ONE MODEL, MANY MINDS: STRATEGY-DIVERSIFIED REASONING IS A CROSS-MODEL ENSEMBLE IN DISGUISE

Anonymous authors
Paper under double-blind review

ABSTRACT

The dominant recipe for test-time scaling is to *sample more*: draw many reasoning chains from one model and aggregate them. Yet chains from the same model under the same prompt share the same failure modes, so their errors are correlated and accuracy saturates well before the compute budget does. Cross-model ensembles break this correlation, since different architectures fail differently, but at the price of running several large models. We ask whether that diversity is genuinely *architectural*, or merely *structural* and reproducible within a single model. By forcing one model through deliberately distinct reasoning strategies (algebraic, geometric, backward, inductive, analogical), we decorrelate its errors most of the way toward switching models. Across four diversity sources, the mean pairwise error correlation follows a clean hierarchy: temperature (0.69) and prompt rephrasing (0.71) keep errors tightly coupled, whereas strategy diversification drops them to 0.60, a third of the way to a five-model ensemble (0.43), without lowering per-attempt accuracy, and so attains the highest answer coverage of any source. At matched compute, the resulting single-model ensemble, **One Model, Many Minds (OMMM)**, matches or beats the compute-matched baselines and rivals far costlier methods such as multi-agent debate. The decorrelation also buys calibration: because the strategies fail independently, the rate at which they agree is a free, well-calibrated confidence signal, and answering only when the strategies agree raises accuracy *on the answered AIME problems* from 37% to 100% (a selective-prediction gain, at reduced coverage). What decorrelates a model's errors is not how *many* times it answers but how *differently* it reasons.

1 INTRODUCTION

Chain-of-thought (CoT) prompting made step-by-step reasoning the default interface to large language models (Wei et al., 2022; Kojima et al., 2022), and the prevailing way to spend *more* test-time compute on it is to *sample more*: draw many chains of thought and aggregate them, as in Self-Consistency (Wang et al., 2023) and Best-of-N (Lightman et al., 2023). This treats each sample as an independent draw at the answer, so that coverage grows and the majority vote sharpens with scale (Brown et al., 2024; Snell et al., 2025). But samples from *one model* at *one prompt* are not independent: sharing the same weights, data, and prompt, they tend to make the *same mistakes*, since a model that misreads a problem geometrically misreads it the same way on resample. Self-consistency therefore saturates within a handful of samples (Sprague et al., 2024): "diminishing returns" names the symptom, while the cause is *error correlation*.

Ensemble theory has long made this precise. The ambiguity decomposition (Krogh & Vedelsby, 1994) writes an ensemble's error as the average member error minus a diversity term that grows as members *decorrelate*. An ensemble of identical predictors has zero diversity and no benefit, and decorrelated members are exactly what make deep ensembles (Lakshminarayanan et al., 2017) and random forests (Breiman, 2001) work. This is also why *cross-model* ensembles of language models help: different architectures, trained differently, have different failure modes, so aggregating or fusing them genuinely improves accuracy (Jiang et al., 2023; Wang et al., 2024). The catch is cost: a cross-model ensemble runs several large models, each with its own deployment.

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

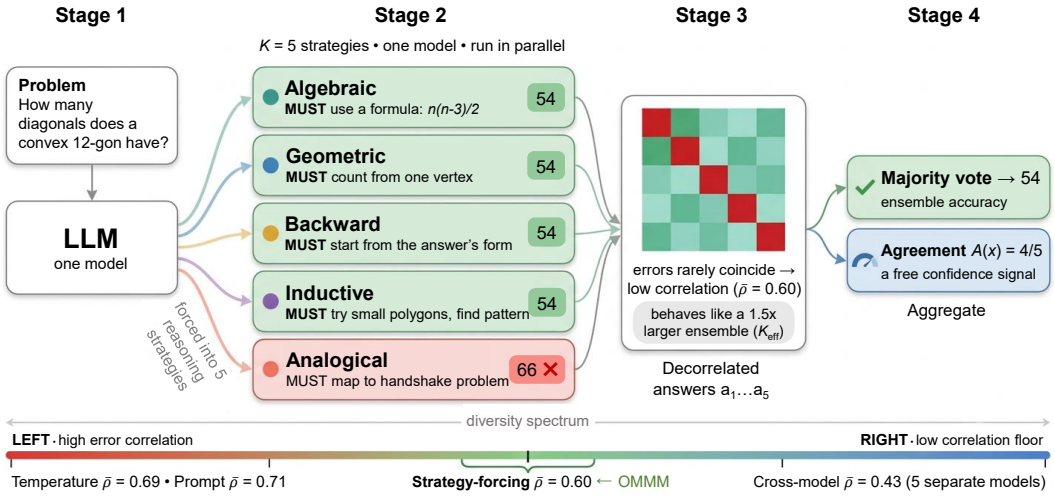

Figure 1: **One Model, Many Minds (OMMM).** A single model solves the same problem under K reasoning strategies that invoke structurally different computations, each enforced by a system prompt with explicit `MUST`/`MUST NOT` constraints. Their errors are therefore *decorrelated* (an algebraic mistake and a geometric mistake rarely coincide), so a majority vote behaves like a much larger ensemble, and cross-strategy *agreement* $A(x)$ doubles as a free confidence signal. The K generations are independent and run in parallel, so the wall-clock cost is that of a single solution.

We ask a simple question: *is the decorrelation that cross-model ensembles buy actually about the models, or about the diversity of computation they induce?* If the latter, we should be able to induce the same decorrelation *inside a single model*, by making it compute differently each time. Our lever is to force the model into structurally distinct *reasoning strategies* (solving the problem algebraically, geometrically, backward, inductively, or by analogy), each enforced by a short system prompt with explicit `MUST`/`MUST NOT` constraints, echoing classical taxonomies of problem solving (Pólya, 1945) and going beyond prior prompt- and perspective-diversity methods that vary the surface form while leaving the underlying computation unchanged (Li et al., 2023; Naik et al., 2023; Zhang et al., 2024). An algebraic error and a geometric error arise from different computations, so they rarely coincide, which is precisely the decorrelation an ensemble needs.

We call this **One Model, Many Minds (OMMM)** and study it through the lens of ensemble diversity. Our central experiment fixes a model and a problem set and measures pairwise error correlation ρ under four diversity sources at a matched budget of five generations: temperature sampling, prompt rephrasing, strategy-forcing (ours), and a true cross-model ensemble. We find a clean ordering (Figure 2): temperature and prompt variation leave errors strongly correlated, while strategy-forcing decorrelates them substantially further, partway toward the cross-model floor, raising the effective ensemble size above that of self-consistency. Because the strategies run independently and in parallel, OMMM's wall-clock cost is that of a single generation. Our core contributions are:

- We diagnose the bottleneck of test-time scaling by giving the first systematic measurement of how LLM error correlation depends on the *source* of diversity (§5): temperature ($\bar{\rho}$=0.69) and prompt (0.71) stay strongly correlated, strategy-forcing drops to 0.60, and cross-model to 0.43, an ordering that explains *why* self-consistency saturates and identifies the source of diversity, more than its amount, as the lever.
- We introduce *strategy-forcing* and show the lever is structural: making one model compute differently decorrelates its errors *without* lowering per-attempt accuracy, unlike a cross-model ensemble, which decorrelates only by admitting weaker models. Strategy-forcing thereby attains the *highest answer coverage of any source*, an effect we ground in the ambiguity decomposition (Krogh & Vedelsby, 1994) (§3.2).
- We convert this decorrelation into ensemble accuracy at single-model cost: compute-matched at K=5, OMMM matches or beats the single-model baselines and Pareto-dominates self-consistency on the hard benchmarks, rivals multi-agent debate and self-correction at 5–10× less compute, and outscores even a five-model cross-model vote (§5).

- We extract a free, calibrated confidence signal from the same decorrelation: where reliable LLM confidence is otherwise hard to elicit (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024), cross-strategy *agreement* is a training-free, verifier-free signal, better calibrated than self-consistency vote frequency, and answering only when the strategies agree raises AIME accuracy on the answered subset from 37% to 100% (selective prediction; §5).

2 Related Work

Sampling-based test-time scaling. CoT (Wei et al., 2022; Kojima et al., 2022) and problem-decomposition prompting (Zhou et al., 2023) elicit multi-step reasoning. Self-Consistency (Wang et al., 2023) samples many such chains and majority-votes, while Best-of-N and tree/graph search (Yao et al., 2023; Besta et al., 2024) expand the space of constructed solutions. They all draw their diversity from a single source, *temperature*: repeated samples of one model under one prompt. Scaling this repeated sampling reliably grows coverage (Brown et al., 2024) and can rival parameter scaling (Snell et al., 2025; Muennighoff et al., 2025), but the payoff from pure resampling is uneven and saturates (Sprague et al., 2024). We diagnose the saturation as *error correlation* and show that the lever is changing the *source* of diversity, whereas increasing the *amount* yields little.

Diversity of prompts and perspectives. DiVeRSe (Li et al., 2023) and diversity-of-thought prompting (Naik et al., 2023) vary prompt templates or exemplars to diversify reasoning, and Self-Contrast (Zhang et al., 2024) solves from several "perspectives" and uses their inconsistencies as a self-revision checklist. These vary the *surface form* of the instruction. OMMM differs on three axes: (i) we force *structurally* different computations (algebraic vs. geometric vs. backward) and show this decorrelates errors far more than surface variation; (ii) we *measure* error correlation and tie it to ensemble theory, where prior work assumes diversity helps; and (iii) we use cross-strategy agreement as a principled *vote* and confidence signal, in place of a heuristic checklist.

Self-correction, debate, and verification. A second family revisits an answer after producing it: self-refinement iterates "critique-then-improve" (Madaan et al., 2023; Shinn et al., 2023), though unaided self-correction can fail to improve or even degrade reasoning (Huang et al., 2023; Kamoi et al., 2024). Multi-agent debate (Du et al., 2024; Liang et al., 2023) runs several model instances that argue toward consensus, but its diversity again derives from temperature-sampled copies of one model. Process and outcome verifiers (Lightman et al., 2023) score solutions to guide selection, and ensembling or fusing *different* models improves over any single one (Jiang et al., 2023; Wang et al., 2024), at the cost of multiple deployments. OMMM needs neither multiple models nor a trained verifier: it obtains decorrelation from a single model's distinct strategies, and its agreement signal is verifier-free. Universal Self-Consistency (Chen et al., 2023) aggregates free-form outputs with the model itself, and we adopt it for open-ended tasks but otherwise vote directly across strategies.

Ensemble diversity, confidence, and calibration. The ambiguity decomposition (Krogh & Vedelsby, 1994) and the bias–variance–covariance view of ensembles (Dietterich, 2000; Kuncheva & Whitaker, 2003) establish that ensemble error falls as members decorrelate, so diversity, alongside individual accuracy, drives the gain, the principle behind deep ensembles (Lakshminarayanan et al., 2017) and random forests (Breiman, 2001). We bring this lens to LLM reasoning: strategy-forcing manufactures decorrelated members from one model, and the correlation hierarchy we measure is that diversity term, made empirical. The same diversity also yields confidence for selective prediction (Geifman & El-Yaniv, 2017), which is otherwise hard to elicit from LLMs: verbalized and answer-probability estimates are frequently miscalibrated (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024), and post-hoc calibration (Guo et al., 2017) needs held-out data. Cross-strategy agreement sidesteps both, providing a training-free confidence signal at no extra cost.

3 One Model, Many Minds

3.1 The Method

Strategy-forcing. For a problem x, OMMM generates K candidate solutions $\{a_1, \ldots, a_K\}$, each produced under a distinct *reasoning strategy* s_k. A strategy is enforced by a system prompt that (i) names the strategy, (ii) states a `MUST` constraint (what the model must do), (iii) states a `MUST NOT`

constraint (what it must avoid), and (iv) gives a one-line sketch. For mathematics we use *algebraic, geometric, backward, inductive, analogical* as the canonical $K{=}5$ set (full prompts in Appendix B). For graduate science (GPQA) we use *first-principles, dimensional analysis, limiting cases, elimination, analogy*, and for code, *top-down, bottom-up, test-first, recursive, iterative*. The strategy set is the only domain-specific component, and the effect does not hinge on its specific members: a disjoint alternative set decorrelates equally well (Appendix G). Crucially, the K generations are mutually independent, so they execute in parallel. OMMM's wall-clock latency is one generation, and its token cost equals K chains, the same budget as self-consistency at K samples.

Aggregation. We extract a final answer from each strategy's output and aggregate by *majority vote* under a domain-aware notion of answer equivalence (numeric tolerance, LaTeX/symbolic equivalence via a math checker, MCQ-letter normalization). For closed-form answers, weighting each answer cluster by its size coincides with majority vote, so the two are identical in our tables. The distinction matters only for open-ended outputs (code, free text), where answers are not directly comparable and we fall back to Universal Self-Consistency (Chen et al., 2023). Ties are broken toward the earliest-appearing cluster.

Agreement as confidence. Define the cross-strategy *agreement* $A(x) = \frac{1}{\binom{K}{2}} \sum_{i<j} \mathbf{1}[a_i \equiv a_j]$, the fraction of strategy pairs that reach the same answer. $A(x){=}1$ means all strategies converge (high confidence), and $A(x)$ near 0 means they disagree (low confidence). Because strategies are decorrelated, their convergence is strong evidence of correctness, giving a training-free, verifier-free confidence estimate at no extra cost (§5).

Compliance. Strategy-forcing only decorrelates if the model actually follows the constraint. We report a *compliance rate* per strategy, the fraction of responses containing the lexical hallmarks of the intended mode (e.g. "equation/solve for" for algebraic, "triangle/coordinate" for geometric), validated against an LLM judge on a subset (Appendix C).

3.2 Why It Works: An Ensemble View

We now make precise why forcing distinct strategies should help, and predict the ordering of error correlation that §5 sets out to test.

Ensemble error and diversity. For an ensemble of K members with individual error rates and pairwise error correlation ρ, the ambiguity decomposition (Krogh & Vedelsby, 1994) states that ensemble error equals average individual error minus a diversity term that increases as ρ decreases. Holding individual accuracy fixed, lowering ρ strictly improves the ensemble: what governs it is not whether the members are individually strong but whether they err *independently*.

Coverage and effective ensemble size. Let p be a member's accuracy. Under independence, the probability that at least one of K members is correct, the *coverage*, is $1-(1-p)^K$. Under positive correlation ρ, the K correlated members behave like $K_{\text{eff}} = K/(1+(K-1)\rho)$ independent ones, so coverage is $\approx 1-(1-p)^{K_{\text{eff}}}$. At $K{=}5$: $\rho{=}0.6$ gives $K_{\text{eff}}{\approx}1.5$ (five samples worth less than two independent draws), whereas $\rho{=}0.2$ gives $K_{\text{eff}}{\approx}2.8$. Reducing ρ is thus worth far more than adding samples at high ρ, which is the quantitative form of self-consistency's saturation.

The correlation hierarchy (hypothesis). We hypothesize an ordering of error correlation by diversity source:

$$\rho(\text{temperature}) \;\approx\; \rho(\text{prompt-variation}) \;>\; \rho(\text{strategy-forcing}) \;>\; \rho(\text{cross-model}).$$

Temperature merely perturbs token selection along the *same* computational path, and prompt variation changes the surface form while the model still applies its default strategy, so both *surface* sources should leave errors strongly correlated. Strategy-forcing changes the *computation itself*, so an algebraic error and a geometric error have disjoint causal pathways and rarely coincide, the same disjointness that decorrelates a cross-model ensemble, which is why strategy-forcing should fall between the surface sources and the cross-model floor. This ordering is our central empirical claim, tested directly in §5 and underpinning the accuracy and confidence results that follow.

Table 1: **Mean pairwise error correlation ρ by diversity source** (lower is better; $K{=}5$, mean over MATH-500, AIME-24, GPQA). **Bold** marks the best *single-model* source, while cross-model is a five-deployment reference floor and is excluded from the single-model comparison. Strategy-forcing decorrelates a single model's errors well below the surface sources, partway to that floor, lifting K_{eff} (Sonnet) from 1.35 to 1.48. The ordering replicates across all four reasoners, and for the reasoning model (Kimi) strategy-forcing (0.44) essentially *matches* the cross-model floor (0.45).

	Mean ρ per reasoner				Sonnet 4.6
Diversity source	Sonnet 4.6	Qwen3-235B	DeepSeek V3.2	Kimi-K2-Thinking	K_{eff}
Temperature	0.69	0.77	0.63	0.57	1.35
Prompt variation	0.71	0.73	0.66	0.61	1.30
Strategy-forcing (ours)	**0.60**	**0.70**	**0.57**	**0.44**	**1.48**
Cross-model	0.43	0.45	0.45	0.45	1.85

4 Experimental Setup

Benchmarks. We evaluate on competition mathematics (**MATH-500**; Hendrycks et al., 2021), olympiad math (**AIME 2024** and **2025**; Mathematical Association of America, 2025), graduate science (**GPQA-Diamond**; Rein et al., 2024), code generation graded by held-out unit tests (**HumanEval**; Chen et al., 2021), and a near-saturated commonsense MCQ control (**ARC-Challenge**; Clark et al., 2018). The core correlation experiment uses MATH-500, AIME, and GPQA, where genuine errors exist to decorrelate.

Models. The correlation experiment fixes the reasoner at Claude Sonnet 4.6 for a clean single-model measurement, and the cross-model condition draws one attempt each from five different deployments (Claude Sonnet 4.6, Qwen3-235B, DeepSeek V3.2, Kimi K2.5, Qwen3-Next-80B). Accuracy experiments span three reasoners (Claude Sonnet 4.6, Qwen3-235B, DeepSeek V3.2), and the model-size ablation adds a Qwen ladder (32B $\rightarrow$ Next-80B $\rightarrow$ 235B). Grading uses symbolic/LaTeX equivalence for math, unit-test execution for code, and exact match for MCQ.

Baselines and protocol. All methods are matched at $K{=}5$ generations per problem unless noted. We compare against Self-Consistency ($K{=}5, 8$), DiVeRSe-style prompt diversity ($K{=}5$), Self-Contrast (Zhang et al., 2024), multi-agent debate (Du et al., 2024), and a five-model cross-model vote. The four diversity conditions in the correlation experiment each produce exactly $K{=}5$ independent attempts: temperature sampling at five temperatures, five prompt rephrasings, five strategies, and five models. Error correlation is the pairwise Phi coefficient (Guttman, 1945) on the binary correct/incorrect outcomes. We report mean off-diagonal ρ, K_{eff}, and Coverage@k, with bootstrap confidence intervals and permutation tests for ρ differences (Appendix D). Full benchmark sizes, decoding settings, compute-matching, extraction, and grading details are in Appendix A.

5 Results

We present the results as a chain from mechanism to payoffs: forcing distinct strategies decorrelates a single model's errors (§5.1), this buys accuracy at parity-or-better with the compute-matched baselines (§5.2), and the cross-strategy agreement it produces is a calibrated confidence signal (§5.3) that turns into a compute saving via agreement-gated cascades (§5.4). Ablations isolating the constraint and the allocation as the drivers are deferred to Appendix F.

5.1 Strategy-Forcing Decorrelates a Single Model's Errors

The correlation hierarchy. Table 1 reports mean pairwise error correlation ρ for the four ways of generating $K{=}5$ attempts on the same problems, with the reasoner fixed at Claude Sonnet 4.6, and Figure 2(a,b) visualizes the same measurement as per-source correlation matrices and their pooled means. The ordering is unambiguous: the two single-model *surface* sources, temperature ($\bar{\rho}{=}0.69$) and prompt rephrasing ($\bar{\rho}{=}0.71$), leave errors strongly correlated (warm matrices), *strategy-forcing* drops correlation substantially to $\bar{\rho}{=}0.60$, and a true five-model ensemble reaches $\bar{\rho}{=}0.43$. Without leaving a single model, strategy-forcing thus lifts the effective ensemble size from temperature's

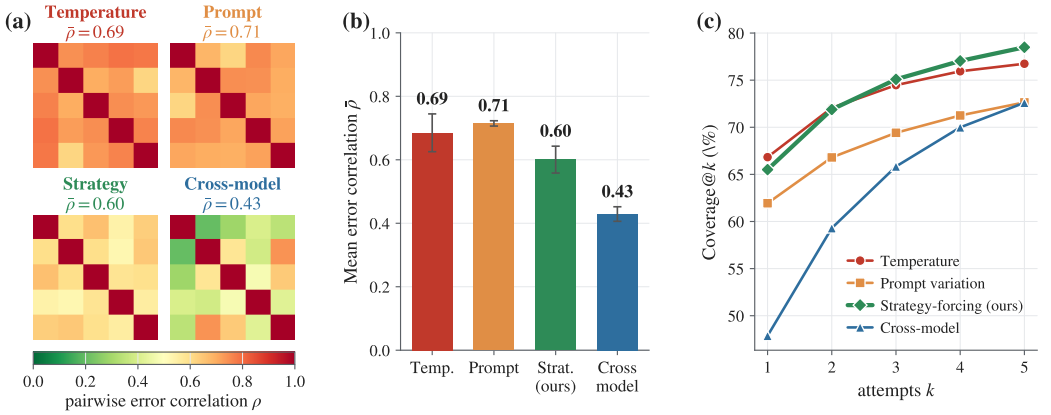

Figure 2: **Strategy-forcing decorrelates a single model's errors. (a)** Pairwise error-correlation matrices (5×5, Claude Sonnet 4.6): temperature and prompt variation stay warm (correlated), strategy-forcing cools toward the cross-model ensemble. **(b)** Mean correlation $\bar{\rho}$ across MATH-500, AIME, GPQA (error bars: standard error over the three benchmarks). **(c)** Coverage@k: strategy-forcing (green) is the *highest of all sources*, since it decorrelates without sacrificing accuracy, whereas the cross-model ensemble does so only by admitting weaker models.

K_{eff}=1.35 to 1.48, a quarter of the way to the cross-model ceiling of 1.85. Temperature and prompt variation are statistically indistinguishable: paraphrasing changes the surface form while leaving the computation intact, so only forcing a new *strategy* changes how the model fails. The ordering holds for all four reasoners (Table 1), and a bootstrap over problems (Appendix D) confirms strategy-forcing decorrelates significantly more than prompt variation and sits significantly above the cross-model floor, with the gap to temperature (consistent in sign everywhere) significant on the large MATH-500 split and for the reasoning model. We lean on the causal analysis below, in place of any single pairwise test, to show the lower correlation is what pays off.

Strategy-forcing has the highest coverage of any source. Decorrelation only helps if it does not come at the cost of individual accuracy. Figure 2(c) plots Coverage@k, the fraction of problems solved by at least one of k attempts, averaged over the three benchmarks. Strategy-forcing is the *best of all four sources*: Coverage@5 reaches 78.5%, above temperature (76.7%), prompt variation (72.6%), and even the cross-model ensemble (72.6%). This is because strategy-forcing keeps high per-attempt accuracy (65.5%, near temperature's 66.8%) *while* decorrelating, whereas the cross-model ensemble's mixed-strength members drag individual accuracy to 47.9%. Its low correlation reflects genuine architectural diversity and is free of weak-model dilution: it persists when the ensemble is restricted to its strongest deployments, where OMMM still leads on coverage at matched size (Appendix H). Strategy-forcing thus attains cross-model-level diversity at single-strong-model quality, the combination an ensemble requires.

The hierarchy is general, and strongest for a reasoning model. The ordering is not specific to Claude Sonnet: Table 1 replicates it on Qwen3-235B and DeepSeek V3.2, and, most tellingly, on Kimi-K2-Thinking, a model that already deliberates internally. On Kimi-K2-Thinking, strategy-forcing reaches $\bar{\rho}$=0.44, essentially *matching* the cross-model floor (0.45): a single reasoning model, forced into different strategies, decorrelates its errors as much as five different models. Externally imposed strategy diversity is therefore not redundant with a model's own long chains of thought, but contributes decorrelation that internal deliberation does not.

Decorrelation *causes* the gain, and strategies cover each other's failures. Two further analyses confirm that the decorrelation is what *drives* the gain and does not merely co-occur with it (Appendices J, K). The first establishes causality: across all 48 (model, benchmark, source) cells, the K_{eff} derived from ρ predicts measured Coverage@K at Pearson r=0.99 (mean abs. error 4.5 pts), and the ensemble gain (coverage minus individual accuracy) falls with $\bar{\rho}$ at a significant slope ($r=-0.53$, $p=10^{-4}$), so lower correlation *produces* larger gains, as the theory dictates. The second establishes

Table 2: **Accuracy (%) at matched compute** (Claude Sonnet 4.6; $K=5$ generations). **Bold** marks the best compute-matched value (top block), and OMMM leads on AIME-2025 (70.0 vs. 60.0) while remaining within noise of the best elsewhere. The higher-compute references (†) and the five-deployment cross-model vote (‡) are shown for context and are excluded from the compute-matched comparison. For Sonnet's closed-form answers the agreement-weighted vote coincides with plain majority vote, and the two diverge only on open-ended ties (Qwen3-235B and DeepSeek V3.2, Appendix E).

Method	MATH	AIME-24	AIME-25	GPQA	HumanEval	ARC
Self-Consistency (5)	95.6	**86.7**	60.0	34.5	98.2	97.3
DiVeRSe-prompts (5)	**96.8**	**86.7**	66.7	**36.0**	97.6	97.8
Self-Contrast	96.2	**86.7**	63.3	30.3	97.6	97.8
OMMM (ours)	95.8	**86.7**	**70.0**	35.9	98.2	97.5
Cross-model vote (5)‡	**96.8**	66.7	60.0	31.3	**98.8**	**98.0**
Higher-compute references (†, 5–10× the tokens)						
Self-Consistency (8)†	96.0	86.7	60.0	36.0	99.4	97.5
Self-Correction†	96.4	86.7	70.0	34.0	98.8	97.5
Self-Correction-multi†	97.0	93.3	70.0	33.0	98.8	98.1
Multi-Agent Debate†	98.2	93.3	83.3	36.0	99.4	97.8

complementarity, showing the strategies are far from noisy copies: on Sonnet math, the chance a strategy is right given another failed is 29–52%, so on problems the best single strategy misses another rescues 59% (32% on Qwen3-235B), and the weakest strategy individually (inductive) is among the best *rescuers*, so low individual accuracy does not imply redundancy.

5.2 Decorrelation Buys Ensemble Accuracy at Matched Compute

Accuracy against eight baselines. Table 2 compares OMMM against eight baselines, all compute-matched at $K=5$ generations except where a baseline is intrinsically more expensive (debate, self-correction). Against the compute-matched single-model baselines (self-consistency, prompt-diversity, Self-Contrast), OMMM is at parity or better on most cells, with its clearest gains on the hard benchmarks (e.g. Qwen3-235B AIME-2024 36.7 vs. self-consistency 33.3). Debate and multi-round self-correction reach higher raw accuracy on the easier benchmarks but spend 5–10× the tokens (Appendix A), and on AIME their edge over OMMM is within noise (Sonnet AIME-2024 debate 93.3 vs. OMMM 86.7, paired bootstrap $p=0.43$). The five-deployment cross-model vote wins only on the near-saturated MATH, HumanEval, and ARC, while on every hard benchmark single-model OMMM *outscores* it (AIME-2024 86.7 vs. 66.7, GPQA 35.9 vs. 31.3).

Coverage headroom and the aggregation bottleneck. A striking pattern in Table 2 is that OMMM's *coverage*@5 far exceeds its majority-vote accuracy: on AIME-2024 (Sonnet) coverage is 96.7% while majority vote returns 86.7%, and on GPQA, 47.0% vs. 35.9%. The decorrelated strategies solve far more problems than the vote captures, so the bottleneck lies in aggregation, and the confidence signal of §5.3 exploits exactly this gap.

Pareto efficiency. Figure 3(b) sweeps the generation budget K for OMMM (more strategies) and self-consistency (more temperature samples) at matched tokens on AIME-2024. OMMM dominates at every budget, from 33.3 vs. 30.0 at $K=2$ to 40.0 vs. 33.3 at $K=10$: the same token budget buys more accuracy spent on diverse strategies than on repeated samples. On the saturated MATH-500 the two curves coincide, where there is no headroom (Figure 6).

5.3 Cross-Strategy Agreement Is a Calibrated Confidence Signal

Agreement among decorrelated strategies signals correctness. Because the strategies are decorrelated, their *agreement* is strong evidence of correctness in a way that agreement among correlated samples is not. Figure 3(a) and Table 3 report selective prediction: abstaining on problems where strategies disagree sharply raises accuracy on the rest. On AIME-2024 (Qwen3-235B) accuracy climbs from 37% over all problems to 100% on the 27% of problems with high agreement ($A \geq 0.6$), and remains 100% on the 13% where all five strategies converge ($A = 1$). On GPQA, it rises from

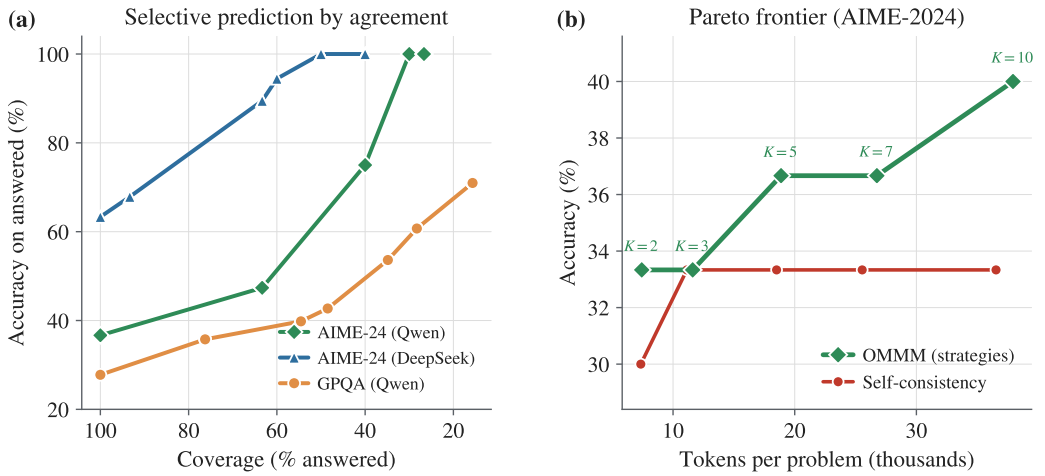

Figure 3: **Two payoffs of decorrelation.** **(a)** Selective prediction: abstaining when strategies disagree (moving right-to-left) sharply raises accuracy on the answered problems, on AIME-2024 from 37% to 100%, with dotted lines showing answer-all accuracy. **(b)** Pareto frontier (AIME-2024): spending a token budget on diverse strategies (OMMM, green) dominates spending it on repeated temperature samples (self-consistency, red) at every budget.

Table 3: **Selective prediction via cross-strategy agreement.** Accuracy on answered problems rises sharply as we restrict to higher agreement, with coverage (fraction answered) in parentheses. Agreement turns OMMM into a reliable selective predictor at no extra cost. The high-agreement AIME slices are small ($n \leq 8$), so the 100% figures carry wide intervals (Wilson 95% CI $[68, 100]$ at $A \geq 0.6$); the result is the consistent direction across all nine cells.

Model	Bench	Answer all	$A \geq 0.6$	$A = 1.0$
Qwen3-235B	AIME-24	37	100 (27%)	100 (13%)
	GPQA	28	61 (28%)	71 (16%)
	MATH	94	98 (91%)	98 (82%)
Sonnet 4.6	AIME-24	87	100 (60%)	100 (47%)
	GPQA	36	53 (55%)	63 (37%)
	MATH	96	98 (97%)	98 (92%)
DeepSeek V3.2	AIME-24	63	100 (50%)	100 (40%)
	GPQA	34	56 (37%)	73 (18%)
	MATH	96	97 (94%)	98 (86%)

28% to 71% on the 16% of unanimous problems. This yields a free, training-free, verifier-free triage rule: answer when the strategies concur, and abstain or escalate when they do not.

Better calibrated than vote frequency where it matters. Agreement is also better *calibrated* than the standard self-consistency confidence (winning-answer vote frequency). On the hard benchmark GPQA, where a confidence signal is actually needed, agreement has markedly lower Expected Calibration Error (ECE, the gap between stated confidence and empirical accuracy) for every model (e.g. Sonnet 0.21 vs. 0.36; DeepSeek 0.13 vs. 0.26), while on near-saturated benchmarks both signals are trivially well-calibrated. Its advantage is thus concentrated exactly where calibration has value (full ECE table and reliability curves in Appendix E).

5.4 Agreement-Gated Cascades Convert Calibration into Compute Savings

The calibration result has a direct system-level use. If agreement reliably flags the problems a single model gets wrong, it can *route* compute: answer cheaply with OMMM when the strategies agree,

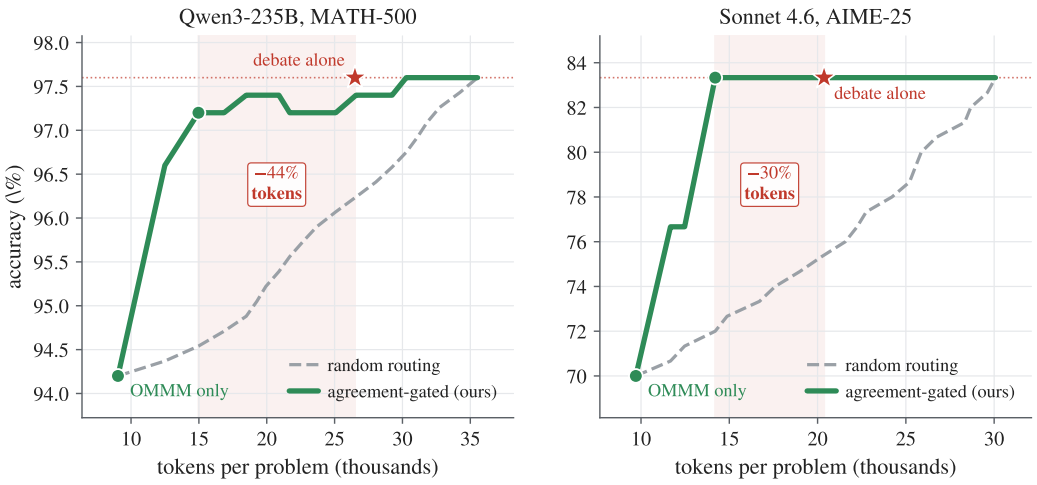

Figure 4: **Agreement-gated cascades convert the confidence signal into compute savings.** Cost–accuracy frontier (green) as the agreement threshold sweeps the escalation fraction: the left endpoint is plain OMMM (escalate nothing), and escalating the lowest-agreement problems first moves rightward toward debate-level accuracy (red star, debate on every problem). The frontier lies well above the random-routing baseline (grey dashed) at every budget.

and escalate only the low-agreement problems to an expensive method, here multi-agent debate. We evaluate this offline on the same problems: below an agreement threshold a problem takes debate's answer at debate's cost, otherwise OMMM's, and sweeping the threshold traces the cost–accuracy frontier of Figure 4.

Routing by agreement reaches debate-level accuracy while paying for debate on only a fraction of problems. On Qwen3-235B / MATH-500, escalating the lowest-agreement 10% matches debate's accuracy (97.2 vs. 97.6) at 56% of debate's token cost, and on Sonnet / AIME-2025, escalating 15% reaches debate's full 83.3% at 70% of its cost. The saving derives from *which* problems are escalated, beyond the mere fact of escalation: at the same fraction, random routing reaches only 94.5% and 72.0%. The effect tracks compliance, and where agreement is poorly calibrated the cascade saves little (Appendix E). OMMM is not more accurate than debate. Its agreement signal simply lets the cheap method stand in wherever the two would agree anyway.

6 Conclusion

Test-time scaling has leaned on *sampling more*, but samples from one model at one prompt fail the same way: their errors are correlated, and that correlation, more than the sample count, is what caps accuracy. We show that the diversity that breaks it is *structural*. Forcing a single model through different reasoning strategies decorrelates its errors substantially more than temperature or prompt variation, partway toward the floor set by a true multi-model ensemble. Compute-matched, OMMM matches or beats the baselines and rivals multi-agent debate and self-correction at a fraction of their compute, and its cross-strategy agreement is a free, calibrated confidence signal. The practical lesson is that to make test-time compute pay, one should ask the model not the same question more times, but to reason *differently*.

Limitations. OMMM needs a hand-designed strategy set per domain, though the choice is forgiving (a disjoint set decorrelates equally well, Appendix G), and its benefit is bounded by compliance: a model that reverts to its default reasoning recovers temperature sampling's high correlation, so OMMM works best on instruction-following frontier models (Appendix C). Where all strategies share one necessary insight, forcing different paths cannot help. Automatic strategy discovery and problem-adaptive routing are natural next steps.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, and interpreting results. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, or proposing or refining hypotheses. Assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability and for identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, AI-suggested interpretations were checked against outputs, and AI-identified references were verified against their source records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

OMMM is implemented entirely through prompting, with no fine-tuning or training. The strategy prompts (Appendix B), aggregation and agreement rules (§3), and all benchmarks (MATH-500, AIME 2024/2025, GPQA, HumanEval, ARC-Challenge) with their splits and problem counts are documented in §4. Grading uses symbolic equivalence for math, unit-test execution for code, and exact match for MCQ. The four diversity conditions, the Phi-coefficient correlation analysis, and the bootstrap/permutation procedures are specified in §4 and Appendix D. Decoding settings, the temperature sweep, compute-matching, and answer extraction are detailed in Appendix A. We release all strategy configurations and prompts. All models are accessed through a hosted inference API, so outputs may shift across provider-side updates. However, the correlation hierarchy and accuracy results replicate on *open-weight* reasoners (Qwen3-235B, Qwen3-Next-80B, Qwen3-32B, DeepSeek V3.2), which can be served locally to reproduce the protocol independently of any provider.

REFERENCES

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

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. In *arXiv preprint arXiv:1803.05457*, 2018.

Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15. Springer, 2000.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Louis Guttman. A basis for analyzing test-retest reliability. *Psychometrika*, 10(4):255–282, 1945.

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 *Advances in Neural Information Processing Systems*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mathematical Association of America. American invitational mathematics examination (AIME) 2024–2025, 2025.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. Diversity of thought improves reasoning abilities of LLMs. *arXiv preprint arXiv:2310.07088*, 2023.

George Pólya. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press, 1945.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-Proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

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

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

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

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

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations (ICLR)*, 2024.

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*, 2023.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A Experimental Setup Details

Benchmarks and sizes. We use the standard test splits: MATH-500 (500 competition problems), AIME 2024 and AIME 2025 (30 problems each), GPQA-Diamond (198 graduate-science questions), HumanEval (164 programming problems), and ARC-Challenge (1,172 commonsense MCQ). GPQA is evaluated in its open-ended (boxed-answer) form instead of as multiple choice, so a strategy must *produce* the answer instead of selecting it. The core correlation experiment uses MATH-500, AIME-2024, and GPQA, the benchmarks with enough genuine errors to make correlation meaningful (ARC and GSM8K-style tasks are near-saturated, leaving little error mass to decorrelate).

Models and decoding. The reasoners are Claude Sonnet 4.6, Qwen3-235B-A22B, and DeepSeek V3.2, and the model-size ladder adds Qwen3-32B and Qwen3-Next-80B-A3B. The cross-model condition and cross-model-vote baseline draw one attempt each from five separate deployments: Claude Sonnet 4.6, Qwen3-235B, DeepSeek V3.2, Kimi K2.5, and Qwen3-Next-80B. Unless noted we decode at temperature 0.7 with a 4096-token cap, while the temperature-diversity condition instead sweeps five temperatures in $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ (the hosted models cap temperature at 1.0). All models are served through a single hosted inference API behind one client interface. Each generation in a $K{=}5$ method is an independent request, so the K attempts are dispatched concurrently and OMMM's wall-clock latency is one generation.

Compute matching. The four diversity conditions and the primary baselines all spend exactly $K{=}5$ generations per problem, so comparisons reflect how a fixed budget is *allocated*. Three baselines are intrinsically more expensive and are marked † in the accuracy tables: Self-Consistency-8 (8 samples), multi-round Self-Correction ($1{+}3$ sequential generations), and multi-agent debate (2 agents $\times 3$ rounds $+$ judge). Empirically these spend 5–$10\times$ the tokens of OMMM-5 on the hard benchmarks. The cross-model vote (‡) spends 5 generations but across five *separate* model deployments.

Prompt-variation and rephrasings. For the prompt-variation condition we generate five meaning-preserving rephrasings of each problem once (using the same reasoner at temperature 0.7), cache them, and then solve each rephrasing under the neutral CoT prompt. The rephrasing instruction explicitly forbids changing the quantities or the answer.

Answer extraction and grading. Each solution is required to end with a `FINAL ANSWER:` line. We extract the answer by regex (boxed expression, final-answer line, or trailing number/ letter as appropriate) and fall back to a cheap LLM extraction only when the regex fails. Grading is by symbolic/LaTeX equivalence for math (fractions, radicals, π, etc.), exact integer match for AIME, sandboxed unit-test execution for HumanEval, and letter/string match for MCQ. The *same* equivalence relation is used for cross-strategy agreement, so two answers that grade as equal also count as agreeing.

Aggregation and agreement. Majority vote clusters answers under the equivalence above and returns a representative of the largest cluster (ties broken toward the earliest cluster). Agreement $A(x)$ is the fraction of the $\binom{K}{2}$ strategy pairs whose answers fall in the same cluster, and the agreement-weighted variant weights clusters by size. For HumanEval, code is graded by executing it against held-out unit tests, which yields only a pass/fail outcome without a distribution over distinct answers. Cross-strategy agreement is informative precisely because wrong strategies scatter across *different* wrong answers (so convergence is evidence of correctness). With only pass/fail, "agreement" collapses into the fraction of strategies that pass, which is simply accuracy and carries no independent confidence information. We therefore omit HumanEval from the calibration analysis.

B Strategy Prompts

Each strategy is enforced by a system prompt naming the strategy, a `MUST` constraint, a `MUST NOT` constraint, and a one-line sketch. We reproduce the mathematics set verbatim, and the science and code sets follow the same template.

Algebraic

```
You solve problems using ONLY algebraic and symbolic manipulation.
- You MUST set up equations, inequalities, or symbolic expressions
before computing.
- You MUST manipulate symbols step-by-step to reach the answer.
- You MUST NOT use geometric intuition, diagrams, or spatial
reasoning.
- You MUST NOT try examples or look for patterns empirically.
If the problem seems geometric, translate it into coordinates and
algebra first.
```

Geometric

```
You solve problems using geometric and spatial reasoning.
- You MUST reason about shapes, distances, areas, transformations, or
coordinate geometry.
- You MUST reference geometric objects (points, lines, regions,
graphs).
- You MUST NOT lead with pure algebraic manipulation.
If the problem seems purely algebraic, find the geometric
interpretation first.
```

Backward / goal-directed

```
You solve problems by working BACKWARD from the answer.
- Your FIRST sentence must describe what form the answer takes.
- You MUST start from the goal and work backward toward the givens.
- You MUST NOT solve forward from the problem statement.
At each step, ask ``what would need to be true for this to hold?''
```

Inductive / empirical

```
You solve problems by trying specific cases and finding patterns.
- You MUST try at least 3 specific small cases before generalizing.
- You MUST explicitly state the pattern and verify it on 1 additional
case.
- You MUST NOT jump to a formula without empirical evidence first.
```

Analogical

```
You solve problems by mapping to a simpler or well-known problem.
- You MUST explicitly name a simpler/known problem this is analogous
to.
- You MUST solve the analogue first and state the mapping (what
corresponds to what).
- You MUST NOT solve the problem directly without identifying an
analogue.
```

The science set (GPQA) is {first-principles, dimensional analysis, limiting cases, elimination, analogy}, and the code set (HumanEval) is {top-down, bottom-up, test-first, recursive, iterative}. For $K > 5$ scaling we extend the math set with {estimation, case-analysis, symmetry, constraint-satisfaction, metacognitive}.

C Strategy Compliance

Strategy-forcing only decorrelates if the model follows the constraint. We measure a keyword-based compliance rate, the fraction of responses containing the lexical hallmarks of the intended mode, and validate it against an LLM judge on a random subset of 40 responses per strategy (the judge sees the strategy's `MUST` requirement and answers YES/NO). Table 4 reports the keyword rate, and the LLM judge agrees with it on 80–98% of cases for the frontier reasoners. Compliance is near-ceiling for Claude Sonnet 4.6 (99.8% mean) and high for Qwen3-235B (97.4%), confirming that the

measured decorrelation reflects genuinely different computations and is not produced by the model ignoring the prompt. The smaller Qwen3-32B is markedly less compliant (89.3% mean, and only 59.4% on the demanding *backward* strategy), a direct mechanism for the weaker OMMM benefit on small models (Table 8): a constraint the model cannot follow yields no new computation, hence no decorrelation.

Table 4: **Keyword-based strategy-compliance rate** (%), math benchmarks. Validated against an LLM judge (per-strategy judge agreement 80–98% on frontier models). Compliance falls on the smallest model, explaining its smaller ensemble benefit.

Strategy	Qwen3-235B	Sonnet 4.6	Qwen3-32B
algebraic	100.0	99.6	99.8
analogical	100.0	100.0	99.8
backward	87.4	100.0	59.4
geometric	99.4	99.8	96.2
inductive	100.0	99.8	91.2
Mean	97.4	99.8	89.3

D Statistical Significance

Each comparison uses an appropriate test. For accuracy differences between two methods on the same problems we use a paired bootstrap (10,000 resamples) for a 95% CI on the difference and a two-sided p-value, plus the exact McNemar test on per-problem agreement. For differences in mean pairwise error correlation ρ between two diversity conditions we resample problems (10,000 bootstrap replicates), recompute each condition's mean off-diagonal ρ, and report the difference, its 95% CI, and a two-sided bootstrap p-value. Calibration differences (ECE) use a paired bootstrap on the ECE difference.

Significance of the correlation hierarchy. Table 5 reports, per reasoner, the bootstrap 95% CI on each source's mean ρ (pooled over MATH-500, AIME-2024, GPQA) and the signed difference of strategy-forcing against each other source. The qualitative ordering $\rho(\text{temp}) \approx \rho(\text{prompt}) > \rho(\text{strategy}) > \rho(\text{cross-model})$ holds in the point estimates for *all four* reasoners. Statistically, the two robust comparisons are (i) strategy-forcing versus the *cross-model floor* (strategy is significantly higher for the three non-reasoning models and statistically indistinguishable for Kimi-Thinking, p=0.84, the "matches the floor" result), and (ii) strategy-forcing versus *prompt* variation (significant for Sonnet, DeepSeek, and Kimi). The strategy-versus-*temperature* gap is consistent in sign everywhere but, pooled, reaches significance only for the reasoning model. Per benchmark it is significant on the large MATH-500 split (Qwen3-235B p=0.003, Kimi $p<0.001$) and underpowered on AIME (n=30). The mechanism does not rest on any single pairwise test: the causal regression of Appendix J, in which ensemble gain falls with $\bar{\rho}$ across all 48 cells ($p=10^{-4}$) and K_{eff} predicts coverage at r=0.99, is the well-powered statement that lower correlation drives the gain.

E Additional Results

Full accuracy table (all three reasoners). Table 6 reports the accuracy comparison across all three reasoners (Claude Sonnet 4.6, Qwen3-235B, DeepSeek V3.2). The pattern in the body holds broadly on the strong instruction-follower (Claude Sonnet 4.6), where OMMM matches or exceeds the compute-matched single-model baselines on most cells. On the two weaker instruction-followers it is more mixed: on the small AIME sets, majority-vote OMMM trails self-consistency or Self-Contrast for Qwen3-235B (AIME-2024 36.7 vs. Self-Contrast 46.7) and DeepSeek (AIME-2024 63.3 vs. self-consistency 70.0). We attribute this to two compounding factors, neither of which indicates a failure of the mechanism. First, forcing a contrived strategy can lower an individual attempt's accuracy when the model follows the constraint only partially, exactly the compliance dependence of Appendix C, and the reason the clean wins concentrate on the high-compliance model. Second, and more importantly, these are precisely the cells where OMMM's *coverage* most exceeds its vote accuracy (DeepSeek AIME-2024: 83.3% coverage vs. 63.3% vote; Qwen GPQA: 41.4% vs.

Table 5: **Bootstrap significance of the correlation hierarchy** (pooled over MATH-500, AIME-2024, and GPQA, with 10,000 resamples). For each reasoner we report the mean ρ and its 95% CI per diversity source, and the signed difference of strategy-forcing minus that source with a two-sided bootstrap p (bold $p<0.05$). Negative Δ means strategy-forcing decorrelates *more* than the source. The cross-model comparison is the key one: strategy-forcing sits significantly above the cross-model floor for the three non-reasoning models, and *matches* it for the reasoning model (Kimi, $p=0.84$).

		Mean ρ (95% CI)			Strategy − source
Reasoner	Source	$\bar{\rho}$	CI low	CI high	Δ (p)
Sonnet 4.6	Temperature	0.69	0.60	0.76	−0.08 (0.15)
	Prompt variation	0.71	0.65	0.78	−0.11 (**0.03**)
	Strategy (ours)	0.60	0.52	0.68	—
	Cross-model	0.43	0.36	0.49	+0.17 (**0.002**)
Qwen3-235B	Temperature	0.77	0.72	0.83	−0.07 (0.09)
	Prompt variation	0.73	0.65	0.79	−0.03 (0.61)
	Strategy (ours)	0.70	0.64	0.76	—
	Cross-model	0.45	0.38	0.51	+0.26 (< **0.001**)
DeepSeek V3.2	Temperature	0.63	0.56	0.70	−0.07 (0.19)
	Prompt variation	0.66	0.59	0.72	−0.09 (0.05)
	Strategy (ours)	0.57	0.49	0.63	—
	Cross-model	0.45	0.38	0.51	+0.12 (**0.02**)
Kimi-Thinking	Temperature	0.57	0.50	0.65	−0.14 (**0.009**)
	Prompt variation	0.61	0.54	0.68	−0.17 (< **0.001**)
	Strategy (ours)	0.44	0.36	0.50	—
	Cross-model	0.45	0.38	0.51	−0.01 (0.84)

27.8%): the decorrelated strategies do solve these problems, but plain majority vote, which weights every strategy equally, discards the solutions. This is an aggregation limitation, and the diversity itself is intact. We probed whether a label-free aggregator could close it (Appendix I): a reliability-weighted vote, with per-strategy weights estimated by leave-one-out over problems, helps on some cells (e.g. DeepSeek AIME-2025 +10 points, HumanEval +2) but is neutral-to-harmful on others, for a mean change of only +0.2 points. Equal-weight majority is a strong baseline, and reliably converting coverage into accuracy appears to require a verifier or held-out labels, which we leave to future work. Finally, the AIME deltas are computed on only 30 problems (a 6.7-point gap is two problems) and are not statistically significant under the paired bootstrap of Appendix D.

Expected Calibration Error. Table 7 reports the full ECE comparison behind the calibration claim of §5.3: cross-strategy agreement versus the self-consistency winning-answer vote frequency, across all models and benchmarks. Agreement is better calibrated on every hard-benchmark cell (the GPQA and AIME columns), while on the near-saturated benchmarks both signals are trivially well-calibrated.

Reliability curves. Figure 5 shows the reliability curves behind the Expected Calibration Error of Table 7. The shape of the gap is informative: on the hard benchmarks self-consistency vote frequency sits *below* the diagonal: it is systematically *over*-confident, assigning high vote shares to answers that are often wrong, because correlated samples pile onto the same plausible-but-incorrect answer and manufacture a large majority. Cross-strategy agreement tracks the diagonal far more closely: when decorrelated strategies nonetheless converge, that convergence is rarely spurious, so a given agreement level corresponds to roughly the empirical accuracy it should. The two signals are comparably, and trivially, calibrated on the near-saturated benchmarks, where almost every answer is correct and any confidence signal appears well-calibrated. Agreement's advantage is concentrated exactly where a confidence estimate is actually consulted, namely the hard problems on which the model is frequently wrong.

Scaling on both benchmarks. Figure 6 plots accuracy against the generation budget K for OMMM and self-consistency, and the two benchmarks exhibit complementary regimes. On the saturated MATH-500 the curves are nearly coincident: when individual accuracy already exceeds 90% there is little error mass to decorrelate, the ambiguity term is small, and how the budget is

Table 6: **Full accuracy (%) at matched compute** across three reasoners. † marks methods that spend more than $K=5$ generations.

Method	MATH	AIME-24	AIME-25	GPQA	HEval	ARC
Qwen3-235B						
Self-Consistency (5)	91.0	33.3	33.3	31.3	93.3	97.4
Self-Consistency (8)†	91.6	30.0	33.3	30.3	92.7	97.6
DiVeRSe-prompts (5)	90.8	30.0	30.0	27.0	95.7	97.7
Self-Contrast	93.6	46.7	40.0	28.3	94.5	97.4
Self-Correction†	95.2	46.7	40.0	31.3	96.3	96.9
Self-Correction-multi†	97.0	73.3	60.0	33.3	96.3	96.7
Multi-Agent Debate†	97.6	83.3	63.3	33.3	95.7	97.0
OMMM (ours)	94.2	36.7	33.3	27.8	94.5	97.7
OMMM-weighted	92.6	36.7	33.3	28.8	94.5	97.7
Sonnet 4.6						
Self-Consistency (5)	95.6	86.7	60.0	34.5	98.2	97.3
Self-Consistency (8)†	96.0	86.7	60.0	36.0	99.4	97.5
DiVeRSe-prompts (5)	96.8	86.7	66.7	36.0	97.6	97.8
Self-Contrast	96.2	86.7	63.3	30.3	97.6	97.8
Self-Correction†	96.4	86.7	70.0	34.0	98.8	97.5
Self-Correction-multi†	97.0	93.3	70.0	33.0	98.8	98.1
Multi-Agent Debate†	98.2	93.3	83.3	36.0	99.4	97.8
OMMM (ours)	95.8	86.7	70.0	35.9	98.2	97.5
OMMM-weighted	95.8	86.7	70.0	35.9	98.2	97.5
DeepSeek V3.2						
Self-Consistency (5)	95.0	70.0	60.0	30.4	94.5	97.1
Self-Consistency (8)†	96.0	70.0	56.7	26.8	96.3	97.4
DiVeRSe-prompts (5)	95.8	66.7	56.7	31.8	95.7	97.3
Self-Contrast	93.2	76.7	63.3	26.0	95.1	97.4
Self-Correction†	94.2	73.3	70.0	30.8	94.5	97.4
Self-Correction-multi†	95.2	80.0	80.0	31.3	93.9	97.3
Multi-Agent Debate†	95.0	80.0	73.3	32.3	97.6	97.0
OMMM (ours)	95.6	63.3	53.3	34.2	95.1	97.4
OMMM-weighted	95.2	63.3	53.3	34.3	95.1	97.4
Multi-model (5 separate deployments)						
Cross-model vote (5)‡	96.8	66.7	60.0	31.3	98.8	98.0

Table 7: **Expected Calibration Error** (lower is better): cross-strategy agreement vs. self-consistency vote frequency. Best per row in **bold**. Agreement wins decisively on the hard GPQA benchmark.

Model	Bench	Agreement (ours)	Vote-freq.
Qwen3-235B	MATH	0.082	**0.046**
	AIME-24	**0.147**	0.180
	GPQA	**0.133**	0.260
Sonnet 4.6	MATH	0.028	**0.023**
	AIME-24	0.230	**0.113**
	GPQA	**0.212**	0.359
DeepSeek V3.2	MATH	0.074	**0.046**
	AIME-24	0.107	0.107
	GPQA	**0.132**	0.255

allocated barely matters, exactly what the coverage-under-correlation model predicts when p is near 1. On AIME-2024, where error mass is abundant, the allocation dominates: OMMM leads self-consistency at every budget and the gap *widens* with K ($+3$ points at $K=2$, $+7$ at $K=10$), because each additional *strategy* contributes a more nearly independent attempt whereas each additional *sample* is largely redundant with the ones already drawn. The practical reading is that strategy di-

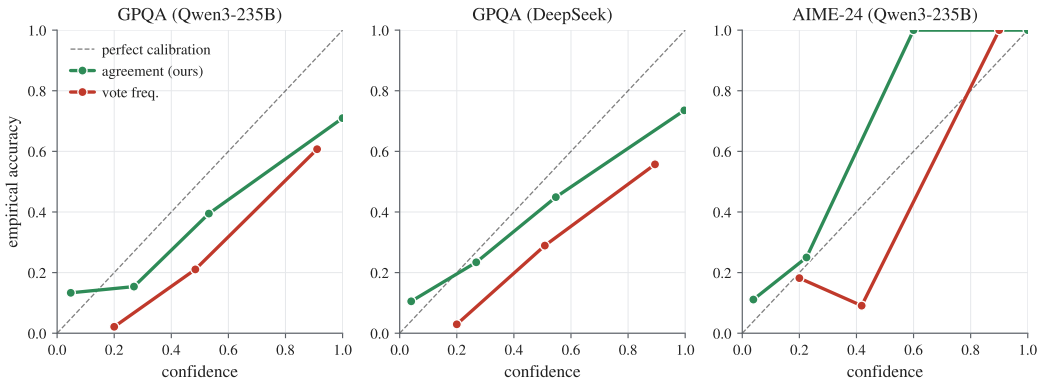

Figure 5: **Reliability curves** for cross-strategy agreement (green) vs. self-consistency vote frequency (red). Agreement tracks the diagonal more closely on the hard benchmarks.

versification helps precisely where test-time compute is otherwise wasted, namely on hard problems with headroom, and neither helps nor hurts where the benchmark is already solved.

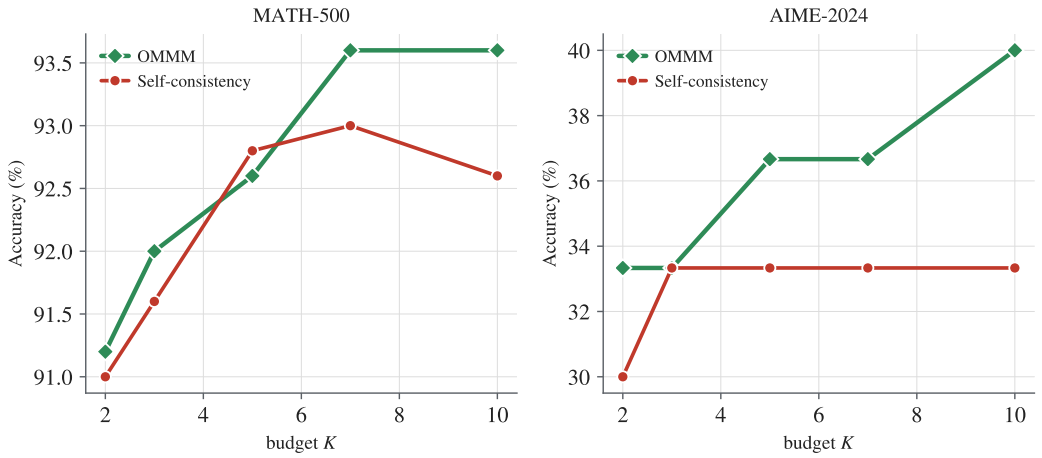

Figure 6: **Accuracy vs. budget** K, OMMM vs. self-consistency.

Does OMMM work for smaller models? Table 8 runs OMMM and self-consistency across a Qwen size ladder. The pattern is governed by two competing forces. On MATH, where all three models are strong ($\geq$ 89%), there is little error mass to decorrelate and the two methods are essentially tied across the ladder. OMMM's small edge appears only at the top (235B: 94.2 vs. 91.0), where high compliance (Appendix C) makes the forced computations genuinely distinct. On GPQA, which has ample error mass at every scale, OMMM helps the *smaller* models most (32B: 26.3 vs. 20.7; Next-80B: 29.3 vs. 26.8) and slightly trails self-consistency only on the 235B cell (27.8 vs. 31.3). This lone dip is *not* a failure of the mechanism: 235B is the most compliant model (Table 4), so the strategies do diverge, and OMMM's GPQA coverage reaches 41.4% while its vote returns only 27.8% (Appendix E). The decorrelated strategies solve the problems, but the equal-weight vote, competing here against self-consistency's stronger default-mode samples, simply fails to extract them. The bottleneck is again aggregation, with the diversity intact. More broadly, the ensemble benefit is set by the interaction of two quantities beyond model size alone: there must be (i) enough error mass for decorrelation to matter and (ii) enough instruction-following for the strategies to actually diverge. Small models can satisfy (i) but partially fail (ii), since the 89.3% compliance and 59.4% *backward*-strategy rate of Qwen3-32B (Table 4) mean some of its "strategies" collapse

back to the default computation, capping the achievable decorrelation. We therefore frame OMMM as a method whose ceiling *rises* with instruction-following, and whose floor is self-consistency: it does not underperform plain sampling by a meaningful margin even where compliance is weak.

Table 8: **Model-size ladder** (OMMM vs. self-consistency at $K=5$). The ensemble benefit is larger for more capable models, which follow the strategy constraints more reliably.

Model	OMMM	Self-Cons. (5)
MATH		
Qwen3-32B	89.6	90.4
Qwen3-Next-80B	94.1	94.2
Qwen3-235B	94.2	91.0
GPQA		
Qwen3-32B	26.3	20.7
Qwen3-Next-80B	29.3	26.8
Qwen3-235B	27.8	31.3

When does OMMM not help? Collecting the regularities above, the gain is governed by three conditions, all of which must hold. **(i) Error mass.** The benchmark must be far enough from saturation that errors remain to decorrelate. On MATH-500, where per-attempt accuracy already exceeds 90%, the ambiguity term is small and OMMM and self-consistency essentially coincide (Table 8, Figure 6). **(ii) Compliance.** The model must follow the strategy constraints. Below roughly 90% compliance some strategies revert to the default computation and recover the high correlation of temperature sampling (Appendix C), which is why the clean wins concentrate on the strong instruction-followers. **(iii) A decomposable problem.** Where a single necessary insight gates the solution, forcing alternative paths cannot manufacture an independent route to it, and decorrelation has nothing to act on. A further caveat is *aggregation*: even when all three hold, plain majority vote can leave coverage on the table (Appendices E, I), so the realized accuracy is a lower bound on what the decorrelated attempts contain. In the regime where the three conditions are met, hard but decomposable benchmarks on instruction-following models, OMMM delivers its largest gains. Where they fail, its floor is self-consistency, which it does not underperform by a meaningful margin.

F Ablations

Two ablations isolate the source of the diversity effect: the explicit constraint (constraint strength) and its allocation across distinct strategies (allocation).

Constraint strength. Table 9 varies the constraint strength: full `MUST`/`MUST NOT` prompts vs. soft ("try to use…") vs. none ("think differently"). Stronger forcing consistently *decorrelates* the strategies further (lower cross-strategy agreement: GPQA 21.5 full vs. 25.0 none) and holds or raises coverage, confirming that the explicit constraint, beyond mere instruction variety, is what diversifies the computation. The downstream accuracy effect is within noise at these sample sizes, so we frame forcing as a *diversity* lever whose accuracy payoff is realized through aggregation and the confidence signal, and not as a direct accuracy boost.

Table 9: **Constraint-strength ablation** (Qwen3-235B). Stronger forcing increases diversity (lower cross-strategy agreement) and maintains coverage.

Bench	Constraint	Acc.	Agreement	Coverage@5
AIME-24	Full (MUST/MUST NOT)	36.7	26.7	40.0
	Soft ("try to")	33.3	31.0	36.7
	None ("think differently")	40.0	37.3	43.3
GPQA	Full (MUST/MUST NOT)	28.8	21.5	41.4
	Soft ("try to")	28.3	25.1	42.4
	None ("think differently")	27.8	24.9	44.9

Allocation. Table 10 compares allocating a fixed budget as five distinct strategies vs. five samples of one strategy vs. a mix. Five distinct strategies yield higher coverage than five samples of a single strategy at the same cost, the allocation-level statement of the correlation result.

Table 10: **Allocation ablation** (Qwen3-235B, AIME-24). Distinct strategies cover more problems than repeated samples of one strategy at matched budget.

Allocation	Acc.	Coverage@5
5 strategies $\times 1$	36.7	40.0
1 strategy $\times 5$	36.7	36.7
3 strategies $\times 2$	36.7	43.3

G Does the Specific Strategy Set Matter?

The canonical strategies (algebraic, geometric, backward, inductive, analogical for math) are one reasonable choice among many. If decorrelation came from these particular five, and not from structural diversity *per se*, the method would be a brittle artifact of prompt engineering. We test this directly by re-running strategy-forcing with a completely *disjoint* $K{=}5$ set on the two hard benchmarks (Claude Sonnet 4.6): for math, {estimation, case-analysis, symmetry, constraint-satisfaction, metacognitive}, and for science, {estimation, conservation, mechanism, symmetry, reference-values}. No strategy is shared with the canonical set.

Table 11 shows the two sets decorrelate almost identically. On GPQA the mean error correlation is $\bar{\rho}{=}0.68$ for both sets (within noise), and on AIME-2024 the alternative set decorrelates slightly *more* (0.43 vs. 0.51). Coverage and per-attempt accuracy track each other closely, and compliance stays high for both ($\geq 92\%$). The decorrelation is therefore a property of forcing structurally distinct computations, independent of the specific strategy names: any set of genuinely different reasoning modes, followed by the model, yields the same effect. This is what the ensemble view predicts: what matters is that the members compute differently, irrespective of which differences they realize.

Table 11: **Strategy-set sensitivity** (Claude Sonnet 4.6, $K{=}5$). A disjoint alternative strategy set decorrelates as well as the canonical set on both hard benchmarks, so the effect is structural diversity itself and does not depend on the specific strategies. $\bar{\rho}$ lower is better, with Coverage@5 and Compliance in $\%$.

Benchmark	Strategy set	$\bar{\rho}$	K_{eff}	Coverage@5	Compliance
AIME-2024	Canonical	0.51	1.64	90.0	100.0
	Alternative	0.43	1.85	100.0	95.3
GPQA	Canonical	0.69	1.33	47.5	88.9
	Alternative	0.68	1.34	49.5	92.0

H Is the Cross-Model Floor an Artifact of Weak Models?

Our cross-model condition draws one attempt each from five deployments of varying strength (individual accuracy 50.9–76.9% pooled over MATH-500, AIME-2024, and GPQA). A natural worry is that its low correlation, and thus the "floor" OMMM is measured against, is manufactured by *quality dilution*: mixing in weaker models trivially decorrelates errors because the weak models fail on different, easier problems. If so, the cross-model floor would be an unfair reference. We test this by recomputing the cross-model metrics using only the *strongest three* deployments (Claude Sonnet 4.6, Kimi-K2.5, and DeepSeek V3.2, the three highest individual accuracies), which removes the two weakest models entirely.

Table 12 shows the decorrelation is *not* an artifact. The strongest-three ensemble has essentially the same mean error correlation as the full five ($\bar{\rho}{=}0.42$ vs. 0.43): dropping the weak models raises the ensemble's individual accuracy ($47.9 \rightarrow 56.3\%$) without raising its correlation, confirming that cross-architecture diversity, and not quality spread, is what decorrelates. The comparison to OMMM

is therefore fair, and at matched ensemble size (k=3) OMMM still leads: it attains higher coverage (76.0 vs. 72.0%) from a *single* model while keeping markedly higher per-attempt accuracy (65.7 vs. 56.3%). Strategy-forcing thus matches a strong-model cross-model ensemble's decorrelation without paying either its multi-deployment cost or its residual quality gap.

Table 12: **Matched-quality cross-model reference** (Claude Sonnet 4.6 as the OMMM reasoner, pooled over MATH-500, AIME-2024, and GPQA). Restricting the cross-model ensemble to its three strongest deployments leaves its correlation unchanged ($\bar{\rho}$=0.42 vs. 0.43), so the low cross-model floor reflects genuine architectural diversity and is free of weak-model dilution. At matched ensemble size (k=3), single-model OMMM still attains higher coverage and substantially higher per-attempt accuracy. (OMMM's $\bar{\rho}$ is a within-model quantity reported in Table 1, so here we compare coverage and individual accuracy at matched k.)

Source	k	$\bar{\rho}$	K_{eff}	Coverage@k	Indiv. acc.
Cross-model (full five)	5	0.43	1.84	72.6	47.9
Cross-model (strongest three)	3	0.42	1.63	72.0	56.3
OMMM / strategy-forcing	3	—	—	**76.0**	**65.7**

I Can a Better Aggregator Close the Coverage Gap?

Across our experiments, OMMM's Coverage@5 exceeds its majority-vote accuracy, sometimes by a wide margin (Table 2, and the losing cells discussed in Appendix E). A natural question is whether a smarter, *label-free* aggregator can convert that headroom into accuracy without the trained verifier that OMMM is designed to avoid. We test the most natural candidate: a *reliability-weighted* vote that weights each strategy's answer by that strategy's accuracy, in place of weighting all strategies equally. To keep the test label-safe, the weights applied to a problem are estimated by leave-one-out over the remaining problems (the problem's own outcome never informs its own aggregation), so no per-problem test label is used and no hyperparameter is tuned on the evaluation set.

Table 13 reports the result, which is negative. Reliability weighting helps on a few cells where one strategy is clearly stronger and the benchmark is large enough to estimate that reliably (DeepSeek AIME-2025 +10.0, HumanEval +1.9, and Qwen GPQA +2.0), but it is neutral or slightly harmful on others: most notably it *worsens* the very small AIME-2024 split for Qwen3-235B (−6.7), where 30 problems are too few to estimate per-strategy reliability without adding variance. The mean change over all 18 cells is only +0.2 points. The conclusion is that equal-weight majority is a strong baseline, and that reliably turning OMMM's coverage into answered accuracy appears to require either a verifier or a modest amount of held-out labelled data, an aggregation problem orthogonal to the decorrelation this paper establishes and a concrete direction for future work.

Table 13: **Reliability-weighted vote vs. plain majority** (%, selected cells, with full results in the released configurations). Weights are per-strategy accuracies estimated leave-one-out over problems, so no test label is used. Coverage (the oracle upper bound) is shown for reference. Reliability weighting helps where one strategy dominates on a large benchmark but adds variance on the small AIME splits, and the mean change across all 18 (model, benchmark) cells is +0.2 points. Equal-weight majority is difficult to surpass without a verifier or held-out labels.

Model	Bench	Majority	Reliability-wt.	Coverage
Qwen3-235B	AIME-24	36.7	30.0	40.0
	GPQA	27.8	29.8	41.4
	MATH	94.2	93.8	96.2
Sonnet 4.6	AIME-25	70.0	66.7	83.3
	GPQA	35.9	35.4	47.0
	HEval	98.2	99.4	100.0
DeepSeek V3.2	AIME-25	53.3	63.3	66.7
	GPQA	34.2	33.7	44.4
	HEval	95.1	97.0	99.4

J CAUSAL LINK: DECORRELATION DRIVES THE GAIN

Section 5.1 claims decorrelation *causes* the ensemble gain. Figure 7 substantiates this across all 48 (model, benchmark, diversity-source) cells of the correlation experiment. **(a)** The ensemble gain (Coverage@K minus mean individual accuracy) falls as mean error correlation $\bar{\rho}$ rises (Pearson $r{=}-0.53$, $p{=}10^{-4}$, slope -0.33): the lower the correlation a diversity source induces, the larger the ensemble benefit, regardless of which source or model produced it. **(b)** The effective ensemble size $K_{\text{eff}}{=}K/(1{+}(K{-}1)\bar{\rho})$ derived from $\bar{\rho}$ predicts the *measured* Coverage@K with $r{=}0.99$ and mean absolute error 4.5 points, validating the coverage-under-correlation model of §3.2 as quantitatively accurate, beyond merely directional.

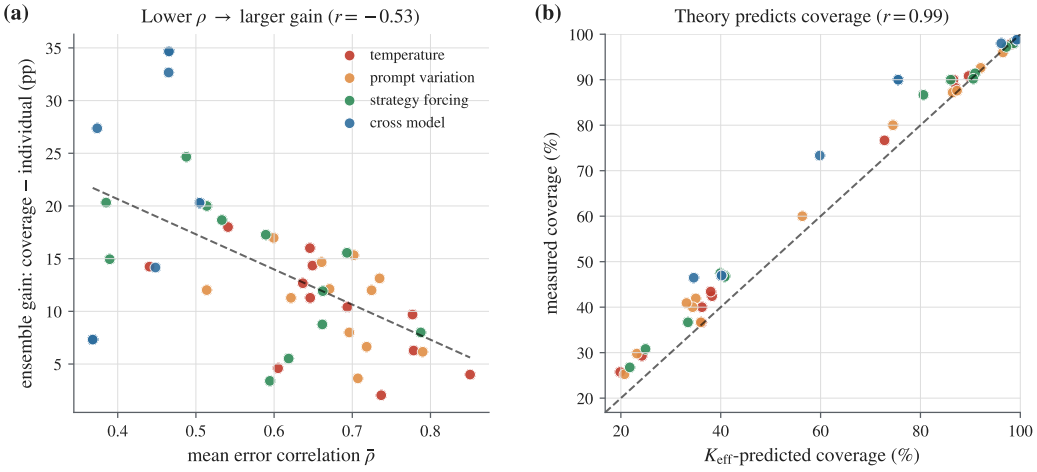

Figure 7: **Decorrelation causes the gain.** **(a)** Ensemble gain vs. mean error correlation $\bar{\rho}$ (each point is a model×benchmark×source cell, and the dashed line is the fit). **(b)** K_{eff}-predicted vs. measured Coverage@K (dashed: identity).

K STRATEGY COMPLEMENTARITY

Figure 8 shows the rescue matrix $R_{ij}{=}P(\text{strategy } j \text{ correct} \mid \text{strategy } i \text{ wrong})$ for Claude Sonnet 4.6 on the math benchmarks. Off-diagonal entries of 29–52% mean a failed strategy is frequently rescued by another, and the weakest individual strategy (inductive) is among the strongest rescuers, the hallmark of genuine complementarity as opposed to redundancy. This is the ambiguity decomposition made concrete at the strategy level: inductive reasoning errs on a different subset of problems than the analytic strategies, so its weak individual accuracy is offset by how *independently* it fails, the accuracy-for-independence trade the ensemble view rewards. In aggregate, on problems the best single strategy misses, some other strategy succeeds 59% of the time (Sonnet) / 32% (Qwen3-235B), and any-strategy coverage exceeds the best single strategy by $+4.1$ points (math) and $+14.6$ (GPQA).

L A WORKED EXAMPLE: CORRELATED VS. DECORRELATED ERRORS

The mechanism is easiest to see on a single problem. Table 14 shows AIME-2024 II-8 (gold answer 127) solved by Claude Sonnet 4.6 under two diversity sources at the same $K{=}5$ budget. Both make two errors, so they are equally accurate per attempt. What differs is *how* the errors fall. Temperature sampling draws the same default solution path five times, and its three wrong attempts collapse onto the *same* incorrect value (55), a correlated failure that outvotes the two correct attempts, so the majority vote returns 55 and is *wrong*. Strategy-forcing makes two errors as well, but they land on *different* values (4 from the geometric strategy, 55 from the backward strategy): the errors do not reinforce one another, the three correct strategies remain the plurality, and the vote returns 127 and

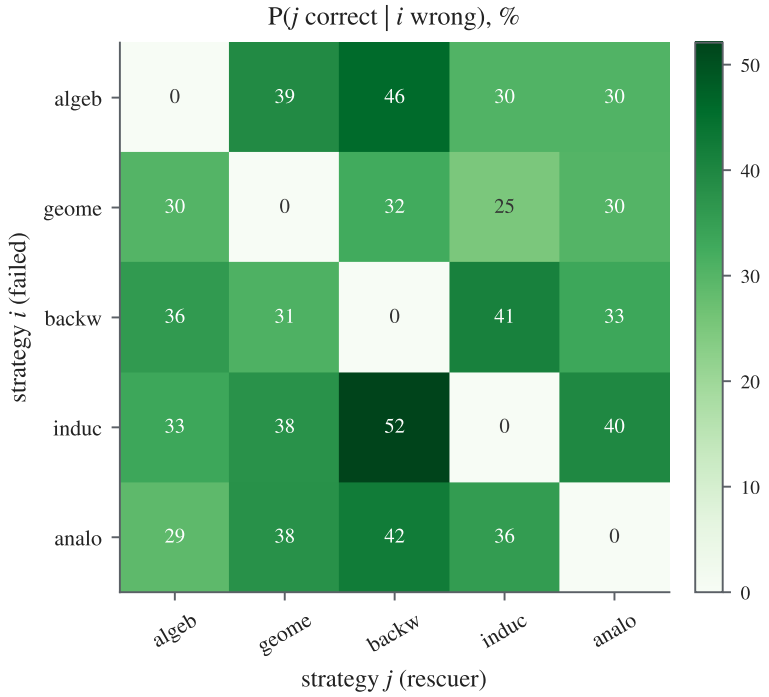

Figure 8: **Strategy complementarity** (Sonnet 4.6, math). $P(j \text{ correct} \mid i \text{ wrong})$: large off-diagonal values show strategies cover each other's distinct failures.

is *right*. The model, the budget, and the per-attempt accuracy are identical, so the only difference is that decorrelated errors fail to form a wrong majority. This is the coverage and K_{eff} argument of §3.2 instantiated on one problem: the same five votes are worth more when the wrong ones disagree.

Table 14: **Correlated vs. decorrelated errors on one problem** (Claude Sonnet 4.6, AIME-2024 II-8, gold 127). Both sources are wrong on two of five attempts. Temperature's errors concentrate on a single wrong value (55, in bold), forming a wrong majority, whereas strategy-forcing's errors scatter (4, 55), so the correct answer keeps the plurality. With equal per-attempt accuracy and opposite vote outcomes, decorrelation alone decides the vote.

Diversity source	Attempt	Answer	Correct
Temperature	T=0.2	127	✓
	T=0.4	**55**	
	T=0.6	127	✓
	T=0.8	**55**	
	T=1.0	**55**	
	Majority vote	**55**	*wrong*
Strategy-forcing	algebraic	127	✓
	geometric	4	
	backward	55	
	inductive	127	✓
	analogical	127	✓
	Majority vote	**127**	*correct*

M STRUCTURAL AND THERMAL DIVERSITY COMPOSE

Strategy-forcing (structural diversity) and temperature sampling (thermal diversity) are *orthogonal* axes, and they compose. Table 15 allocates a generation budget as K strategies $\times M$ samples each. Pure thermal scaling saturates (on Sonnet AIME-24, 1 strategy $\times 10$ samples stays at 86.7, no better than $\times 5$), whereas adding thermal samples *on top of* the five strategies keeps climbing: 5×1=86.7 $\rightarrow$ 5×2=90.0 $\rightarrow$ 5×3=96.7. Structural diversity opens headroom that thermal sampling then fills, and neither alone reaches the combination. This suggests a two-dimensional test-time scaling law: scale strategies first (steeper), then samples within each.

Table 15: **Structural $\times$ thermal diversity** (K strategies $\times M$ samples, for $K \cdot M$ total generations). Thermal samples added on top of strategies keep improving where pure resampling saturates.

Allocation	Gens	MATH	AIME-24
Qwen3-235B			
1 strat $\times 5$	5	93.6	36.7
1 strat $\times 10$	10	93.8	40.0
5 strat $\times 1$ (OMMM)	5	92.6	36.7
5 strat $\times 2$	10	93.6	40.0
5 strat $\times 3$	15	93.6	40.0
Sonnet 4.6			
1 strat $\times 5$	5	96.0	86.7
1 strat $\times 10$	10	96.2	86.7
5 strat $\times 1$ (OMMM)	5	95.8	86.7
5 strat $\times 2$	10	96.4	90.0
5 strat $\times 3$	15	96.6	96.7

Scaling on GPQA. The crossover on the hard science benchmark (Table 16, Qwen3-235B) is the cleanest illustration of the saturation argument. At small budgets self-consistency leads: with only two attempts, two temperature samples of the model's *default* mode are individually stronger than two of the forced strategies, some of which (e.g. a contrived analogy) are weaker in isolation. Thermal sampling, however, saturates almost immediately, since its third, fourth, and fifth samples are near-duplicates of the first, so its accuracy actually drifts down (27.3 $\rightarrow$ 24.2), whereas each additional strategy keeps adding a distinct, decorrelated attempt, so OMMM climbs (22.7 $\rightarrow$ 28.8) and overtakes by K=5. This is the same mechanism as the AIME scaling curve (Figure 6), viewed from the other side: where self-consistency's curve flattens, OMMM's is still climbing, because the binding constraint on the ensemble is the diversity of its members, more than their number. The practical implication is a simple allocation rule: spend the first one or two generations on the model's strongest default mode, then spend every further generation on a *new* strategy in place of a repeat sample.

Table 16: **Accuracy (%) vs. budget K on GPQA** (Qwen3-235B, 198 problems). Self-consistency leads at small K but saturates and drifts down as added samples duplicate the default mode, while OMMM keeps climbing as each new strategy adds a decorrelated attempt, overtaking by K=5. **Bold** marks the better method at each budget.

Method	K=2	K=3	K=5
Self-Consistency	**27.3**	**25.3**	24.2
OMMM (ours)	22.7	24.2	**28.8**