)$ from the single-agent sweep, Eq. 15
Require: $\hat{\rho} = \mathrm{corr}(X_i, X_j)$, $X_i = \mathbf{1}$ [agent i answers correctly at round 0]
Require: tie penalty c_w
Require: deployment window W, protocol (q, m, c), cap $n_{\max}$
for $n \in [1, n_{\max}]$ **continuous relaxation do**
  $g \leftarrow \mathrm{clip}\big(W - q - (n-1)m - c, ; g_{\min}, ; g_{\mathrm{cap}}\big)$
  $\beta_{\mathrm{eff}} \leftarrow p_{\max}, \phi(g; g_0, \kappa)$
  $\delta \leftarrow (1 + c_w)\beta_{\mathrm{eff}} - c_w$
  $n_{\mathrm{eff}} \leftarrow n/(1 + \hat{\rho}(n-1))$
  $S(n) \leftarrow \delta\sqrt{n_{\mathrm{eff}}}$
end for
$\hat{n}^* \leftarrow \lfloor \arg\max_{n \in [1, n_{\max}]} S(n) \rfloor$
return $\hat{n}^*$

Algorithm 2 Grid measurement: empirical argmax and the quadratic surface

Require: calibration budget B in output-token equivalents; sizes $\mathcal{N}$; windows $\mathcal{W}$; question pool $\mathcal{Q}$
Require: per-cell price $c(n, W) = \mathrm{in}(n, W) + \mathrm{out}(n, W)$, ▷ symmetric pricing
$u \leftarrow \sum_{(n,W) \in \mathcal{N} \times \mathcal{W}} c(n, W)$; $K \leftarrow \min\{|\mathcal{Q}|, \lfloor B/u \rfloor\}$ ▷ requires $B \geq u$
draw $\mathcal{D} \subseteq \mathcal{Q}$ uniformly without replacement, $|\mathcal{D}| = K$
for $(n, W) \in \mathcal{N} \times \mathcal{W}$ **do**
  run the n-agent debate on $\mathcal{D}$ at W and record $\hat{A}(n, W)$
end for
$\hat{n}_{\mathrm{emp}}(W) \leftarrow \min \arg\max_{n \in \mathcal{N}} \hat{A}(n, W)$ for each $W \in \mathcal{W}$ ▷ **empirical argmax**
$\psi(n, W) \leftarrow (1, n, W, n^2, W^2, nW)^\top$ ▷ requires $|\mathcal{N}|\,|\mathcal{W}| \geq 6$
$\hat{\theta} \leftarrow \arg\min_{\theta \in \mathbb{R}^6} \sum_{(n,W) \in \mathcal{N} \times \mathcal{W}} \big(\hat{A}(n, W) - \theta^\top \psi(n, W)\big)^2$
$\hat{n}_{\mathrm{quad}}(W) \leftarrow \min \arg\max_{n \in \mathcal{N}} \hat{\theta}^\top \psi(n, W)$ for each $W \in \mathcal{W}$ ▷ **quadratic surface**
return $\hat{n}_{\mathrm{emp}}(W),\ \hat{n}_{\mathrm{quad}}(W)$

Algorithm 3 Successive halving

Require: calibration budget B; sizes $\mathcal{N}$; windows $\mathcal{W}$; question pool $\mathcal{Q}$; price $c(n, W)$

$R \leftarrow \lceil \log_2 |\mathcal{N}| \rceil$; $\quad B_W \leftarrow B/|\mathcal{W}|$
for $W \in \mathcal{W}$ **do**
  $\mathcal{S} \leftarrow \mathcal{N}$; $\quad b \leftarrow B_W$; $\quad \mathcal{P} \leftarrow \mathcal{Q}$; $\quad$ ran $\leftarrow$ false
  $k \leftarrow \max\{1, \lfloor B_W / (R \sum_{n \in \mathcal{S}} c(n, W)) \rfloor\}$
  while $|\mathcal{S}| > 1$ **do**
    $C \leftarrow k \sum_{n \in \mathcal{S}} c(n, W)$
    if $C > b$ **or** $k > |\mathcal{P}|$ **then**
      break
    end if
    draw k questions from $\mathcal{P}$ without replacement; remove them from $\mathcal{P}$
    run every $n \in \mathcal{S}$ at W on these questions; update $\hat{A}(n, W)$ as the running mean over all questions seen by n
    $b \leftarrow b - C$; $\quad$ ran $\leftarrow$ true
    $\mathcal{S} \leftarrow$ the $\lceil |\mathcal{S}|/2 \rceil$ arms with highest $\hat{A}(\cdot, W)$, ties to smaller n; $\quad k \leftarrow 2k$
  end while
  $\hat{n}(W) \leftarrow \min \arg\max_{n \in \mathcal{S}} \hat{A}(n, W)$ if ran, else $\min \mathcal{N}$ ▷ budget too small for one round
end for
return $\hat{n}(W)$

Algorithm 4 Closed-form aggregation baselines: learned β_{eff} and star exponent

Require: calibration budget B; probe caps $\mathcal{G}$ ($|\mathcal{G}|$=7, 64–4096 log-spaced); question pool $\mathcal{Q}$
Require: price $c(g) = r \cdot \text{in}(g) + \text{out}(g)$, r=1; measured $\hat{\rho}$; tie penalty c_w
Require: protocol (q, m, c); sizes $\mathcal{N}$; windows $\mathcal{W}$; room limits $g_{\min}, g_{\text{cap}}$

$u \leftarrow \sum_{g \in \mathcal{G}} c(g)$; $\quad K \leftarrow \min\{|\mathcal{Q}|, \lfloor B/u \rfloor\}$ ▷ requires $B \geq u$
draw $\mathcal{D} \subseteq \mathcal{Q}$ uniformly without replacement, $|\mathcal{D}| = K$
run one agent on $\mathcal{D}$ at every $g \in \mathcal{G}$; record accuracy $\hat{A}(g)$ ▷ no debate is run
$\hat{\beta}_{\text{gbt}} \leftarrow \text{GBT}\big(\{(g, \hat{A}(g)) : g \in \mathcal{G}\}\big)$ ▷ 100 trees, depth 3, learning rate 0.1
$(\hat{p}_{\max}, \hat{g}_0, \hat{\kappa}) \leftarrow \arg\min_{p_{\max}, g_0, \kappa} \sum_{g \in \mathcal{G}} \big(\hat{A}(g) - p_{\max}\big(1 - (1 + g/g_0)^{-\kappa}\big)\big)^2$ ▷ POW of Table 1
$\hat{\beta}_{\text{pow}}(g) \leftarrow \hat{p}_{\max}\big(1 - (1 + g/\hat{g}_0)^{-\hat{\kappa}}\big)$
for $W \in \mathcal{W}$, $n \in \mathcal{N}$ **do**
  $g \leftarrow \text{clip}\big(W - q - (n-1)m - c, g_{\min}, g_{\text{cap}}\big)$; $\quad n_{\text{eff}} \leftarrow n/\big(1 + \hat{\rho}(n-1)\big)$
  for $x \in \{\text{gbt}, \text{pow}\}$ **do**
    $\delta_x \leftarrow (1 + c_w)\hat{\beta}_x(g) - c_w$; $\quad S_x(n, W) \leftarrow \delta_x \sqrt{n_{\text{eff}}}$
  end for
end for
return $\hat{n}_x(W) = \min \arg\max_{n \in \mathcal{N}} S_x(n, W)$ for each $x \in \{\text{gbt}, \text{pow}\}$

Algorithm 5 Difficulty mixture, with our window accounting

Require: calibration budget B; probe caps $\mathcal{G}$; question pool $\mathcal{Q}$; price $c(g)$
Require: samples per cap $S{=}8$; simulation draws $L{=}400$; clip level $\epsilon{=}10^{-6}$
Require: sizes $\mathcal{N}$; windows $\mathcal{W}$; protocol (q, m, c); room limits $g_{\min}, g_{\text{cap}}$; measured $\hat{\rho}$; tie penalty c_w

$u \leftarrow S \sum_{g \in \mathcal{G}} c(g)$; $\quad K \leftarrow \min\{|\mathcal{Q}|, \lfloor B/u \rfloor\}$ ▷ requires $B \geq u$
draw $\mathcal{D} \subseteq \mathcal{Q}$ uniformly without replacement, $|\mathcal{D}| = K$
run one agent S times on each $i \in \mathcal{D}$ at every $g \in \mathcal{G}$; $\hat{p}_i(g) \leftarrow$ fraction correct
for $W \in \mathcal{W}$, $n \in \mathcal{N}$ **do**
  $g \leftarrow \text{clip}\big(W - q - (n-1)m - c,\ g_{\min},\ g_{\text{cap}}\big)$
  for $i \in \mathcal{D}$ **do**
    $p_i \leftarrow \text{clip}\big(\text{linear interpolation of } \hat{p}_i \text{ in } \log g,\ \epsilon,\ 1-\epsilon\big)$
    for $\ell = 1, \ldots, L$ **do**
      $\zeta, \varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \text{Normal}(0,1)$; $\quad z_j \leftarrow \sqrt{\hat{\rho}}\,\zeta + \sqrt{1-\hat{\rho}}\,\varepsilon_j$
      $h \leftarrow \#\{j : z_j < \Phi^{-1}(p_i)\}$; $\quad r \sim \text{Bin}(n-h,\ c_w)$
      $\text{win}_{i\ell} \leftarrow \mathbf{1}[h > r] + \frac{1}{2}\,\mathbf{1}[h = r]$
    end for
  end for
  $\hat{A}(n, W) \leftarrow \frac{1}{KL} \sum_{i \in \mathcal{D}} \sum_{\ell=1}^{L} \text{win}_{i\ell}$
end for
return $\hat{n}(W) = \min \arg\max_{n \in \mathcal{N}} \hat{A}(n, W)$

Algorithm 6 Predicting the best size and model subset on a heterogeneous set

Require: per-model $(p_{\max,i}, g_{0,i}, \kappa_i)$ from the single-agent sweep, Eq. 15, $i = 1, \ldots, K$
Require: within/between round-0 correlations $\hat{\rho}_w, \hat{\rho}_b$; tie penalty c_w
Require: deployment window W, protocol (q, m, c), cap $n_{\max}$

for $n \in \{K, 2K, \ldots, n_{\max}\}$ **do**
  $g \leftarrow \text{clip}\big(W - q - (n-1)m - c,\ g_{\min},\ g_{\text{cap}}\big)$
  $\delta_i \leftarrow (1 + c_w)\, p_{\max,i}\, \phi_i(g; g_{0,i}, \kappa_i) - c_w$ **for** $i = 1, \ldots, K$
  $\mathcal{S} \leftarrow \big\{\, \{1, \ldots, K\} \,\big\}$
  if $\min_i \delta_i \leq 0 < \max_i \delta_i$ **then** ▷ edges mixed in sign: a subset may win
    $\mathcal{S} \leftarrow \mathcal{S} \cup \big\{\, \{i\} : \delta_i > 0 \,\big\}$
  end if
  for $S \in \mathcal{S}$ **do**
    $\bar{\delta}_S \leftarrow |S|^{-1} \sum_{i \in S} \delta_i$
    $n_{\text{eff}} \leftarrow n \big/ \big(1 + \hat{\rho}_w(n/|S| - 1) + \hat{\rho}_b(n - n/|S|)\big)$
    $\mathcal{V}(n, S) \leftarrow \text{sign}(\bar{\delta}_S)\, n_{\text{eff}}\, \bar{\delta}_S^2$
  end for
end for
$(\hat{n}^*, \hat{S}) \leftarrow \arg\max_{(n,S)} \mathcal{V}(n, S)$
return $\hat{n}^*,\ \hat{S}$

Table 24: Every cell behind Table 2 at the 1.0M budget: held-out regret (percentage points) against the per-cell **Oracle** in the last column. Columns are SH successive halving, EA empirical argmax, QS quadratic surface, LA learned ability, SE star exponent, DM difficulty mixture, TB ours.

Model	**Dataset**	QS	EA	SH	$n{=}n_{\max}$	$n{=}1$	DM	SE	LA	**TB**	**Oracle**
Llama3.2-3B	GSM8K	5.20	7.04	5.84	19.22	13.08	7.81	5.70	2.26	0.41	79.34
	MATH500	2.76	5.57	10.13	15.38	8.23	8.23	3.31	3.93	3.48	43.43
	MMLU-Pro	4.52	3.33	6.21	14.74	4.42	4.43	3.71	3.60	4.42	37.22
	CSQA	4.35	3.59	6.32	22.09	5.80	2.77	2.56	1.08	0.87	69.87
Qwen2.5-7B	GSM8K	4.14	6.57	11.20	39.78	7.83	0.91	1.81	1.40	2.30	92.10
	MATH500	3.97	3.45	20.37	38.52	2.89	2.89	1.28	1.38	1.12	70.46
	MMLU-Pro	0.69	1.20	11.52	27.29	0.78	0.78	0.69	1.08	0.96	49.43
	CSQA	1.03	1.48	5.13	36.25	1.37	0.70	0.85	0.71	0.78	82.07
Falcon3-7B	GSM8K	2.35	6.34	8.64	33.32	7.89	0.77	1.58	0.95	0.29	90.88
	MATH500	3.62	3.12	15.99	32.95	3.35	3.71	1.19	1.31	1.03	63.92
	MMLU-Pro	2.07	2.55	8.18	22.72	2.76	2.76	1.98	1.33	1.25	51.66
	CSQA	1.67	1.45	3.90	20.14	1.48	1.22	1.13	0.71	1.15	74.71
Gemma2-9B	GSM8K	2.17	4.17	1.45	15.53	5.87	3.15	0.93	0.26	0.24	86.54
	MATH500	5.99	4.32	11.77	16.00	6.55	6.55	3.83	2.30	2.70	49.92
	MMLU-Pro	1.74	1.63	8.02	22.31	2.03	2.03	1.91	1.69	1.71	52.76
	CSQA	2.03	2.63	1.30	13.11	2.83	2.43	1.57	0.83	0.96	75.88
Gemma3-12B	GSM8K	1.16	1.61	8.83	36.79	1.67	1.54	0.54	2.08	0.98	93.84
	MATH500	0.76	1.28	23.51	37.77	0.99	0.95	0.43	0.50	0.60	73.67
	MMLU-Pro	2.60	1.91	10.60	33.35	1.42	1.42	0.97	2.80	2.69	59.49
	CSQA	4.57	0.84	2.61	38.90	0.05	1.29	2.94	3.41	3.35	77.68

Table 25: Every cell behind Table 2 at the 2.0M budget. Columns and entries as in Table 24.

Model	**Dataset**	QS	EA	SH	$n{=}n_{\max}$	$n{=}1$	DM	SE	LA	**TB**	**Oracle**
Llama3.2-3B	GSM8K	5.04	4.26	3.40	19.22	13.08	5.15	5.70	1.80	0.41	79.34
	MATH500	4.83	6.22	6.31	15.38	8.23	8.23	3.54	2.69	1.80	43.43
	MMLU-Pro	2.58	3.40	9.52	14.74	4.42	4.42	4.42	4.05	4.42	37.22
	CSQA	4.54	4.04	2.84	22.09	5.80	3.67	2.50	1.03	0.87	69.87
Qwen2.5-7B	GSM8K	4.81	4.03	1.44	39.78	7.83	1.06	1.80	1.39	0.46	92.10
	MATH500	1.95	2.34	6.05	38.52	2.89	2.89	1.28	1.31	1.10	70.46
	MMLU-Pro	0.75	1.55	10.47	27.29	0.78	0.83	0.78	0.88	0.96	49.43
	CSQA	1.45	1.58	1.76	36.25	1.37	0.92	0.85	0.60	0.79	82.07
Falcon3-7B	GSM8K	5.34	4.32	1.53	33.32	7.89	1.02	1.58	0.69	0.29	90.88
	MATH500	2.66	2.43	10.63	32.95	3.35	3.35	1.19	1.37	1.04	63.92
	MMLU-Pro	2.28	2.40	12.13	22.72	2.76	2.87	1.46	1.52	0.89	51.66
	CSQA	1.52	1.92	1.26	20.14	1.48	1.41	1.19	0.62	1.15	74.71
Gemma2-9B	GSM8K	2.05	3.98	2.89	15.53	5.87	4.03	0.93	0.19	0.22	86.54
	MATH500	2.98	2.90	4.05	16.00	6.55	6.55	3.89	2.30	2.66	49.92
	MMLU-Pro	1.63	2.54	5.43	22.31	2.03	2.03	1.51	0.87	0.77	52.76
	CSQA	1.61	2.77	1.16	13.11	2.83	2.83	1.57	1.05	0.94	75.88
Gemma3-12B	GSM8K	1.20	1.67	0.42	36.79	1.67	1.52	0.55	1.47	0.37	93.84
	MATH500	0.40	2.57	11.89	37.77	0.99	0.99	0.43	0.50	0.60	73.67
	MMLU-Pro	1.34	1.76	14.34	33.35	1.42	1.44	1.07	2.06	1.57	59.49
	CSQA	0.69	0.47	8.70	38.90	0.05	1.57	2.94	3.35	3.35	77.68

Table 26: Every cell behind Table 19 at the 4.0M budget. Columns and entries as in Table 24.

Model	**Dataset**	QS	EA	SH	$n{=}n_{\max}$	$n{=}1$	DM	SE	LA	**TB**	**Oracle**
Llama3.2-3B	GSM8K	3.22	4.34	2.16	19.22	13.08	4.30	5.70	1.59	0.41	79.34
	MATH500	3.93	1.85	3.50	15.38	8.23	8.23	3.20	3.04	1.92	43.43
	MMLU-Pro	2.60	2.17	2.15	14.74	4.42	4.42	4.42	4.32	4.42	37.22
	CSQA	3.51	3.36	1.76	22.09	5.80	2.77	2.49	0.96	0.87	69.87
Qwen2.5-7B	GSM8K	4.57	2.69	2.43	39.78	7.83	0.99	1.80	1.98	0.43	92.10
	MATH500	2.10	2.51	1.77	38.52	2.89	2.89	1.28	1.51	1.09	70.46
	MMLU-Pro	1.65	2.16	7.11	27.29	0.78	0.78	0.71	1.32	0.98	49.43
	CSQA	1.25	1.96	0.92	36.25	1.37	0.80	0.85	0.57	0.76	82.07
Falcon3-7B	GSM8K	5.42	2.89	1.44	33.32	7.89	1.19	1.58	0.91	0.29	90.88
	MATH500	1.85	1.75	6.34	32.95	3.35	3.35	1.19	1.31	0.94	63.92
	MMLU-Pro	1.46	3.97	8.85	22.72	2.76	2.76	1.50	1.46	0.92	51.66
	CSQA	1.96	1.45	0.92	20.14	1.48	1.30	1.13	0.48	1.22	74.71
Gemma2-9B	GSM8K	1.32	1.68	0.95	15.53	5.87	4.66	0.83	0.21	0.20	86.54
	MATH500	3.11	1.65	2.27	16.00	6.55	6.55	3.06	2.02	2.53	49.92
	MMLU-Pro	1.52	2.08	1.78	22.31	2.03	2.03	1.43	1.20	0.88	52.76
	CSQA	2.15	2.19	1.25	13.11	2.83	2.83	1.57	0.89	0.95	75.88
Gemma3-12B	GSM8K	1.12	0.84	0.96	36.79	1.67	1.48	0.57	1.58	0.37	93.84
	MATH500	0.44	1.14	8.33	37.77	0.99	0.99	0.43	0.50	0.60	73.67
	MMLU-Pro	1.08	1.41	4.19	33.35	1.42	1.42	1.00	2.06	1.51	59.49
	CSQA	0.69	1.51	1.61	38.90	0.05	1.40	2.94	3.35	3.35	77.68

Table 27: Every cell behind Table 19 at the 8.0M budget. Columns and entries as in Table 24.

Model	**Dataset**	QS	EA	SH	$n{=}n_{\max}$	$n{=}1$	DM	SE	LA	**TB**	**Oracle**
Llama3.2-3B	GSM8K	5.11	2.08	2.49	19.22	13.08	4.30	5.70	1.59	0.41	79.34
	MATH500	3.77	2.74	2.76	15.38	8.23	8.23	3.20	3.04	2.62	43.43
	MMLU-Pro	2.74	1.99	3.65	14.74	4.42	4.42	4.42	4.32	4.42	37.22
	CSQA	4.80	2.80	2.42	22.09	5.80	2.77	2.49	0.96	0.87	69.87
Qwen2.5-7B	GSM8K	4.92	1.80	1.09	39.78	7.83	0.99	1.80	1.98	0.43	92.10
	MATH500	2.08	1.76	1.03	38.52	2.89	2.89	1.28	1.51	1.10	70.46
	MMLU-Pro	0.54	1.85	2.06	27.29	0.78	0.78	0.71	1.21	1.01	49.43
	CSQA	0.99	0.99	1.33	36.25	1.37	0.80	0.85	0.57	0.76	82.07
Falcon3-7B	GSM8K	4.79	1.88	1.26	33.32	7.89	1.19	1.58	0.91	0.29	90.88
	MATH500	2.15	1.70	1.23	32.95	3.35	3.35	1.19	1.31	0.94	63.92
	MMLU-Pro	2.35	1.93	1.32	22.72	2.76	2.76	1.46	1.44	1.03	51.66
	CSQA	1.70	1.04	1.39	20.14	1.48	1.30	1.13	0.48	1.22	74.71
Gemma2-9B	GSM8K	1.07	1.28	1.29	15.53	5.87	4.66	0.83	0.21	0.20	86.54
	MATH500	1.70	1.22	1.39	16.00	6.55	6.55	2.72	1.57	2.66	49.92
	MMLU-Pro	1.58	1.12	1.23	22.31	2.03	2.03	1.47	0.94	0.86	52.76
	CSQA	2.09	1.85	1.32	13.11	2.83	2.83	1.57	0.89	0.95	75.88
Gemma3-12B	GSM8K	1.22	0.83	0.23	36.79	1.67	1.48	0.57	1.58	0.37	93.84
	MATH500	0.47	1.61	1.05	37.77	0.99	0.99	0.43	0.50	0.60	73.67
	MMLU-Pro	1.32	1.45	2.25	33.35	1.42	1.42	1.07	2.06	1.81	59.49
	CSQA	0.78	1.48	2.15	38.90	0.05	1.40	2.94	3.35	3.35	77.68

Table 28: Every cell behind Table 19 at the 16.0M budget. Columns and entries as in Table 24.

Model	**Dataset**	QS	EA	SH	$n{=}n_{\max}$	$n{=}1$	DM	SE	LA	**TB**	**Oracle**
Llama3.2-3B	GSM8K	3.68	1.13	0.75	19.22	13.08	4.30	5.70	1.59	0.41	79.34
	MATH500	4.34	1.84	1.41	15.38	8.23	8.23	3.20	3.04	2.62	43.43
	MMLU-Pro	2.35	1.08	1.67	14.74	4.42	4.42	4.42	4.32	4.42	37.22
	CSQA	4.21	1.86	1.53	22.09	5.80	2.77	2.49	0.96	0.87	69.87
Qwen2.5-7B	GSM8K	4.74	1.37	0.94	39.78	7.83	0.99	1.80	1.98	0.43	92.10
	MATH500	2.17	1.39	0.99	38.52	2.89	2.89	1.28	1.51	1.10	70.46
	MMLU-Pro	0.54	2.29	1.19	27.29	0.78	0.78	0.71	1.21	1.01	49.43
	CSQA	1.01	0.94	0.98	36.25	1.37	0.80	0.85	0.57	0.76	82.07
Falcon3-7B	GSM8K	5.13	1.43	0.69	33.32	7.89	1.19	1.58	0.91	0.29	90.88
	MATH500	2.28	1.46	1.04	32.95	3.35	3.35	1.19	1.31	0.94	63.92
	MMLU-Pro	2.05	0.69	1.33	22.72	2.76	2.76	1.46	1.44	1.03	51.66
	CSQA	1.71	1.20	1.00	20.14	1.48	1.30	1.13	0.48	1.22	74.71
Gemma2-9B	GSM8K	1.90	1.35	0.60	15.53	5.87	4.66	0.83	0.21	0.20	86.54
	MATH500	1.97	1.10	0.82	16.00	6.55	6.55	2.72	1.57	2.66	49.92
	MMLU-Pro	1.71	1.63	1.67	22.31	2.03	2.03	1.47	0.94	0.86	52.76
	CSQA	2.17	1.51	1.23	13.11	2.83	2.83	1.57	0.89	0.95	75.88
Gemma3-12B	GSM8K	1.31	1.20	0.59	36.79	1.67	1.48	0.57	1.58	0.37	93.84
	MATH500	0.47	1.01	0.66	37.77	0.99	0.99	0.43	0.50	0.60	73.67
	MMLU-Pro	1.46	1.35	2.23	33.35	1.42	1.42	1.07	2.06	1.81	59.49
	CSQA	0.67	1.22	1.11	38.90	0.05	1.40	2.94	3.35	3.35	77.68

Table 29: Every cell behind Table 13 at the 1.0M budget: held-out regret (percentage points) against the widened per-cell oracle (set-best or any single member). Abbreviations as in Table 24, $n{=}3$–8 the fixed-size rules.

Model set	**Dataset**	SH	QS	EA	$n{=}8$	$n{=}6$	$n{=}3$	$n{=}4$	TB w/o gate	DM	SE	LA	**TB**
Qwen2.5-7B+Gemma3-12B+Llama3.2-3B	GSM8K	0.47	0.59	0.80	–	15.18	0.93	–	1.50	0.17	0.42	0.13	0.13
	MATH500	3.65	2.52	3.16	–	19.65	3.54	–	1.98	3.54	2.10	1.27	1.03
	MMLU-Pro	6.61	6.31	3.48	–	11.78	5.59	–	3.39	5.59	3.42	3.02	3.39
	CSQA	1.15	0.91	1.13	–	11.99	1.47	–	8.19	1.10	0.62	0.50	0.50
Qwen2.5-7B+Falcon3-7B+Gemma2-9B	GSM8K	2.02	0.79	0.62	–	13.06	1.17	–	1.63	0.00	0.40	0.02	0.02
	MATH500	5.10	3.98	4.22	–	20.73	4.14	–	1.91	4.13	3.51	3.38	2.00
	MMLU-Pro	0.94	0.99	0.89	–	13.85	1.51	–	2.98	1.51	0.88	0.80	2.23
	CSQA	4.30	3.87	1.78	–	8.56	1.56	–	2.79	1.81	1.60	1.75	1.75
Qwen2.5-7B+Falcon3-7B+Gemma2-9B+Gemma3-12B	GSM8K	1.82	1.28	1.30	22.17	–	–	1.30	1.60	1.21	1.15	1.09	1.09
	MATH500	5.40	4.23	5.05	26.67	–	–	4.40	5.07	4.23	3.80	3.80	4.31
	MMLU-Pro	9.44	0.98	2.24	18.13	–	–	1.48	6.75	1.38	1.37	0.86	4.77
	CSQA	1.25	0.76	0.98	15.90	–	–	1.00	0.70	0.70	0.76	0.70	0.70
Qwen2.5-7B+Falcon3-7B+Gemma3-12B+Llama3.2-3B	GSM8K	0.93	0.38	0.52	22.04	–	–	0.52	0.77	0.10	0.31	0.12	0.12
	MATH500	6.65	3.67	3.98	24.86	–	–	3.98	4.06	3.27	2.98	2.98	2.77
	MMLU-Pro	5.66	1.65	3.11	13.23	–	–	1.96	4.10	1.74	1.39	1.30	4.10
	CSQA	1.40	0.88	1.02	17.72	–	–	1.04	3.12	0.66	0.70	0.62	0.62

Table 30: Every cell behind Table 13 at the 2.0M budget. Columns and entries as in Table 29.

Model set	**Dataset**	SH	QS	EA	$n{=}8$	$n{=}6$	$n{=}3$	$n{=}4$	TB w/o gate	DM	SE	LA	**TB**
Qwen2.5-7B+Gemma3-12B+Llama3.2-3B	GSM8K	1.03	0.59	0.64	–	15.18	0.93	–	1.01	0.24	0.42	0.13	0.17
	MATH500	2.17	2.65	2.11	–	19.65	3.54	–	1.48	3.54	2.10	1.27	0.83
	MMLU-Pro	6.03	3.42	3.66	–	11.78	5.59	–	1.38	4.91	3.42	2.99	1.38
	CSQA	0.39	0.96	0.58	–	11.99	1.47	–	3.79	0.74	0.62	0.50	1.60
Qwen2.5-7B+Falcon3-7B+Gemma2-9B	GSM8K	0.56	0.83	0.87	–	13.06	1.17	–	1.89	0.22	0.40	0.02	0.02
	MATH500	3.98	3.94	3.52	–	20.73	4.14	–	0.83	4.14	3.51	3.43	2.43
	MMLU-Pro	0.23	1.08	0.69	–	13.85	1.51	–	5.20	1.25	0.88	0.18	1.97
	CSQA	1.97	1.40	2.10	–	8.56	1.56	–	1.96	1.54	1.60	1.75	1.75
Qwen2.5-7B+Falcon3-7B+Gemma2-9B+Gemma3-12B	GSM8K	1.58	1.28	1.20	22.17	–	–	1.30	1.10	1.09	1.15	1.09	1.02
	MATH500	5.18	4.23	4.12	26.67	–	–	4.40	3.26	3.94	3.80	3.80	3.95
	MMLU-Pro	3.30	1.38	1.81	18.13	–	–	1.48	3.81	1.27	1.10	0.57	2.52
	CSQA	1.03	0.75	0.97	15.90	–	–	1.00	3.34	0.68	0.76	0.70	0.70
Qwen2.5-7B+Falcon3-7B+Gemma3-12B+Llama3.2-3B	GSM8K	0.53	0.38	0.44	22.04	–	–	0.52	0.68	0.20	0.31	0.12	0.12
	MATH500	5.68	3.98	3.98	24.86	–	–	3.98	2.24	3.25	2.98	2.98	2.24
	MMLU-Pro	3.71	1.60	1.60	13.23	–	–	1.96	2.36	1.96	1.34	1.09	2.36
	CSQA	1.43	0.88	0.91	17.72	–	–	1.04	3.13	0.74	0.70	0.62	0.62

Table 31: Every cell behind Table 12 at the 1.0M budget: held-out regret (percentage points) against the widened per-cell oracle (set-best or any single member). Abbreviations as in Table 24.

Model pair	**Dataset**	SH	QS	EA	$n{=}n_{\max}$	$n{=}2$	DM	SE	LA	TB w/o gate	**TB**
Qwen2.5-7B+Gemma3-12B	GSM8K	6.34	1.78	2.34	32.07	2.34	0.26	0.80	0.61	0.47	0.51
	MATH500	12.75	2.80	3.46	36.14	3.99	3.17	0.84	0.80	2.28	0.84
	MMLU-Pro	10.71	3.47	3.26	28.72	5.65	4.71	4.36	3.85	1.83	1.75
	CSQA	6.05	2.37	2.67	26.34	2.67	1.64	1.69	0.97	0.77	0.77
Qwen2.5-7B+Llama3.2-3B	GSM8K	7.42	4.82	5.04	31.07	6.82	1.34	2.17	1.09	0.37	1.03
	MATH500	15.36	8.67	9.42	33.24	7.74	6.67	4.96	4.59	0.27	1.77
	MMLU-Pro	12.87	8.52	4.97	20.67	3.70	3.30	2.57	2.19	2.35	2.35
	CSQA	7.86	6.04	8.94	26.28	9.31	3.77	4.93	2.83	0.29	2.34
Falcon3-7B+Gemma2-9B	GSM8K	6.86	2.89	3.49	16.41	3.59	0.65	1.62	0.56	1.33	0.35
	MATH500	9.44	4.84	6.05	25.46	6.90	7.25	3.84	3.67	0.27	3.87
	MMLU-Pro	11.61	1.50	1.65	19.72	1.99	1.84	1.41	1.53	4.48	3.35
	CSQA	3.95	1.48	2.02	9.63	2.32	0.65	0.94	0.89	2.14	0.63
Gemma3-12B+Llama3.2-3B	GSM8K	5.94	1.66	2.28	22.43	2.28	1.59	1.15	0.67	0.71	0.67
	MATH500	10.49	2.37	2.83	32.39	3.15	2.99	1.59	2.66	0.28	0.28
	MMLU-Pro	12.53	1.85	2.60	19.88	2.98	2.41	1.71	1.19	4.63	3.73
	CSQA	0.89	4.66	4.56	16.69	5.21	1.77	1.94	0.69	4.01	0.69

Table 32: Every cell behind Table 12 at the 2.0M budget. Columns and entries as in Table 31.

Model pair	**Dataset**	SH	QS	EA	$n{=}n_{\max}$	$n{=}2$	DM	SE	LA	TB w/o gate	**TB**
Qwen2.5-7B+Gemma3-12B	GSM8K	2.88	1.58	2.05	32.07	2.34	0.54	0.89	0.56	1.19	0.51
	MATH500	5.57	2.87	3.23	36.14	3.99	3.99	0.84	0.95	2.89	0.61
	MMLU-Pro	8.15	11.57	4.28	28.72	5.65	4.91	1.82	2.15	1.06	1.33
	CSQA	2.40	2.40	2.26	26.34	2.67	1.78	1.69	0.86	0.77	0.77
Qwen2.5-7B+Llama3.2-3B	GSM8K	3.14	3.77	3.80	31.07	6.82	1.61	1.61	1.03	0.60	1.16
	MATH500	12.61	6.69	6.83	33.24	7.74	7.51	4.96	4.51	0.22	1.42
	MMLU-Pro	7.40	10.67	5.57	20.67	3.70	3.45	1.70	1.51	3.37	3.37
	CSQA	4.34	5.37	4.53	26.28	9.31	4.80	4.93	2.34	0.29	2.34
Falcon3-7B+Gemma2-9B	GSM8K	1.81	2.87	3.33	16.41	3.59	1.06	1.62	0.44	0.41	0.41
	MATH500	12.32	5.05	5.54	25.46	6.90	6.90	3.84	3.56	0.57	3.73
	MMLU-Pro	3.85	1.66	3.99	19.72	1.99	1.44	1.42	1.84	3.14	3.02
	CSQA	1.29	1.59	1.70	9.63	2.32	1.27	0.94	0.49	2.93	0.63
Gemma3-12B+Llama3.2-3B	GSM8K	3.47	1.77	2.18	22.43	2.28	1.12	1.12	0.67	0.77	0.67
	MATH500	8.69	2.45	3.15	32.39	3.15	3.15	1.59	2.59	0.28	0.87
	MMLU-Pro	9.42	10.49	5.33	19.88	2.98	2.67	1.36	1.22	1.17	1.17
	CSQA	0.43	1.51	2.08	16.69	5.21	1.43	1.94	0.77	3.23	0.69