When Does Combining Language Models Help?

A Co-Failure Ceiling on Routing, Voting, and Mixture-of-Agents Across 67 Frontier Models

Josef Chen
KAIKAKU
josef@kaikaku.ai

June 26, 2026

Abstract

Multi-model LLM systems, including routing, voting, cascades, fusion, and mixture-of-agents (MoA), are increasingly used to push accuracy beyond any single model. We show that the achievable gain has a ceiling, fixed by a quantity other than the one the field usually reports. For any policy whose output is one of the member models' answers, such as a router, a majority vote, or a cascade, accuracy cannot exceed $1-\beta$, where β is the rate at which every model is wrong on the same query. Practice instead reports the average pairwise error correlation ρ, and we prove that ρ cannot identify β: error laws with identical marginals and identical pairwise correlations can still differ in β. A Clopper–Pearson bound on β turns one graded, held-out query set into a \$0 certificate on the largest gain any such policy could deliver, before a router is trained. Across 67 frontier models from 21 providers, among them GPT-5.5, Claude Opus 4.8, Gemini 3.1 Pro, Grok-4.3, DeepSeek V4, Qwen3.7-Max, and Kimi K2.7, a correctly (tetrachoric) calibrated single-factor model still underprices the all-wrong tail, by a margin that widens with the size of the pool: about 2.5 times on open-ended mathematics (90% CI 1.7 to 3.4, $k=17$), holding under the full 67-model Gaussian copula ($\beta=0.052$ versus a predicted 0.023) and recurring on execution-graded code ($\beta=0.079$). At matched quality, a diverse low-ρ ensemble beats a high-ρ Self-MoA one. Re-asking the same GPQA-Diamond questions in free-response rather than multiple-choice form reopens the tail ($\beta=0.127$; a five-judge LLM panel, κ from 0.73 to 0.92), locating the effect in task format rather than subject. On our pool, and on tasks where answers can be checked, combining models rarely beats the single best model without a strong query-level routing signal; the gains come from models that fail on different questions rather than from adding more of them.

1 Introduction

The single-model era is closing. Enterprises serve production traffic across hundreds of models from dozens of providers, selecting per workload on cost, latency, reliability, and capability; a routing layer increasingly mediates spend and governance, and carries the provider risk [2, 12, 31]. The operative question has shifted from which model is best to *how a buyer should allocate a token and dollar budget across a heterogeneous, correlated, rapidly depreciating pool.* Practitioners answer it with a single diagnostic: the pairwise error correlation ρ between models, low values signalling that diversity will pay.

Our central finding is that this diagnostic is the wrong one. What bounds orchestration is β, the rate at which *all* models fail on the same query: no router, vote, or cascade can exceed accuracy $1-\beta$, and ρ cannot see β. The gap is not academic, because on today's open-ended tasks the

strongest models increasingly fail together. A Clopper–Pearson bound on β, measurable from one graded query set, says in advance how much room any such policy has to beat the single best model.

What we concede. The equicorrelated variance floor is classical portfolio and ensemble theory [18, 26, 34, 36, 38] and, in its Gaussian-copula form for language-model ensembles, Turkmen et al. [35]; the oracle upper envelope and the optimality of routing and cascading are due to Dekoninck et al. [3]; and our tools (linear-programming duality [1], Clopper–Pearson intervals, the Gaussian copula, the single-factor probit) are standard. We claim no new routing algorithm. The contribution is their specialization to priced inference orchestration and the market-scale measurement.

Contributions.

1. *The orchestration ceiling and a finite-sample certificate* (§5, Prop. 1): no router, vote, or cascade can exceed $1-\beta$; the oracle gain localizes as $\Pr[\text{single-best wrong}]-\beta$; and a Clopper–Pearson bound turns one query sample into a certificate on the largest gain any such policy can deliver.
2. *Why pairwise ρ underprices co-failure* (Prop. 2): under tail dependence the single-factor estimate of β from ρ is downward-biased, with the bias diverging in pool size and driven by a common-mode atom rather than tail dependence as such.
3. *A market-scale measurement* (§4–5): on 67 models from 21 provider families, oracle routing gain is positive yet a learned router realizes almost none of it; the β/ρ gap and its growth with pool size are measured directly; and two regimes, ceiling-bound (open-ended math) and realizability-bound (science), appear across domains, though the decisive all-models-wrong counts are small (§5).
4. *Supporting economics* (App. A): budget-constrained routing as a priced assignment with a single shadow price (Prop. 4); a cost-aware diversification limit (Props. 6, 7); and cascade calibration boundaries (Prop. 8, Cor. 1). These specialize standard tools and are deferred to the appendix, as is an observational option value of breadth under churn (App. E).

2 Related Work

Routing and cascades. Learned routers select one model per query [4, 31]; cascades escalate from cheap to strong on low confidence [2, 25]. Dekoninck et al. [3] unify the two and prove optimality of a routing strategy and an optimal cascade; Jitkrittum et al. [14] characterize the optimal two-model deferral rule and show confidence-based deferral can be sub-optimal when the downstream model's errors are unmodeled, directly relevant to our calibration-and-edge condition. Hu et al. [12] benchmark routing on pre-computed outcomes. These works treat the oracle as an empirical ceiling and the cascade threshold as a tuned hyperparameter; we add the dollar stopping rule, the calibration dominance boundary, and the volume ceiling they leave implicit.

Ensembling and fusion. Jiang et al. [13] rank-and-fuse generations; Wang et al. [37] aggregate in layers; Li et al. [23] sample-and-vote; Li et al. [24] show that sampling the single best model (Self-MoA) often beats heterogeneous fusion, attributing this to a quality–diversity trade-off (mixing helps only when members are of similar quality). Classical ensemble theory supplies the variance decomposition [18, 36], its modern unification [38], and the caution that diversity does not guarantee

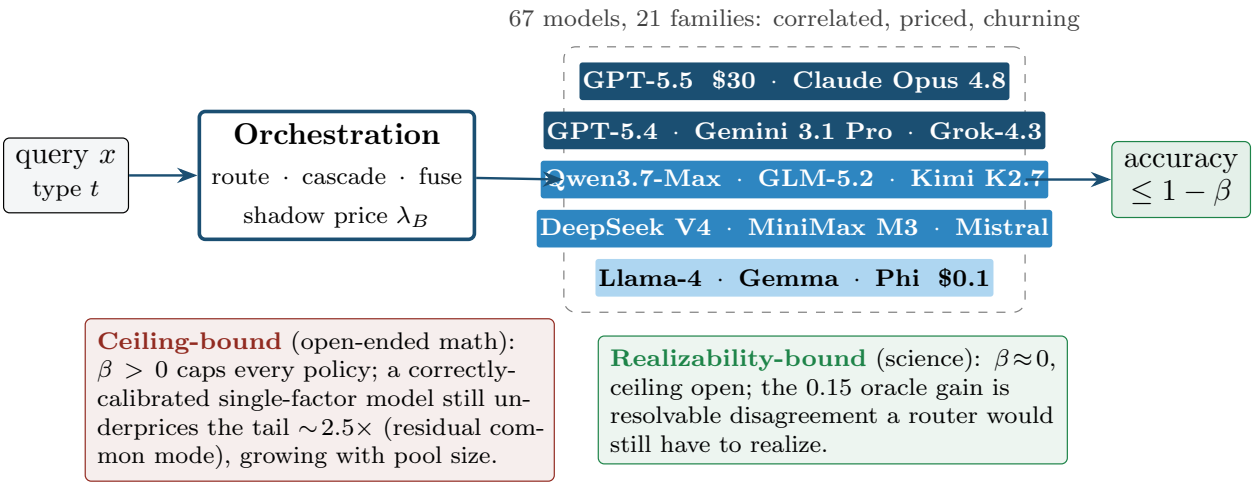

Figure 1: **Orchestration is allocation.** A query is routed, cascaded, or fused over a priced, correlated, fast-churning pool of 67 frontier-to-cheap models; the optimal policy is a per-type bang-per-buck rule with a single shadow price λ_B on the inference dollar (Prop. 4). No selection policy can exceed the ceiling $1-\beta$ set by the rate β at which *all* models fail at once (Prop. 1). The field reports pairwise error correlation ρ to decide whether to orchestrate; ρ is blind to β, and headroom is foreclosed in two opposite regimes that ρ cannot tell apart.

accuracy gains [20]. The accuracy *ceiling* we use is also classical: it is Kuncheva's *Oracle* combiner (correct iff some member is correct, i.e. $1-\beta$ [19]) and the majority-vote-accuracy limits of Kuncheva et al. [21], both predating LLMs and presuming labels. Our contribution is not the ceiling but (i) that it bounds *every* selection policy, debate and self-consistency included, whose outputs are a.s. member answers, and (ii) its conversion into a labels-free, finite-sample \$0 *certificate* (Prop. 1); an Oracle needs the labels a certificate does not.

Error correlation and inference economics. Kim et al. [15] document, over 350+ models, that *pairwise* errors are substantially correlated and rise with accuracy and shared provider, but they measure the bivariate agree-when-wrong statistic on multiple-choice tasks and stop at qualitative implications; they neither define the joint all-models-wrong rate β nor bound any selection policy by it. We show that pairwise correlation provably cannot identify β for $m \geq 3$ (Prop. 3), and that β, not ρ, is the ceiling. This monoculture of errors traces to Kleinberg and Raghavan [16]. Closest is Turkmen et al. [35], who also fit a tetrachoric Gaussian copula to binary LLM errors and derive an equicorrelated ensemble error *floor*; we reach the *opposite* conclusion and show why: that very Gaussian floor *underprices* the empirical co-failure tail, because the driver is a common-mode atom the copula cannot represent (zero lower tail dependence; Prop. 2), which we confirm on the full 67-model matrix and against a Clayton control. The copula they use to bound the tail is the one we measure to be loose. Erol et al. [8] formalize dollars-per-correct via production theory, our primary metric.

Finance and real options. Mean–variance allocation [26], cost-aware diversification [34], the value of information [11], selective prediction [10], switching costs [17], and real-options volatility comparative statics [5, 28, 29] are the antecedents we specialize and, where standard, concede.

3 Problem Formulation

Queries x carry a latent type $t = T(x) \sim D$, with type prior $p(t) = \Pr[T = t]$. A pool $M = \{1, \ldots, m\}$ of models has quality $q_i(t) \in [0, 1]$ (expected per-query utility on type t) and price $c_i \geq 0$ (dollars/query); write $\bar{q}_i = \mathbb{E}_t q_i(t)$. A (possibly stochastic) routing policy $\pi \colon T \to \Delta(M)$ has value $V(\pi) = \mathbb{E}_t \sum_i \pi(i \mid t)\, q_i(t)$ and cost $K(\pi) = \mathbb{E}_t \sum_i \pi(i \mid t)\, c_i$. We also consider fusion (query several, combine) and cascading (escalate on low confidence). The buyer's objective is dollars-per-correct [8], or quality subject to a budget; we make it explicit in each section.

Economic scaffolding (deferred to App. A). Treating orchestration as allocation yields a compact economic layer that we use but do not foreground; the tools are standard and the empirical results below do not depend on it. Three facts, stated and proved in App. A: under a dollar budget, routing is a priced assignment with a single shadow price λ_B (Prop. 4); cost-aware fusion has a diversification limit $k^*(\rho, c)$ that shrinks as error correlation rises (Props. 6, 7), above the classical equicorrelated variance floor we concede (Prop. 5); and a calibrated cascade collapses to random mixing exactly as the verifier AUC falls to 1/2, with a price-independent escalation ceiling $1 - a_L/a_H$ (Prop. 8, Cor. 1).

4 Experimental Setup

We executed a pre-registered program on a pool of 15 current models across 9 provider families: frontier (Claude Opus 4.8, GPT-5.1, Gemini 3.1 Pro, Kimi K2.7), mid (Claude Sonnet 4.6, GPT-5-mini, Gemini 3.5 Flash, Qwen3-235B, Mistral-Large, MiniMax M2.7, DeepSeek V3.2), and cheap (Claude Haiku 4.5, GPT-5-nano, Gemini 3.1 Flash-Lite, Llama-4-Maverick); exact dated snapshots and prices are frozen in the registry (App. C). The pillar experiments use five benchmarks: a saturated mix (GSM8K, MMLU, ARC-Challenge, MATH-500) and a harder set (MMLU-Pro), with 100–200 queries per dataset. For the *market-scale* realizability measurement (§5) we expand the pool to 67 models across 21 provider families—the live OpenRouter catalog from the current frontier down to small open-weights (GLM, Qwen, DeepSeek, MiniMax, Nemotron, Llama-3.x, Mistral, Gemma, Phi, Granite, and others), chat/instruct only, with live-verified prices (full named roster, App. D)—and add hard domains that probe co-failure: open-ended competition mathematics on two benchmarks (MATH-500 and the harder MATH-Hard Level-5; plus AIME-2024/2025, whose release post-dates the older models' training cutoff) and graduate-level science (GPQA-Diamond, Physics/Chemistry/Biology). Grading is programmatic throughout: exact-match arithmetic, multiple-choice and boxed-letter extraction, and boxed/integer answer matching, so no LLM judge is used. Costs are metered per call against the OpenRouter account usage endpoint; OpenRouter is an aggregator, so this is account-level usage, not per-provider reconciliation. We itemize per-run metered cost in App. C: the core pillar experiments total ≈$47, the market-scale realizability and two-regime measurement ≈$111, and the two third-domain experiments (code, open-ended GPQA) ≈$110, for ≈$270 of reported-experiment cost; total account usage including all exploratory and superseded iteration was ≈$560 (approximate; see App. C). We report the itemized experiment figures rather than presenting account-level usage as experiment cost. Baselines: single-cheapest, single-best (selected in-sample; the optimistic bias *understates* the oracle gain $G = V^o - a_{\text{sb}}$, so our small-G claim is conservative, though it flatters the learned-router comparison), random and random-mixing-at-matched-budget, cost-matched Self-MoA [24], a partition-conditioned oracle and a partition-free per-query oracle, a confidence cascade, majority vote, and heterogeneous fusion.

5 Results

All correctness is scored by an answer-anchored grader; an earlier first-letter extractor systematically mis-scored verbose models (e.g. Llama-4-Maverick by +0.26 accuracy, mean $|\Delta| = 0.05$ across the pool), so we re-graded all cached model outputs at no additional inference cost and report the corrected numbers throughout.

Quantity	Saturated multi-domain mix	Hard single-domain (MMLU-Pro)
Single-best	0.923 (Opus 4.8)	0.850 (Sonnet 4.6)
Oracle (per-query)	0.967	0.970
Oracle gain G (95% CI)	**0.044** $[0.027, 0.062]$	**0.120** $[0.075, 0.155]$
mean off-diagonal ρ	0.464	0.382
within-family ρ	**0.528**	**0.402**
cross-family ρ	0.459	0.380

Table 1: Pillar A in two regimes (re-graded; G with 2000-resample query-bootstrap 95% CIs, N=120–200). $G > 0$ with both CIs excluding zero confirms Q is not row-dominated (Lem. 1), and G is larger on the harder, more dispersed regime. Within-family ρ exceeds cross-family in both regimes, with a larger gap on the multi-domain mix (0.069 vs 0.022): family specialization shows most across domains, consistent with the shared-provider correlation of Kim et al. [15].

Pillar A (confirmed, both regimes). Oracle gain $G > 0$ with bootstrap CIs excluding zero in both regimes (0.044 saturated, 0.120 hard; Table 1): routing helps, *modestly because the frontier agrees*, and more on the harder, more dispersed regime, the dispersion signature the theory predicts. Within>cross family ρ holds in both regimes (gap larger on the multi-domain mix). The cost–quality frontier is populated by cheap models (Fig. 5). A deployable learned router captures essentially none of G, and this holds against router strength: a held-out TF-IDF+domain logistic attains 0.906 vs. single-best 0.901 on the mix (9% of G, 95% CI $[-0.67, 0.50]$), and, to rule out a weak-baseline artifact, three substantially stronger routers fare no better. A gradient-boosted per-model correctness predictor on word+char TF-IDF features captures -0.09 of G; a direct multiclass best-model predictor captures -1.27 (it actively hurts); and a deployment-realistic *LLM-as-router* (GPT-5-mini shown each query and a capsule of every model's strengths, asked to pick the best) routes to single-best on 100% of queries and captures exactly 0 of G (`router_strong.py`, `router_llm.py`). All four routers are evaluated on the 15-model saturated mix, the pool whose per-query prompts we logged; the market-scale (67-model) and GPQA matrices store outcomes but not prompts, so no router was trained there and the market-scale routing statement rests on the certificate of Prop. 1, not an end-to-end routing run. This scope limit we state plainly rather than paper over. Against the cost-aware oracle (optimal) frontier all sit far below the upper bound (Fig. 9). The realizable routing gain is thus near zero *not because the router is weak* but because the prompt carries little signal about which model will be the one that is right when the frontier disagrees: the small oracle bound is itself largely unreachable.

Tail co-failure: a realizability certificate and an empirical finding (§5). Why is the realizable gain near zero, and the oracle gain itself small? Both are governed by how often the pool fails together. The next proposition makes the ceiling exact and turns it into a $0 pre-deployment test; we then report what the tail looks like on the frontier.

Proposition 1 (Ceiling, gain localization, and a realizability certificate)**.** *Let* $\beta = \Pr_t[\textit{all } m \textit{ wrong}]$, $a_{\mathrm{sb}} = \max_i \bar{q}_i$, *and* $i^\star = \arg\max_i \bar{q}_i$. (i) Ceiling. *Any selection policy—a router, a (weighted) vote, or a cascade, whose output is almost surely one of the members' answers—has accuracy at most* $1-\beta$, *attained by the per-query oracle, so the maximum gain over single-best is exactly* $\Delta^{\mathrm{ceil}} = (1-\beta) - a_{\mathrm{sb}}$. (ii) Gain localization. $G = V^o - a_{\mathrm{sb}} = \Pr_t[\textit{single-best wrong}] - \beta$, *supported entirely on the* resolvable *mass (non-unanimous, single-best wrong); the co-failure tail* β *contributes nothing to* G. (iii) Certificate. *From* n *i.i.d. queries with* K *all-wrong, let* $\beta_{\mathrm{lo}}(K, n, \delta)$ *be the Clopper–Pearson lower confidence limit; then with probability* $\geq 1-\delta$ *every selection policy obeys* $\mathrm{Acc} - a_{\mathrm{sb}} \leq (1-\beta_{\mathrm{lo}}) - a_{\mathrm{sb}}$. *If this certified bound falls below the orchestration overhead, no policy in the class can pay for itself—a \$0 test (*a_{sb} *replaceable by its own confidence bound).*

Parts (i)–(ii) are elementary identities, with one-line proofs (App. B): on the all-wrong event every member is wrong, so any selector is wrong; and G rearranges $V^o = 1-\beta$. We state them not as deep results but because part (iii) turns them into a pre-deployment \$0 instrument, and part (ii) corrects a tempting misstatement: a *small* β does not by itself imply orchestration cannot help; it implies a high ceiling. The binding quantity is $\Delta^{\mathrm{ceil}} = (1-\beta) - a_{\mathrm{sb}}$, which is small on the frontier only because a_{sb} already sits near the ceiling; the certificate is most informative exactly there.

The empirical finding (measured, not a law). The statistic the field reports, mean pairwise error correlation ρ, systematically underprices this tail. Fitting a single-factor Gaussian copula to the measured ρ predicts an all-wrong rate β_{sf} far below the observed β (Table 2), and the gap persists when the pool is restricted to one model per provider family (not a same-vendor artifact). The implied tail correlation that reproduces β far exceeds the pairwise value—the body-vs-tail base-correlation smile of Gaussian-copula portfolio-credit (CDO) models, which we invoke as a known analogy, not a new object. Because β rests on few all-wrong events, we report it with exact Clopper–Pearson intervals; the underpricing factor is the headline, but its magnitude carries real uncertainty, which the market-scale measurement below narrows.

	Saturated mix	Hard (MMLU-Pro)
all-models-wrong rate β (95% CP)	0.033 [0.019, 0.054]	0.030 [0.011, 0.064]
mean pairwise ρ (*naive Pearson-of-indicators*)	0.464	0.382
β predicted by single-factor copula at that ρ	0.0011	0.0050
naive-Pearson underpricing (**overstated**; cf. tetrachoric below)	30× [17, 48]	6× [2, 13]
implied (tail) correlation reproducing β	0.88	0.64
realizable router gain (fraction of G)	0.09 (CI spans 0)	< 0

Table 2: Tail co-failure on the 15-model frontier pool (recomputed from the logged re-graded matrices via `realizability.py`, \$0; β with exact Clopper–Pearson 95% intervals, n=480/200, all-wrong counts k=16/6). The 6–30× figures use the *naive Pearson-of-indicators* calibration and are **overstated**—we retain them only to show the raw gap; the correctly *tetrachoric*-calibrated residual is single-digit (≈2.5× on the market-scale MATH-500 tail, Fig. 2), the order-of-magnitude difference being the calibration artifact we diagnose in §5. The wide intervals reflect the few all-wrong events, which the market-scale measurement narrows. The ceiling $1-\beta$ and the near-zero realizable router gain both follow from this tail, not from ρ (Prop. 1).

The mispricing is a large-pool phenomenon (market scale). We expand to a 67-model, 21-family market pool, the live OpenRouter frontier (GPT-5.5, Claude Opus 4.8, Gemini 3.1 Pro, Grok-4.3, GLM-5.2, Qwen3.7-Max, DeepSeek V4, Kimi K2.7, MiniMax M3) down to small open-weights, over the hard benchmarks (Fig. 2). The load-bearing benchmark is *MATH-500*, the one

domain with enough co-failure events to estimate β at all: over the full 67-model pool (n=330 fully-covered queries) all models miss the same problem $\beta = 0.052$ of the time, but this rests on only k=17 all-wrong events (Clopper–Pearson $[0.030, 0.081]$, a wide interval). The single-factor copula must be calibrated with the *tetrachoric* (latent) correlation, not the Pearson correlation of 0/1 correctness indicators (Prop. 7), a century-old psychometric point [30, 32] we claim no novelty for, but which the LLM-evaluation literature routinely elides. The measured tetrachoric $\bar{\rho} = 0.78$ predicts $\beta_{\text{sf}} = 0.021$, so the observed tail is $\approx$**2.5**$\times$ fatter (bootstrap 90% CI 1.7–3.4$\times$ over queries, jointly propagating the all-wrong count and the fitted $\bar{\rho}$; `ratio_uncertainty.json`) than even a correctly-calibrated single-factor model: a real but *modest* residual common-mode excess (the implied $\rho_{\text{eff}} = 0.89$ exceeds the measured 0.78). **The residual is not an artifact of the single-factor restriction.** Fitting the *full* 67$\times$67 pairwise-tetrachoric correlation matrix Σ (every pair calibrated to its own joint wrong-rate, projected to the nearest PSD matrix) and Monte-Carlo-integrating the all-wrong event under the resulting Gaussian copula still predicts only $\beta_{\text{full-}\Sigma} = 0.023$ (4×10^5 draws; `residual_decomp.py`), leaving the empirical 0.052 a **2.25**$\times$ excess beyond the *nearest-PSD Gaussian copula*—the finite-pool signature of a common-mode atom (Props. 2, 3), which a Gaussian copula cannot represent (zero lower tail dependence), not single-factor misspecification. Two caveats keep this exact: the empirical tetrachoric matrix is indefinite (26 negative eigenvalues) and the PSD projection *lowers* the mean calibrated correlation (0.78$\rightarrow$0.74)—which, if anything, *inflates* this ratio; and a non-Gaussian copula with *genuine* lower-tail dependence does no better on the real data. An exchangeable Clayton copula ($\lambda_L = 0.69$), calibrated to the same mean pairwise co-failure on the full 67-model matrix, still predicts $\beta = 0.026$ versus the empirical 0.052, a 1.96$\times$ residual (6.3$\times$ on MATH-Hard; `clayton_real.py`). So the gap is *not* an artifact of the Gaussian's zero tail dependence. The residual therefore lies beyond any exchangeable pairwise-calibrated copula we can fit, Gaussian *or* tail-dependent: the signature of a common-mode atom that no pairwise statistic represents. We flag the calibration trap explicitly, because an earlier version of this analysis fell into it: the *naive* Pearson-of-indicators calibration ($\bar{\rho} = 0.53$) gives $\beta_{\text{sf}} = 0.0016$ and a spurious 32$\times$: an order-of-magnitude artifact of the wrong correlation transform, not a co-failure effect (`realizability_tetrachoric.json`). **The excess is a pool-*size* effect, not a composition accident.** Resampling pool composition (random k-model subsets, 60 per k) the tetrachoric ratio rises monotonically from 1.0 at k=2 to a median 2.5 (5–95% band $[2.1, 2.7]$) at k=67, with *every* subset showing a populated tail (`residual_decomp.json`): isolating size, not which models, as the driver, exactly as Prop. 2 predicts; the newest frontier (GPT-5.5 and peers) still co-fails. **The finding replicates on a second, harder open-ended math benchmark**, though both are the *same task family* at two difficulties, not independent domains. On MATH-Hard (Level-5 MATH; 67 models, n=298) the co-failure tail is again populated ($\beta = 0.044$, k=13 all-wrong, CP$[0.023, 0.073]$), and the tetrachoric-calibrated single-factor model underprices it by a point 8.3$\times$ ($\bar{\rho}_{\text{tet}} = 0.69$; bootstrap 90% CI 4.5–16$\times$). We deliberately do *not* read this larger point ratio as stronger co-failure. MATH-Hard's β (0.044) is in fact *lower* than MATH-500's (0.052); the higher ratio is a *denominator* effect: its lower fitted $\bar{\rho}_{\text{tet}}$ shrinks the single-factor baseline β_{sf}. Matched to MATH-500's $\bar{\rho} = 0.78$, MATH-Hard's ratio is 3.3$\times$ (`ratio_uncertainty.json`), comparable to MATH-500's. The honest reading is therefore a *consistent single-digit* residual common-mode excess ($\approx$2.5–3.3$\times$ at matched ρ; point ratios 2.5–8.3$\times$, both with wide CIs over k=17 and 13 events), replicated *within* open-ended math but not across domains. The thinner MMLU-Pro tail (one all-wrong event in 124) we read only as directional, and on multiple-choice GPQA the tail vanishes ($\beta \approx 0$). **The co-failure regime generalizes to a third, structurally independent domain.** On execution-graded competitive programming (`code_contests`: 63 hard problems, rating 1900–3500, each graded against its *private* + *generated* stress tests under an enforced, Python-fair time limit; §F), the tail is populated ($\beta = 0.079$, k=5 all-wrong over 63, CP$[0.026, 0.176]$). The same naive-Pearson trap recurs

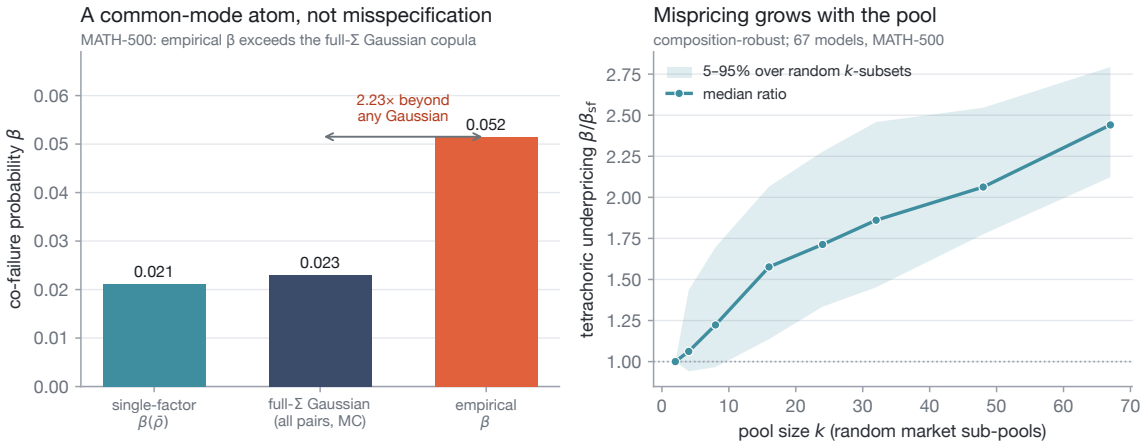

Figure 2: The co-failure residual is a common-mode atom, not copula misspecification (MATH-500, 67 models, k=17/330). *Left*: three predictions of the all-models-wrong rate against the empirical $\beta = 0.052$—the one-parameter single-factor copula ($\beta(\bar{\rho}) = 0.021$), and the *full* 67×67 pairwise-tetrachoric Gaussian copula Monte-Carlo'd over all pairs ($\beta_{\text{full-}\Sigma} = 0.023$). The empirical tail exceeds even the nearest-PSD full-Σ Gaussian fit by 2.25×, whose lower tail is asymptotically independent (Props. 2, 3). (The discredited Pearson-of-indicators calibration would put $\beta_{\text{sf}} = 0.0016$, a spurious 32×; we exclude it.) *Right*: resampling pool *composition* (random k-model subsets, 60 per k), the tetrachoric underpricing rises monotonically from 1.0 at k=2 to median 2.5 (5–95% band $[2.1, 2.7]$) at k=67—size, not which models, drives it. Computed by `residual_decomp.py`.

(Pearson $\bar{\rho}$=0.27 implies a spurious 17×), the tetrachoric single-factor model underprices by **3.1**×, and even the full-Σ Gaussian copula leaves a 1.7× residual (`residual_decomp.json`): the same common-mode signature as math. With the harder problems the underpricing is now *statistically resolved*: bootstrap 90% CI $[1.5, 6.2]$, excluding 1. The honest caveats remain: k=5 is still a small event base, 18 models (not 67), and a strict-but-not-official judge (App. F). The signature thus holds across *three* structurally disjoint open-ended domains (two math families and execution-graded code) and *vanishes* on multiple-choice: co-failure ($\beta > 0$), the Pearson trap, a full-Σ residual, and a correlation-excluding-1 tetrachoric ratio. The open-ended-versus-multiple-choice split is a cross-domain phenomenon, not a math artifact.

The monotone growth is not an artifact of the scan; it is forced by any positive tail dependence, which we now prove.

Proposition 2 (Pairwise ρ underprices co-failure, with bias growing in pool size)**.** *Model errors by a common-shock mixture: each query is* co-hard *with probability π (all m models err together) and otherwise each model errs independently with probability α_0. The marginal error rate is $\alpha = \pi + (1-\pi)\alpha_0$ and the pairwise error correlation is $\bar{\rho} = [\pi + (1-\pi)\alpha_0^2 - \alpha^2]/[\alpha(1-\alpha)] > 0$. The true co-failure rate is $\beta(m) = \pi + (1-\pi)\alpha_0^m$. Let $\beta_{\text{sf}}(m)$ be the all-wrong rate of the single-factor Gaussian copula calibrated to $(\alpha, \bar{\rho})$. Then (i) $\beta(m) \downarrow \pi > 0$ while $\beta_{\text{sf}}(m) \downarrow 0$, because a Gaussian copula with $\bar{\rho} < 1$ has zero lower tail dependence [7, 33]; hence (ii) the underpricing ratio $\beta(m)/\beta_{\text{sf}}(m) \to \infty$ and is eventually strictly increasing in m, while equalling 1 at $m = 2$. Mean pairwise ρ is a sufficient statistic for the* bivariate *error law but discards the higher-order tail dependence that governs joint failure of a large pool; it is exact for pairs and increasingly inadequate as the pool grows.*

The classical ingredient—a Gaussian/elliptical copula has zero lower tail dependence [33], so a

common-mode atom (Marshall–Olkin-type shared-failure component; 27) cannot be represented by any pairwise calibration—is not ours; the body–vs–tail underpricing it implies is the base-correlation "smile" familiar from Gaussian-copula credit models [6, 22]. What we own is the *transfer* to LLM orchestration: the co-failure instantiation, the pool-size-divergence framing, and the empirical measurement that it does not vanish on the real frontier. It grounds the paper's flagship empirical claim: the measured curve (ratio ≈ 1 at k=2, growing monotonically with k) has the monotone-divergent shape this predicts. The exchangeable, homogeneous model licenses the mechanism and the sign—any positive tail dependence forces the divergence—while the heterogeneous magnitude (the specific $1.3\times \to 2.5\times$ tetrachoric on MATH-500) is empirical. It also says precisely *why* a buyer cannot price orchestration from the reported statistic: ρ certifies pairwise substitutability, not the tail in which orchestration either pays or does not.

The driver is a common mode, not tail dependence *per se*. It is tempting to attribute the effect to the lower tail-dependence coefficient λ_L, but that is false, and the distinction is the substance of the result. What pairwise ρ misses is a *common-mode atom*: a positive-probability event of joint failure, $\beta_\infty := \lim_m \beta(m) > 0$, that the pairwise correlation underweights. Smooth tail dependence that is *already reflected* in ρ does not produce the effect. We verified the dichotomy in a logged simulation (`copula_dichotomy.py`): an exchangeable Clayton copula (Archimedean, $\lambda_L = 2^{-1/\theta}$) yields only *bounded* underpricing under the same binary-ρ pipeline—4.0× at m=53 for λ_L=0.71, and *smaller* (2.6×) for larger λ_L=0.84, because higher λ_L raises the pairwise ρ the single-factor model then absorbs—whereas a common-shock mixture with the same marginals but a rare shared-failure atom (β_∞=0.05) drives the ratio past $10^7\times$. The orchestration-relevant object is thus the multivariate co-failure floor β_∞, which no single pairwise number can identify; this both sharpens the certificate of Prop. 1 and explains why the gap *widens* with pool size (pairwise ρ saturates while β_∞ does not). We state the underlying non-identification exactly.

Proposition 3 (Non-identification of the co-failure floor; specialization of a classical Fréchet-class fact)**.** *For $m \geq 3$ the co-failure rate $\beta = \Pr[\text{all } m \text{ wrong}]$ is* not *a function of the pairwise error law: there exist joint distributions over $\{0,1\}^m$ with identical one- and two-dimensional marginals—hence identical marginal error rates and identical pairwise correlations, Pearson* and *tetrachoric—yet different β. Consequently no statistic computed from pairwise correlations, a single-factor copula included, can identify β or the large-pool floor β_∞; the common-shock atom and a matched-$\bar{\rho}$ Gaussian dependence (Prop. 2) are precisely the two extremes of this ambiguity ($\beta_\infty > 0$ versus $\beta_\infty = 0$ at identical pairwise law).*

We claim no novelty for the mathematics: the proof (App. B) is the classical fact that low-order marginals underdetermine a joint on $\{0,1\}^m$ for $m \geq 3$ (the Fréchet class has positive dimension; 9), instantiated for the co-failure cell. The consequence for practice is that pairwise ρ cannot, even in principle, see the quantity that caps orchestration, so a direct estimate of β (with the certificate of Prop. 1), not any function of ρ, is the right instrument.

Two regimes across three domains: either the ceiling binds, or it does not. The gain-localization identity $G = \Pr[\text{single-best wrong}] - \beta$ says orchestration headroom can be foreclosed in two opposite ways, realized across our hard benchmarks (Table 3). On open-ended mathematics (MATH-500) the **ceiling binds**: a real co-failure tail ($\beta = 0.052$, k=17) caps every policy at $1 - \beta$, which a correctly tetrachoric-calibrated single-factor model still underprices $\sim 2.5\times$, and little gain is even available. On graduate-level science (GPQA-Diamond, 52-model complete-coverage subset) the picture **inverts**: all-models-wrong is indistinguishable from zero (0 all-wrong on 130 covered queries; 95% Clopper–Pearson upper bound 0.03), so the ceiling is effectively open—and yet the

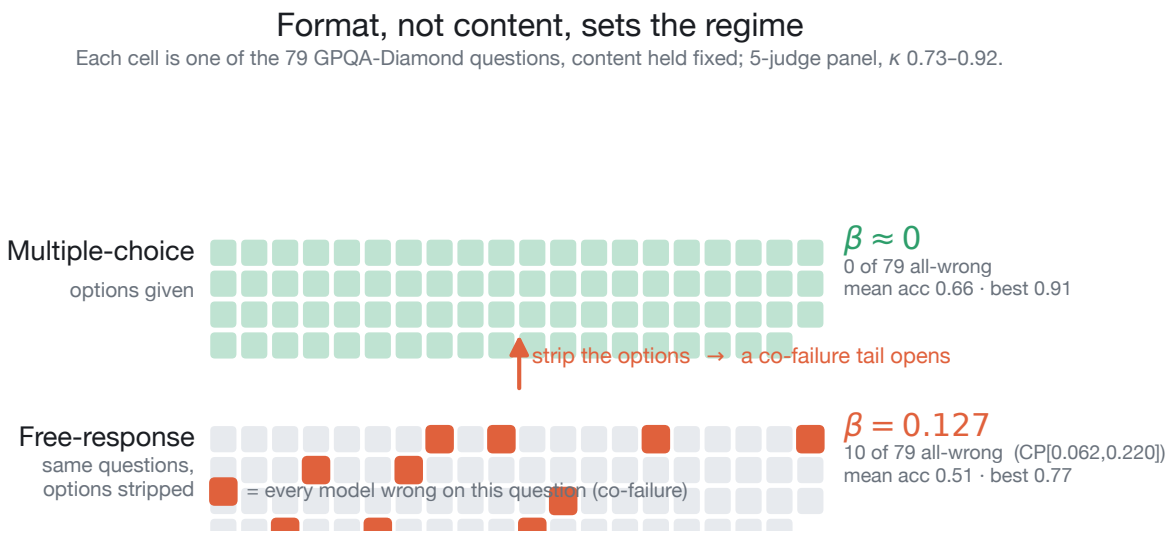

Figure 3: **Format, not content, sets the regime.** The same 79 GPQA-Diamond questions, asked multiple-choice (top) and free-response (bottom; options stripped, 5-judge panel, κ 0.73–0.92). Each cell is one question; an orange cell is one on which *every* model is wrong. Changing only the format opens a co-failure block of 10/79 (β=0.127, CP$[0.062, 0.220]$) where multiple-choice had none ($\beta \approx 0$), while mean accuracy falls $0.66 \to 0.51$. Recomputed from the committed open-ended outcome matrix (`matrix_marketGPQAOPEN`).

oracle gain is *larger*, $G = 0.15$, of which the identity certifies that essentially every point is resolvable disagreement, not co-failure. We did not train a router here—the market/GPQA matrices log no prompts—so whether one could capture this gain is open; what the certificate shows is that the gap is bounded by *routing regret*, not by the tail, the opposite of the math regime. Execution-graded code (App. F) sits in the *ceiling-bound* regime alongside math ($\beta = 0.079$), so the pattern holds across three structurally disjoint open-ended domains: co-failure tracks *open-endedness*. A multiple-choice guess floor over a broad pool tends to make joint failure rare, though not impossible (MMLU-Pro shows a lone all-wrong event). So the operative question is not "how correlated are the models" but "which regime is this workload in," which pairwise ρ cannot answer and the certificate can (Fig. 4).

The regime is set by format, not content—a content-controlled test. We re-ran the *same* GPQA-Diamond questions as free-response (options stripped), graded by a 5-judge LLM panel (pairwise $\kappa = 0.73$–0.92, substantial-to-near-perfect agreement; App. G). Holding the science content fixed and changing only the *format* **flips the regime**: on the models and questions common to both runs, mean accuracy falls $0.66 \to 0.51$ (best-model $0.91 \to 0.77$), and a co-failure tail *opens*—$\beta = 0.127$ (k=10 over 79 fully-judged questions, CP$[0.062, 0.220]$), comparable to math and code, where the multiple-choice version had $\beta \approx 0$. So co-failure tracks *open-endedness* itself, not subject matter, and it appears even on an LLM-judged generative task, not only programmatically-verifiable ones—the strongest evidence that the open-ended/multiple-choice split is real and not a domain confound (Fig. 3).

Pillar B (two complementary findings). *(1) Naive diversity is a liability.* On all $\binom{15}{3} = 455$ three-model triplets, regressing the unweighted majority-vote gain over the best member on ρ with accuracy-headroom control and model-clustered (leave-one-model-out jackknife) inference, the mean vote gain is negative (-0.10 hard, -0.02 saturated, robust across the jackknife): mixing

	MATH-500 (math)	code_contests (code)	GPQA-Diamond (science)
models / queries	67 / 330	18 / 63	52 / 130
single-best	0.836	0.825	0.846
per-query oracle	0.948	0.921	1.000
oracle gain G	0.112	0.096	**0.154**
all-models-wrong β	**0.052** (k=17)	**0.079** (k=5)	< 0.03 (0/130)
mean pairwise ρ	0.53	0.27	0.25
ρ underprices β by (tetrachoric)	**2.5**×	**3.1**× (CI [1.5, 6.2])	— ($\beta \approx 0$)
regime	**ceiling-bound**	**ceiling-bound**	**realizability-bound**

Table 3: Three structurally disjoint open-ended domains, two opposite ways orchestration headroom is foreclosed (2026 frontier market pool: MATH-500 on the full 67 models; execution-graded code on 18 models / 63 reference-certified problems (App. F; β CP $[0.026, 0.176]$, underpricing CI $[1.5, 6.2]$ now excludes 1, though k=5 is still small); GPQA on the 52-model complete-coverage subset—the newest reasoning models' GPQA cells were incomplete at the budget cap, β there indistinguishable from 0). Where the ceiling binds (math), pairwise ρ underprices the binding co-failure tail; where it is slack (science), a large oracle gain is pure resolvable disagreement a deployable router still misses. Pairwise ρ identifies neither regime.

unequal-quality models lets diverse-but-weaker members outvote a strong one. The finer prediction that the gain rises with $1 - \rho$ has a positive point estimate (slope +0.13 hard) but is not significant under model-clustering (jackknife 95% CI $[-0.07, +0.34]$; the naive i.i.d. bootstrap CI is $\approx 3\times$ too narrow), so we report it as suggestive (Fig. 7). This refutes the "more diverse $\Rightarrow$ better fusion" intuition and matches Kuncheva and Whitaker [20], Li et al. [24]. *(2) At matched quality, the diversification mechanism is supported in one regime.* We test the diversification theorem in its valid regime by contrasting Self-MoA (distinct samples of the single best model; high intra-model error correlation ρ=0.80) against heterogeneous fusion over an accuracy-matched 6-model band (members 0.74–0.865; lower inter-model correlation ρ=0.42, estimated on a disjoint sample split), with S=9 so Self-MoA scales on distinct draws rather than recycling a fixed budget (Fig. 8). At the *information-fair* comparison (k=3, equal distinct draws per side), the low-correlation heterogeneous ensemble beats Self-MoA at k=3. Across 60 resamplings of the sample-to-split partition the gain averages +0.027 (range +0.010 to +0.050; positive in all 60), so the *direction* is robust while the magnitude is partition-sensitive. The +0.055 a single favourable partition gives (query-bootstrap CI $[+0.025, +0.090]$) is the upper end of this spread, and an alternative aggregation gives +0.025 near the mean; we therefore report the partition-averaged +0.027 and treat the mechanism as supported in one regime, not established (§7). This is consistent with the diversification limit's core prediction: lower inter-model error correlation buys larger diversifiable gains at matched quality. The advantage also shrinks as ρ rises, the direction $k^*(\rho)$ predicts. On the higher-correlation MATH-500 regime (ρ_{inter}=0.59) it falls to +0.020 (not significant). Across up to 502 matched-quality sub-bands of a 9-model band, the accuracy-controlled gain-vs-ρ slope is robustly negative (the diversification-limit sign, reversing the unmatched pool's positive slope) and stable across 6- and 9-model pools. It is not significant under model-clustering even with 9 clusters: the sign is real but small against model-level variance (Fig. 10). We report the contrast in the Self-MoA frame, where ρ_{intra} and ρ_{inter} are distinct estimands. The sensitivity λ is now pinned down as a decision-rule Jacobian (Prop. 7); fitting it across ρ levels on data, to predict k^* out of sample, remains future work.

Pillar C (confirmed, with caveats). With L =GPT-5-nano ($a_L = 0.748$ via 5-sample self-consistency), H =Opus 4.8 ($a_H = 0.921$), and a self-consistency verifier (AUC = 0.899; near-binary

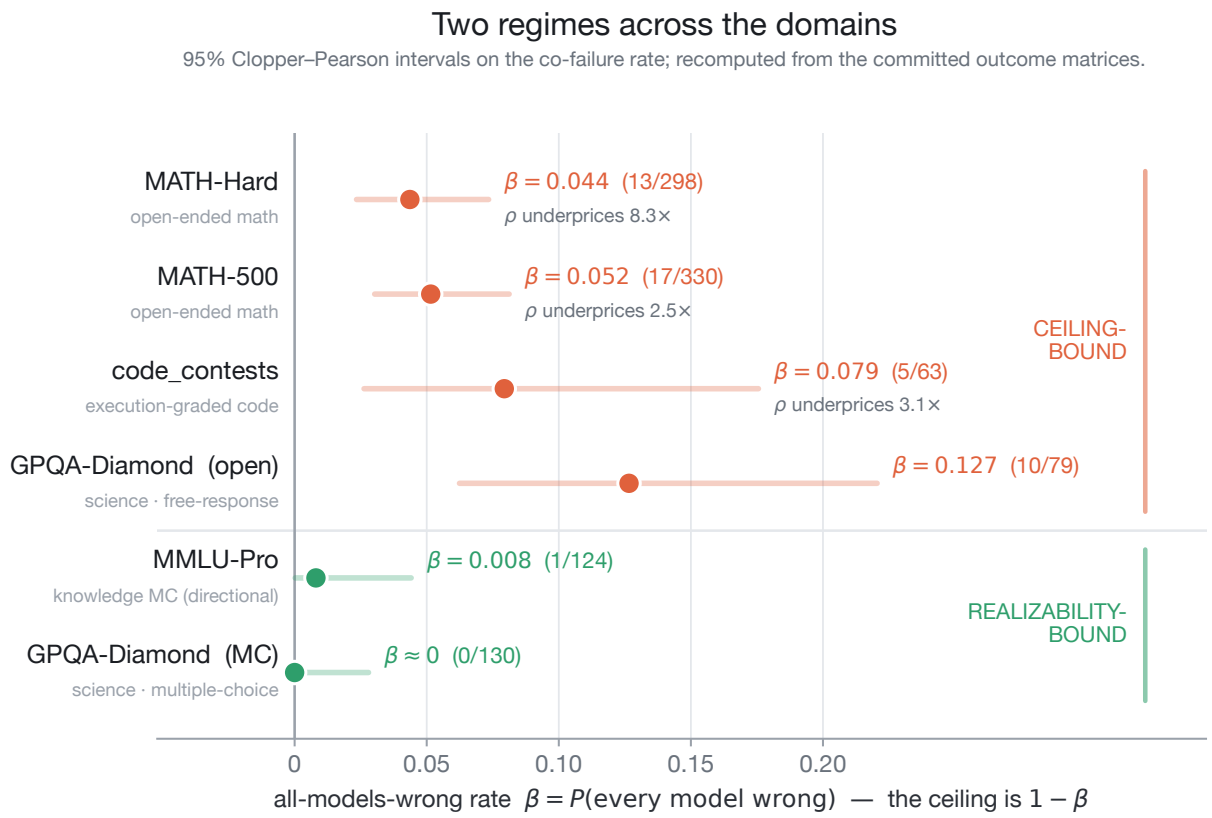

Figure 4: **Two regimes across the domains.** All-models-wrong rate β per domain with 95% Clopper–Pearson intervals, recomputed from the committed outcome matrices. *Ceiling-bound* domains (open-ended math and code, free-response GPQA) carry a co-failure tail $\beta > 0$ that caps every selection policy at $1 - \beta$ and that pairwise ρ underprices by 2.5–8.3× (tetrachoric); *realizability-bound* domains (multiple-choice GPQA, MMLU-Pro) have $\beta \approx 0$, so the oracle gain is pure resolvable disagreement a router could in principle capture. The same GPQA-Diamond content sits in *both* regimes depending only on answer format (Fig. 3).

at k_c=5), the collapse identity (2) is confirmed directly (Fig. 6): averaged over 20 noise-injection seeds, the cascade's advantage over random mixing falls monotonically toward zero (0.121→0.012, seed std ≤ 0.005) as the verifier degrades, while the injected-noise verifier AUC falls 0.899→0.510 (non-monotone at low injection, as expected; we therefore plot the advantage against AUC, not against injection level). The volume ceiling is $1 - a_L/a_H = 0.188$. At the unconstrained optimum the cascade collapses to the L corner (it merely matches L on dollars-per-correct); its dominance over H-only holds in the quality-constrained band $q \in (a_L, a_H]$. A control replacing H with Mistral-Large removes dominance, but because it lowers both a_H and H's price it confounds the tail-edge effect with price, so we report it as suggestive (cf. 14). We replace the in-sample optimism with a 5-fold **held-out** evaluation: choosing the deferral threshold on the train folds, the cascade still beats random-mixing-at-matched-budget by +0.114 accuracy on the held-out fold (cross-fold sd 0.010), with held-out confidence AUC 0.899—the dominance is not an in-sample artifact (`cascade_heldout.py`). The one remaining cascade gap is the optimal two-model deferral upper bound of Jitkrittum et al. [14], which requires H's confidence (unlogged in our matrices); we flag it rather than claim it.

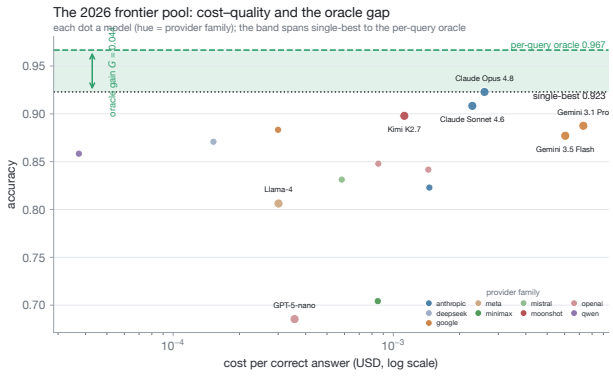

Figure 5: Pillar A cost–quality frontier (regraded): the per-query oracle (green) sits just above single-best (black); cheap models populate the frontier.

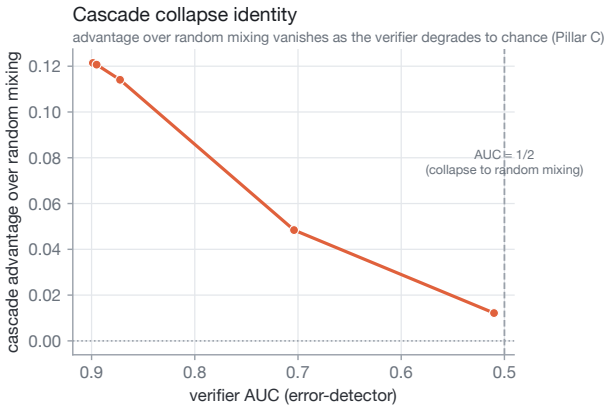

Figure 6: Pillar C: the cascade's advantage over random-mixing-at-matched-budget collapses to zero as the verifier's AUC falls to $\frac{1}{2}$, confirming Eq. (2).

Optionality under churn (secondary). A separate observational study on the 2024–2026 release timeline (frontier cost per unit capability fell $\approx 14\times$) is not load-bearing for the β/ρ result and is deferred to App. E.

6 Discussion

One allocation problem on two timescales. The results are not separate vignettes but one allocation problem viewed at two timescales. Within a release epoch, prices and the pool are fixed and the buyer solves the static allocation of App. A.1–A.3: a budget-priced assignment with value $V(B)$ and shadow price λ_B (Prop. 4), capped by the realizability ceiling (Prop. 1). Across epochs, frontier releases arrive and the buyer holds an option on the next pool; the option value of breadth (App. E) is the continuation value attached to that stage problem. We do not claim a single solved Bellman system—the churn primitives (ν, Γ, v) are not derived from the stage problem's—so this is a decomposition that organizes the results, explaining why the static claims hold within an epoch and the option value across epochs; the continuation algebra itself is standard renewal/real-options machinery, which we concede. With that: routing value is a first-moment selection effect that scales with dispersion rather than capability (App. A.1); fusion value is a second-moment effect bounded by systematic error and, empirically, realized only under accuracy-matched combination (App. A.2); cascade value is a decision-theoretic effect equal to the integrated AUC lift of the verifier (App. A.3). Because all three shrink as the frontier converges and errors correlate (Cor. 2), the value of a routing layer tracks market churn and heterogeneity, not the absolute capability of the best model (App. E). The empirical signature is already visible: on the 2026 frontier, oracle gains are small and *naive* fusion is a net liability precisely because today's best models agree—yet once member quality is matched, lower error correlation still buys a significant gain, so the lever is failure-mode heterogeneity, not count.

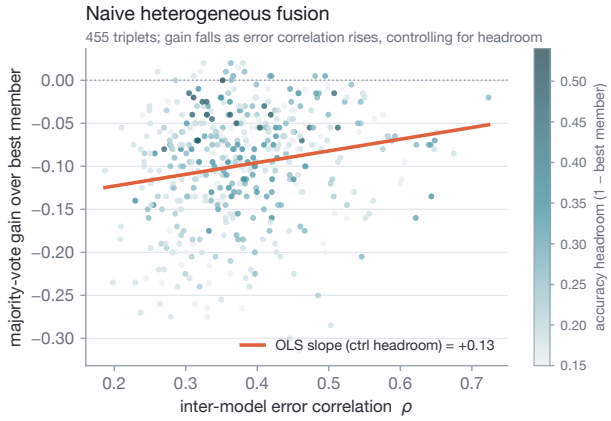

Figure 7: Pillar B, 455 triplets: unweighted majority-vote gain over the best member vs. ρ; gains are mostly negative (robust), the positive slope not significant under model-clustering.

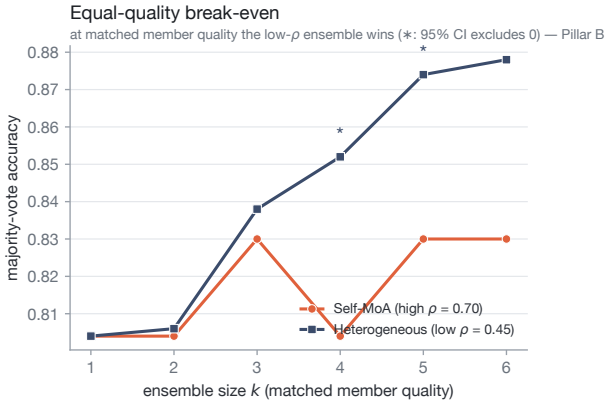

Figure 8: Equal-quality break-even (S=9, matched 6-model band): low-ρ heterogeneous fusion (ρ=0.42) vs. high-ρ Self-MoA (ρ=0.80), both on distinct draws so Self-MoA scales without recycling samples. Under the pre-registered distinct-draw aggregation the heterogeneous ensemble beats Self-MoA from the information-fair k=3 onward (∗: query-bootstrap 95% CI excludes zero), supporting the diversification mechanism at matched quality in this regime.

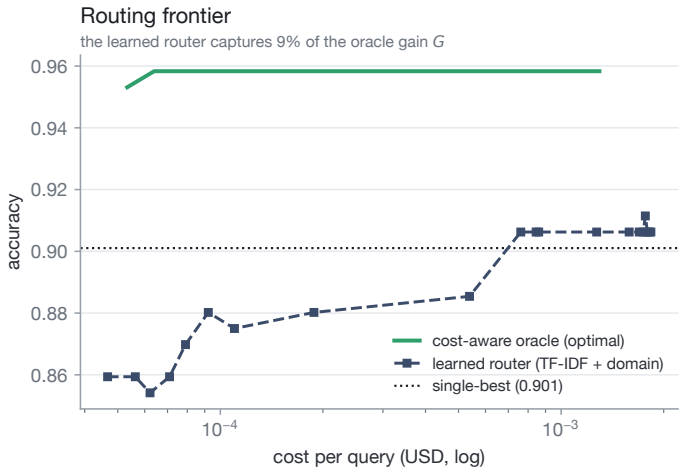

Figure 9: Pillar A, realizable routing: a held-out learned router (TF-IDF+domain) vs. the cost-aware oracle (optimal) frontier and single-best, on the multi-domain mix. The learned router barely exceeds single-best (captures ∼9% of G, CI spans zero) and lies well below the optimal frontier; realizable routing gain is near zero on the 2026 frontier.

7 Limitations

Programmatic grading covers verifiable tasks only and is sensitive to answer-extraction heuristics that can mildly penalize verbose models; open-ended quality would reintroduce judge bias. Saturated benchmarks inflate ρ, mitigated but not removed by the hard regime. The static-price assumptions of App. A.1–A.3 are in tension with the churn of App. E; those claims are restricted to within-release-epoch validity. The equal-quality assumption is empirically load-bearing (naive heterogeneous voting hurts); the diversification mechanism is supported at matched quality (§5). Its sensitivity λ is now derived as a decision-rule Jacobian (Prop. 7), but the empirical fit of λ across ρ levels and out-of-sample prediction of k^* remain open, and the matched-quality test rests on one provider-matched band; an alternative aggregation pipeline gives a smaller, non-significant gain. Inference on the unconditional ρ-slope is inconclusive under model-clustering; G and the block-ρ gap are reported without seed replication. The churn study (App. E) is stylized and observational. We instantiate a deployable learned router and the cost-aware oracle (optimal routing) frontier from logged outcomes (§5, Fig. 9): the router captures $\sim$0 of G, answering the routing question in the negative on this pool. What remains is the optimal cascade-routing policy of Dekoninck et al. [3] as a cascade-side upper bound (the cascade result is still measured against a naive confidence cascade, not the cascade optimum). Relatedly, our cascade verifier scores only the cheap model (L), whereas the optimal deferral rule conditions on both models [14], so our single-model AUC verifier is a practical, provably-dominated choice. The cascade threshold is now validated out-of-sample by a 5-fold held-out evaluation (App. A.3); what remains on the cascade side is only the optimal two-model deferral upper bound above. **External validity** is now supported across three structurally disjoint open-ended domains—two math families and execution-graded code, where the co-failure signature (populated β, the Pearson trap, a full-Σ residual, and a tetrachoric underpricing CI that excludes 1) replicates and inverts on multiple-choice (App. F)—though the code *magnitude* still rests on k=5 events under a strict-but-not-official judge on an 18-model pool, so the point ratio carries real uncertainty. A precise code magnitude (official hidden-test judge at scale on the full pool, with logged prompts for an in-domain router) is the one remaining sharpening. The remaining items for a top-venue submission are therefore: a tight-ratio code replication; ≥ 3 seeds and held-out model selection on the matched-quality result, and its extension across multiple ρ levels to fit λ and predict k^* out of sample; the optimal cascade-routing baseline above; and a price-controlled tail-edge experiment. (The option value of breadth is already measured, not stylized, on a real generational timeline, §5.) The authors' own orchestration workflow ran tiered but single-provider with unmetered cost/latency and is reported only as a transparency note, never as evidence.

8 Conclusion

The field decides whether to orchestrate by reading one number, pairwise error correlation, and that number is blind to the joint failures that set the ceiling. Treating orchestration as allocation over a correlated, priced, churning pool replaces it with the right object (β, the co-failure tail), a \$0 certificate on the achievable gain, and an economics of when the gain is reachable: a calibrated diversification limit, a cascade calibration boundary, and an option value of breadth that loads on churn. Empirically, headroom is foreclosed two ways: a binding co-failure ceiling on open-ended tasks, where ρ underprices the tail, and a slack ceiling on others, where a large oracle gain is resolvable disagreement no deployable router yet captures. Both say the same thing for practice—on our pool and verifiable tasks, and absent a strong query-level routing signal: *on open-ended tasks* the best models increasingly fail alike, so the lever is failure-mode dispersion and market churn, not

peak capability or model count. Pairwise correlation will not tell a buyer which lever they hold. Whether this holds on open-ended generative tasks beyond our verifiable benchmarks is open.

References

[1] Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

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

[3] Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. *arXiv preprint arXiv:2410.10347*, 2024.

[4] Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations (ICLR)*, 2024.

[5] Avinash K. Dixit and Robert S. Pindyck. *Investment Under Uncertainty*. Princeton University Press, 1994.

[6] Catherine Donnelly and Paul Embrechts. The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *ASTIN Bulletin*, 40(1):1–33, 2010.

[7] Paul Embrechts, Alexander J. McNeil, and Daniel Straumann. Correlation and dependence in risk management: properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, 2002.

[8] Mehmet Hamza Erol, Batu El, Mirac Suzgun, et al. Cost-of-pass: An economic framework for evaluating language models. *arXiv preprint arXiv:2504.13359*, 2025.

[9] Roberto Fontana and Patrizia Semeraro. Representation of multivariate bernoulli distributions with a given set of specified moments. *Journal of Multivariate Analysis*, 168:290–303, 2018.

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

[11] Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

[12] Qitian Jason Hu, Jacob Bieker, et al. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024.

[13] 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. arXiv:2306.02561.

[14] Wittawat Jitkrittum, Neha Gupta, Aditya Krishna Menon, Harikrishna Narasimhan, Ankit Singh Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2307.02764.

[15] Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2506.07962.

[16] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[17] Paul Klemperer. Competition when consumers have switching costs. *The Review of Economic Studies*, 62(4):515–539, 1995.

[18] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 7, pages 231–238, 1995.

[19] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[20] 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.

[21] Ludmila I. Kuncheva, Christopher J. Whitaker, Catherine A. Shipp, and Robert P. W. Duin. Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, 6(1): 22–31, 2003.

[22] David X. Li. On default correlation: A copula function approach. *Journal of Fixed Income*, 9 (4):43–54, 2000.

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

[24] Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025.

[25] Aman Madaan, Pranjal Aggarwal, et al. AutoMix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2023.

[26] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[27] Albert W. Marshall and Ingram Olkin. A multivariate exponential distribution. *Journal of the American Statistical Association*, 62(317):30–44, 1967.

[28] Robert McDonald and Daniel Siegel. The value of waiting to invest. *The Quarterly Journal of Economics*, 101(4):707–727, 1986.

[29] Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2):125–144, 1976.

[30] Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

[31] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.18665.

[32] Karl Pearson. Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society A*, 195:1–47, 1900.

[33] Masaaki Sibuya. Bivariate extreme statistics, i. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210, 1959.

[34] Meir Statman. How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, 22(3):353–363, 1987. doi: 10.2307/2330969.

[35] Yigit Turkmen, Baturalp Buyukates, and Melih Bastopcu. Don't always pick the highest-performing model: An information-theoretic view of LLM ensemble selection. *arXiv preprint arXiv:2602.08003*, 2026.

[36] Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks (ICNN)*, 1996.

[37] 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.

[38] Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24(359):1–49, 2023. arXiv:2301.03962.

A Economic scaffolding: routing, diversification, and cascades

This appendix gives the economic scaffolding deferred from the main text (§3); its tools are standard, and the empirical claims do not rest on it.

A.1 Routing as priced assignment

The envelope and dispersion-scaling material below is background, claimed by no one: the oracle envelope and the value-of-information gap are textbook [11], and routing optimality is proven by Dekoninck et al. [3]. Proposition 4 makes "orchestration is allocation" literal: under a dollar budget, routing is a priced assignment with an explicit shadow price.

Lemma 1 (Envelope and dominance condition; background)**.** *For every routing policy π, $V(\pi) \le V^o := \mathbb{E}_t \max_i q_i(t)$, attained by the oracle $\pi^o(t) = \arg\max_i q_i(t)$. Let the oracle gain be $G := V^o - \max_i \bar{q}_i \ge 0$. Then $G > 0$ iff no model is uniformly best across types D-almost everywhere; equivalently, the quality matrix $Q = [q_i(t)]$ is not row-dominated under D.*

Proposition 4 (Budget-constrained routing: shadow price and bang-per-buck rule)**.** *Fix a dollar budget B and let $V(B) = \max_\pi\{V(\pi) : K(\pi) \le B\}$ over routing policies. This is a linear program whose dual collapses to a* single scalar *price $\lambda_B \ge 0$ on the budget,*

$$V(B) = \min_{\lambda \ge 0} \Big\{ \lambda B + \mathbb{E}_t \max_i \big[q_i(t) - \lambda c_i \big] \Big\},$$

because the per-type simplex constraints are absorbed pointwise. Every optimal policy routes type t to $\arg\max_i [q_i(t) - \lambda_B c_i]$*, a per-type* bang-per-buck *rule, mixing only where the budget binds. The value $V(B)$ is nondecreasing, concave, and piecewise linear, with $V'(B) = \lambda_B$ wherever differentiable; once B exceeds the oracle's cost, $\lambda_B = 0$ and the rule is the unconstrained per-query oracle.*

This specializes standard linear-programming duality [1] to inference routing. The step is small but earns the allocation framing: orchestration under a budget is a priced assignment, and λ_B, the *shadow price of the inference dollar*, is the price a budget-aware buyer trades quality against (the same dollars-per-correct currency the cost-aware rules of §A.2–A.3 optimize). With several resources (e.g. dollars and latency) λ_B becomes a vector and the score is $q_i(t) - \lambda_B \cdot c_i$. We verified the dual price, the bang-per-buck characterization, and concavity numerically.

Dispersion scaling (heuristic). If, purely as an illustration, the cell qualities $q_i(t)$ were i.i.d. $\mathcal{N}(\mu, s^2)$ across the $m \times |T|$ cells with $|T|$ large relative to $\ln m$, then $V^o \approx \mu + s\sqrt{2 \ln m}$ while the best row mean concentrates at μ, giving $G \sim s\sqrt{2 \ln m}$: linear in cross-model within-type dispersion s, only logarithmic in pool size m. We flag this as heuristic: real errors are block-correlated (§5), which lowers the rate, but the qualitative message (value from *how* models differ, not how many) is what the empirics test. A learned router with expected routing regret R attains $V^o - R$ and beats single-best iff $R < G$. As G is monotone in partition fineness, we report a partition-free per-query oracle as the true upper bound.

A.2 The cost-aware diversification limit

Let model errors satisfy $\mathrm{Var}(e_i) = \sigma^2$ and $\mathrm{Corr}(e_i, e_j) = \rho$, with systematic variance $\tau^2 = \rho\sigma^2$ and idiosyncratic variance $(1-\rho)\sigma^2$.

Proposition 5 (Variance floor (classical))**.** *Equal-weight fusion of k equicorrelated models has mean-squared error $V(k) = \sigma^2(\rho + \frac{1-\rho}{k}) \to \rho\sigma^2$. Under a symmetric single-factor probit with per-model error rate α, unweighted majority vote has infinite-ensemble error floor $\Phi(-\Phi^{-1}(1-\alpha)/\sqrt{\rho})$, with limits 0 ($\rho \to 0$) and α ($\rho \to 1$).*

The continuous floor is the equicorrelated portfolio-variance limit [26, 34] and the bias–variance–covariance/diversity decomposition [18, 36, 38]; the copula floor is Turkmen et al. [35]. We claim none of these.

Proposition 6 (Cost-aware break-even)**.** *Let λ be the local sensitivity of expected correctness to fused-estimate variance for the operative decision rule, $\lambda := -\partial(\text{expected correct})/\partial V$, evaluated near the operating point (a derived Jacobian for a threshold rule, not a risk-aversion parameter; under strict risk neutrality λ is the local curvature through which second moments enter the first-moment objective). With per-model cost c, the largest index whose addition pays for itself is*

$$k^*(\rho, c) = \tfrac{1}{2}\Big(-1 + \sqrt{1 + 4\lambda\sigma^2(1-\rho)/c}\Big), \qquad k^* \sim \sqrt{\lambda\sigma^2(1-\rho)/c}. \tag{1}$$

Hence $\partial k^*/\partial\rho < 0$, $\partial k^*/\partial c < 0$, and $\rho \to 1 \Rightarrow k^* \to 0$. The optimal ensemble cardinality is the nearest feasible integer to $k^*{+}1$.

Proposition 7 (Calibration of λ and a sign condition)**.** *Under the symmetric single-factor probit of Prop. 5 (latent $S_i = m + \sqrt{\rho}\,U + \sqrt{1-\rho}\,\xi_i$, $m = \Phi^{-1}(1-\alpha)$, correct iff $S_i > 0$), equal-weight fusion has fused score $Z_k \sim \mathcal{N}(m, V(k))$, and the risk-neutral majority/mean-vote rule has expected correctness $P(V) = \Phi(m/\sqrt{V})$ in the single sufficient statistic $V = V(k)$. The sensitivity λ in (1) is then not a free parameter but the explicit Jacobian*

$$\lambda(V) = -\frac{\partial P}{\partial V} = \frac{m}{2V^{3/2}}\,\varphi\Big(\frac{m}{\sqrt{V}}\Big), \qquad m = \Phi^{-1}(1-\alpha),$$

evaluated at the operating point $V = V(k)$. Its sign is $\operatorname{sgn}\lambda = \operatorname{sgn}(\frac{1}{2} - \alpha)$: when each member beats chance ($\alpha < \frac{1}{2}$), $\lambda > 0$ and P strictly decreases in V, so reducing V by adding members strictly

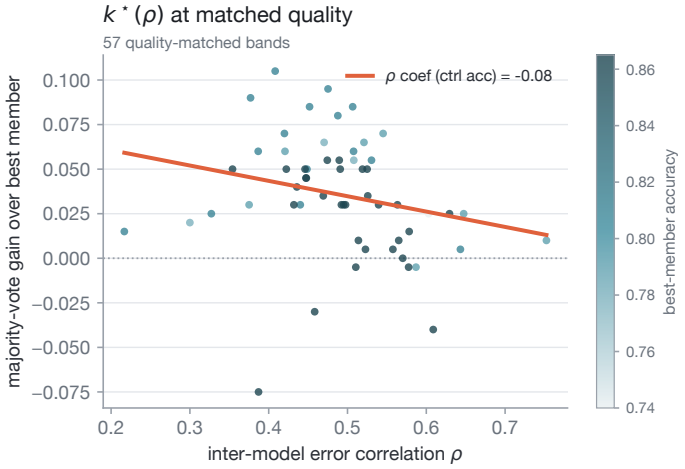

Figure 10: Pillar B, $k^*(\rho)$ at matched quality: across 57 sub-bands of the matched 6-model pool (MMLU-Pro), majority-vote gain over the best member vs. inter-model ρ, controlling for member accuracy. The slope is negative (diversification-limit direction; CI spans zero), reversing the positive slope in the unmatched pool (Fig. 7).

raises expected correctness and k^* is well defined; at $\alpha > \frac{1}{2}$, $\lambda < 0$ and the diversification framing fails (adding members drives the majority more *wrong, and* (1) *returns no real k^*).*

This pins down the economic bridge the cost-aware rule rests on and delimits where it is valid. λ is an operating-point Jacobian (it depends on V, hence weakly on k); we evaluate it at the marginal index, where the local-linearization error $O(V(k)^{-2})$ is negligible. A finite-sample estimator plugs the tetrachoric $\hat{\rho}$ and per-model $\hat{\alpha}$ into $\lambda(V(k))$; the Pearson coefficient of 0/1 indicators biases ρ (hence λ) downward. We verified $\lambda = -\partial P/\partial V$ and the sign condition numerically.

Query-conditional ρ and block covariance. Classical ρ is a global constant; making $\rho(t)$ type-conditional makes k^* vary within a single workload and resolves the diversity-unreliability result [20] by conditioning. Because errors are block-structured and rise with accuracy [15], the headline object is the block covariance Σ: equal weights are no longer optimal (the minimum-variance weights are $w \propto \Sigma^{-1}\mathbf{1}$), and the floor is the undiversifiable common-factor component of Σ. The equicorrelation form is the special case. We caution that the portfolio analogy is a first-moment-plus-local-curvature analogy, not a claim of risk aversion.

A.3 Cascade calibration economics

A cheap model L (price c_L, accuracy a_L) answers first with confidence s; the query escalates to a strong model H when $s < \tau$. Let $\beta = \Pr[s < \tau]$ be the escalation budget, $w(\beta) = \Pr[Y_L = 0 \mid$ escalated β-tail$]$ the tail error rate, and $a_H(\tau)$ the strong model's accuracy on the deferred tail. Then $C(\beta) = c_L + \beta c_H$ and $Q(\beta) = a_L + \beta[a_H(\tau) - 1 + w(\beta)]$.

Proposition 8 (Collapse, ceiling, dominance)**.** *Against random mixing (escalate a random β-fraction; tail error $1 - a_L$, tail accuracy a_H) at equal budget,*

$$Q(\beta) - Q_{\text{mix}}(\beta) = \beta\big[w(\beta) - (1 - a_L)\big] + \beta\big[a_H(\tau) - a_H\big]. \tag{2}$$

If $a_H(\tau) = a_H$ (no tail adverse selection), the integrated advantage $\int_0^1 [w(\beta) - (1 - a_L)]\, d\beta \geq 0$ iff the verifier AUC $\geq \frac{1}{2}$*, with equality iff* AUC $= \frac{1}{2}$ *(the cascade is then random mixing); the*

pointwise inequality additionally requires w non-increasing (a threshold-rational verifier). Under perfect calibration the escalation ratio $\beta_{\text{cas}}/\beta_{\text{mix}} \to 1 - a_L/a_H$: relative to random mixing, a calibrated cascade needs only a fraction $1 - a_L/a_H$ as many strong-model calls at a fixed quality floor, independent of price. Within a one-parameter family of verifiers ordered by AUC, $\beta_{\text{cas}}(q)$ is decreasing in AUC, so a critical AUC^* *exists below which (given $c_L/c_H < q/a_H$) no threshold meets floor $q \in (a_L, a_H]$ at lower dollars-per-correct than the strong model alone.*

Corollary 1 (Calibration-and-edge: necessary condition)**.** *Cascade dominance over H-only requires both* $\text{AUC}(s) > \text{AUC}^*$ *and a positive conditional edge $a_H(\tau) > a_L(\tau)$ on the deferred tail. If H is adversely weak where L defers, calibration alone is insufficient (consistent with 14).*

B Proofs

Proof of Lemma 1. For any π and t, $\sum_i \pi(i \mid t) q_i(t) \le \max_i q_i(t)$; take $\mathbb{E}_t$ to get $V(\pi) \le V^o$, attained by π^o. Write $G = \mathbb{E}_t[\max_i q_i(t) - q_{i^*}(t)]$ with $i^* = \arg\max_i \bar{q}_i$; the integrand is nonnegative. If model j is uniformly best D-a.e. then $\bar{q}_j = V^o$, so $G = 0$; conversely, if no model is uniformly best, the set where i^* is suboptimal has positive measure and the integrand is strictly positive there, so $G > 0$. The dispersion-scaling remark uses the extreme-value asymptotic $\mathbb{E}[\max \text{ of } m \text{ i.i.d. } \mathcal{N}(0,1)] \sim \sqrt{2 \ln m}$ and concentration of the best row mean at μ when $|T| \gg \ln m$; it is heuristic under correlation. □

Proof of Prop. 5. With $e_i = b + \varepsilon_i$, $\text{Var}(b) = \rho\sigma^2$, $\text{Var}(\varepsilon_i) = (1-\rho)\sigma^2$ uncorrelated, the equal-weight average has $\text{Var}(\bar{e}) = \rho\sigma^2 + (1-\rho)\sigma^2/k \to \rho\sigma^2$. For binary correctness under $S_i = m + \sqrt{\rho}\,U + \sqrt{1-\rho}\,\xi_i$ with $m = \Phi^{-1}(1-\alpha)$, conditioning on U gives correctness probability $\Phi((m + \sqrt{\rho}U)/\sqrt{1-\rho})$; by the law of large numbers the majority is correct iff $U > -m/\sqrt{\rho}$, so the floor is $\Phi(-m/\sqrt{\rho})$; the limits are immediate. □

Proof of Prop. 6. The marginal variance reduction of the $(k{+}1)$-th model is $\Delta V(k) = \sigma^2(1-\rho)/[k(k+1)]$. Converting to expected correctness through the local sensitivity λ, adding a model is worthwhile iff $\lambda\,\Delta V(k) \ge c$, i.e. $k(k+1) \le R$ with $R = \lambda\sigma^2(1-\rho)/c$; solving $k(k+1) = R$ gives (1). The comparative statics follow by inspection; $R \to 0$ as $\rho \to 1$. k^* is the last paying index, so the optimal cardinality is $\approx k^* + 1$. □

Proof of Prop. 8 and Cor. 1. Escalated mass β pays c_H on top of c_L, giving $C(\beta) = c_L + \beta c_H$; of L's correct mass a_L, the part in the escalated tail $\beta(1 - w(\beta))$ is replaced by H scoring $a_H(\tau)$ there, giving $Q(\beta) = a_L + \beta[a_H(\tau) - 1 + w(\beta)]$. Random mixing escalates a random β-fraction with tail error $1 - a_L$ and tail accuracy a_H, giving $Q_{\text{mix}} = a_L + \beta(a_H - a_L)$; subtracting yields (2). With $a_H(\tau) = a_H$, the residual is $\beta[w(\beta) - (1 - a_L)]$, whose integral is nonnegative iff s ranks errors above successes on average ($\text{AUC} \ge \frac{1}{2}$), with equality iff $\text{AUC} = \frac{1}{2}$; the pointwise sign needs w non-increasing. Under perfect calibration the escalated tail is exactly L's errors until exhausted, so $Q(\beta) = a_L + \beta a_H$ for $\beta \le 1 - a_L$, whence $\beta_{\text{cas}}(q) = (q - a_L)/a_H$ and $\beta_{\text{cas}}/\beta_{\text{mix}} \to 1 - a_L/a_H$. Beating H-only (c_H/a_H per correct) at floor q requires $c_L + \beta_{\text{cas}}(q) c_H \le (q/a_H) c_H$, feasible only if $c_L/c_H < q/a_H$; within a one-parameter AUC-ordered family β_{cas} decreases in AUC, so AUC^* exists. If $a_H(\tau) \le a_L(\tau)$ on the deferred tail, escalation cannot raise quality there, so no τ achieves q: dominance requires both AUC above threshold and a positive tail edge. We prove necessity; sufficiency additionally needs reachability of q and the cost ordering. □

Proof of Prop. 9. The captured value per arrival is $v\,\mathbb{E}[\max(\Gamma, 0)] = v\,\text{Eg}$ by the Gaussian truncation identity. A captured lead is held for an $\text{Exp}(\nu)$ time and discounted at r (factor $1/(r+\nu)$); the

Poisson arrival stream contributes a present-value multiplier ν/r, giving $V_R = v\,\mathrm{Eg}\cdot\nu/(r+\nu)\cdot(1/r)$. Then $\partial\mathrm{Eg}/\partial\eta = \varphi(\mu/\eta) > 0$ and $\partial[\nu/(r+\nu)]/\partial\nu > 0$; adding a separable m(level) leaves both partials unchanged. □

Proof of Prop. 4. (P_B) is linear in $\{\pi(i \mid t)\}$. Dualize only the budget row with multiplier $\lambda \geq 0$, keeping the per-type simplices as the domain; the Lagrangian separates across t into inner problems $\max_{\pi(\cdot|t)\in\Delta}\sum_i \pi(i \mid t)(q_i(t) - \lambda c_i) = \max_i(q_i(t) - \lambda c_i) =: g_t(\lambda)$. Weak duality gives $V(B) \leq \lambda B + \mathbb{E}_t g_t(\lambda)$ for every $\lambda \geq 0$; LP strong duality (Slater holds for $B > \underline{B}$) gives equality at some λ_B. Complementary slackness places every optimal $\pi^\star(\cdot \mid t)$ on $\arg\max_i(q_i(t) - \lambda_B c_i)$, mixing only to exhaust the budget. As a minimum of affine functions of B, V is concave; it is nondecreasing (enlarging B relaxes a constraint) and piecewise linear (parametric LP). By the envelope theorem $V'(B) = \lambda_B$ where differentiable, with subdifferential $[\lambda_B^-, \lambda_B^+]$ at the finitely many kinks; once $B \geq K(\pi^o)$ the budget is slack, $\lambda_B = 0$, and the rule is the per-query oracle. □

Proof of Prop. 7. Under the single-factor probit, $Z_k = \frac{1}{k}\sum_i S_i = m + \sqrt{\rho}\,U + \frac{\sqrt{1-\rho}}{k}\sum_i \xi_i \sim \mathcal{N}(m, V(k))$, $V(k) = \rho + (1-\rho)/k$ (σ^2=1). The mean-vote rule is correct iff $Z_k > 0$, so $P(V) = \Pr[Z_k > 0] = \Phi(m/\sqrt{V})$. Then $\partial P/\partial V = \varphi(m/\sqrt{V})\cdot m\cdot(-\frac{1}{2})V^{-3/2}$, so $\lambda := -\partial P/\partial V = (m/2V^{3/2})\varphi(m/\sqrt{V})$. As $\varphi > 0, V > 0$, $\operatorname{sgn}\lambda = \operatorname{sgn} m = \operatorname{sgn}(\Phi^{-1}(1-\alpha)) = \operatorname{sgn}(\frac{1}{2} - \alpha)$. Substituting λ into $\lambda\,\Delta V(k) \geq c$ with $\Delta V(k) = (1-\rho)/[k(k+1)]$ reproduces (1). □

Proof of Prop. 1. (i) On the all-wrong event every member answer differs from the label; a selection policy outputs one member answer, hence is wrong there, so $\mathrm{Acc}(\pi) \leq \Pr[\text{not all wrong}] = 1 - \beta$, attained by the per-query oracle. (ii) $V^o = \mathbb{E}_t\mathbf{1}\{\exists i : Y_i{=}1\} = 1 - \beta$ and $a_{\mathrm{sb}} = \Pr[Y_{i^\star}{=}1]$, so $G = V^o - a_{\mathrm{sb}} = \Pr[Y_{i^\star}{=}0] - \beta = \Pr[\text{single-best wrong}] - \beta$, the non-unanimous single-best-wrong mass. (iii) $K = \sum_j \mathbf{1}\{\text{all wrong on } t_j\} \sim \mathrm{Binomial}(n, \beta)$ since the t_j are i.i.d. and "all wrong" is a fixed event; the Clopper–Pearson lower limit obeys $\Pr[\beta \geq \beta_{\mathrm{lo}}] \geq 1 - \delta$, so with probability $\geq 1 - \delta$, $(1-\beta) \leq (1-\beta_{\mathrm{lo}})$ and every selection policy obeys $\mathrm{Acc} - a_{\mathrm{sb}} \leq (1-\beta) - a_{\mathrm{sb}} \leq (1-\beta_{\mathrm{lo}}) - a_{\mathrm{sb}}$. A union bound replaces a_{sb} by an upper confidence bound at level δ', with total error $\leq \delta + \delta'$. □

Proof of Prop. 2. $\beta(m) = \pi + (1-\pi)\alpha_0^m$ is strictly decreasing to π since $0 < \alpha_0 < 1$. Calibrate the single-factor Gaussian copula to reproduce the marginal α and the pairwise co-error $\pi + (1-\pi)\alpha_0^2$, fixing a latent correlation $\bar\rho < 1$; then $\beta_{\mathrm{sf}}(2) = \pi + (1-\pi)\alpha_0^2 = \beta(2)$, so the ratio is 1 at $m = 2$. For $m \geq 2$, conditioning on the common factor Z, $\beta_{\mathrm{sf}}(m) = \int \varphi(z)\,\Phi\big((t - \sqrt{\bar\rho}\,z)/\sqrt{1-\bar\rho}\big)^m\,dz$ with $t = \Phi^{-1}(\alpha)$. For each fixed z the conditional error probability $\Phi(\cdot) \in (0,1)$ since $\bar\rho < 1$, so the integrand $\to 0$ pointwise; dominated convergence gives $\beta_{\mathrm{sf}}(m) \to 0$ (this is the zero lower-tail-dependence of the Gaussian copula, 7). Since $\beta(m) \to \pi > 0$ and $\beta_{\mathrm{sf}}(m) \to 0$, the ratio $\beta(m)/\beta_{\mathrm{sf}}(m) \to \infty$. For monotonicity, compare successive decay rates: $\beta(m{+}1)/\beta(m) \to 1$ because $\beta(m) \to \pi > 0$, whereas $\beta_{\mathrm{sf}}(m{+}1)/\beta_{\mathrm{sf}}(m)$ tends to the conditional error probability at the dominating factor value, which is < 1 for $\bar\rho < 1$; hence the ratio's successive increments are eventually positive and the underpricing ratio is eventually strictly increasing (numerically, strict from m=2 in all settings we tested). □

Proof of Prop. 3. Take exchangeable error triples on $\{0,1\}^3$ and let q_j be the probability of exactly j errors ($\sum_j q_j = 1$). Every pair has the same 2×2 table, determined by the marginal error rate $p_1 = \frac{1}{3}\sum_j j\,q_j$ and the pairwise co-error $p_2 = \Pr[X_i{=}X_j{=}1] = \frac{1}{3}(q_2 + 3q_3)$; the triple-failure cell $\beta = q_3$ is left free in the resulting one-parameter Fréchet class. Concretely, fix $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$: both $(q_0, q_1, q_2, q_3) = (\frac{1}{4}, 0, \frac{3}{4}, 0)$ and $(0, \frac{3}{4}, 0, \frac{1}{4})$ reproduce *every* one- and two-dimensional marginal—hence every pairwise Pearson and tetrachoric correlation, which are functions of that common bivariate table—yet have $\beta = 0$ and $\beta = \frac{1}{4}$. So β is not a function of the pairwise law. Padding with independent coordinates extends the example to any $m \geq 3$, and the common-shock mixture

($\beta_\infty = \pi > 0$) versus a matched-$\bar{\rho}$ Gaussian copula ($\beta_\infty = 0$ by zero lower tail dependence) realize the two extremes as $m \to \infty$, both consistent with the same pairwise $\bar{\rho}$. □

C Reproducibility

Data and code availability. We release the full per-cell outcome matrices (every model's graded correctness, token counts, and metered cost on every query), the model registry with dated snapshots and live prices, the programmatic graders, and all analysis scripts, so that every number in the paper—including the β/ρ tables, the pool-size scaling, and the two-regime comparison—can be regenerated end to end. **Runnable certificate.** The realizability certificate (Prop. 1) ships as a standalone tool, `beta_certificate.py`: from a pool's logged outcomes—or merely the all-wrong count K/n and the single-best accuracy—it returns the Clopper–Pearson-certified \$0 lower bound on the maximum gain any router, vote, or cascade could deliver over single-best, requiring no pairwise ρ (Prop. 3). The full-Σ residual decomposition (`residual_decomp.py`) and the propagated ratio CIs (`bootstrap_ratio.py`) are likewise released. All runs are logged with exact model identifiers, dated snapshots, prices, seeds, and per-call costs. **Spend.** We meter each call against the OpenRouter account usage endpoint (an aggregator, not a per-provider source); per-run figures are sums of logged per-cell costs (cache-independent), itemized in the released `cost_registry.csv`. The core pillar experiments (A–D) cost ≈\$47 (matching the dated `/key` ground-truth total); the market-scale measurement adds ≈\$111 (the 53- and refreshed 67-model realizability runs, the truncation re-runs, and the GPQA second-domain run), and the third-domain experiments—execution-graded code (App. F) and the open-ended GPQA panel (App. G)—add ≈\$110, for ≈\$270 of reported-experiment cost. Total account usage including all exploratory and superseded iteration was ≈\$560 at submission (the live usage meter was transiently clobbered by concurrent runs late in the project; this figure is reconstructed from the last clean reading plus the logged third-domain run costs, and we flag it as approximate). We never present account-level usage as experiment cost. Programmatic graders and the registry are versioned; illustrative parameter values are labelled and never reported as measurements. **Provenance.** The market-scale matrix is reconstructed from the response cache (`reconstruct.py`) after a hard-first run was stopped at ≈41% (the saturated benchmarks, least informative for the tail, were not needed); the canonical graded, truncation-corrected matrix is `matrix_marketE2` (67-model frontier refresh). The empirical claims correspond to run identifiers `matrix_stageA2`, `matrix_hardA`, `matrix_churnD` (App. E churn study), `fusion_eqq2` (matched quality), and the market-scale reconstruction. The tail-co-failure numbers regenerate via `realizability.py --tag marketE2`. *Selection used for the headline:* the MATH-500 figures ($\beta = 0.052$, tetrachoric underpricing 2.5×, k=17, n=330; `realizability_tetrachoric.json`) are computed on the common-coverage subset of all 67 models (queries answered by every model, run id `matrix_marketE3`); GPQA on its 52-model complete-coverage subset (`matrix_marketE2`). `realizability.py`'s default ≥ 0.95-coverage filter surfaces a slightly smaller, also-valid model subset with the same order-of-magnitude underpricing; both are correct. The truncation control via `detruncate.py`; the cascade collapse (20 seeds) via `cascade.py`; the matched-quality partition-robustness (+0.027, positive in all 60 partitions) via `eqq_robustness.py`; and Pillar-B clustered inference via the model-jackknife in `rho_fusion_test.py`. All references were audited with scite Smart Citations: every entry resolves to a real indexed work, none carries a retraction, correction, or editorial concern, and no load-bearing citation has any contrasting smart-citations (the only contrasting citations in the bibliography are two on Statman [34], from the unrelated optimal-holding-count debate).

D The market-scale model pool

The 67-model, 21-family market pool used in §5 and the two-regime measurement, with live OpenRouter prices (snapshot 2026-06-19). All are chat/instruct models; pure reasoning/"thinking" variants are excluded so finite-token programmatic grading is clean. Prices are USD per million tokens.

model	family	tier	$/Mtok in	$/Mtok out
openai/gpt-5.5	openai	frontier	5	30
anthropic/claude-opus-4.8	anthropic	frontier	5	25
anthropic/claude-sonnet-4.6	anthropic	frontier	3	15
openai/gpt-5.4	openai	frontier	2.5	15
openai/gpt-5.2	openai	frontier	1.75	14
google/gemini-3.1-pro-preview	google	frontier	2	12
openai/gpt-5.1	openai	mid	1.25	10
google/gemini-3.5-flash	google	mid	1.5	9
ai21/jamba-large-1.7	ai21	mid	2	8
mistralai/mistral-medium-3-5	mistral	mid	1.5	7.5
qwen/qwen3.6-max-preview	qwen	mid	1.04	6.24
writer/palmyra-x5	writer	mid	0.6	6
anthropic/claude-haiku-4.5	anthropic	mid	1	5
z-ai/glm-5.2	zai	mid	1.2	4.1
qwen/qwen3-max	qwen	mid	0.78	3.9
qwen/qwen3.7-max	qwen	mid	1.25	3.75
moonshotai/kimi-k2.7-code	moonshot	mid	0.74	3.5
qwen/qwen3-coder-plus	qwen	mid	0.65	3.25
z-ai/glm-5.1	zai	mid	0.98	3.08
nousresearch/hermes-4-405b	nous	mid	1	3
x-ai/grok-4.3	xai	mid	1.25	2.5
moonshotai/kimi-k2-0905	moonshot	mid	0.6	2.5
qwen/qwen3.5-397b-a17b	qwen	mid	0.385	2.45
nvidia/nemotron-3-ultra-550b-a55b	nvidia	mid	0.5	2.2
qwen/qwen3.5-122b-a10b	qwen	mid	0.26	2.08
openai/gpt-5-mini	openai	mid	0.25	2
z-ai/glm-5	zai	mid	0.6	1.92
z-ai/glm-4.7	zai	mid	0.4	1.75
z-ai/glm-4.6	zai	mid	0.43	1.74
mistralai/mistral-large-2512	mistral	mid	0.5	1.5
google/gemini-3.1-flash-lite	google	mid	0.25	1.5
qwen/qwen3.7-plus	qwen	mid	0.32	1.28
minimax/minimax-m3	minimax	cheap	0.3	1.2
stepfun/step-3.7-flash	stepfun	cheap	0.2	1.15
qwen/qwen3-next-80b-a3b-instruct	qwen	cheap	0.09	1.1
minimax/minimax-m2	minimax	cheap	0.255	1
minimax/minimax-m2.7	minimax	cheap	0.25	1
deepseek/deepseek-v3.1-terminus	deepseek	cheap	0.27	0.95
minimax/minimax-m2.5	minimax	cheap	0.15	0.9
deepseek/deepseek-v4-pro	deepseek	cheap	0.435	0.87
xiaomi/mimo-v2.5-pro	xiaomi	cheap	0.435	0.87
deepseek/deepseek-chat-v3.1	deepseek	cheap	0.21	0.79
qwen/qwen-plus-2025-07-28	qwen	cheap	0.26	0.78
google/gemma-2-27b-it	google	cheap	0.65	0.65
meta-llama/llama-4-maverick	meta	cheap	0.15	0.6
upstage/solar-pro-3	upstage	cheap	0.15	0.6
nvidia/nemotron-3-super-120b-a12b	nvidia	cheap	0.09	0.45
nvidia/llama-3.3-nemotron-super-49b-v1.5	nvidia	cheap	0.4	0.4
meta-llama/llama-3.1-70b-instruct	meta	cheap	0.4	0.4
nousresearch/hermes-4-70b	nous	cheap	0.13	0.4
openai/gpt-5-nano	openai	cheap	0.05	0.4
microsoft/phi-4-mini-instruct	microsoft	cheap	0.08	0.35
deepseek/deepseek-v3.2	deepseek	cheap	0.2288	0.3432
meta-llama/llama-3.2-3b-instruct	meta	cheap	0.0509	0.335
meta-llama/llama-3.3-70b-instruct	meta	cheap	0.1	0.32
nvidia/nemotron-3-nano-30b-a3b	nvidia	cheap	0.05	0.2

model	family	tier	$/Mtok in	$/Mtok out
deepseek/deepseek-v4-flash	deepseek	cheap	0.09	0.18
openai/gpt-oss-120b	openai	cheap	0.039	0.18
openai/gpt-oss-20b	openai	cheap	0.029	0.14
google/gemma-3n-e4b-it	google	cheap	0.06	0.12
ibm-granite/granite-4.0-h-micro	ibm	cheap	0.017	0.112
qwen/qwen3-235b-a22b-2507	qwen	cheap	0.09	0.1
ibm-granite/granite-4.1-8b	ibm	cheap	0.05	0.1
mistralai/mistral-small-24b-instruct-2501	mistral	cheap	0.05	0.08
meta-llama/llama-3.1-8b-instruct	meta	cheap	0.02	0.03
mistralai/mistral-nemo	mistral	cheap	0.02	0.03
inclusionai/ling-2.6-flash	inclusionai	cheap	0.01	0.03

E Optionality under churn (secondary)

Deferred from the main thread because it is observational and not load-bearing for the β/ρ result. Frontier releases arrive as a Poisson process of intensity ν (the churn rate); each offers a relative capability gap $\Gamma = \mu + \eta Z$, $Z \sim \mathcal{N}(0,1)$. Broad access (ROUTE) captures $v \max(\Gamma, 0)$ per arrival, held until the next arrival and discounted at r; committing (self-host) earns premium δ but pays switch cost K.

Proposition 9 (Additive option value)**.** *With* $\mathrm{Eg} = \mu\Phi(\mu/\eta) + \eta\,\varphi(\mu/\eta)$ *(φ the standard normal density), the option value of broad access is* $V_R = \frac{v}{r} \cdot \frac{\nu}{r+\nu} \cdot \mathrm{Eg}(\mu, \eta)$*, and the build-vs-route threshold is* $\delta^* = rK + v\frac{\nu}{r+\nu}\mathrm{Eg}$*. If buyers additionally value absolute capability via a separable $m(\cdot)$ independent of (ν, η), total value is $m(level) + V_R$, so $\partial V/\partial\nu > 0$ and $\partial V/\partial\eta > 0$ regardless of $dm/dlevel$.*

Corollary 2 (Convergence collapse)**.** *If error correlation rises toward* 1 *as models improve (an extrapolation of Kim et al. [15], who document rising, not unit, correlation), the diversification gain shrinks; jointly with a falling common error rate, the orchestration value G and V_R contract. We verify the premise's sign, not the unit limit.*

The comparative statics are standard real-options results [5, 28, 29]; the separability of m is an assumption, not a finding, and the contribution is only the additive form under it. **Measured on a real timeline:** on the release sequence of an 18-model generational pool (Claude-3-Haiku Mar 2024 to Gemini-3.1-Pro Feb 2026), the best achievable dollars-per-correct dropped $\approx$14–15$\times$ (clean, since cost is metered not graded), echoing Erol et al. [8]; and the capability option value of broad access is regime-dependent—+0.33 accuracy by 2026 on hard MMLU-Pro versus +0.01 on saturated GSM8K. This is an observational single-path study, not a controlled comparative static; the dispersion→capture association is directionally positive but noisy, so the grading-independent cost-churn result is the robust part.

F The third domain: execution-graded competitive code

The binding external-validity question is whether co-failure is a math artifact or tracks open-ended generation. We test it on a *structurally independent* domain, competitive programming, where generation is open-ended but grading is programmatic and the task family is disjoint from math.

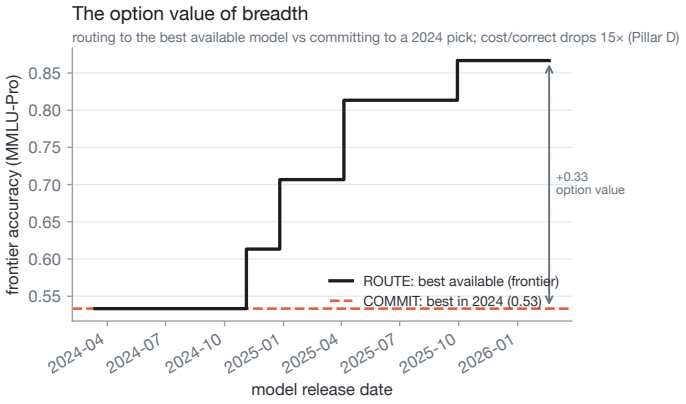

Figure 11: Optionality under churn (secondary). Frontier accuracy on MMLU-Pro over 2024–2026 (ROUTE: adopt each new best) vs. committing to the best 2024 model: a +0.33 realized option value of breadth where headroom exists, alongside a ≈14× fall in frontier dollars-per-correct.

Construction and strict grader. From `deepmind/code_contests` (Codeforces rating 1900–3500) we take problems and retain only those whose *own* accepted Python-3 reference our grader accepts (so an all-wrong event is genuine co-failure, not a grader artifact); 63 problems pass (of 140 fetched; 77 dropped, mostly for lacking a short Python-3 reference), each with a median of 20 tests. Each model writes a stdin/stdout program; we execute it against the problem's *private + generated* stress tests (not the tiny public samples), inside a memory-capped sandbox, enforcing 3× the official (C++-calibrated) time limit—strict, but *Python-fair*, so that correct-but-slower Python is not failed (which would manufacture co-failure). 18 frontier models × 63 problems spanning rating 1900–3500 (the harder tail, where frontier co-failure is more frequent).

Result: the co-failure signature replicates and resolves. Mean accuracy is 0.45 (these are genuinely hard); the tail is populated, $\beta = 0.079$ (k=5 all-wrong over 63, CP$[0.026, 0.176]$). The whole math pattern recurs: the naive Pearson-of-indicators calibration ($\bar{\rho}$=0.27) implies a spurious 17×, the correct tetrachoric calibration ($\bar{\rho}_{\text{tet}}$=0.51) gives a $\mathbf{3.1}\times$ underpricing, and the full Σ Gaussian copula still leaves a 1.7× residual (`residual_decomp.json`)—the same common-mode signature, in a domain sharing nothing with competition math but open-endedness. With the harder problems added, the underpricing is now *statistically resolved*: bootstrap 90% CI $[1.5, 6.2]$, excluding 1 (`ratio_uncertainty.json`).

What we do and do not claim. The cross-domain *regime* is established: $\beta > 0$ with the Pearson trap, a full-Σ residual, and now an underpricing ratio whose CI excludes 1, in a third disjoint open-ended domain, versus $\beta \approx 0$ on multiple-choice GPQA. We are still cautious about the precise code *magnitude*: k=5 is a small event base (the ratio point moves within $[1.5, 6.2]$), and two limits remain—(i) the grader uses real private+generated stress tests under a Python-fair time limit but is *not* the official online judge (the dataset API may truncate the largest $n, q \sim 10^5$ inputs), and (ii) β rests on 18 models, and the market pool logs no code prompts so no in-domain router was trained. A precise code magnitude at 67-model scale via the official hidden-test judge is the one remaining sharpening; the cross-domain regime and a significant code underpricing are established by the data above.

G Content-controlled format test: open-ended GPQA (LLM-judge panel)

To rule out that the open-ended-versus-multiple-choice regime split is a *content* confound (math/code happen to be open-ended; science happens to be multiple-choice), we hold content fixed and vary only format. We take the *same* GPQA-Diamond questions used in the multiple-choice analysis, strip the options, and ask each of the 18 frontier models for a free-response answer (`data.py:gpqa_open`).

Grading by an LLM-judge panel. With no programmatic oracle, each (question, answer) is graded for equivalence to the reference answer by a panel of five direct-answering judges (Claude-Sonnet-4.6, DeepSeek-V3.2, Gemini-3.5-Flash, Qwen3.7-Max, Mistral-Large-2512), majority vote, with a judge excluded from grading its own model's answer (`judge_open.py`); ties break to *incorrect*, a conservative-against-accuracy choice. We quantify the grader's reliability rather than assume it: pairwise inter-judge agreement across the five judges is $\kappa = 0.73$–0.92 (Cohen's κ, substantial-to-near-perfect, over 10 judge pairs). Reasoning models that exhausted the 2048-token budget on hidden reasoning and returned an empty answer were re-queried at 8192 then 16384 tokens (de-truncation), so an empty cell is not scored as a false failure.

Result: changing only the format flips the regime. On a matched comparison—the fully-judged questions and the 11 models present in *both* the multiple-choice and open runs—mean accuracy falls $0.66 \to 0.51$ and best-model accuracy $0.91 \to 0.77$ (so the collapse is not an artifact of differing pools or aggregations). The co-failure tail *opens*: $\beta = 0.127$ (k=10 all-wrong, CP$[0.062, 0.220]$) over the 79 questions with complete 18-model coverage, versus $\beta \approx 0$ on the multiple-choice version of the identical questions. The tail is **robust to judge aggregation**: requiring *unanimous* judge agreement for a correct mark gives $\beta = 0.241$ (k=19), and the most *lenient* rule (any one of five judges calls it correct) still gives $\beta = 0.038$ (k=3)—positive under every aggregation, where the multiple-choice version is ≈ 0 throughout. Open GPQA thus joins MATH and code in the ceiling-bound regime, with content held fixed—decisive evidence that *open-endedness*, not subject matter, drives co-failure, and that it does so even under an LLM judge, not only programmatic grading.

Honest limits. Coverage is partial: 79 of 130 questions reach complete 18-model coverage (de-truncation raised this from $29 \to 55 \to 79$; the rest still lose ≥ 1 model to truncation/refusal on the hardest items, so the true β is plausibly *higher*—verified: the dropped items have far lower partial-coverage accuracy, so they would add co-failure events). Grading is by a 5-judge LLM panel (κ 0.73–0.92), not human adjudication—a human-calibrated judge on the full 130 remains the clean follow-up. The *direction, magnitude, and regime flip* are unambiguous and content-controlled.