Diversity Combining for Multi-Path LLM Reasoning

Guangsheng Yu
University of Technology Sydney

Litianyi Zhang
The University of Sydney

Qin Wang
CSIRO

Xu Wang
University of Technology Sydney

Mingyuan Li
ELLIS Institute Finland
University of Turku

Shaoxiong Ji
ELLIS Institute Finland
University of Turku

Ren Ping Liu
University of Technology Sydney

Massimo Piccardi
University of Technology Sydney

Abstract

Multi-path reasoning methods such as self-consistency (SC) sample K reasoning paths and choose the most frequent answer. However, their gains quickly plateau as K increases, and locating this saturation point currently requires fitting a scaling model to accuracies measured at several values of K or stopping online per query. We formalize multi-path LLM reasoning as a diversity combining problem from wireless communications: each path is a noisy channel observation, and the pairwise correlation of path correctness caps the design-effect effective sample size of the vote at a finite ceiling. Generalized least squares (GLS) analysis shows that, for paths with equal gains, the optimal linear combiner of latent embeddings is uniform exactly when every row of the path covariance has the same sum, as under exchangeability, supporting majority vote as the natural default in standard SC while leaving room for weighting or pruning under heterogeneous prompt-template branches. Across 5 models and 12 benchmarks, prompt-template diversity reduces path correlation in 55 of 60 model-benchmark cells, with a mean relative reduction of 16% and reductions in every task domain. We derive an Adaptive-K rule that uses a four-path pilot to select K^*, retaining 96–103% of MV@K=32 accuracy across Math, QA, and NLU. Our code is hosted at `https://github.com/OniReimu/DiversityCombining`.

1 Introduction

Multi-path reasoning has become a standard technique for improving LLM answer reliability. Self-consistency Wang et al. [2023] samples K chain-of-thought paths and returns the majority-vote answer; extensions include weighted voting Taubenfeld et al. [2025], tree-structured search Yao et al. [2023], and token-level embedding aggregation Xu et al. [2025].

However, a common empirical observation lacks a satisfactory theoretical explanation: accuracy gains from additional reasoning paths diminish rapidly, and beyond a model-dependent threshold K^*, adding more paths yields negligible improvement. Existing analyses attribute this saturation to answer-space coverage or sampling temperature, yet these explanations remain qualitative. The quantitative predictor of compound-inference scaling Chen et al. [2024] fits a query-difficulty model to accuracies measured at a few small values of K and extrapolates to larger budgets, and online stopping rules Aggarwal et al. [2023], Li et al. [2024] decide per query without a model-level K^*. They also do not establish a formal relationship between the observable path agreement rate and the underlying correlation structure, which is what a saturation formula requires.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

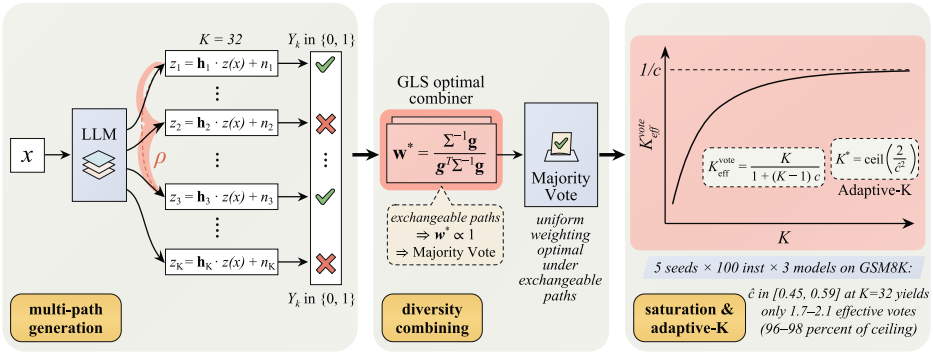


Figure 1: Diversity combining for multi-path LLM reasoning. The LLM generates K correlated paths $\mathbf{z}_k = h_k \mathbf{z}(x) + \mathbf{n}_k$ (h_k: reasoning quality; $\mathbf{n}_k$: errors; ρ: pairwise correlation), collapsed to correctness indicators Y_k. **Aggregation**: under exchangeability, the GLS-optimal linear combiner is uniform (Corollary 4.5), supporting uniform majority vote as the natural default. **Diagnostic**: $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)c)$ saturates at $1/c$; Adaptive-K estimates K^* from a $K{=}4$ pilot.

The classical Condorcet jury theorem Condorcet [1785] predicts that majority voting among independent voters, each correct with probability above $1/2$, converges to certainty as $K \to \infty$, a guarantee that dependence among voters can break Berg [1993]. Compound LLM inference indeed saturates much earlier Chen et al. [2024], indicating that path correctness is correlated across questions. While ensemble diversity measures Kuncheva and Whitaker [2003], Jeffares et al. [2023] quantify disagreement, they do not provide a closed-form saturation bound for LLM reasoning. We address this gap via an analogy to *diversity combining* over *multipath channels*: a receiver observes K noisy signal copies, and correlation among the copies reduces the combining gain they deliver Tse and Viswanath [2005], so that the effective rank of the correlation matrix, rather than K, counts the copies that act as independent observations. We show that the correctness of reasoning paths sampled from the same model and prompt is positively correlated across questions, and that this correlation limits the gain from increasing K. Fig. 1 illustrates the framework. The main contributions are:

- **LLM-specific saturation diagnostic.** Adapting the classical design-effect Kish [1965] and participation-ratio formulations to multi-path LLM reasoning, we cast reasoning paths as correlated branches and obtain the vote-level diagnostic $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)c)$ with ceiling $1/c$, where c is the pairwise correctness correlation, a vote-level overdispersion measurable from outputs alone. Across three model families on GSM8K, $K_{\text{eff}}^{\text{vote}}$ at $K{=}32$ reaches 96–98% of this ceiling (§4, §5).
- **Uniform-weighting condition.** Applying classical GLS analysis to the latent-embedding combiner, we show that for paths with equal gains the optimal linear weights are uniform exactly when every row of the path covariance has the same sum, a condition the equicorrelated model satisfies, supporting majority vote as the natural default in standard SC and leaving room for weighting or pruning when prompt-template branches become heterogeneous (§4, §I).
- **Broad decorrelation.** Across 5 models and 12 benchmarks, prompt-template perturbation reduces path correlation in 55 of 60 model-benchmark cells (mean $\Delta\rho{=}-16\%$, median -15%). The reduction appears in all six task domains, and its size varies with the model-task cell rather than with the openness of the answer space (§5.2.4, Fig. 3).
- **Adaptive-K design rule.** The diagnostic yields a closed-form operating point K^* (Eq. 12) from a fixed $K{=}4$ pilot, with no held-out tuning and no per-instance scorer. Cross-domain validation (Math, QA, NLU) shows this rule retains 96–103% of accuracy at K^* (§5.2.5); positioning relative to online-stopping and scaling-law approaches is summarized in Table 6.

2 Background and Related Work

Multi-path LLM reasoning. Self-consistency Wang et al. [2023] generates K reasoning paths by sampling at temperature $\tau > 0$ and returns the plurality answer. Extensions include tree-structured search Yao et al. [2023] and token-level embedding aggregation Xu et al. [2025], Zhang et al. [2025]. A parallel line on efficient reasoning Han et al. [2025], Wen et al. [2025], Sui et al. [2025], Snell

et al. [2025] studies when and how to reduce multi-path compute, showing that adaptive scaling outperforms brute-force path multiplication. Feng et al. Feng et al. [2026] derive power-law scaling of SC from mode estimation and allocate samples across questions dynamically. Concurrently, Bay and Yearick Bay and Yearick [2026] describe a modal ceiling and a correlation ceiling of test-time scaling and summarize the cutoff by an effective number of samples. We derive this effective number in closed form from the pairwise correctness correlation and set K^* from a four-path pilot. Training-based approaches change how paths are generated. Global forking tokens Jia et al. [2026] train models toward diverse yet correct reasoning modes, and Native Parallel Reasoner Wu et al. [2026] trains models to reason in parallel branches within a single response. Our diagnostic measures the path correlation that such train-time interventions act on (Appendix C).

Diversity combining in multipath channels. In wireless communications, a receiver observes K correlated copies of a transmitted signal through distinct paths with gain and additive noise; classical combiners (selection, maximal-ratio, MMSE) trade off complexity against correlation awareness Tse and Viswanath [2005]. For an equally-correlated model with coefficient ρ, the latent effective rank scales as $K/(1+(K-1)\rho^2)$ Simon and Alouini [2004]; the corresponding vote-level effective sample size $K/(1+(K-1)c)$ divides K by the Kish design effect $1+(K-1)c$ Kish [1965].

Positioning. Existing efficient-SC work treats path-budget reduction as either online stopping by vote agreement or quality (Adaptive-Consistency Aggarwal et al. [2023], ESC Li et al. [2024], RASC Wan et al. [2025]), confidence-weighted aggregation (CISC Taubenfeld et al. [2025], Sharma and Chopra [2025], Kuang et al. [2026]), or compound-inference scaling-law fitting (Chen et al. [2024], large-scale repeated sampling Brown et al. [2024], Snell et al. [2025]). We organize these under a single *upstream* diagnostic: pairwise correctness correlation c limits effective diversity to $K/(1+(K-1)c)$, supplying a closed-form ceiling $1/c$ and a pilot-estimated K^*. Because $\hat{c}$ equals the corrected between-instance variance of per-instance accuracy divided by $\bar{p}(1-\bar{p})$, it summarizes in one vote-level statistic the difficulty heterogeneity that query-difficulty scaling models fit as a mixture. Table 6 (Appendix C) summarizes the per-method differences.

3 System Model

A summary of notation used throughout the paper is provided in Table 4 (Appendix A).

3.1 The Reasoning Channel

We use a communications-inspired *effective* model as an analytically tractable abstraction, not as a claim about the internal mechanism of language generation. Reasoning paths generated by the same model under the same prompt share that model's competence on each problem, so their correctness is correlated across problems even when the paths for a given problem are sampled independently. In this view, the ground-truth answer is the latent target, each reasoning path is a branch observation, and the final aggregator is a receiver-side combiner. The value of the model lies in the predictions it enables (saturation law, majority-vote calibration) and the correlation diagnostic it motivates, which we validate empirically in §5. The saturation law, the beta-binomial prediction and Adaptive-K use observable correctness alone. Among the theoretical results only the latent-rank part of Theorem 4.1 uses (1), and the GLS results use the separate linear model of §4.3 (Appendix B).

Consider an LLM generating K reasoning paths for a problem with ground-truth answer $x \in \mathcal{A}$, where $\mathcal{A}$ is a finite answer space. Let $\mathbf{z}(x) \in \mathbb{R}^D$ denote the latent embedding of the correct answer. Each reasoning path k produces a latent representation:

$$\mathbf{z}_k = h_k\, \mathbf{z}(x) + \mathbf{n}_k, \quad k = 1, \ldots, K, \tag{1}$$

where $h_k > 0$ is a real-valued channel gain capturing the reasoning quality of path k, and $\mathbf{n}_k \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_D)$ represents reasoning noise (hallucination, arithmetic errors).

The final answer is decoded from a combined representation $\hat{\mathbf{z}}$ as $\hat{x} = \arg\min_{a \in \mathcal{A}} \|\hat{\mathbf{z}} - \mathbf{z}(a)\|^2$. This decoding step is part of the analytical abstraction only, and no latent-embedding decoding is performed. The implemented pipeline extracts discrete answers from generated text. Deployment returns the plurality vote over these answers (Algorithm 1), and the reported accuracies are the binary majority vote on their gold-scored correctness (§5.1). The empirical GLS-R combiner (§4.3) is the only construct defined on embeddings, and it requires hidden-state access.

Assumption 3.1 (Exchangeable paths)**.** Channel gains $h_1, \ldots, h_K$ are identically distributed with $\mathbb{E}[h_k] = \mu_h > 0$ and $\mathrm{Var}(h_k) = \sigma_h^2$. Noise vectors $\mathbf{n}_k$ are i.i.d. across paths and independent of h_k.

This assumption is more likely to hold when all K paths are generated by the same model with the same prompt and sampling temperature.

Remark 3.2 (Evaluation-induced binary collapse)*.* For open-form tasks (HotpotQA, TriviaQA, DROP), we operate on the binary-collapsed correctness indicator $Y_k = \mathbf{1}\{a_k = x\}$, reducing any task to $\mathcal{A} = \{0, 1\}$ so that the majority-vote (Theorem 4.2) and beta-binomial (9) results apply regardless of the original output format. Scope, thresholding, and GLS exclusion are discussed in Appendix E.

3.2 Path Correlation

Since all paths originate from the same model and prompt, the channel gains $h_1, \ldots, h_K$ are correlated. We model this correlation through an equally-correlated structure for the *path correlation matrix* $\mathbf{R} \in \mathbb{R}^{K \times K}$:

$$R_{ij} = \begin{cases} \sigma_h^2 & i = j, \\ \sigma_h^2 \rho & i \neq j, \end{cases} \tag{2}$$

where $\rho \in [0, 1]$ is the pairwise path correlation coefficient.

Estimating path correlation from data. The latent correlation ρ is not directly observable. We estimate path dependence through the *correctness correlation*

$$\hat{c}_{jk} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(j)} Y_i^{(k)} - \bar{p}^2}{\bar{p}(1 - \bar{p})}, \tag{3}$$

where $Y_i^{(k)} = \mathbf{1}\{a_i^{(k)} = x_i\}$ is the correctness indicator for path k on instance i, and $\bar{p} = \frac{1}{nK}\sum_{i,k} Y_i^{(k)}$ is the empirical mean accuracy. This is the natural estimator for the vote-level correlation that directly governs majority-vote behavior through $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)c)$. In the saturation and Adaptive-K experiments, we report $\hat{c} = \binom{K}{2}^{-1} \sum_{j<k} \hat{c}_{jk}$, averaged over all path pairs, and the cross-benchmark comparison reports the Pearson estimate $\hat{\rho}$ defined in §5. Pooled over instances, $\hat{c}$ equals the finite-K-corrected between-instance variance of per-instance accuracy divided by $\bar{p}(1 - \bar{p})$. In identically prompted SC the K paths of an instance are sampled independently, so $\hat{c}$ is a marginal intraclass correlation, $K_{\text{eff}}^{\text{vote}}$ is the corresponding design-effect effective sample size, and $1/c$ is its asymptotic ceiling. Under prompt templates the K slots follow different distributions, and we use that setting to test whether $\hat{c}$ moves with the sampling configuration. We use $\hat{c}$ rather than alternatives such as Cohen's kappa because $\hat{c}$ is the quantity that enters the effective sample size formula (see Appendix F for comparison).

Assumption 3.3 (Gaussian decision surrogate)**.** For analytical tractability, we model the binary correctness indicators $Y_1, \ldots, Y_K$ as arising from thresholding jointly Gaussian latent decision scores $M_1, \ldots, M_K$: $Y_k = \mathbf{1}\{M_k > 0\}$, with $\Pr(M_k > 0) = p$ and $\mathrm{Corr}(M_j, M_k) = \rho_d$. In the binary case $\mathcal{A} = \{x, a\}$, the nearest-neighbor decision margin $M_k = h_k \langle \mathbf{z}(x), \boldsymbol{\Delta} \rangle + \langle \mathbf{n}_k, \boldsymbol{\Delta} \rangle - b$ with $\boldsymbol{\Delta} = \mathbf{z}(x) - \mathbf{z}(a)$ and $b = (\|\mathbf{z}(x)\|^2 - \|\mathbf{z}(a)\|^2)/2$ is linear in the observation $\mathbf{z}_k$. When gains are jointly Gaussian, the margins are jointly Gaussian with $\rho_d = \kappa\rho$, where $\kappa = \sigma_h^2 \langle \mathbf{z}(x), \boldsymbol{\Delta} \rangle^2 / (\sigma_h^2 \langle \mathbf{z}(x), \boldsymbol{\Delta} \rangle^2 + \sigma_n^2 \|\boldsymbol{\Delta}\|^2) \in [0, 1]$ whenever the margin variance is positive, because the path noise is independent across paths and only inflates the variances. The noise therefore attenuates the gain correlation, and for $\rho > 0$ and $\sigma_h^2 \langle \mathbf{z}(x), \boldsymbol{\Delta} \rangle^2 > 0$ the equality $\rho_d = \rho$ holds only in the noiseless limit. For $|\mathcal{A}| > 2$, the nearest-neighbor decision boundary is polyhedral and ρ_d depends on both ρ and the answer-space geometry; we treat ρ_d as a surrogate parameter calibrated through the observable $\hat{c}$.

3.3 Two Notions of Effective Diversity

We distinguish two notions that must not be conflated.

Latent covariance effective rank. The participation ratio of the gain covariance matrix $\mathbf{R}$ measures the effective rank of the gain structure:

$$K_{\text{eff}}^{\text{rank}}(\mathbf{R}) = \frac{(\mathrm{tr}\,\mathbf{R})^2}{\mathrm{tr}(\mathbf{R}^2)}. \tag{4}$$

The participation ratio is scale-invariant, so $K_{\text{eff}}^{\text{rank}}$ depends only on ρ, not on σ_h^2.

Vote-level effective sample size. For binary correctness indicators $Y_k = \mathbf{1}\{a_k = x\}$ with pairwise correlation $c = \text{Corr}(Y_i, Y_j)$, the standard design-effect formula gives:

$$K_{\text{eff}}^{\text{vote}} = \frac{K}{1 + (K-1)c}. \tag{5}$$

The latent effective rank characterizes covariance geometry; the vote-level effective sample size is the more direct object for predicting majority-vote accuracy.

Proposition 3.4 (Bridge between latent and vote-level diversity)**.** *Both effective diversity measures share the form $K/(1 + (K-1)\gamma)$ with $\gamma = \rho^2$ (latent rank) or $\gamma = c$ (vote level), where $\gamma \mapsto K/(1 + (K-1)\gamma)$ is monotone decreasing. Under the Gaussian decision surrogate (Assumption 3.3) with $p \in (0, 1)$, the binary correctness correlation $c = \text{Corr}(Y_j, Y_k)$ is a strictly monotone increasing function of the decision-layer correlation ρ_d (proof in Appendix W). The relationship between c and ρ^2 depends on both p and ρ_d; the two quantities are not directly comparable in general.* Assumption used: *Assumption 3.3 only.*

Practical consequence. Since the latent ρ is unobservable, the saturation and Adaptive-K experiments use $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)\hat{c})$ with the directly measured $\hat{c}$, and the cross-benchmark comparison substitutes $\hat{\rho}$ for $\hat{c}$. Proposition 3.4 ensures that higher latent dependence produces higher vote-level dependence, so the two effective-diversity measures move in the same direction (design-effect derivation Kish [1965] in Appendix D).

4 Theoretical Analysis

4.1 Effective Rank of Correlated Paths

Theorem 4.1 (Effective rank under equicorrelation)**.** *For K equally-correlated paths with correlation coefficient ρ as in* (2)*,*

$$K_{\textit{eff}}^{\textit{rank}} = \frac{K}{1 + (K-1)\rho^2}, \qquad K_{\textit{eff}}^{\textit{rank}} \to \frac{1}{\rho^2} \text{ as } K \to \infty. \tag{6}$$

The vote-level analog $K_{\textit{eff}}^{\textit{vote}} = K/(1+(K-1)c)$ saturates at $1/c$. Proof via eigenvalue decomposition of $\mathbf{R}$ is in Appendix G. Assumptions used: *for $K_{\textit{eff}}^{\textit{rank}}$,* (1) *with Assumption 3.1 and the equicorrelated matrix* (2)*. For $K_{\textit{eff}}^{\textit{vote}}$, equicorrelated correctness indicators.*

4.2 Majority Vote Accuracy

We analyze majority vote through a binary-collapsed correctness process. Let $Y_k = \mathbf{1}\{a_k = x\}$ where a_k is the extracted answer from path k.

Theorem 4.2 (Majority vote under independence)**.** *If $Y_1, \ldots, Y_K$ are i.i.d. Bernoulli(p), for odd K:*

$$P_{MV}(K, p) = \sum_{j=(K+1)/2}^{K} \binom{K}{j} p^j (1-p)^{K-j}. \tag{7}$$

For even K with random tie-breaking, the tie term $\binom{K}{K/2} p^{K/2}(1-p)^{K/2}$ contributes a factor of $1/2$. Assumption used: *i.i.d. correctness indicators (the independence reference).*

This formula is exact for the binary-collapsed model. In the original multiclass answer space with $|\mathcal{A}| > 2$, plurality self-consistency depends on the full wrong-answer distribution and the binomial formula is a surrogate.

Remark 4.3 (Multiclass plurality heuristic). Let $|\mathcal{A}| = M > 2$ and let p denote the probability that a single path produces the correct answer. Under uniform wrong-answer fragmentation among $M-1$ alternatives, the effective pairwise accuracy for correct-vs-most-popular-wrong is $p' = p(M-1)/(pM - 2p + 1)$. Since $p' > p$ for $M > 2$, plurality voting can gain substantially over the binary-collapsed prediction. For $M=4$ and $p=0.5$, $p'=0.75$ (see Appendix X for the derivation).

The binary-collapsed accuracy is in any case a lower bound, $P_{\text{PV}}^{(M)} \geq P_{\text{MV}}^{(2)}$ on the same draws, for any wrong-answer distribution and any dependence among paths. A correct answer holding more than $K/2$ votes leaves fewer than $K/2$ for every other answer and therefore wins the plurality, and at an even-K tie of $K/2$ votes it wins with probability at least $1/2$ under random tie-breaking.

Non-uniform wrong-answer fragmentation (where some wrong answers are more popular) reduces the plurality advantage.

Correlated majority vote. When paths are correlated, we model the joint correctness distribution using the classical beta-binomial (BB) model Skellam [1948]. Assume $\Theta \sim \text{Beta}(\alpha, \beta)$ and $Y_k \mid \Theta \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\Theta)$. Then Y_k are exchangeable with:

$$p = \frac{\alpha}{\alpha+\beta}, \quad c = \frac{1}{\alpha+\beta+1}. \tag{8}$$

The count $S_K = \sum_k Y_k$ follows a beta-binomial distribution, and the correlated majority-vote accuracy is:

$$P_{\text{MV}}^{\text{corr}} = \sum_{j>K/2} \binom{K}{j} \frac{B(j+\alpha, K-j+\beta)}{B(\alpha,\beta)}, \tag{9}$$

where $B(\cdot,\cdot)$ is the beta function. For even K with random tie-breaking, the $j=K/2$ term contributes half its probability mass: add $\frac{1}{2}\binom{K}{K/2} B(K/2+\alpha,\ K/2+\beta)/B(\alpha,\beta)$ to (9).

4.3 GLS-Optimal Combining

Equal-weight combining ignores cross-path redundancy. We now analyze a generalized linear aggregation model that shares the branch-combining structure of §3 but is not a direct corollary of Eq. (1). The GLS results (Theorem 4.4, Corollaries 4.5–4.7) are the classical generalized least squares solution Aitken [1936] and hold for any linear estimation model satisfying $\mathbf{e}_k = g_k \mathbf{s} + \boldsymbol{\varepsilon}_k$, independent of the channel model. Assume each path embedding follows this model for $k = 1, \ldots, K$, where $\mathbf{s} \in \mathbb{R}^D$ is the latent answer embedding, $g_k > 0$ is a path-quality gain, and for each feature dimension d, the branch error vector $\boldsymbol{\varepsilon}^{(d)} = (\varepsilon_{1d}, \ldots, \varepsilon_{Kd})^\top$ has mean zero and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$.

For a linear estimator $\hat{\mathbf{s}}(\mathbf{w}) = \sum_k w_k \mathbf{e}_k$, unbiasedness requires $\mathbf{g}^\top \mathbf{w} = 1$. The risk is $\mathbb{E}\|\hat{\mathbf{s}} - \mathbf{s}\|^2 = D \cdot \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$.

Theorem 4.4 (GLS-optimal combiner)**.** *Let $\boldsymbol{\Sigma}$ be positive definite. Among all linear unbiased estimators satisfying $\mathbf{g}^\top \mathbf{w} = 1$, the minimum-MSE weights are:*

$$\mathbf{w}_* = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{g}}{\mathbf{g}^\top \boldsymbol{\Sigma}^{-1} \mathbf{g}}, \qquad \textit{with optimal risk } \text{MSE}(\mathbf{w}_*) = D/(\mathbf{g}^\top \boldsymbol{\Sigma}^{-1}\mathbf{g}). \tag{10}$$

Assumption used: *the linear embedding model of §4.3, independent of* (1)*.*

The Lagrangian derivation and strict-improvement conditions are given in Appendix H.

Corollary 4.5 (When uniform weighting is optimal)**.** *With equal gains $\mathbf{g} = \mathbf{1}$, $\mathbf{w}_* = (1/K)\mathbf{1}$ if and only if $\boldsymbol{\Sigma}\mathbf{1} \propto \mathbf{1}$, i.e., every row of $\boldsymbol{\Sigma}$ has the same sum, because $\mathbf{w}_* \propto \boldsymbol{\Sigma}^{-1}\mathbf{1}$ and $\boldsymbol{\Sigma}^{-1}\mathbf{1} \propto \mathbf{1} \Leftrightarrow \boldsymbol{\Sigma}\mathbf{1} \propto \mathbf{1}$. The exchangeable model $\boldsymbol{\Sigma} = \sigma^2[(1-\rho)\mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top]$ satisfies this condition, and so does any covariance whose unequal variances and correlations balance within each row. Other special cases (white noise ⇒ Maximum Ratio Combining (MRC), and equal gains ⇒ covariance-aware averaging) are given in Appendix H.* Assumption used: *as Theorem 4.4, with equal gains.*

Remark 4.6 (GLS→majority vote). Corollary 4.5 establishes that, with equal path gains, uniform weighting is optimal among linear combiners operating on the latent embeddings exactly when the path covariance has equal row sums. Majority vote operates on the decoded discrete answers, and the two can disagree even under the exchangeable model, because decoding the uniform average of the embeddings need not return the most frequent decoded answer. We interpret the GLS result as theoretical grounding for the empirical effectiveness of majority vote, not as a claim that MV is Bayes-optimal among all aggregation rules. More precisely, GLS minimizes mean-squared error over continuous latent estimates, whereas majority vote selects a discrete answer evaluated by classification accuracy. These are different optimization problems, so the GLS optimality does not transfer to majority vote.

Corollary 4.7 (Gain over uniform averaging)**.** $\mathrm{MSE}(\mathbf{w}_*) \le \mathrm{MSE}(\mathbf{w}_{unif})$*, with equality iff* $\boldsymbol{\Sigma}\mathbf{1} \propto \mathbf{1}$*. Heterogeneous path variances or correlations therefore improve on uniform averaging only when they make the row sums of* $\boldsymbol{\Sigma}$ *unequal (proof in Appendix R).* Assumption used: *as Theorem 4.4, with equal gains* $\mathbf{g} = \mathbf{1}$ *so that uniform weights are feasible.*

Empirical GLS-R. In practice, the branch covariance $\boldsymbol{\Sigma}$ is estimated from a calibration set. Given T problems with path embeddings, we form centered residuals $\tilde{\mathbf{E}}_t \in \mathbb{R}^{K\times D}$ and estimate $\hat{\boldsymbol{\Sigma}} = (TD)^{-1}\sum_t \tilde{\mathbf{E}}_t\tilde{\mathbf{E}}_t^\top$. The regularized empirical GLS weights are:

$$\hat{\mathbf{w}}_\lambda = \frac{(\hat{\boldsymbol{\Sigma}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{g}}}{\hat{\mathbf{g}}^\top(\hat{\boldsymbol{\Sigma}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{g}}}. \tag{11}$$

5 Experiments

The experiments *validate* the framework's predictions (saturation ceiling, BB calibration, uniform-weighting optimality) and *characterize* how prompt-template perturbations probe the correlation structure of multi-path reasoning across 12 benchmarks spanning six domains.

5.1 Experimental Settings

Models. We evaluate five instruction-tuned models across three architecture families: Qwen2.5-0.5B/7B/32B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3. Saturation analysis (§5.2.1) uses Qwen-7B, Llama-8B, Mistral-7B at $K \in \{4, 8, 16, 32\}$; cross-benchmark analysis (§5.2.4) uses all five at K=8. A reasoning model, Qwen3.5-9B in thinking mode, is evaluated separately (Appendix N).

Benchmarks. We evaluate on 12 benchmarks spanning six task domains (Math, QA, Sci/MC, Code, Commonsense, NLU); the full list with citations and the per-benchmark *effective evaluated output space* (numeric, open text, bounded choice, binary, code pass/fail) is in Appendix U, with the answer-space label appearing as a column of Table 7. Saturation experiments (§5.2.1–§5.2.2) use n=100 instances per seed on GSM8K, pooled across 5 seeds (500 per cell). Cross-benchmark experiments (§5.2.4) use the same n=100 instances per seed, pooled across 5 seeds for GSM8K, MATH, TriviaQA, DROP, CruxEval and MBPP and across 3 seeds for the other six benchmarks (500 or 300 instances per cell).

Methods and metrics. Saturation analysis (§5.2.1–§5.2.2) runs standard self-consistency (SC, temperature τ=0.7) at $K \in \{4, 8, 16, 32\}$ and reports the correctness correlation $\hat{c}$ from (3), the *modeled* quantity entering the design-effect formula. Cross-benchmark comparisons (§5.2.4) contrast SC (K=8, the first eight of 32 sampled paths) against prompt-template SC (PT, K=8 structurally distinct templates per task, such as algebraic vs. estimation for math and chain-of-thought vs. extract-then-answer for QA, with the full set in Appendix V), reporting $\hat{\rho}$, the *measured* mean pairwise Pearson correlation of the K-path binary correctness vectors (the two estimators agree to within 0.004 under SC and to within 0.06 under PT, where per-template accuracies differ, and with $\hat{c}$ the same 55 cells decorrelate), with $\Delta\rho = (\hat{\rho}_{\mathrm{PT}} - \hat{\rho}_{\mathrm{SC}})/|\hat{\rho}_{\mathrm{SC}}|$. Binary correctness comes from task-specific scorers on the extracted final answer, identical in both arms: numeric equivalence for GSM8K, canonical exact match for MATH, alias exact match for TriviaQA and DROP, token-level F1 ≥ 0.5 for HotpotQA, the option letter for the multiple-choice tasks, yes/no for BoolQ, literal equality for CruxEval, and passing every reference assertion in a sandbox for MBPP. Throughout, MV@K is the binary majority vote on these indicators: an instance scores 1 when more than $K/2$ of its paths are correct and $1/2$ at an exact tie, matching the decision rule of (9).

Implementation details. All experiments use Hugging Face Transformers with bfloat16 precision on NVIDIA H100 GPUs. Maximum generation length is set per benchmark and shared by both arms: 256 tokens for BoolQ, 512 for HotpotQA, TriviaQA, DROP, HellaSwag and WinoGrande, 1024 for GSM8K, ARC-Challenge, CruxEval and MBPP, 2048 for MATH and 4096 for MMLU. The saturation and Adaptive-K experiments with non-reasoning models use the same budgets for GSM8K, HotpotQA and BoolQ, and the reasoning model of Appendix N uses 16,384 tokens.

5.2 Experimental Results

5.2.1 Diversity Saturation in Self-Consistency

We use GSM8K as the primary saturation benchmark because it is the principled stress test for the diagnostic: closed-form numeric answers admit a clean binary correctness collapse and isolate the correlation parameter $\hat{c}$ from partial-credit or open-form scoring confounds; the cross-task results in Table 3 (HotpotQA $\hat{c}$=0.61, BoolQ $\hat{c}$=0.79) further show that the saturation rule remains predictive in higher-correlation regimes spanning open-form QA and binary NLU at $K=\{4, 8, 16, 32\}$ for those rows; a per-K saturation sweep across additional open-form benchmarks is left for future work.

Table 1 reports the effective diversity analysis for Qwen2.5-7B-Instruct on GSM8K. As K increases from 4 to 32, the correctness correlation $\hat{c}$ remains stable near 0.59, indicating persistent inter-path correlation. The vote-level effective diversity $K_{\text{eff}}^{\text{vote}}$ saturates at 1.67 at K=32, reaching 98% of its theoretical ceiling $1/\hat{c} = 1.71$.

Table 1: Effective diversity analysis for Qwen2.5-7B on GSM8K (100 instances, 5 seeds pooled). Agree is the mean pairwise agreement $\binom{K}{2}^{-1} \sum_{j<k} \Pr(Y^{(j)}=Y^{(k)})$. Correctness correlation $\hat{c}$ is computed via (3). $K_{\text{eff}}^{\text{vote}}$ is the design-effect (5), and Ceiling $= 1/\hat{c}$. $\bar{p}$ is the mean per-path accuracy, reported as per-seed mean $\pm$ std over 5 seeds, and is not the majority-vote accuracy MV@K of Table 2. The K=1 row is a single sampled path.

K	Agree	$\hat{c}$	$K_{\text{eff}}^{\text{vote}}$	Ceiling	% Ceil.	$\bar{p}$ (%) ↑	Tokens ↓
1	–	–	1.0	–	–	78.0 ± 2.5	323
4	0.868	0.604	1.42	1.66	86%	79.0 ± 1.3	1293
8	0.866	0.593	1.55	1.69	92%	79.2 ± 0.8	2592
16	0.862	0.579	1.65	1.73	96%	79.4 ± 0.8	5179
32	0.864	0.586	1.67	1.71	98%	79.2 ± 0.4	10327

The rapid ceiling approach confirms the design-effect prediction. $K_{\text{eff}}^{\text{vote}} \ll K$ and grows only from 1.55 to 1.67 between K=8 and K=32, so paths beyond K=8 add little independent evidence to the vote.

5.2.2 Cross-Architecture Validation

Table 2: Saturation and beta-binomial calibration on GSM8K (Qwen2.5-7B, Llama-3.1-8B, Mistral-7B, 100 instances, 5 seeds pooled per cell). $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)\hat{c})$. All models reach 96–98% of the ceiling $1/\hat{c}$ at K=32. MV@K denotes the binary majority-vote accuracy at K paths (§5.1), which is distinct from the mean per-path accuracy $\bar{p}$ of Table 1. BB (beta-binomial) and Binom (independence-binomial) columns are in-sample MV@32 predictions, and both estimators are formally defined in the next subsection (Eq. (9)).

		$K_{\text{eff}}^{\text{vote}}$ at $K=$					Predicted MV@32	
Model	$\hat{c}_{K=4}$	4	8	16	32	MV@32	BB	Binom
Qwen2.5-7B	0.60	1.42	1.55	1.65	1.67	81.9%	81.1%	100.0%
Llama-3.1-8B	0.53	1.55	1.72	1.86	1.94	79.3%	77.8%	99.9%
Mistral-7B	0.45	1.71	1.89	2.00	2.13	42.4%	41.3%	20.6%

Table 2 validates the vote-level saturation law across three model families spanning base accuracies of 43–79%. Mistral has the lowest correctness correlation ($\hat{c}$=0.45), yielding the highest ceiling ($1/\hat{c}$=2.21) and the highest $K_{\text{eff}}^{\text{vote}}$ at K=32 (2.13). The saturation law is model-agnostic: it depends only on c, not on the absolute accuracy level.

5.2.3 Beta-Binomial Calibration

We validate the correlated majority-vote model from (9) (derivation in Appendix P) by fitting beta-binomial parameters (α, β) via method of moments (MoM) from the observed vote-count distribution, then comparing the predicted MV accuracy against empirical values. We use MoM

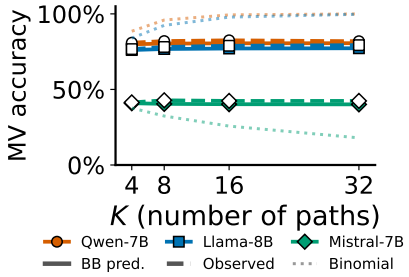


Figure 2: Held-out BB prediction on GSM8K: (α, β) fitted from $K{=}4$. BB tracks within 3.3–4.8 pp at $K{=}32$ and the binomial diverges by 18–24 pp.

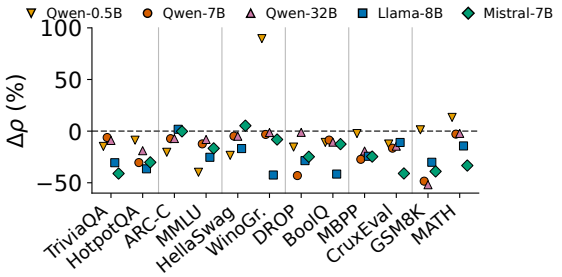


Figure 3: Per-cell PT decorrelation $\Delta\rho$ for all 60 model-benchmark cells, grouped by domain. The dashed line marks no change.

rather than maximum-likelihood or GLMM estimation because it yields closed-form (α, β) from (p, c) alone, matching the pilot-based diagnostic use case; at $n{=}100$, MoM and MLE estimates are comparable for the overdispersion levels observed here.

Table 2 shows that the independence-assuming binomial diverges catastrophically at $K{=}32$ (e.g., 100.0% predicted vs. 81.9% actual for Qwen; 99.9% vs. 79.3% for Llama; 20.6% vs. 42.4% for Mistral), while the beta-binomial predicts within 0.8–1.5 pp of the observed MV accuracy.

Held-out prediction test. When (α, β) are fitted from a $K{=}4$ pilot on one half of the GSM8K instances and used to predict $K \in \{8, 16, 32\}$ on the other half, BB absolute error at $K{=}32$ is 3.3–4.8 pp across all three models (averaged over both halves), whereas the binomial diverges by 18–24 pp (Fig. 2).

5.2.4 Prompt-Template Diversity Across 12 Benchmarks

The value of prompt-template diversity lies in the structure it reveals: its success criterion is whether the diagnostic variable moves under a controlled prompt change on a fixed question set. This subsection measures that *decorrelation*, which raises the saturation ceiling $1/c$. Decorrelation by itself does not raise base path accuracy $\bar{p}$, and observed MV accuracy changes are correspondingly small and mixed in sign (Appendix M). Temperature-only sampling yields exchangeable correctness patterns with limited decorrelation (Corollary 4.5); to break exchangeability, we assign structurally distinct prompt templates to each of the $K{=}8$ slots and measure $\Delta\rho$ against SC on the *same model* across all 12 benchmarks. No cell falls below the 2% base-accuracy cut, and PT reduces $\hat{\rho}$ in 55 of the 60 model-benchmark cells (Fig. 3, with per-benchmark means in Table 7 and all cells in Table 8, Appendix L), with mean $\Delta\rho{=}{-}16\%$ (median -15%) and mean $\Delta K_{\text{eff}}^{\text{vote}}{=}{+}0.29$. Four of the five exceptions are small, with $\Delta\rho$ between $+1\%$ and $+13\%$ (Llama-8B on ARC-C, Mistral-7B on HellaSwag, Qwen-0.5B on GSM8K and MATH). The fifth, Qwen-0.5B on WinoGrande ($\Delta\rho{=}{+}90\%$), sits at chance accuracy ($\bar{p}{=}49\%$ on two options). There the templates push the slots toward the same option letter (B in 77% of PT answers against a balanced gold), and the shared position bias correlates the slots.

Decorrelation is broad, not answer-space-gated. Domain medians of $\Delta\rho$ range from -5% (Commonsense) to -25% (QA), and Math (median -22%) decorrelates about as much as QA. A Mann-Whitney test of the Math cells against the QA cells ($n{=}10$ each, the unit being the model-benchmark cell) finds no difference ($p{=}0.73$), and answer diversity under SC does not predict $\Delta\rho$ across benchmarks ($r{=}{-}0.34$, $p{=}0.28$, $n{=}12$). Path correlation therefore moves under a controlled prompt change on every task type, and the size of the move is a property of the model-task cell. Per-cell bootstrap intervals are reported in Appendix K. This decorrelation is not visible from accuracy curves or aggregation weights alone, which makes path correlation the diagnostic, and we use prompt-template perturbation as a structural probe of c rather than as a replacement for SC.

Heterogeneous slot regime. When prompt templates create heterogeneous per-slot accuracy, the exchangeability of Corollary 4.5 breaks and the GLS analysis leaves room for non-uniform weighting. Consistent with Corollary 4.7, the gain of in-sample accuracy-weighted MV over uniform MV on PT

cells tracks the slot-accuracy coefficient of variation (CV, r=0.85 over 60 cells), averaging $+4.4$ pp in the seven cells with CV > 0.225 and $+1.1$ pp in the rest. Full results and the corresponding SC comparison are in Appendix I.

5.2.5 Adaptive-K

The saturation law motivates a compute-allocation rule. The marginal diversity gain from adding one path is $\partial K_{\text{eff}}^{\text{vote}}/\partial K = (1-c)/(1+(K-1)c)^2$, which decreases in K. Setting this marginal gain to a threshold ε and solving yields:

$$K^* = \left\lceil \frac{\sqrt{(1-c)/\varepsilon} - 1}{c} + 1 \right\rceil. \tag{12}$$

For the default threshold ε=0.025 (2.5% marginal gain) and a K=4 pilot estimate $\hat{c}$, the shorthand $\lceil 2/\hat{c}^2 \rceil$ lies within one path of (12), and never above it, for typical $\hat{c} \in [0.5, 0.7]$. Algorithm 1 (Appendix S) lists the full procedure.

Table 3: Adaptive-K compute savings across three domains. $\hat{c}$ is estimated from K=4 SC (n=100 instances, 5 seeds pooled for GSM8K and 3 seeds pooled for HotpotQA and BoolQ). Retained = MV accuracy at K^* as a percentage of MV at K=32. Net cost = $(K^*+4)/32$ accounts for the K=4 pilot in addition to the operating-point paths.

Task	Model	$\hat{c}_{K=4}$	K^*	MV@K^*	MV@32	Retained	Net cost
GSM8K (Math)	Qwen-7B	0.60	6	81.6%	81.9%	100%	31%
GSM8K (Math)	Llama-8B	0.53	8	78.1%	79.3%	98%	38%
GSM8K (Math)	Mistral-7B	0.45	10	43.6%	42.4%	103%†	44%
HotpotQA (QA)	Llama-8B	0.61	6	50.2%	52.2%	96%	31%
BoolQ (NLU)	Llama-8B	0.79	4	80.2%	80.3%	100%	25%

†Retention above 100% is consistent with finite-sample variance: the paired bootstrap 95% CI of MV@K^* −MV@32 contains zero in every cell.

Table 3 shows that a K=4 pilot reliably predicts saturation across Math, QA, and NLU: K^* retains 96–103% of MV@K=32 at net inference cost 25–44% of full K=32 SC, with higher-correlation tasks saturating earlier (the single >100% cell is annotated by the dagger). The rule is also stable to ε: varying it from 0.01 to 0.1 changes K^* but preserves 97–101% accuracy on Qwen-7B GSM8K (Appendix T).

6 Conclusion

This paper formalized multi-path LLM reasoning as a diversity combining problem. The design-effect formula $K_{\text{eff}}^{\text{vote}} = K/(1+(K-1)c)$ predicts that 32 paths yield a design-effect effective sample size of only 1.7–2.1, a ceiling confirmed across three model families. GLS analysis shows that, with equal path gains, the optimal linear combiner of latent embeddings is uniform exactly when the path covariance has equal row sums, as under exchangeability, supporting uniform majority vote as the natural default in standard SC and motivating slot-pruning and weighted-MV variants in the heterogeneous-template regime. Prompt-template diversity reduces path correlation in 55 of 60 model-benchmark cells across 12 benchmarks and all six task domains (mean -16%). An Adaptive-K rule derived from the saturation law retains 96–103% of accuracy across Math, QA, and NLU domains from a four-path pilot.

Acknowledgments and Disclosure of Funding

The authors received no specific funding for this work and declare no competing interests.

References

Pranjal Aggarwal, Aman Madaan, Yiming Yang, and Mausam. Let's sample step by step: Adaptive-consistency for efficient reasoning and coding with LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

A. C. Aitken. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1936. doi: 10.1017/S0370164600014346.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Yong Yi Bay and Kathleen A. Yearick. When More Sampling Hurts: The Modal Ceiling and Correlation Ceiling of Test-Time Scaling. *arXiv preprint arXiv:2606.28661*, 2026.

Sven Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1): 87–95, 1993.

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.

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 37 (NeurIPS)*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

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

Alex Gu, Baptiste Roziere, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

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*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021b.

Alan Jeffares, Tennison Liu, Jonathan Crabbé, and Mihaela van der Schaar. Joint training of deep ensembles fails due to learner collusion. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Sheng Jia, Xiao Wang, and Shiva Prasad Kasiviswanathan. Training large language models to reason in parallel with global forking tokens. In *International Conference on Learning Representations*, 2026.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

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

Peng Kuang, Yanli Wang, Xiaoyu Han, Yaowenqi Liu, Kaidi Xu, and Haohan Wang. Optimal aggregation of LLM and PRM signals for efficient test-time scaling. In *International Conference on Learning Representations*, 2026.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3-4):351–360, 1954. doi: 10.1093/biomet/41.3-4.351.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Aman Sharma and Paras Chopra. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute. *arXiv preprint arXiv:2511.02309*, 2025.

Marvin K. Simon and Mohamed-Slim Alouini. *Digital Communication over Fading Channels*. Wiley-IEEE Press, 2nd edition, 2004.

J. G. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, 10(2):257–261, 1948.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 3613–3635, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Hao Wen, Xinrui Wu, Yi Sun, Feifei Zhang, Liye Chen, Jie Wang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. BudgetThinker: Empowering budget-aware LLM reasoning with control tokens. *arXiv preprint arXiv:2508.17196*, 2025.

Tong Wu, Yang Liu, Jun Bai, Zixia Jia, Shuyi Zhang, Ziyong Lin, Yanting Wang, Song-Chun Zhu, and Zilong Zheng. Native parallel reasoner: Reasoning in parallelism via self-distilled reinforcement learning. In *International Conference on Machine Learning*, 2026.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

A Notation

Table 4: Notation used throughout the paper.

Symbol	Meaning
$x \in \mathcal{A}$	Ground-truth answer in discrete answer space.
$\mathbf{z}(x) \in \mathbb{R}^D$	Latent embedding of answer x.
K	Number of reasoning paths.
h_k	Channel gain of path k ($h_k > 0$).
$\mathbf{n}_k$	Additive noise vector of path k.
$\rho \in [0, 1]$	Latent channel gain correlation (in $\mathbf{R}$).
p	Single-path accuracy $\Pr(Y_k = 1)$.
$K_{\text{eff}}^{\text{rank}}$	Latent effective rank (participation ratio).
$K_{\text{eff}}^{\text{vote}}$	Vote-level effective sample size (design effect).
c	Pairwise correctness correlation $\mathrm{Corr}(Y_i, Y_j)$.
$\mathbf{R}$	Gain covariance matrix ($K \times K$).
$\mathbf{\Sigma}$	Branch error covariance ($K \times K$).
$\mathbf{g}$	Path quality gain vector.

Upper: problem and channel variables. Lower: analysis quantities.

B Assumption Map

Table 5 records which assumptions each result uses. The results the experiments test sit on the correlated-Bernoulli voting layer and are computed from observable correctness alone. The latent channel model (1) enters only the latent-rank part of Theorem 4.1, and the GLS results rest on the separate linear embedding model of §4.3.

Table 5: Assumption dependencies of each result. "Observable" marks results computed from output-level correctness alone.

Result	Assumptions used	Observable
$K_{\text{eff}}^{\text{vote}}$, Eq. (5)	Equicorrelated correctness indicators (design effect Kish [1965])	✓
Theorem 4.2	i.i.d. Bernoulli correctness (independence reference)	✓
Beta-binomial, Eq. (9)	$\Theta \sim \text{Beta}(\alpha, \beta)$, $Y_k \mid \Theta$ i.i.d. Bernoulli(Θ) Skellam [1948]	✓
Adaptive-K, Eq. (12)	$K_{\text{eff}}^{\text{vote}}$ and a marginal-gain threshold ε	✓
Theorem 4.1, latent rank	Eq. (1) with Assumption 3.1 and the equicorrelated matrix (2)	✗
Proposition 3.4	Gaussian decision surrogate (Assumption 3.3)	✗
Theorem 4.4	Linear embedding model $\mathbf{e}_k = g_k \mathbf{s} + \boldsymbol{\varepsilon}_k$ Aitken [1936]	✗
Corollaries 4.5–4.7	As Theorem 4.4 with $\mathbf{g} = \mathbf{1}$ (and symmetric $\boldsymbol{\Sigma}$ for Corollary 4.5)	✗
GLS-R, Eq. (11)	Access to path embeddings	✗

C Positioning: Analytical Coverage of Prior Work

Table 6: Analytical coverage of multi-path reasoning studies. Existing methods focus on aggregation strategies, online stopping, or empirical scaling laws; this work provides an upstream correlation-based diagnostic framework. ✓ = yes, ✗ = no, ∼ = partial. Superscripts indicate the predictive mechanism: $^{\text{online}}$ stops sampling per query based on observed agreement/quality; $^{\text{difficulty}}$ predicts K from a query-difficulty mixture model fit to scaling curves; $^{\text{correlation}}$ predicts K^* from a closed-form ceiling derived from inter-path correctness correlation. Compound-inference scaling Chen et al. [2024] fits query-difficulty heterogeneity as a mixture over scaling curves. Our $\hat{c}$ equals the corrected between-instance variance of per-instance accuracy divided by $\bar{p}(1-\bar{p})$ (§3.2), so it summarizes the same heterogeneity in one vote-level statistic without decomposing its sources.

Study	Formal saturation bound	Aggregation optimality analysis	Predicts K^*	Cross-task characterization	Answer-space analysis
SC Wang et al. [2023]	✗	✗	✗	✗	✗
CISC Taubenfeld et al. [2025]	✗	✗	✗	✗	✗
Entropy Voting Sharma and Chopra [2025]	✗	✗	✗	✗	✗
Optimal Agg. Kuang et al. [2026]	✗	∼	✗	✗	✗
SoftCoT Xu et al. [2025], Zhang et al. [2025]	✗	✗	✗	✗	✗
Test-time scaling Snell et al. [2025]	✗	✗	∼	✗	✗
Adaptive-Consistency Aggarwal et al. [2023]	✗	✗	∼$^{\text{online}}$	✗	✗
ESC Li et al. [2024]	✗	✗	∼$^{\text{online}}$	✗	✗
RASC Wan et al. [2025]	✗	∼	∼$^{\text{online}}$	✗	✗
Compound-inf. scaling Chen et al. [2024]	∼	✗	✓$^{\text{difficulty}}$	∼	✗
LLM Monkeys Brown et al. [2024]	✗	✗	✗	∼	✗
This work	✓	✓	✓$^{\text{correlation}}$	✓	✓

D Design-Effect Derivation

The vote-level effective sample size in (5) follows from the standard design-effect formula for correlated binary variables. For exchangeable $Y_1, \ldots, Y_K$ with $\mathbb{E}[Y_k] = p$ and $\text{Corr}(Y_j, Y_k) = c$ for

$j \neq k$, each variance is $p(1-p)$ and each of the $K(K-1)$ off-diagonal covariances is $c\,p(1-p)$, so

$$\mathrm{Var}(\bar{Y}) = \frac{1}{K^2}\Big[Kp(1-p) + K(K-1)\,c\,p(1-p)\Big] = p(1-p)\,\frac{1+(K-1)c}{K}, \tag{13}$$

so the effective sample size relative to K i.i.d. draws is $K_{\text{eff}}^{\text{vote}} = K/(1+(K-1)c)$ Kish [1965], the number of independent draws whose mean has the same variance $p(1-p)/K_{\text{eff}}^{\text{vote}}$. Eq. (3) is the sample analog of $\mathrm{Corr}(Y_j, Y_k)$ pooled over instances, and averaging it over path pairs gives the $\hat{c}$ that enters this formula. Monotonicity (Proposition 3.4) guarantees qualitative consistency (higher ρ_d implies lower ceiling), not numerical equivalence between c and ρ^2; the two quantities enter different-level formulas. The quantitative accuracy of $\hat{c}$ as a predictor of observed MV accuracy is validated empirically in §5 (Fig. 2).

E Binary Collapse Details and Scope

The formal system model in §3 assumes a finite answer space $\mathcal{A}$ with nearest-neighbor decoding, which directly applies to closed-form (numeric) and multiple-choice tasks. For open-form QA tasks, $Y_k = \mathbf{1}\{a_k = x\}$ is computed by alias exact match for TriviaQA and DROP and by token-level F1 ≥ 0.5 for HotpotQA. Under this reduction, all theoretical results that depend on binary correctness (Theorem 4.2, the design-effect formula (5), and the beta-binomial model (9)) apply regardless of the original output format. The GLS analysis (Theorem 4.4) operates on latent embeddings and does not depend on binary collapse; see §4.3.

F Kappa vs. Correctness Correlation

An alternative estimator of path agreement is Cohen's kappa $\kappa = (p_{\text{agree}} - p_{\text{chance}})/(1 - p_{\text{chance}})$, which measures answer-identity agreement corrected for chance. Since κ conflates "both correct" and "both wrong with the same wrong answer," it differs from c in general. We use $\hat{c}$ in the saturation and Adaptive-K analyses as it is the quantity that enters the effective sample size formula (5).

G Proof of Theorem 4.1 (Effective Rank under Equicorrelation)

The eigenvalues of the equally-correlated matrix $\mathbf{R}$ with diagonal σ_h^2 and off-diagonal $\sigma_h^2\rho$ are:

$$\lambda_1 = \sigma_h^2(1+(K-1)\rho), \quad \text{[multiplicity 1]} \tag{14}$$

$$\lambda_j = \sigma_h^2(1-\rho), \quad j = 2, \ldots, K. \tag{15}$$

The trace and squared trace are:

$$\mathrm{tr}\,\mathbf{R} = K\sigma_h^2, \tag{16}$$

$$\mathrm{tr}(\mathbf{R}^2) = \sigma_h^4\big[(1+(K-1)\rho)^2 + (K-1)(1-\rho)^2\big]. \tag{17}$$

Expanding the denominator:

$$\begin{aligned}
&(1+(K-1)\rho)^2 + (K-1)(1-\rho)^2 \\
&= 1 + 2(K-1)\rho + (K-1)^2\rho^2 \\
&\quad + (K-1) - 2(K-1)\rho + (K-1)\rho^2 \\
&= K + K(K-1)\rho^2.
\end{aligned} \tag{18}$$

Therefore $K_{\text{eff}}^{\text{rank}} = K^2\sigma_h^4/[\sigma_h^4(K + K(K-1)\rho^2)] = K/(1+(K-1)\rho^2)$. Taking $K \to \infty$ yields $K_{\text{eff}}^{\text{rank}} \to 1/\rho^2$.

H Proof of Theorem 4.4 and Special Cases

Proof of Theorem 4.4. Form the Lagrangian $\mathcal{L}(\mathbf{w}, \nu) = \mathbf{w}^\top\boldsymbol{\Sigma}\mathbf{w} + \nu(\mathbf{g}^\top\mathbf{w} - 1)$. Stationarity gives $2\boldsymbol{\Sigma}\mathbf{w} + \nu\mathbf{g} = \mathbf{0}$, hence $\mathbf{w} = -(\nu/2)\boldsymbol{\Sigma}^{-1}\mathbf{g}$. Substituting into the constraint $\mathbf{g}^\top\mathbf{w} = 1$ yields the stated formula $\mathbf{w}_* = \boldsymbol{\Sigma}^{-1}\mathbf{g}/(\mathbf{g}^\top\boldsymbol{\Sigma}^{-1}\mathbf{g})$.

Additional special cases of Corollary 4.5. Beyond the equicorrelated symmetric case stated in the main text, the GLS combiner recovers:

1. **White noise** ($\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$): $\mathbf{w}_* \propto \mathbf{g}$, i.e., quality-weighted combining (the MRC principle).
2. **Equal gains** ($\mathbf{g} = \mathbf{1}$): $\mathbf{w}_* = \boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1})$, i.e., covariance-aware averaging.

I Heterogeneous Slot Regime: Weighted Voting

This appendix collects the empirical follow-ups to Corollary 4.7 for prompt-template (PT) data, where heterogeneous per-slot accuracy breaks the symmetric exchangeable regime.

When does uniform voting fail? We compute accuracy-weighted voting, with each slot weighted by its accuracy on the evaluation set, across all 60 PT cells and measure the gain over uniform MV as a function of the slot-accuracy coefficient of variation (CV). The correlation between CV and weighted-MV gain is r=0.85 (p<10^{-4}). In the 7 cells with heterogeneous slot accuracies (CV > 0.225), weighted voting gains $+4.4$ pp on average over uniform MV, and in the 53 near-homogeneous cells (CV ≤ 0.225) the gain is $+1.1$ pp. The gain is larger in the low-accuracy regime. Cells with MV accuracy below 30% gain $+2.2$ pp on average (mean CV 0.19), and cells at or above 30% gain $+1.2$ pp (mean CV 0.07). Of the low-accuracy cells, 31% have CV > 0.225, against 5% of the high-accuracy cells.

Weighting in the SC regime. Slot weights estimated on the evaluation set give an optimistic, in-sample estimate of what slot weighting can achieve, and deployment would require held-out calibration. Under standard SC every slot CV is at most 0.10, and the same in-sample accuracy weighting changes MV accuracy by between -0.7 and $+1.8$ pp across the 60 cells. We read this as an empirical observation that slot-level accuracy weighting has little room to improve on uniform MV in the near-exchangeable SC regime. It is not a bound on other weighting schemes, and confidence-based methods such as CISC Taubenfeld et al. [2025] use per-path signals that slot-level weights do not capture.

J Scope and Boundaries

What the framework enables. The saturation formula (5) uses the observable $\hat{c}$ directly at the vote level; Proposition 3.4 justifies treating $\hat{c}$ as a monotone proxy for the unobservable latent dependence. The GLS derivation establishes that, with equal path gains, uniform weighting is optimal among linear combiners of latent embeddings when the path covariance has equal row sums, as under exchangeability (Theorem 4.4, Corollary 4.5). Majority vote on decoded answers is a different, discrete rule, so this result motivates uniform voting without establishing its optimality (Remark 4.6). The channel model motivates path correlation as the diagnostic variable, and prompt-template perturbation moves it in all six task domains (§5.2.4).

Practical applicability. The framework is designed as an offline diagnostic tool for multi-path reasoning pipelines. Applying it to a new model-task pair requires (1) a calibration run of K=4 SC paths on $n \approx 50$–100 instances to estimate $\hat{c}$ (Appendix Q confirms n=50 yields $\hat{c}$ with CV $\leq 14.2\%$ and K^* std ≤ 2.0), and (2) evaluating the closed-form expressions for $K_{\text{eff}}^{\text{vote}}$ and K^*. The calibration run costs $4n$ path generations once per model-task pair, which is 12.5% of a K=32 budget on the calibration instances and is amortized over all later queries, which need no additional inference. With n=100, the one-time pilot costs 400 paths, and serving N queries at K^* costs $400 + K^*N$ paths against $32N$ at fixed K=32. On the five cells of Table 3 the pilot is repaid after 15–19 queries, and at N=1000 the total cost including the pilot is 13.8–32.5% of the fixed-K=32 budget. Calibration uses gold labels on the n pilot instances only, and deployment is label-free. The estimate belongs to one (model, prompt, task) configuration: a model update, a prompt change or a distribution shift defines a new configuration and triggers recalibration, and transfer of an operating point without recalibration is outside the claimed regime. For API-based deployments where hidden-state access is unavailable, the framework remains applicable: $\hat{c}$ is estimated from output-level correctness patterns alone, and the Adaptive-K rule operates entirely on observed vote statistics.

Equicorrelated approximation quality. The theoretical results assume Assumption 3.1 with a common correlation ρ. Prompt-template diversity breaks this symmetry: different templates produce heterogeneous slot accuracy. To quantify the approximation, we compute the variance-based $K_{\text{eff}}^{\text{block}} = \bar{p}(1-\bar{p})/\operatorname{Var}(S_K/K)$ from the full $K\times K$ correlation matrix across all 60 PT cells and

compare to the equicorrelated $K_{\text{eff}}^{\text{vote}}$. The mean ratio $K_{\text{eff}}^{\text{block}}/K_{\text{eff}}^{\text{vote}}$ is 1.01 (std 0.03, maximum 1.15 on Qwen-7B GSM8K), so the equicorrelated formula is accurate to within a few percent on average and slightly conservative. The binary-collapsed correctness model loses information about the wrong-answer distribution in multiclass settings, which $K_{\text{eff}}^{\text{vote}}$ does not capture.

K Statistical Precision of $\Delta\rho$ Estimates

Bootstrap confidence intervals (10,000 paired resamples of the pooled (seed, instance) rows per cell, with the same rows drawn for both arms) at the pooled sample size (n=500 for GSM8K, MATH, TriviaQA, DROP, CruxEval and MBPP, n=300 for the other six benchmarks) give per-cell 95% CI widths with median 16 pp (IQR 12–22 pp). All ten QA cells exclude zero (TriviaQA across all five models lies within $[-47\%, -3\%]$). For MATH, the Qwen-7B and Qwen-32B cells include zero (the Qwen-32B interval is $[-7\%, +3\%]$), consistent with their small effects. Across the 60 cells, 38 have CIs that exclude zero at the 95% level, 37 of them below zero, and the remaining 22 intervals include zero and do not resolve the sign of the effect. Per-cell $\Delta\rho$ estimates span $[-52\%, +90\%]$, and the positive extreme is the chance-level Qwen-0.5B WinoGrande cell discussed in §5.2.4. No cell has $\bar{p} < 2\%$, so none is excluded.

L Per-Cell Prompt-Template Results

Table 7 reports the per-benchmark means across the five models, and Table 8 lists all 60 model-benchmark cells behind them and Fig. 3.

Table 7: Prompt-template diversity across 12 benchmarks (K=8, 100 instances per seed, pooled over 5 seeds for GSM8K, MATH, TriviaQA, DROP, CruxEval and MBPP and over 3 seeds for the other six benchmarks). Five models evaluated: Qwen2.5-0.5B/7B/32B, Llama-3.1-8B, Mistral-7B. The *Answer space* column gives the effective evaluated output space (post-extraction value compared against gold), used as the relevant axis for the correlation analysis. $\Delta\rho$ is the *relative* change in pairwise correlation, $(\hat{\rho}_{\text{PT}} - \hat{\rho}_{\text{SC}})/\hat{\rho}_{\text{SC}}$, expressed as a percentage (so -20.3 on TriviaQA means correlation drops by 20.3% of its SC value, not by 20.3 percentage points). No cell falls below the 2% base-accuracy cut, so n=5 models for every benchmark, and each row reports the mean across them. The WinoGrande mean is driven by the chance-level Qwen-0.5B cell (median -3.2%). Per-cell values are in Table 8.

Domain	Benchmark	Answer space	n	Mean $\Delta\rho$ (%) ↓	Mean $\Delta K_{\text{eff}}^{\text{vote}}$ ↑	Mean $\hat{\rho}_{\text{SC}}$	Mean $\hat{\rho}_{\text{PT}}$
QA	TriviaQA	open text	5	-20.3	$+0.3$	0.70	0.57
	HotpotQA	open text	5	-25.0	$+0.4$	0.71	0.53
Science/MC	ARC-C	bounded (4)	5	-6.7	$+0.1$	0.54	0.51
	MMLU	bounded (4)	5	-20.5	$+0.4$	0.41	0.35
Commonsense	HellaSwag	bounded (4)	5	-9.0	$+0.2$	0.47	0.44
	WinoGrande	binary (2)	5	$+6.9$	$+0.0$	0.41	0.42
NLU	DROP	open text	5	-22.6	$+0.4$	0.59	0.45
	BoolQ	binary (y/n)	5	-16.9	$+0.2$	0.78	0.65
Code	MBPP	code pass/fail	5	-19.6	$+0.3$	0.62	0.49
	CruxEval	code pass/fail	5	-19.1	$+0.3$	0.53	0.42
Math	GSM8K	numeric	5	-33.6	$+0.7$	0.48	0.31
	MATH	numeric	5	-7.9	$+0.2$	0.51	0.47

M Prompt-Template MV Accuracy Deltas

Table 9 reports per-domain MV accuracy under SC and prompt-template (PT), and the difference $\Delta\text{Acc} = \text{MV}_{\text{PT}} - \text{MV}_{\text{SC}}$, averaged over the model-benchmark cells of each domain (K=8, seeds pooled as in Table 7). ΔAcc is small (within $\pm$2 pp on average) and mixed in sign across the six domains, with within-domain standard deviations comparable to or larger than the means. This pattern is consistent with the $K_{\text{eff}}^{\text{vote}}$ analysis. PT enlarges the saturation ceiling $1/c$ but does not by itself raise base per-path accuracy $\bar{p}$, so the operational MV accuracy remains near the SC level. The

diagnostic value of PT in this paper is therefore the *structure* it reveals (Section 5.2.4, Table 7), not a uniform accuracy improvement over SC.

N Reasoning Model

We apply the saturation and Adaptive-K protocol to Qwen3.5-9B in thinking mode on GSM8K, with 3 seeds (42, 123, 456) $\times$ 100 instances, K=32 paths per instance sampled at τ=0.7 and top-p 0.95, and a generation cap of 16,384 tokens. Each K is evaluated on the first K of the 32 paths. 5.4% of paths reach the cap, and these are scored as incorrect. The model is close to its accuracy ceiling on this task, with mean per-path accuracy 92.9%. The K=4 pilot gives $\hat{c}$=0.33, and at K=32 the measured $\hat{c}$=0.38 gives a ceiling $1/\hat{c}$=2.67 with $K_{\text{eff}}^{\text{vote}}$=2.53.

Table 10 reports the held-out beta-binomial test with the protocol of §5.2.3. The parameters (α, β) are fitted from the K=4 pilot on one half of the instances and used to predict MV@K on the other half. The beta-binomial error is 0.8–1.7 pp across K, whereas the independence model is off by 2.8–3.2 pp. Adaptive-K selects K^*=14 and retains 99.8% of MV@32 (paired bootstrap 95% CI of MV@K^*−MV@32: $[-0.5, 0.0]$ pp) at a net cost of 56% of the fixed K=32 budget. Harder benchmarks such as AIME need a generation budget above 16,384 tokens and are left to future work.

O Limitations and Future Work

Limitations. The Adaptive-K rule uses a global $\hat{c}$ estimate and does not adapt to per-instance difficulty variation; incorporating instance-level confidence into K^* is a natural refinement. The GLS analysis (Theorem 4.4) requires access to hidden-state embeddings; for API-only deployments the vote-level framework remains applicable with the observable $\hat{c}$ (Proposition 3.4), and the Adaptive-K rule operates entirely on output statistics. Per-cell $\Delta\rho$ estimates at the pooled sample size (n=300 or 500) have 95% CI widths with median 16 pp (IQR 12–22 pp) (Appendix K), and the intervals of individual small-effect cells, such as MATH on Qwen-7B and Qwen-32B, include zero and would need larger n to resolve the sign. Adaptive-K fixes one budget per (model, prompt, task) configuration before deployment, while online-stopping methods Aggarwal et al. [2023], Li et al. [2024], Wan et al. [2025] act per query during sampling. A matched-compute comparison with them remains open.

Future work. Optimal prompt-template design under a decorrelation–accuracy trade-off, and extensions to block-covariance or non-exchangeable regimes, remain open.

P Beta-Binomial Correlated Voting Model

Under the beta-binomial model $\Theta \sim \text{Beta}(\alpha, \beta)$, $Y_k \mid \Theta \sim \text{Bernoulli}(\Theta)$, the moments are $p = \mathbb{E}[\Theta] = \alpha/(\alpha+\beta)$ and, for $j \neq k$, $\text{Cov}(Y_j, Y_k) = \text{Var}(\Theta) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Dividing by $\text{Var}(Y_k) = p(1-p) = \alpha\beta/(\alpha+\beta)^2$ gives $c = 1/(\alpha+\beta+1)$, so $\alpha+\beta = (1-c)/c$, and the method-of-moments estimates follow from $\alpha = p(\alpha+\beta)$ and $\beta = (1-p)(\alpha+\beta)$:

$$\alpha = \frac{p(1-c)}{c}, \quad \beta = \frac{(1-p)(1-c)}{c}, \tag{19}$$

where $p = \mathbb{E}[Y_k]$ and $c = \text{Corr}(Y_i, Y_j) = 1/(\alpha+\beta+1)$. The probability mass function of the count $S_K = \sum_k Y_k$ is:

$$P(S_K = j) = \binom{K}{j} \frac{B(j+\alpha, K-j+\beta)}{B(\alpha, \beta)}. \tag{20}$$

The correlated majority-vote accuracy is:

$$P_{\text{MV}}^{\text{corr}} = \sum_{j > K/2} P(S_K{=}j). \tag{21}$$

For even K with random tie-breaking, add $\frac{1}{2}\, P(S_K{=}K/2)$.

Q Pilot Size Sensitivity

Table 11 shows the stability of $\hat{c}$ and K^* estimates as a function of pilot size n, computed via 500 bootstrap resamples of instances (with all seed replicates of an instance kept together) from the pooled 5-seed $K{=}4$ SC data on GSM8K. At $n{=}50$, $\hat{c}$ has a coefficient of variation of at most 14.2% across all three models and K^* standard deviation is 1.5–2.0, sufficient for practical guidance. At $n{=}25$, $\hat{c}$ CV reaches 20.5–21.5% and K^* variance widens (std 3.1–6.8), motivating $n{\geq}50$ as the minimum recommended pilot size.

R Proof of Corollary 4.7

Let $\mathbf{w}_u = (1/K)\mathbf{1}$. Since $\mathbf{w}_*$ minimizes $\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}$ subject to $\mathbf{g}^\top \mathbf{w} = 1$, and $\mathbf{w}_u$ is feasible when $\mathbf{g} = \mathbf{1}$, the optimal value cannot exceed the feasible value: $\mathrm{MSE}(\mathbf{w}_*) \leq \mathrm{MSE}(\mathbf{w}_u)$. Equality holds when $\mathbf{w}_u$ itself satisfies the KKT conditions, i.e., when $\boldsymbol{\Sigma}\mathbf{w}_u \propto \mathbf{g}$. For $\mathbf{g} = \mathbf{1}$, this requires $\boldsymbol{\Sigma}\mathbf{1} \propto \mathbf{1}$, i.e., all row sums of $\boldsymbol{\Sigma}$ are equal.

S Adaptive-K Algorithm

Algorithm 1 states the procedure behind §5.2.5. The pilot stage is the only stage that uses gold labels, and only on the n pilot instances. The deployment stage is label-free. The pilot's compute is $4n$ sampled paths once per (model, prompt, task) configuration, which the net-cost column of Table 3 already charges.

Algorithm 1 Adaptive-K path-budget selection

Require: model M, task instances $\mathcal{D}$, pilot size n (default 50–100), threshold $\varepsilon > 0$ (default 0.025), maximum budget $K_{\max}$ (default 32)
Ensure: operating point K^* and final answers
1: **Pilot:** for each of n pilot instances, sample $K_0{=}4$ paths from M
2: Extract discrete answers and score them against gold to obtain $Y_i^{(k)}$
3: Estimate $\bar{p}$ and $\hat{c}$ from the $Y_i^{(k)}$ by (3)
4: **if** $\bar{p} \in \{0, 1\}$ **then**
5: $\quad K^* \leftarrow K_{\max}$ {degenerate pilot}
6: **else**
7: $\quad \hat{c} \leftarrow \mathrm{clip}(\hat{c}, 0.05, 0.99)$; $K^* \leftarrow \min\left(\max(\lceil(\sqrt{(1{-}\hat{c})/\varepsilon} - 1)/\hat{c} + 1\rceil, 1), K_{\max}\right)$ by (12)
8: **end if**
9: **Deployment:** for each new instance, sample K^* paths, extract answers and return the plurality vote

T Adaptive-K Threshold Sensitivity

Table 12 shows that the Adaptive-K rule is robust to the threshold ε: across a $10\times$ range ($\varepsilon \in [0.01, 0.1]$), retained accuracy stays within 97–103% of MV@$K{=}32$ for all three models.

On retention above 100%. Several cells in Tables 3 and 12 show MV@K^* exceeding MV@$K{=}32$ (e.g., Mistral-7B GSM8K at 103%, $\varepsilon{=}0.025$). This is consistent with finite-sample variance: a paired bootstrap over instances puts the 95% CI of MV@$K^*{-}$MV@32 around zero in all five cells of Table 3, so a gap of this size is within sampling noise. The saturation prediction is therefore a guideline for the operating point rather than a strict upper bound on observed accuracy.

Net compute including pilot. The net cost in Table 3 includes the pilot. A deployment that estimates $\hat{c}$ from a fresh $K{=}4$ pilot pays for those four pilot paths in addition to the K^* operating-point paths, so the net inference cost relative to full $K{=}32$ self-consistency is $(K^* + 4)/32$. For our five cells this is 25–44% (vs. a nominal $K^*/32$ of 13–31%); in amortized deployments where $\hat{c}$ is reused across many queries the pilot cost vanishes and the nominal figure applies.

U Benchmark Details

We evaluate on 12 benchmarks spanning six task domains: *Math*: GSM8K Cobbe et al. [2021], MATH Hendrycks et al. [2021b]; *QA*: HotpotQA Yang et al. [2018], TriviaQA Joshi et al. [2017]; *Science/MC*: ARC-Challenge Clark et al. [2018], MMLU Hendrycks et al. [2021a]; *Code*: MBPP Austin et al. [2021], CruxEval Gu et al. [2024]; *Commonsense*: HellaSwag Zellers et al. [2019], WinoGrande Sakaguchi et al. [2020]; *Natural language understanding (NLU)*: BoolQ Clark et al. [2019], DROP Dua et al. [2019].

These benchmarks span a range of *answer-space structures*, classified by the *effective evaluated output space* (the post-extraction value compared against the gold answer): *numeric* (GSM8K, MATH), *open text* (HotpotQA, TriviaQA, DROP; scored by extractive matching), *bounded choice* of 3–4 options (ARC, MMLU, HellaSwag), *binary* (BoolQ as yes/no, WinoGrande as 2-way), and *code pass/fail* (MBPP, CruxEval; scored by execution outcome). For code, the grading space is the binary pass/fail signal returned by the test suite, which is the relevant axis for the correlation analysis. The per-benchmark labels appear in the *Answer space* column of Table 7.

Licenses. Qwen2.5-0.5B/7B/32B-Instruct, Qwen3.5-9B and Mistral-7B-Instruct-v0.3 are released under Apache-2.0, and Llama-3.1-8B-Instruct under the Llama 3.1 Community License. GSM8K, MATH, MMLU, HellaSwag and CRUXEval are released under MIT, and TriviaQA under Apache-2.0. HotpotQA, ARC and DROP are released under CC BY-SA 4.0, BoolQ under CC BY-SA 3.0, MBPP under CC BY 4.0, and WinoGrande under CC BY. All assets are used for evaluation only. The released generation records contain the benchmarks' gold answers and the models' outputs, and their dataset card states these license terms.

V Prompt Templates

Math tasks (GSM8K, MATH). $K{=}8$ templates:

1. Standard CoT: "Solve step by step."
2. Algebra: "Use algebraic equations. Define variables, write equations, solve."
3. Estimate+Precise: "First estimate, then solve precisely."
4. Decompose: "Break into sub-problems, solve each, combine."
5. Work Backwards: "Work backwards from a hypothetical answer, then solve forward."
6. Verify: "Solve, then verify by substituting back."
7. Concise: "Show key steps only."
8. Explain: "Explain as if teaching a student."

QA tasks (HotpotQA, TriviaQA). $K{=}8$ templates:

1. Standard: "Answer based on the context."
2. Chain-of-thought: "Reason step by step."
3. Extract-then-answer: "Identify key facts, then answer."
4. Direct: "Give a short, direct answer."
5. Verify: "Answer, then verify against context."
6. Decompose: "Break into sub-questions, answer each, combine."
7. Concise: "Answer in as few words as possible."
8. Explain: "Explain reasoning thoroughly, then give final answer."

Science/MC tasks (ARC-Challenge, MMLU). $K{=}8$ templates:

1. Direct: "Answer the following multiple choice question."
2. Step-by-step: "Think step by step, then select the best answer."
3. Elimination: "Eliminate wrong answers first, then choose."

4. Letter only: “Give the answer directly with just the letter.”
5. Explain options: “Explain why each option is right or wrong, then select.”
6. Scientific: “Use your scientific knowledge to answer.”
7. Real-world: “Consider real-world examples to determine the answer.”
8. Teacher: “What would a teacher say is the correct answer?”

Code generation (MBPP). $K{=}8$ templates:

1. Standard: “Complete the following Python function.”
2. Algorithmic: “Think about the approach step by step, then complete.”
3. Test-driven: “Consider what test cases to handle, then implement.”
4. Concise: “Write the most concise implementation possible.”
5. Defensive: “Write a robust implementation with edge case handling.”
6. Efficient: “Choose the most efficient algorithm and implement.”
7. Readable: “Write clean, readable code with meaningful variable names.”
8. Alternative: “Think of an alternative approach, then implement.”

Code understanding (CruxEval). $K{=}8$ templates:

1. Direct: “What is the output of the following Python code?”
2. Trace: “Trace through this code step by step, then give the output.”
3. Mental execution: “Execute this function mentally and predict the result.”
4. Output only: “Give only the output value, nothing else.”
5. Return value: “What does the function return?”
6. Logic: “Analyze the code logic, then predict the output.”
7. Edge cases: “Think about edge cases, then give the output.”
8. Simulate: “Simulate a Python interpreter running this code.”

Commonsense (HellaSwag). $K{=}8$ templates:

1. Most likely: “Choose the most likely continuation of the scenario.”
2. Step-by-step: “Think step by step about what happens next, then select.”
3. Elimination: “Eliminate unlikely continuations first, then choose.”
4. Letter only: “Give just the letter of the most likely continuation.”
5. Common sense: “Consider real-world common sense to determine the continuation.”
6. Visualization: “Visualize the scenario, then choose what would happen next.”
7. Cause-effect: “Think about cause and effect to select the continuation.”
8. Natural: “What would naturally follow in this situation?”

Commonsense (WinoGrande). $K{=}8$ templates:

1. Standard: “Which option best fills the blank?”
2. Context clues: “Think about context clues step by step.”
3. Elimination: “Consider both options and eliminate the wrong one.”
4. Binary: “Give just A or B.”
5. Real-world: “Use real-world knowledge to decide.”
6. Careful reading: “Read carefully, focus on meaning, and select.”
7. Logical: “Which option makes the sentence logically coherent?”
8. Semantic: “Think about what makes grammatical and semantic sense.”

Reading comprehension (BoolQ). $K{=}8$ templates:

1. Standard: "Based on the passage, answer Yes or No."
2. Chain-of-thought: "Read carefully and reason step by step, then answer."
3. Evidence first: "Find the relevant evidence, then answer."
4. Direct: "Answer only Yes or No, nothing else."
5. Quote: "First quote the relevant passage part, then answer."
6. Support/contradict: "Does the passage support or contradict the question?"
7. Stated vs. implied: "What does the passage actually say vs. imply?"
8. Explicit or inferred: "Is the answer explicitly stated or inferred?"

Discrete reasoning (DROP). $K{=}8$ templates:

1. Direct: "Read the passage and answer the question."
2. Step-by-step: "Think step by step, count or compare as needed."
3. Identify facts: "Find the relevant numbers or facts, then answer."
4. Short answer: "Give a short, direct answer."
5. Show reasoning: "Show your arithmetic or reasoning, then answer."
6. Focus details: "Focus on the specific details asked about."
7. Extract-compute: "Extract relevant information, then compute the answer."
8. Trace quantities: "Trace the quantities mentioned to answer."

W Proof of Proposition 3.4 (Monotonicity of c in ρ_d)

Under the Gaussian decision surrogate (Assumption 3.3), each path's correctness is determined by thresholding a Gaussian decision score: $Y_k = \mathbf{1}\{M_k > 0\}$ with $\Pr(M_k > 0) = p$ and $\mathrm{Corr}(M_j, M_k) = \rho_d$.

Consider two paths j, k. The joint correctness probability is:

$$\Pr(Y_j{=}1, Y_k{=}1) = \Phi_2\big(\Phi^{-1}(p),\, \Phi^{-1}(p);\, \rho_d\big)\,, \tag{22}$$

where $\Phi_2(\cdot, \cdot; \rho_d)$ is the bivariate standard normal CDF with correlation ρ_d, and $\Phi^{-1}(p)$ is the threshold corresponding to single-path accuracy p. This is a direct application of the thresholded-Gaussian representation in Assumption 3.3.

The correctness correlation is then:

$$c(\rho_d) = \frac{\Phi_2(\Phi^{-1}(p), \Phi^{-1}(p); \rho_d) - p^2}{p(1-p)}. \tag{23}$$

To show monotonicity, we use the known result that the bivariate normal CDF $\Phi_2(a, b; \rho)$ is strictly increasing in ρ for all fixed a, b, because its derivative in ρ equals the bivariate normal density $\phi_2(a, b; \rho) > 0$ (Plackett's identity Plackett [1954]). Since $p \in (0, 1)$ is fixed, $\Phi^{-1}(p)$ is finite, and both p^2 and $p(1-p)$ are constants with respect to ρ_d. Therefore $c(\rho_d) = [\Phi_2(\cdot; \rho_d) - \text{const}]/\text{const}$ inherits strict monotonicity: $\partial c / \partial \rho_d > 0$ for all $\rho_d \in (0, 1)$.

This confirms that higher decision-layer correlation always produces higher binary correctness correlation, justifying the use of $\hat{c}$ as a monotone proxy for the unobservable ρ_d.

X Derivation for Remark 4.3 (Multiclass Plurality Heuristic)

Consider $|\mathcal{A}| = M > 2$ answer classes. Each path produces the correct answer with probability p and an incorrect answer with probability $1 - p$. Under uniform fragmentation, wrong answers are spread equally among $M{-}1$ alternatives, so each wrong answer has probability $(1{-}p)/(M{-}1)$.

For plurality vote, the correct answer wins if it receives more votes than every individual wrong answer. We reduce this to a pairwise contest: the correct answer (p) competes against the single most popular wrong alternative ($(1-p)/(M-1)$).

The effective binary accuracy for this pairwise contest is:

$$p' = \frac{p}{p + (1-p)/(M-1)} = \frac{p(M-1)}{p(M-1) + (1-p)} = \frac{p(M-1)}{pM - 2p + 1}. \tag{24}$$

For M=4 and p=0.5: $p' = (0.5 \times 3)/(0.5 \times 4 - 1 + 1) = 1.5/2 = 0.75$. More generally, $p' > p$ whenever $M > 2$. This pairwise reduction indicates the size of the plurality advantage under uniform fragmentation and is a heuristic. The ordering itself holds without it. On every draw in which the binary majority vote is correct, the correct answer holds more than $K/2$ votes, so no other answer can match it and plurality vote is also correct. When the correct answer holds exactly $K/2$ votes (even K), the binary vote scores $1/2$, and plurality vote either wins outright, if the remaining votes are split, or ties with one wrong answer and scores $1/2$. Hence $P_{\text{PV}}^{(M)} \geq P_{\text{MV}}^{(2)}$ for any joint distribution of the K answers.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state four contributions that are delivered in the paper: (i) the saturation law $K_{\text{eff}}^{\text{vote}} = K/(1 + (K-1)c)$ (§4, Theorem 4.1, Appendix G); (ii) the condition, equal row sums of the path covariance with equal path gains, under which uniform weights are the optimal linear combiner (Theorem 4.4, Appendix H); (iii) prompt-template decorrelation measured on all 60 cells of five models $\times$ twelve benchmarks (§5.2.4, Table 7, Appendix L); and (iv) an Adaptive-K rule validated on Math/QA/NLU (§5.2.5, Table 3).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix O discusses limitations explicitly: the Adaptive-K rule uses a global $\hat{c}$ without per-instance adaptation; the GLS analysis requires hidden-state embeddings, whereas the vote-level framework and Adaptive-K use output statistics only; a matched-compute comparison with online-stopping methods remains open; and per-cell $\Delta\rho$ estimates at the pooled sample size (n=300 or 500) have wide 95% CIs for small-effect cells (Appendix K).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each formal result states its assumptions inline: the design-effect / saturation results rest on the exchangeable-paths model (Assumption 3.1); Theorem 4.4 requires only a positive-definite $\boldsymbol{\Sigma}$ in the linear aggregation model of §4.3 and is independent of the channel model; Proposition 3.4 relies on the Gaussian decision surrogate (Assumption 3.3). Full proofs appear in the appendix: Appendix D and Theorem 4.1 in Appendix G; Theorem 4.4 in Appendix H; Corollary 4.7 in Appendix R; Proposition 3.4 in Appendix W; the binary-collapse scope in Appendix E; and the multiclass plurality heuristic in Appendix X.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §5 specifies models (Qwen2.5-0.5B/7B/32B, Llama-3.1-8B, Mistral-7B-Instruct-v0.3, Qwen3.5-9B), benchmarks, the five-seed and three-seed protocols with their benchmark assignment, and sample sizes; Appendix V gives the $K{=}8$ prompt

templates verbatim; Appendix Q and T document pilot-size and threshold sensitivity. Code reproducing every table and figure is released at `https://github.com/OniReimu/DiversityCombining`, and the generation records it reads are released at `https://huggingface.co/datasets/UTSCybeR/DiversityCombining`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is released at `https://github.com/OniReimu/DiversityCombining`, including experiment drivers, aggregation and analysis scripts (`scripts/reproduce_tables.py` for the saturation, BB and Adaptive-K results including Figure 2, `scripts/reproduce_cross_benchmark.py` for the cross-benchmark $\Delta\rho$ matrix, and `scripts/plot_drho_cells.py` for Figure 3), and a `README` with environment setup, exact commands, and a per-artifact mapping to paper tables and figures. The generation records behind every reported number are released at `https://huggingface.co/datasets/UTSCybeR/DiversityCombining`. Underlying benchmarks (GSM8K, HotpotQA, BoolQ, *etc.*) are loaded directly from Hugging Face Datasets with explicit split identifiers in `diversity_combining/evaluation/data_loader.py`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper performs no model training; all experiments are inference-time evaluations of open-weight LLMs. §5 reports the decoding hyperparameters (temperature τ=0.7, $K \in \{4, 8, 16, 32\}$, per-benchmark maximum generation lengths of 256 to 4096 tokens for the cross-benchmark experiments, 1024/512/256 tokens for GSM8K/HotpotQA/BoolQ in the saturation and Adaptive-K experiments, and 16,384 tokens for the reasoning model, bf16 on NVIDIA H100), the per-seed and pooled sample sizes, and the held-out pilot/eval split protocol for beta-binomial calibration. The prompt-template slot set is reproduced verbatim in Appendix V, and the exact seed values and pooling procedure are documented in the release `README` together with runnable commands.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Saturation cells pool five independent seeds, except the HotpotQA and BoolQ Adaptive-K cells and the reasoning-model study, which pool three, and cross-benchmark cells pool five or three seeds (§5); per-seed SC accuracies are reported as mean $\pm$ standard deviation (Table 1). Per-cell $\Delta\rho$ estimates use 10,000-replicate paired bootstrap 95% CIs (Appendix K), and the QA-vs-math comparison uses a Mann-Whitney U test (§5.2.4).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §5 reports the compute workers (NVIDIA H100 GPUs, bf16 precision) and the model scales involved (0.5B–32B parameters). No training is performed; all experiments are inference sweeps, and the release `README` provides an order-of-magnitude estimate for the full reproduction cost. Per-experiment wall-clock is not tabulated individually, as drivers are resume-safe and partial reproductions (e.g., a single model on GSM8K) scale proportionally with the pooled sample size.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses only publicly released open-weight LLMs and widely adopted academic benchmarks; no human subjects, private data, or models with known high misuse potential are involved. The released code and all prose comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper contributes a diagnostic analytical framework for existing multi-path reasoning methods rather than a new capability, model, or dataset. The main positive effect is incidental compute and energy savings: the Adaptive-K rule retains 96–103% of $K{=}32$ accuracy at a smaller predicted K^*, reducing inference cost. We identify no plausible path to novel misuse beyond risks already present in the underlying open-weight LLMs.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We release evaluation and analysis code (experiment drivers plus aggregation and plotting scripts) and the sampled outputs of public open-weight models on public benchmarks. No new pre-trained models, generative systems, or scraped datasets are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models (Qwen2.5, Qwen3.5, Llama-3.1, Mistral-7B-Instruct-v0.3) and benchmarks (GSM8K, MATH, HotpotQA, TriviaQA, ARC, MMLU, HellaSwag, WinoGrande, BoolQ, DROP, MBPP, CruxEval) are cited at first use in §5 and accessed via their canonical Hugging Face identifiers at runtime. The released generation records contain benchmark gold answers and model outputs, and their dataset card states the license of each source. The license of each asset is listed in Appendix U.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets are the analysis and evaluation code released at `https://github.com/OniReimu/DiversityCombining` and the generation records released at `https://huggingface.co/datasets/UTSCybeR/DiversityCombining`, documented by a dataset card with field descriptions, generation settings and licenses. The release includes a `README` with setup, usage, and per-artifact reproduction instructions, and is structured as a standard Python package (`pyproject.toml`, module docstrings, explicit CLI flags).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects research; all experiments are automated evaluations of LLMs on public benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in the research; IRB review is therefore not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are the evaluation target of this paper (we study the statistics of multi-path reasoning *on* LLMs), not a tool embedded in the core methodology. The analytical framework itself (saturation law, GLS optimality, beta-binomial calibrator, Adaptive-K) is derived independently of any specific LLM.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Table 8: All 60 model-benchmark cells (K=8). n is the number of pooled instances, $\bar{p}$ the mean per-path accuracy (%), $\hat{\rho}$ the mean pairwise correlation of slot correctness, $\Delta\rho$ the relative change (%) with its paired bootstrap 95% CI, and MV the majority-vote accuracy (%) with ties scored as 0.5.

Benchmark	Model	n	$\bar{p}_{\text{SC}}$	$\bar{p}_{\text{PT}}$	$\hat{\rho}_{\text{SC}}$	$\hat{\rho}_{\text{PT}}$	$\Delta\rho$	95% CI	MV_{SC}	MV_{PT}
TriviaQA	Qwen-0.5B	500	37.3	34.8	0.65	0.55	−14.8	[−19.4, −10.0]	36.8	34.6
	Qwen-7B	500	62.0	61.1	0.87	0.81	−6.2	[−9.3, −3.1]	62.4	62.9
	Qwen-32B	500	67.1	63.8	0.84	0.76	−9.0	[−12.5, −5.8]	67.4	63.4
	Llama-8B	500	30.1	30.9	0.52	0.36	−30.6	[−36.9, −24.0]	29.1	28.7
	Mistral-7B	500	28.6	25.9	0.61	0.36	−41.1	[−46.7, −35.4]	26.9	23.3
HotpotQA	Qwen-0.5B	300	27.7	24.3	0.68	0.62	−8.8	[−15.8, −1.7]	27.2	21.0
	Qwen-7B	300	66.5	69.5	0.86	0.60	−30.4	[−36.7, −24.4]	66.7	70.8
	Qwen-32B	300	80.4	80.3	0.79	0.64	−18.9	[−25.6, −13.1]	80.2	83.5
	Llama-8B	300	50.9	50.6	0.59	0.37	−36.5	[−43.8, −28.7]	51.0	54.8
	Mistral-7B	300	30.2	31.3	0.60	0.42	−30.2	[−36.8, −23.4]	30.0	29.5
ARC-C	Qwen-0.5B	300	33.1	32.8	0.22	0.18	−20.5	[−39.8, +4.2]	24.8	22.5
	Qwen-7B	300	91.0	90.4	0.70	0.65	−7.3	[−17.9, +4.7]	92.5	92.0
	Qwen-32B	300	92.6	92.9	0.76	0.71	−7.2	[−15.4, +0.7]	93.7	94.0
	Llama-8B	300	76.2	76.8	0.51	0.52	+1.8	[−8.2, +13.0]	77.8	79.3
	Mistral-7B	300	80.4	74.9	0.52	0.52	−0.2	[−11.6, +13.7]	83.0	76.5
MMLU	Qwen-0.5B	300	25.3	24.0	0.07	0.04	−39.8	[−77.0, +25.4]	11.2	8.5
	Qwen-7B	300	70.8	69.4	0.63	0.55	−12.4	[−19.4, −4.9]	72.7	71.7
	Qwen-32B	300	85.2	84.0	0.65	0.60	−8.1	[−16.1, +0.4]	86.3	85.2
	Llama-8B	300	41.7	40.8	0.38	0.28	−25.3	[−35.9, −13.4]	41.3	39.3
	Mistral-7B	300	35.6	33.4	0.33	0.27	−16.7	[−31.5, +0.4]	32.7	27.7
HellaSwag	Qwen-0.5B	300	32.2	32.5	0.12	0.09	−23.4	[−49.5, +16.0]	22.2	23.5
	Qwen-7B	300	76.5	74.0	0.62	0.59	−4.8	[−12.6, +3.7]	79.2	75.3
	Qwen-32B	300	82.0	81.4	0.67	0.64	−4.9	[−11.9, +2.0]	82.8	82.7
	Llama-8B	300	59.0	59.1	0.46	0.38	−16.9	[−25.7, −7.6]	61.8	61.8
	Mistral-7B	300	61.5	60.8	0.48	0.50	+5.3	[−5.5, +17.7]	63.8	61.7
WinoGrande	Qwen-0.5B	300	49.2	44.8	0.23	0.43	+89.6	[+59.1, +129.5]	50.5	45.7
	Qwen-7B	300	76.1	75.8	0.51	0.49	−3.2	[−12.4, +6.9]	79.2	78.7
	Qwen-32B	300	84.6	83.4	0.66	0.65	−1.4	[−9.3, +7.0]	86.2	84.7
	Llama-8B	300	62.8	62.3	0.26	0.15	−42.5	[−56.4, −27.1]	66.0	68.7
	Mistral-7B	300	63.6	60.1	0.40	0.37	−8.1	[−21.8, +8.0]	66.8	60.8
DROP	Qwen-0.5B	500	15.6	15.2	0.48	0.40	−15.4	[−24.3, −5.6]	14.0	11.8
	Qwen-7B	500	56.0	64.8	0.76	0.43	−43.0	[−48.5, −37.6]	54.6	69.2
	Qwen-32B	500	78.1	79.5	0.63	0.62	−1.2	[−7.0, +5.1]	81.1	82.3
	Llama-8B	500	35.5	34.2	0.43	0.31	−28.6	[−36.6, −20.1]	34.7	30.9
	Mistral-7B	500	12.0	10.4	0.64	0.48	−24.8	[−33.9, −16.5]	10.8	7.1
BoolQ	Qwen-0.5B	300	60.6	52.5	0.49	0.44	−10.9	[−19.4, −1.5]	64.2	53.2
	Qwen-7B	300	81.7	83.0	0.80	0.73	−8.7	[−16.5, −0.6]	81.0	83.5
	Qwen-32B	300	85.0	85.9	0.82	0.73	−10.7	[−17.0, −5.0]	84.7	87.2
	Llama-8B	300	80.0	78.5	0.80	0.47	−41.6	[−48.3, −34.9]	79.7	80.7
	Mistral-7B	300	83.8	81.9	0.98	0.86	−12.6	[−17.1, −8.6]	83.8	81.8
MBPP	Qwen-0.5B	500	31.9	32.1	0.41	0.40	−2.4	[−11.7, +7.9]	31.5	29.8
	Qwen-7B	500	80.8	77.5	0.82	0.59	−27.3	[−32.7, −22.0]	81.2	80.4
	Qwen-32B	500	89.7	87.7	0.74	0.59	−19.3	[−27.3, −11.2]	91.4	90.4
	Llama-8B	500	74.1	69.3	0.71	0.54	−24.3	[−29.2, −19.5]	75.3	73.1
	Mistral-7B	500	37.1	35.7	0.45	0.34	−24.6	[−32.2, −16.4]	39.1	36.7
CruxEval	Qwen-0.5B	500	12.2	10.1	0.62	0.54	−12.5	[−21.7, −3.4]	10.2	8.7
	Qwen-7B	500	58.9	57.6	0.56	0.46	−16.5	[−21.7, −11.0]	59.0	61.4
	Qwen-32B	500	83.0	80.5	0.57	0.49	−14.6	[−21.3, −7.9]	84.1	85.1
	Llama-8B	500	16.2	29.8	0.24	0.21	−11.0	[−27.9, +11.7]	9.6	25.3
	Mistral-7B	500	34.1	32.7	0.66	0.39	−41.0	[−46.3, −35.3]	34.5	32.9
GSM8K	Qwen-0.5B	500	33.3	31.1	0.37	0.37	+1.3	[−9.4, +13.1]	31.0	27.3
	Qwen-7B	500	79.2	70.7	0.59	0.31	−48.4	[−54.6, −42.0]	82.0	80.9
	Qwen-32B	500	80.5	73.3	0.45	0.22	−51.9	[−59.0, −44.5]	83.9	84.1
	Llama-8B	500	75.1	69.0	0.52	0.36	−30.1	[−36.6, −23.3]	78.1	73.6
	Mistral-7B	500	42.8	38.0	0.46	0.28	−39.0	[−46.0, −31.8]	42.8	37.4
MATH	Qwen-0.5B	500	21.1	24.0	0.33	0.37	+13.3	[−1.0, +29.7]	16.0	19.9
	Qwen-7B	500	75.8	74.6	0.63	0.61	−2.9	[−8.2, +2.6]	75.9	75.8
	Qwen-32B	500	81.1	80.8	0.67	0.66	−2.2	[−7.4, +3.2]	83.8	84.0
	Llama-8B	500	45.1	43.4	0.57	0.49	−14.3	[−20.4, −8.1]	45.9	43.3
	Mistral-7B	500	14.1	11.9	0.35	0.23	−33.3	[−45.3, −19.5]	10.3	6.5

Table 9: Per-domain MV accuracy under SC and PT (K=8, seeds pooled as in Table 7). n = number of model-benchmark cells (none falls below the 2% base-accuracy cut). $\pm$ denotes within-domain standard deviation across cells.

Domain	n	$\mathrm{MV}_{\mathrm{SC}}$	$\mathrm{MV}_{\mathrm{PT}}$	ΔAcc (pp)
QA	10	47.8%	47.3%	-0.5 ± 3.5
Sci/MC	10	61.6%	59.7%	-1.9 ± 2.4
Commonsense	10	65.8%	64.3%	-1.5 ± 2.7
NLU	10	58.9%	58.8%	-0.1 ± 6.5
Code	10	51.6%	52.4%	$+0.8 \pm 5.4$
Math	10	55.0%	53.3%	-1.7 ± 2.8

Table 10: Qwen3.5-9B (thinking) on GSM8K, 3 seeds $\times$ 100 instances. Observed binary MV and held-out absolute prediction errors (pp), averaged over the two instance halves.

K	Observed MV	Beta-binomial error	Independence error
4	95.8%	0.8	2.8
8	96.8%	1.2	3.1
16	96.8%	1.3	3.2
32	96.8%	1.7	3.2

Table 11: Pilot size sensitivity on GSM8K (K=4, 500 bootstrap resamples of instances from pooled 5-seed data).

Model	n	$\hat{c}$ mean	$\hat{c}$ std	K^* mean	K^* std
Qwen-7B	25	0.571	0.122	7.9	6.77
	50	0.589	0.083	6.9	1.69
	100	0.596	0.055	6.6	1.02
	200	0.605	0.039	6.5	0.72
Llama-8B	25	0.512	0.105	8.7	3.14
	50	0.522	0.068	8.1	1.53
	100	0.526	0.048	8.0	1.06
	200	0.525	0.034	7.9	0.77
Mistral-7B	25	0.437	0.090	10.6	3.10
	50	0.438	0.062	10.3	1.98
	100	0.447	0.044	9.9	1.30
	200	0.446	0.031	9.9	0.92

Table 12: Sensitivity of Adaptive-K to threshold ε on GSM8K (n=100 instances, 5 seeds pooled). Retained = MV@K^* as a percentage of MV@K=32.

Model	ε	K^*	MV@K^*	Retained
Qwen-7B	0.01	10	82.4%	101%
	0.025	6	81.6%	100%
	0.05	5	81.4%	99%
	0.1	3	79.8%	97%
Llama-8B	0.01	13	78.8%	99%
	0.025	8	78.1%	98%
	0.05	5	76.8%	97%
	0.1	4	76.7%	97%
Mistral-7B	0.01	16	42.4%	100%
	0.025	10	43.6%	103%
	0.05	7	43.6%	103%
	0.1	5	41.6%	98%