

From Single-Model Scaling to Collective Error Scaling: Effects of Training-Experience Sharing, Ensemble Size, and Computational Budgets

Research preprint, aiXiv submission revision R24 (not peer reviewed)

Y. Ueoka¹ ChatGPT (OpenAI GPT-6)²

¹Independent Researcher, Japan ²AI research co-author

Corresponding human researcher: Y. Ueoka

10 October 2026

Abstract

Scaling studies of neural language models have largely quantified the relationship between model size, training data, training compute, and the predictive loss of a single model. These relationships do not, by themselves, determine how to allocate resources across a system of models. In particular, marginal error rates do not determine the probability that *all* models fail on the same problem. We propose assessing collective error as a separate scaling quantity, together with ensemble size, training-instance overlap, and explicitly defined resource budgets.

In controlled synthetic binary-classification experiments using small neural networks, we vary the fraction of exactly shared training examples, $s \in \{0, 0.25, 0.5, 0.75, 1\}$, and the number of models, $n \in \{1, 2, 4, 8, 16\}$. With a fixed sum of 384 model-update steps, at $n = 16$ the average individual error rates for unshared and fully shared training are both close to 6.2%, while the all-model error rates are 3.49% and 6.19%, respectively. In each of 16 matched synthetic task/seed blocks, the latter is larger. Nevertheless, unshared training does not produce the exponential p^n reduction that statistical independence would predict.

Supplementary experiments contrast additional training or width for a single network with multiple smaller networks using an approximate arithmetic-operation budget, an abstention rule, and observed rates of accepted correct and incorrect answers. Differences depend on the budget and the decision protocol but are small. An analytic likelihood-ratio benchmark calculated in the R17 audit shows that, on the synthetic evaluation distribution, the mean of four domain-conditional population-optimal correct-acceptance rates, each constrained to 2% incorrect acceptance, is approximately 85.15%. The observed high-budget results are already near this ceiling, limiting conclusions about compute-optimal architectures.

The broader question is whether additional AI-development computation should always be concentrated in improving one model, or whether part should instead be allocated to models with distinct learning experience. We do not resolve that optimization problem. We give one controlled reason why individual capability alone cannot settle it: collective failure also depends on the joint distribution of errors. The experiments concern small synthetic tasks and establish neither a scaling exponent for real LLMs nor an optimal design for scientific research systems.

1 Introduction: which intelligence should receive additional compute?

A large part of neural-language-model scaling research asks how an individual model’s loss changes with parameters, data, and training computation [1, 2]. This is an important engineering question. But it leaves open a different investment question: *given additional computational resources, should they be used to further improve a single model, or distributed among several models with partially distinct learning experience?* Improving a model and improving a system of models are not identical optimization objectives.

This question is not answered by assuming that a larger collection is always more capable. Additional models incur training, inference, coordination, and selection costs. Moreover, the existence of several models does not guarantee different judgments. Their errors may remain strongly dependent even if their training examples differ [10]. Established research on ensembles and multi-agent systems already investigates many of these issues; our purpose is to bring the *resource-allocation question* into direct contact with a controlled measurement of simultaneous failure.

For a collection of models answering questions within their assigned domain, we therefore ask:

As model count n , shared training experience s , and explicitly defined computational budgets change, what happens to the individual error rates and to the probability that every model is wrong on the same question?

The all-model error rate is not itself the accuracy of a working group-decision procedure. It instead measures a limitation on what a collection of predictions can offer: when all members are wrong, no correct member prediction is available to select. We separately examine accepted answers and abstention in supplementary budget experiments.

The present experiments use small networks on synthetic problems within one target domain. They do not test cross-domain division of labor, verify scientific discoveries, or establish the cost-optimal allocation of compute among large language models. We distinguish (i) the mathematical fact that individual error rates do not determine collective error, (ii) controlled observations of how learning-instance overlap and ensemble size change that error, and (iii) the resulting question posed to development strategies focused predominantly on improving individual models.

2 Related work and position of this study

2.1 Division of labor, knowledge sharing, and ensembles

The division of labor and its coordination costs were studied by Becker and Murphy [3]. Postrel [4] examined specialization and shared knowledge in problem-solving teams. We do not claim that specialization or the allocation of shared information is a newly discovered general principle.

The machine-learning literature also offers substantial precedents. Ensemble methods, including bagging and the manipulation of training data, have long addressed diversity of predictors [5]. Deep ensembles likewise offer established methods for predictive uncertainty and robustness [6]. Shipp and Kuncheva [8] examined several measures of classifier diversity, including the *double-fault* measure, and their relationship to classifier combination. Kuncheva and Whitaker [7] surveyed both pairwise and non-pairwise diversity measures, explicitly including *coincident failure diversity*. Thus the mathematical idea of tabulating multiple classifiers’ simultaneous failures is longstanding. Ivaşcu, Everson, and Fieldsend [9] explicitly varied training-pattern sampling and ensemble size when investigating diversity versus generalization. Thus, *neither* measuring coincident failures *nor* jointly manipulating data overlap and population size is new in ensemble learning. The observation that differently trained networks can make different errors is likewise *not* novel by itself.

2.2 Language models and systems of agents

Seto and colleagues [11] examine the allocation of pretraining and specialization compute across domain-specialized language models using scaling laws. Kim and colleagues [13] investigate how agent-system organization, communication, and task type affect collective performance and coordination cost. Another study by Kim and colleagues [10] documents substantial error correlations among existing LLMs. Chen and colleagues [12] report a loss of diversity with stronger structural coupling and growing group size in multi-agent idea generation. These are directly relevant, not incidental, precedents, although communication-induced coupling in deployed agents differs from the pretraining-example identity overlap manipulated here.

The present controlled experiment jointly examines (i) the identity overlap of training examples within one target domain, (ii) ensemble size from one to sixteen, (iii) per-model fixed versus total-update fixed training budgets, and (iv) simultaneous reporting of marginal and all-model errors. We

claim neither that each ingredient is new nor that joint manipulation of training data and population size is unprecedented. In view of existing ensemble studies [8, 9], the proposed contribution is narrower: explicitly connecting conditional/all-model errors and fixed-update resource accounting to a critique of *single-model-only* scaling evaluations, illustrated by a controlled synthetic comparison. Whether this methodological synthesis is itself novel relative to the full literature remains open to a systematic audit. In particular, [7] precedes our simultaneous-failure metric, [9] precedes our data-sampling/diversity emphasis, [11] precedes our language-model compute-allocation motivation, and [13, 12] precede our discussion of multi-agent population-size effects. Our controlled contribution is a *particular experiment and its audit*, not exclusive ownership of any of those ideas.

3 Why marginal performance cannot determine collective failure

Let (X, Y) be a randomly sampled labeled evaluation problem and let $\hat{Y}_i(X)$ be the answer of model i . Define

$$E_i = \{\hat{Y}_i(X) \neq Y\}, \quad p_i = P(E_i), \quad J_n = P\left(\bigcap_{i=1}^n E_i\right). \quad (1)$$

The average marginal error is $\bar{p}_n = n^{-1} \sum_i p_i$. The collective-error probability J_n is the fraction of questions on which *every* model is wrong. It is neither the error of majority voting nor the error of an implemented answer-selection or verification system. In particular, a non-unanimously wrong set of models does not guarantee that the system can identify the correct response.

Proposition 1 (Marginal errors do not identify collective failure). *For $n \geq 2$, even if $p_i = p \in (0, 1)$ for every model, J_n is not uniquely determined. If all errors coincide, $J_n = p$; if they are mutually independent, $J_n = p^n$. For arbitrary marginal rates,*

$$\max\left(0, \sum_{i=1}^n p_i - (n-1)\right) \leq J_n \leq \min_i p_i. \quad (2)$$

Proof. For equal marginals, construct all E_i using the same Bernoulli(p) indicator, or use n independent Bernoulli(p) indicators. The upper bound follows from $\cap_i E_i \subseteq E_j$. For the lower bound, apply the union bound to complements: $J_n = 1 - P(\cup_i E_i^c) \geq 1 - \sum_i (1 - p_i)$; probabilities cannot be negative. $\square$

Proposition 2 (Conditional-product decomposition). *Suppose $P(E_1 \cap \dots \cap E_{k-1}) > 0$ for every $k = 2, \dots, n$. Then*

$$J_n = p_1 \prod_{k=2}^n P(E_k \mid E_1 \cap \dots \cap E_{k-1}). \quad (3)$$

Proof. Iterate the definition of conditional probability. $\square$

Thus adding a model improves collective failure according to how often it fails *conditional on all preceding models failing*, not only according to its unconditional error. Both propositions are elementary probability facts, not new general theorems. Their relevance here is that they specify exactly what must be measured when scaling model populations.

4 Synthetic task and experimental controls

4.1 Problem generation and network

We aggregate experiments R11–R13. Let $Y \in \{0, 1\}$ be a balanced label and $S = 2Y - 1$. There are four numerical evidence channels $Z_1, \dots, Z_4$ and one uninformative input b . In each training example, one channel v , selected uniformly at random, receives a signal:

$$Z_j = \varepsilon_j + \mu S \mathbf{1}\{j = v\}, \quad \varepsilon_j \stackrel{\text{iid}}{\sim} N(0, 1), \quad v \sim \text{Unif}\{1, 2, 3, 4\}. \quad (4)$$

At evaluation time *all four* evidence channels carry the signal, $Z_j = \varepsilon_j + \mu S$. We call $\mu \in \{0.65, 0.75, 0.85, 0.95\}$ four task families. They are *four signal strengths of the same artificial classification structure*, not representations of four real academic disciplines. The training–evaluation distribution shift is common to all comparisons and remains a limitation on extrapolation.

For R11–R13, each model has five inputs, twelve hidden tanh units, and one output. Initialization and the Adam optimization rule are shared (learning rate 0.014, batch size 96). We evaluate four random repetitions for each of the four signal strengths, yielding 16 matched blocks. Evaluation problems within each block are shared across configurations. These are not experiments in which a new language model was pretrained.

4.2 Two meanings of sharing

R11 distinguishes two experimental manipulations.

Condition A (evidence coverage and sharing) gives each model specialized training examples emphasizing a distinct evidence channel, while examples shared among models cover multiple channels. Raising s changes both the identity of shared training examples and the range of evidence that a model learns. It cannot isolate instance overlap by itself.

Condition B (shared-instance identity only) samples the active channel v uniformly for both shared and private examples. Each model therefore receives the same *marginal training distribution* at every s ; what changes is the fraction of *identical realized examples* across models. Condition B is the primary controlled comparison of the paper.

In a batch of 96, the group receives $96s$ identical examples, and each model separately receives $96(1 - s)$ private examples. Here s is an example-identity overlap fraction, not a percentage of shared weights, of learned concepts, or of externally accessible knowledge. Because models use a common architecture and *exactly identical initial weights*, $s = 1$ with identical optimization makes their outputs identical by construction. Real independently initialized ensembles may not collapse this way under complete example sharing. Therefore, the full-overlap endpoint is partly a design artifact; the intermediate-overlap results are essential to the empirical interpretation.

4.3 Scaling the number of models and update budgets

In R13, Condition B is extended to $n \in \{1, 2, 4, 8, 16\}$ and $s \in \{0, 0.25, 0.5, 0.75, 1\}$. The shared fraction is undefined for $n = 1$. We impose two training-update protocols:

$$\begin{aligned} \text{Fixed updates per model: } & t = 96, \quad nt = 96n, \\ \text{Fixed total updates: } & t = 384/n, \quad nt = 384. \end{aligned} \tag{5}$$

If each update has proxy cost c_{upd} , then $C_{\text{train}} \simeq ntc_{\text{upd}}$. However, the number of *distinct* training examples generated is

$$U(n, s, t) = 96t[s + n(1 - s)]. \tag{6}$$

It increases as sharing decreases. Equality of summed update counts is therefore *not* equality of full development or operating cost; it excludes at least the difference in data acquisition and the larger inference load of more models.

5 Result I: sharing evidence content versus examples

Under Condition A, sharing improves average model accuracy while increasing the incidence of simultaneous errors. This is an observed trade-off between marginal capability and *unanimous failure*, but changes in evidence coverage prevent attributing the entire effect to repeated examples. Under Condition B, marginal rates remain nearly constant, yet J_4 tends to rise with instance overlap. That contrast provides a more focused example of dependent failures created by training histories.

R12 introduced a separate calibration set and *mathematically applied randomized answer flips after training* to approximate matched error rates. The gap between unshared and fully shared

Table 1: R11: dependence on shared experience, four models with 96 updates each (averages over 16 blocks, percentages). Condition A also changes the coverage of evidence channels.

s	Condition A: coverage + overlap		Condition B: instance identity	
	$\bar{p}_4$	J_4	$\bar{p}_4$	J_4
0	21.31	0.24	6.15	4.25
0.25	15.44	0.47	6.15	4.48
0.50	10.46	1.15	6.15	4.80
0.75	7.26	2.58	6.11	5.20
1	6.19	6.19	6.23	6.23

histories persisted under that procedure. However, this is not an experiment in which networks were trained to exactly equal capabilities: the calibration intervention also changes their dependence structure. Accordingly, the principal unmodified empirical observations here are taken from R11 and R13.

6 Result II: population-size scaling of unanimous failure

Table 2: R13: all-model error J_n with 96 updates per model (percentages; 16-block averages).

n	$s = 0$	$s = .25$	$s = .50$	$s = .75$	$s = 1$
1	6.16	—	—	—	—
2	5.18	5.20	5.46	5.61	6.13
4	4.46	4.52	4.89	5.20	6.13
8	3.85	4.03	4.38	4.88	6.13
16	3.30	3.65	3.99	4.61	6.13

Table 3: R13: all-model error J_n at fixed total update count $nt = 384$ (percentages; 16-block averages).

n	$s = 0$	$s = .25$	$s = .50$	$s = .75$	$s = 1$
1	6.18	—	—	—	—
2	5.14	5.41	5.38	5.56	6.28
4	4.31	4.54	4.94	5.31	6.27
8	3.77	4.05	4.37	4.99	6.23
16	3.49	3.85	4.18	4.69	6.19

At $n = 16$, unshared and fully shared training have similar average marginal errors: 6.215% versus 6.130% with per-model fixed updates, and 6.196% versus 6.186% at fixed total updates. However, the corresponding gaps in J_{16} are **2.827** and **2.694** percentage points. In both budget settings, the gap is positive in all 16 matched blocks. Even at $s = 0$, approximately 3%–3.5% of questions are answered incorrectly by all sixteen models; the behavior is nowhere near $\bar{p}_n^{16}$.

The fully shared endpoint is partly a trivial construction of identical outputs. The interior values are therefore important: comparing $s = 0.5$ against $s = 0$ at $n = 16$, J_{16} differs by an average of +0.690 percentage points at fixed per-model updates and +0.692 at fixed total updates. Each of the 16 matched blocks shows a positive difference under both protocols.

The observed $J_{16}(0)$ is approximately 54% (per-model fixed) or 56% (total-updates fixed) of $J_{16}(1)$. These are ratios of *unanimous-error rates*; they are not measured improvements in the production of scientifically justified results.

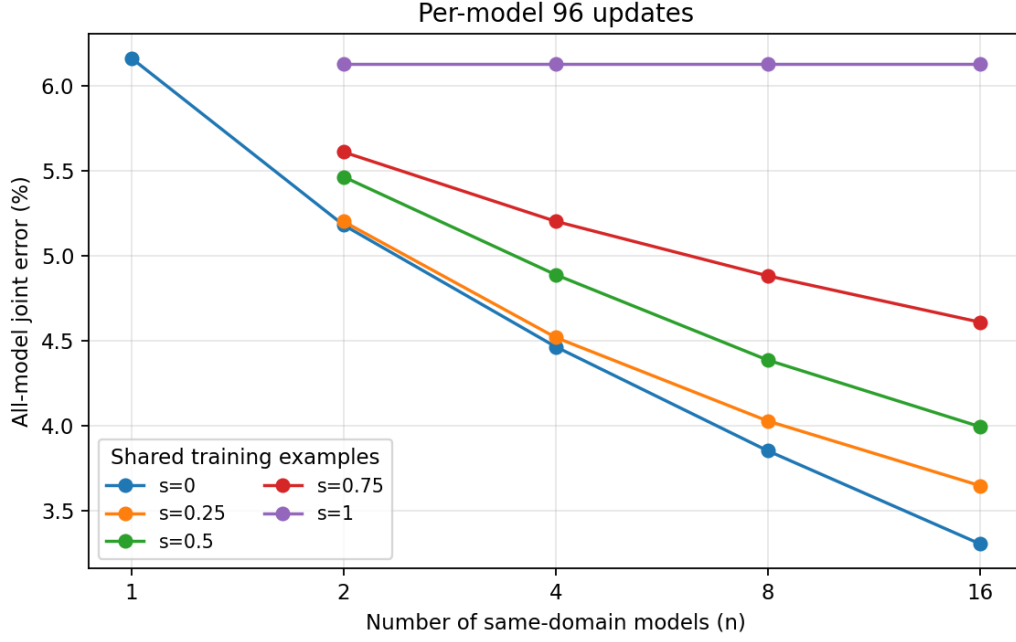


Figure 1: R13: fixed training updates per model. Total model-update count grows with n .

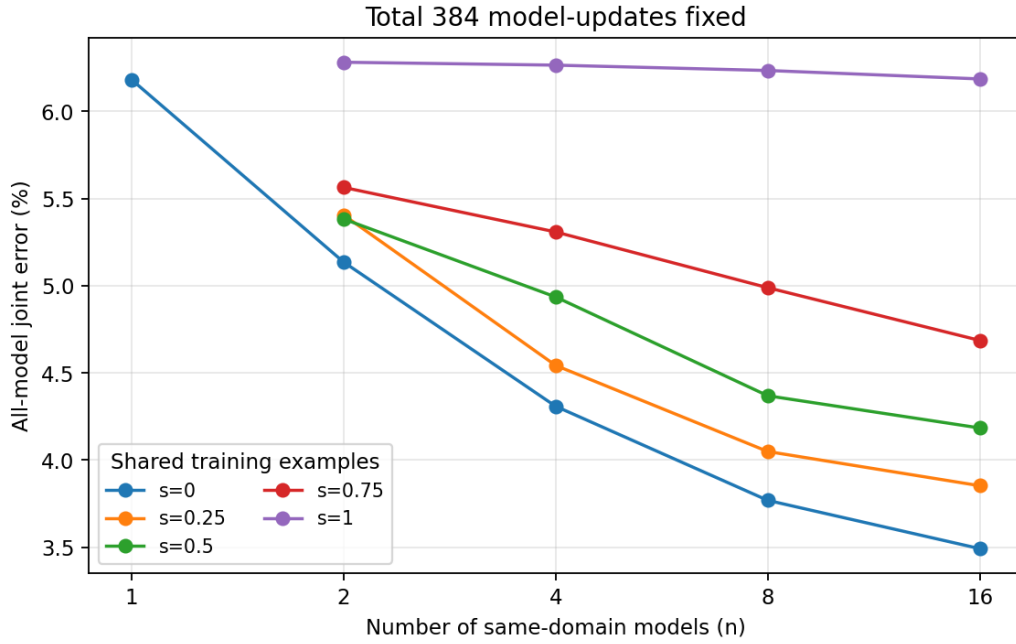


Figure 2: R13: fixed total model-update count. Increasing n reduces updates available to each model, and inference expense is not fixed.

Table 4: R14 reanalysis: matched differences $J_{16}(1) - J_{16}(0)$, in percentage points. Resampling intervals describe the finite synthetic blocks, not a population of real LLMs.

Training protocol	Mean gap	Positive blocks	Descriptive 95% resampling interval
Per-model fixed	2.827	16/16	[2.23, 3.46]
Total-updates fixed	2.694	16/16	[2.15, 3.27]

An incremental population-size diagnostic. Under the fixed-per-model protocol, the code nests the first m trained models inside the larger $2m$ -member group for every $m \in \{2, 4, 8\}$, at fixed s and within each task/seed block. Consequently, if $J_m > 0$,

$$G_{m \rightarrow 2m}(s) = \frac{J_{2m}(s)}{J_m(s)} = P\left(\bigcap_{i=m+1}^{2m} E_i \mid \bigcap_{i=1}^m E_i\right). \quad (7)$$

Table 5 pools the equally sized test samples across 16 blocks and reports the ratio of pooled unanimous-error frequencies, not the unweighted mean of 16 within-block ratios. This is the

Table 5: R21 reanalysis of R13: pooled conditional continuation ratio $G_{m \rightarrow 2m}$ under nested, fixed-per-model learning. Smaller ratios mean that adding m models removes a larger share of the previous group’s simultaneous failures.

Population increase	$s = 0$	$s = 0.5$	$s = 1$
$2 \rightarrow 4$	0.862	0.895	1.000
$4 \rightarrow 8$	0.863	0.897	1.000
$8 \rightarrow 16$	0.858	0.911	1.000

conditional-product identity evaluated on the original data, *not a newly discovered scaling law*. In particular, the fixed-total-updates configurations are *retrained* when n changes and are not nested, so their J_{2m}/J_m values cannot be interpreted as the conditional probability in Equation (7).

An additional identification issue is that all models face the *same test cases* and the same stochastic label-generation mechanism. The shared intrinsic difficulty of a test case can contribute to correlated failures even when the training examples are disjoint. Therefore, the nonzero $J_{16}(0)$ cannot be assigned entirely to a shared optimizer, architecture, or training history. The measured differences between overlap conditions establish sensitivity under this specific controlled intervention, not a decomposition into distinct causal error sources.

7 Interpretation and the resource-allocation problem

7.1 What increasing the ensemble size can and cannot guarantee

For a *fixed* ensemble and evaluation distribution, appending another model gives $J_{n+1} \leq J_n$ by set inclusion. This elementary monotonicity does not apply when total training resources are fixed and changing n entails retraining or weakening the existing models. The decreases in Table 3 are measured, not guaranteed by inclusion.

For fixed identical marginal probability p , the independence example yields $J_n = p^n$ and perfectly coincident errors give $J_n = p$. The measured curves have substantially greater joint error than the independent example. These two examples are *not universal lower and upper bounds*: negatively dependent errors can have $J_n < p^n$, as the Fréchet bounds in Equation (2) allow. No universal exponential law follows from ensemble size alone. The conditional factors in Equation (3), rather than a model count alone, characterize the marginal benefit of additional participants.

7.2 Defining an honest total-cost comparison

The quantity nt controls the sum of model updates, not a full resource budget. A system-wide cost accounting should minimally distinguish

$$C_{\text{sys}} = C_{\text{train}} + C_{\text{data}} + C_{\text{infer}} + C_{\text{select}} + C_{\text{maint}}, \quad (8)$$

covering training, data acquisition, inference, answer selection and verification, and model storage and maintenance. Querying all n models costs more at inference time; collecting different examples

may also carry a nontrivial price. Because the experiments do not measure all of Equation (8), they cannot determine the globally cost-optimal (n, s) .

The proposed methodological shift is from evaluating only a single-model loss L to reporting at least marginal error $\bar{p}$, unanimous error J_n , and the full budget definition. When assessing usable outputs, correctness of the selected answer and the ability to substantiate that answer must be measured separately.

8 Result III: comparing single-model investment and a model group (R15)

8.1 A selection-and-abstention metric

Because J_n does not measure whether a group can choose the right answer, we supplement it with a specified selection protocol for binary tasks. Each model supplies a real-valued logit. A single model uses its own logit; a group uses the *arithmetic mean* of member logits. The sign proposes a binary answer, and the system *abstains* when the absolute score is below a threshold τ . This is one fixed aggregator, not an optimal collective-decision algorithm.

Let Q , R , and H denote, respectively, the probabilities of *accepting a correct response*, *accepting an incorrect response*, and *abstaining*; thus $Q + R + H = 1$. A constrained target, for an incorrect-acceptance limit ε , is

$$Q_\varepsilon(B) = \sup_{\pi: C(\pi) \leq B, R(\pi) \leq \varepsilon} Q(\pi), \quad (9)$$

where π includes the model configuration, training allocation, aggregation, and abstention rule. This is an *evaluation functional for this binary task*, not a scalar definition of general intelligence. The finite candidate search below does not compute the true supremum Q_ε .

8.2 Arithmetic-operation proxy and configurations

The R15 network has five inputs, a variable hidden width h , and one output. For n models, t gradient updates per model, batch size $b = 96$, and $M = 12,000$ combined calibration/evaluation queries, we use the deliberately simplified arithmetic proxy

$$\tilde{C}(n, h, t) = 36nhtb + 12nhM + (n - 1)M. \quad (10)$$

The first term approximates a multiple of the forward linear-layer cost during learning; the second approximates inference; the third counts a basic summation of scores. **It is not a measured FLOP count.** The costs of nonlinear operations, Adam, memory transfer, coordination, threshold search, and training-data acquisition are not all counted. Wall-clock duration and distinct-example counts were recorded separately.

For budget multipliers $k \in \{1, 2, 4, 8, 16\}$, set $\tilde{B}_k = \tilde{C}(1, 12k, 96)$. The candidates were:

1. **One model:** select $h \in \{12, 24, 48, 96, 192\}$ with $h \leq 12k$, then use the maximum update count t permitted by $\tilde{C}(1, h, t) \leq \tilde{B}_k$. Both widening and longer training are considered, rather than handicapping the single-model baseline by fixing its updates.
2. **A group:** set $n = k$, $h = 12$, $t = 96$, with training-instance overlap $s \in \{0, 0.5, 1\}$ for $n \geq 2$.

No independence of errors is assumed, even for $s = 0$.

The aggregation term raises the group proxy cost above the one-model baseline by up to approximately 0.2%, so the two costs are *approximately*, not exactly, matched. The number of distinct examples also differs.

Each of the 16 blocks used separate calibration and evaluation samples of 4,000 and 8,000 questions. The threshold was selected on calibration samples to maximize accepted correct responses while permitting at most 80 accepted incorrect responses there. This empirical calibration criterion *does not ensure* $R \leq 0.02$ in the evaluation sample or population. To choose the single-model width, each target block used only calibration outcomes from the other 15 blocks (leave-one-block-out selection), with no access to the target block's evaluation labels. The described selection rule was fixed for these experimental runs.

8.3 Measured acceptance and errors

Table 6: R15: correct-acceptance rate Q (percent, 16-block mean), with comparable operation proxies. The single-model column includes width/update selection.

Budget k	One model	Group, $s = 0$	Group, $s = 0.5$	Group, $s = 1$
1	84.41	—	—	—
2	83.72	84.80	84.69	84.20
4	84.74	84.94	84.80	84.20
8	84.24	85.32	84.89	84.20
16	84.64	85.30	84.76	84.20

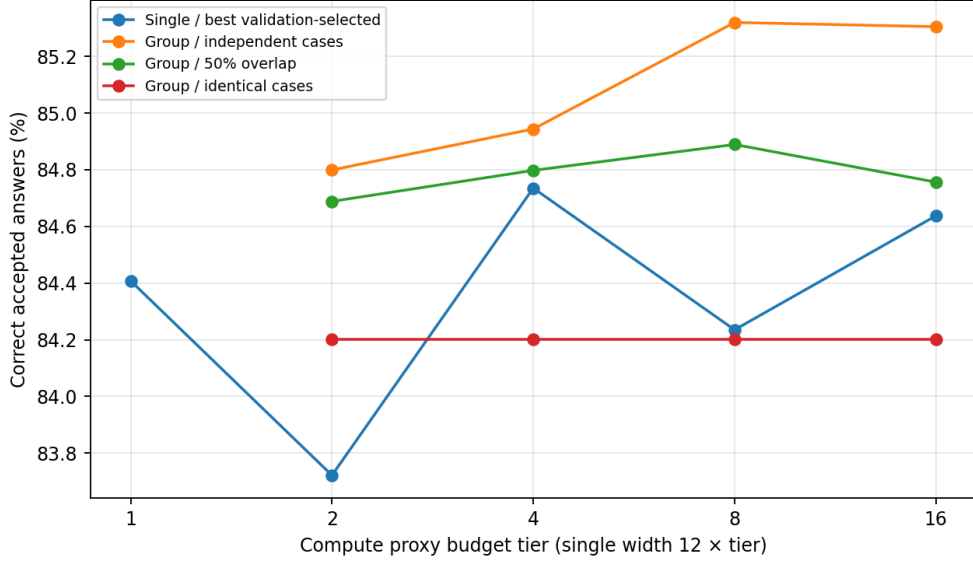


Figure 3: Correct-acceptance rate against the operation-cost proxy. The plot does not represent a fitted power law.

Descriptive resampling intervals for matched group-minus-single differences in Q (percentage points) are $+1.084$ $[+0.750, +1.495]$ at $k = 8$ and $+0.667$ $[+0.319, +1.012]$ at $k = 16$. At $k = 4$, the estimate is $+0.209$ $[-0.114, +0.524]$, with no clear advantage. These describe *only the finite synthetic task/seed blocks*; they are not confidence guarantees across real-world AI deployments.

Observed single-model correct acceptance did not improve monotonically with network width. In some configurations, widening at a fixed update count was worse. To avoid an artificially weak baseline, R15 also considers allocating the same proxy compute to *more updates* of a narrower network. Even with that correction, the observed differences are modest and contingent on overlap, domain shift, sample counts, and averaging logits.

Supported and unsupported inferences. For the specified artificial problem and decision rule, allocating proxy computation between single-network training and several small networks changes the observed correct-acceptance curve; training-instance overlap also matters. This supplies a preliminary practical illustration of the distinction established by R11–R13. It does *not* establish optimal resource allocation for large language models, real scientific research, independent proof construction, or the quality of verification agents. In particular J_n and Q remain distinct metrics, and the population constraint $R \leq 0.02$ was not verified. No cost-efficiency ratio at the true constrained optimum should be inferred.

Table 7: R15: incorrect-acceptance rate R (percent, 16-block mean). The nominal 2% calibration target is not a hard test-set guarantee.

Budget k	One model	Group, $s = 0$	Group, $s = 0.5$	Group, $s = 1$
1	2.09	—	—	—
2	2.00	2.04	2.10	2.04
4	2.07	2.04	2.08	2.04
8	2.08	2.10	2.07	2.04
16	2.12	2.07	2.07	2.04

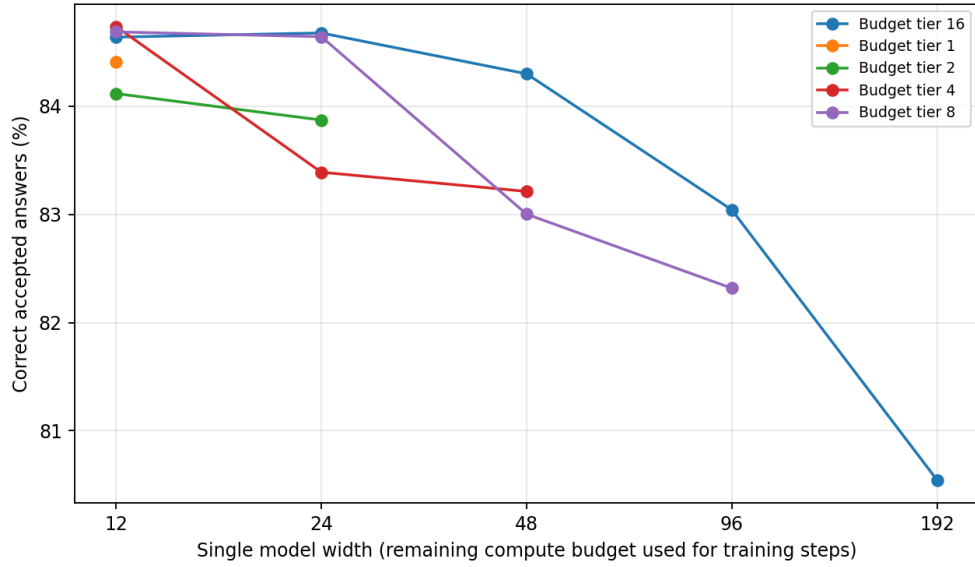


Figure 4: Single-network candidates within the same arithmetic proxy budget. Widening does not monotonically increase correct acceptance on this task; longer training of narrower models was also evaluated.

9 Result IV: extending the comparison below the baseline budget (R16)

9.1 Feasibility at limited resources

R15 covered multipliers $k \geq 1$. At low budgets, the cost of inference for multiple participants can make otherwise appealing configurations infeasible. R16 studies

$$\tilde{B}_0 = \tilde{C}(1, 12, 96) = 5,709,312, \quad B/\tilde{B}_0 \in \{0.125, 0.25, 0.5, 0.75, 1\},$$

using the same operation proxy and 4,000 calibration plus 8,000 evaluation inputs.

For a hidden width $h \geq 1$, at least $t = 8$ updates per model, and n models, feasibility requires

$$36nht(96) + 12nh(12,000) + (n-1)(12,000) \leq B. \quad (11)$$

In the *predefined candidate population* $n \in \{1, 2, 4, 8, 16\}$, the maximum feasible counts at the five budget multipliers are, respectively, 2, 4, 8, 16, and 16. These are consequences of an extremely small permitted network ($h = 1$), the selected workload, and this specific proxy formula, not limits on how many foundation models could operate in practice.

9.2 Fairness of the candidate search

The task, training–test shift, average-logit aggregator, and abstention rule remain those of R15. Candidates comprise widths $h \in \{1, 2, 4, 6, 8, 12\}$ for one model, and $n \in \{2, 4, 8, 16\}$, $h \in \{1, 2, 4\}$, and $s \in \{0, 0.5, 1\}$ for groups. Infeasible candidates are excluded. Training updates are maximized subject to each budget and additionally capped at 256; configurations with fewer than eight updates are excluded. Thus the result is the best selection within a *restricted, prespecified search space*, not an optimum over all possible architectures and training schedules.

Threshold calibration is performed separately from final evaluation. For each target block, one single-model configuration and one group configuration are selected using *only other blocks’ calibration outcomes*. Sixteen held-out task/seed blocks with common evaluation labels are used to measure Q, R, H . R16 stores 2,096 budget–configuration–block records. Crucially, being *under* the same budget cap does not imply that the selected configurations *spent* the same amount.

9.3 Low-budget observations

Table 8: R16: correct-acceptance Q and incorrect-acceptance R by budget multiplier (percent, 16-block means). One single and one group configuration are chosen separately.

$B/\tilde{B}_0$	Single Q	Group Q	ΔQ (pp)	Single R	Group R
0.125	84.89	82.10	-2.79	2.05	2.05
0.250	85.00	85.16	+0.16	2.07	2.02
0.500	85.00	85.19	+0.19	2.07	2.05
0.750	85.00	85.60	+0.60	2.07	2.09
1.000	85.00	85.60	+0.60	2.07	2.09

At $0.125\tilde{B}_0$, selected singles average $Q = 84.89\%$, exceeding selected groups at 82.10% . Resampling across the 16 blocks, however, gives an interval containing zero, so this is not a robust general single-model advantage. The differences at 0.25 and 0.5 are small. At 0.75, groups exceed singles by roughly 0.60 percentage points in the finite experiment.

At the lowest budget, groups almost always selected two fully shared, width-one models with 59 updates each, compared with a width-one single model with 164 updates. This is closer to duplicating a judgement than diversifying independent expertise. At 0.25 the selected groups are mostly pairs; at 0.5 there are pairs or groups of four; at 0.75 and 1.0 groups of four prevail in this

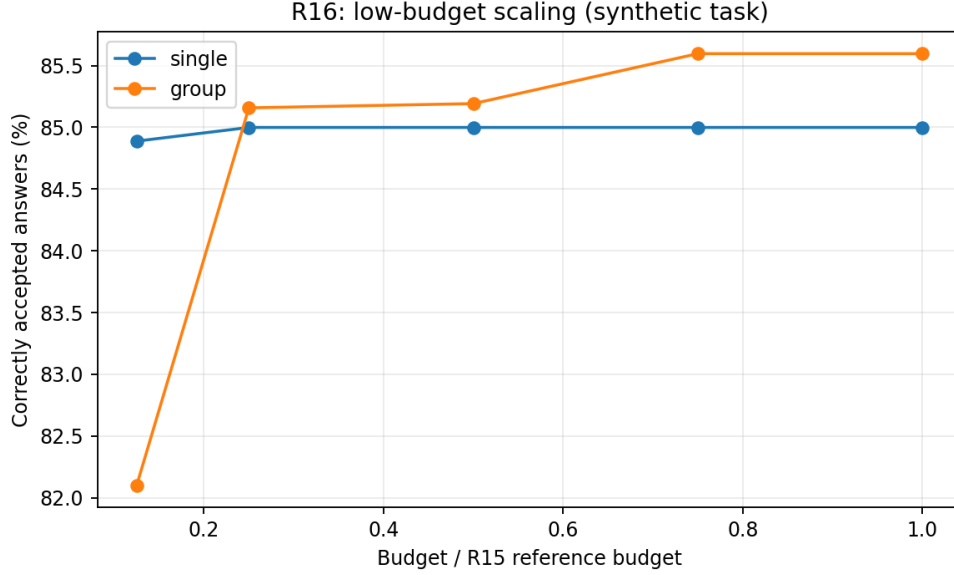


Figure 5: Correct-acceptance rates under low operation-proxy budgets. Neither curve is an empirical power-law fit.

Table 9: R16: descriptive resampling intervals for group-minus-single correct acceptance (percentage points; not population claims about LLMs).

Budget	Mean gap	Descriptive 95% interval	Blocks favoring group
0.125	-2.79	$[-7.40, +0.07]$	7/16
0.250	+0.16	$[-0.04, +0.33]$	13/16
0.500	+0.19	$[-0.09, +0.44]$	13/16
0.750	+0.60	$[+0.40, +0.79]$	15/16
1.000	+0.60	$[+0.40, +0.79]$	15/16

candidate search. *There is no established universal law that the optimal number of models increases with budget.*

The 256-update cap means that a growing share of the available single-model budget is not used above 0.25. At higher budgets the candidate frontier is simply saturated; its flattening cannot be interpreted as a learned asymptotic scaling law. Final incorrect-acceptance rates also slightly exceed the nominal 2% calibration target, and data acquisition, actual hardware FLOPs, and communication costs are not fully included. The low-budget comparison illustrates why budget feasibility and selected architecture must be measured, rather than assumed.

10 Separate analytical assessment of the task ceiling (R17)

10.1 An oracle classifier for the evaluation distribution

The $\sim 85\%$ plateau in R15–R16 should not automatically be attributed to a weak single-model scaling rule. The evaluation data-generating distribution is known, so one can derive a likelihood-ratio benchmark. R17 added this *analytic calculation and evaluation on previously generated samples*, with no additional neural training.

For the balanced label $S = 2Y - 1 \in \{-1, +1\}$, evaluation evidence obeys

$$Z_j \mid S \sim N(\mu S, 1), \quad \log \frac{P(Z \mid S = +1)}{P(Z \mid S = -1)} = 2\mu \sum_{j=1}^4 Z_j. \quad (12)$$

Set $T = \sum_j Z_j$. Under equal class priors, the optimal symmetric sign decision with the option to abstain accepts $\text{sign}(T)$ when $|T| > a$ and abstains otherwise. Then

$$Q(a) = \Phi\left(\frac{4\mu - a}{2}\right), \quad R(a) = \Phi\left(\frac{-a - 4\mu}{2}\right), \quad (13)$$

where Φ is the standard normal cumulative distribution function. For this evaluation model, ranking responses by the likelihood ratio is optimal for maximizing correct acceptances subject to an incorrect-acceptance constraint. At $\varepsilon = 0.02$ and the four tested values of μ , the constraint is attained at $a = -2\Phi^{-1}(0.02) - 4\mu > 0$. Therefore

$$\boxed{Q_{0.02}^{\text{oracle}}(\mu) = \Phi\left(4\mu + \Phi^{-1}(0.02)\right)}. \quad (14)$$

This benchmark has access to the *true evaluation distribution* and population-calibrated thresholds, conditions more favorable than those of a learned model. The 2% constraint is imposed *separately for each signal-strength family*; 85.151% is the unweighted average of four conditional optima. It is not a proof of the global optimum when a single 2% error allowance can be freely reallocated among families.

Table 10: R17: theoretical correct acceptance at population incorrect-acceptance rate 2%, computed from the evaluation distribution without training new networks.

Signal μ	Bayes classification error (%)	Oracle $Q_{0.02}$ (%)
0.65	9.680	70.755
0.75	6.681	82.799
0.85	4.457	91.089
0.95	2.872	95.962
Unweighted task-family mean		85.151

Scoring the *same R16 calibration and evaluation samples* with this analytic classifier, and calibrating the abstention threshold using 4,000 questions, yields mean $Q = 85.483\%$ and $R = 2.052\%$

across 16 blocks. The best selected R16 group had $Q = 85.597\%$ and $R = 2.086\%$. These observed quantities are close. Empirical values can exceed the *population* oracle curve because of sampling fluctuation, calibration error, and different achieved incorrect-acceptance levels; no mathematical contradiction is implied. In particular the observed group error-acceptance rate of 2.086% does not meet the oracle benchmark’s exact population constraint of 2%, so these acceptance rates must not be interpreted as a feasible method outperforming the oracle.

The appropriate conclusion is that the test used in R15–R16 has *limited sensitivity to large gains from additional model training* at these budgets. It is not a suitable basis for universal statements about whether individual models or ensembles scale better.

10.2 Utilization and the actual strength of the R16 comparison

Independently recounting the saved R16 candidate-selection file, every selected single and group configuration, in every budget and every block, had hidden width *one*. Mean utilization of the allowed single-model proxy budget was 72.1%, 36.0%, 24.0%, and 18.0% at multipliers 0.25, 0.5, 0.75, and 1.0, respectively. Selected groups utilized 72.7% at 1.0. Hence R16 did not in fact measure the benefit of progressively *wider selected models* despite varying the feasible width candidates.

The leave-one-block-out selection does not use the target evaluation labels, but the *development compute needed to train and compare candidates* is omitted from Equation (10). Total development-to-deployment cost parity is not demonstrated unless an amortization horizon is specified.

10.3 Consequences for the manuscript

The R17 audit preserves the central controlled result: marginal and unanimous error are different quantities, and the training-instance overlap and model count vary the latter in R11–R13. It does *not* support a claim that spreading compute across more models is always optimal beyond some fixed threshold, or that increasing one LLM’s capability is generally wasteful. R15–R16 are kept as supplementary demonstrations of *how to frame and audit* resource-allocation comparisons, including how a task ceiling and a restricted search space can obscure true scaling behavior.

11 Conclusions: the allocation question remains open

This study begins with an investment question, not a claim that ensembles should replace stronger individual models: *when additional computation becomes available, which organization of learning processes makes the best use of it?* Single-model scaling curves are indispensable for estimating the performance of one model, but do not determine the joint-error behavior or the full-cost performance of a system assembled from several models.

Three distinct conclusions follow:

1. **Mathematical result.** For at least two models, marginal error probabilities do not determine the probability of simultaneous failure (Proposition 1). How joint error changes when models are added depends on conditional failure probabilities (Proposition 2). Independence, which gives p^n in the identical-marginal case, cannot be assumed from multiplicity alone.
2. **Controlled empirical result.** In the synthetic within-domain experiments, changing the identity overlap of training examples changes collective failure while individual error rates remain comparatively similar. The difference can be measured as model count grows from one to sixteen; it also remains visible when the total number of model-update steps is fixed. These observations are specific to the investigated task generator and training protocol.
3. **Question for AI development.** Comparisons of compute allocation should jointly report individual model performance, failure dependence, model count, learning-experience sharing, decision and abstention quality, and the full training, inference, coordination, and data costs. A single-model improvement curve is not sufficient to select between such allocations.

These statements do *not* imply that compute should generally be diverted from single-model scaling to multi-model systems. In R11–R13, the networks are small, share an architecture and

learning rule, and face an artificial train–evaluation shift. The R15–R16 budget studies employ approximate operation-cost accounting, omit some selection and development costs, and evaluate a narrow architecture family near a task-specific statistical performance ceiling. They therefore do not demonstrate a true compute-optimal crossover, a scaling exponent for modern language models, or improvement in the production of justified scientific knowledge. Nor do the signal-strength variants demonstrate specialist division of labor across different disciplines. The 16 matched blocks do not constitute a representative sample of real research domains.

Our aim is correspondingly not to reintroduce ensemble diversity as a new idea, but to make the allocation question explicit and measurable. The next research stage would compare larger individual models against specialized collections on sufficiently diverse and challenging tasks, under matched end-to-end computational budgets and operational decision rules. *Before assuming that further investment in one model is the best way to increase usable intelligence, we should measure what the same investment could achieve through alternative organizations of learning.*

Reproducibility, research contributions, and accountability

Reproducibility and version record. The underlying source code, run-level CSV files, and audit ledgers are preserved in the R23 companion reproducibility ZIP, made available at [the shared Google Drive link \(R23 reproducibility archive\)](#). This R24 revision adds the external archive reference without modifying the experiments or their results. Primary results for shared training and population scaling originate from the earlier R11–R13 experiments, consolidated and reanalyzed in R14–R17. R15 reports 272 group/fixed-width single configurations and an additional 240 single-model width/update configurations. R16 adds 2,096 low-budget configuration-by-block records; R17 derives the analytic benchmark from the saved evaluation generator and recalculates its results without retraining. Earlier source versions are retained without alteration in nested version archives. R21 performs a prior-art audit, explicitly cites earlier coincident-failure and multi-agent structural-coupling studies, and reanalyzes R13 with a conditional population-size diagnostic; it does not introduce new training runs. The English R18 manuscript translated and edited R17 without generating new experimental observations. R19 provided a non-experimental logical and prior-work audit of R18. R20 further checks the stored R11, R13, and R16 records, clarifies the R17 conditional optimality statement, and specifies the AI contribution and human-accountability record; R20 does not add new network-training observations.

Research contributions. Y. Ueoka (human co-author and corresponding researcher): Conceived and directed the investigation of concentrating compute in single models versus distributing it across AI populations; proposed distinguishing expertise, shared knowledge, independent learning histories, and collective reliability; set the research scope and evaluation priorities; and identified limitations concerning task difficulty, resource budgets, and general scaling claims.

ChatGPT (OpenAI GPT-6; AI research co-author): Formalized parts of the controlled experimental design; implemented and executed Python/PyTorch experiments and analyses; calculated paired comparisons, simultaneous-error statistics, and population-size diagnostics; examined probability relations and the R17 Gaussian decision benchmark; and carried out internal numerical, logical, computational-budget, and literature audits. The corresponding source code and research records accompany this preprint.

AI identity, responsibility, and verification. The designation “AI research co-author” identifies the model’s substantive contributions and distinguishes it from a human researcher; it does not imply human identity, independent legal responsibility, or institutional affiliation with OpenAI. Y. Ueoka is the corresponding human researcher and is responsible for deciding whether to release the manuscript and for reviewing the evidence and claims. Repeated script-based checks and ChatGPT-assisted auditing are *not independent human replication*; no external independent audit is claimed. No experimental verification by Nagi, Akari, or Sui is asserted. The change from R22 to R23 concerns venue-facing attribution and presentation only and introduces no new experiments

or numerical results. The R24 update adds a public-facing link to the R23 reproducibility archive and makes no change to the experiments or analyses.

References

- [1] J. Kaplan et al. (2020), “Scaling Laws for Neural Language Models,” <https://arxiv.org/abs/2001.08361>.
- [2] J. Hoffmann et al. (2022), “Training Compute-Optimal Large Language Models,” <https://arxiv.org/abs/2203.15556>.
- [3] G. S. Becker and K. M. Murphy (1992), “The Division of Labor, Coordination Costs, and Knowledge,” *Quarterly Journal of Economics*, 107(4), 1137–1160, <https://doi.org/10.2307/2118383>.
- [4] S. Postrel (2002), “Islands of Shared Knowledge: Specialization and Mutual Understanding in Problem-Solving Teams,” *Organization Science*, 13(3), 303–320, <https://doi.org/10.1287/orsc.13.3.303.2773>.
- [5] T. G. Dietterich (2000), “Ensemble Methods in Machine Learning,” *Multiple Classifier Systems*, https://doi.org/10.1007/3-540-45014-9_1.
- [6] B. Lakshminarayanan, A. Pritzel, and C. Blundell (2017), “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,” *NeurIPS*, https://papers.nips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.
- [7] L. I. Kuncheva and C. J. Whitaker (2003), “Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy,” *Machine Learning*, 51, 181–207, <https://doi.org/10.1023/A:1022859003006>.
- [8] C. A. Shipp and L. I. Kuncheva (2002), “Relationships Between Combination Methods and Measures of Diversity in Combining Classifiers,” *Information Fusion*, 3(2), 135–148, [https://doi.org/10.1016/S1566-2535\(02\)00051-9](https://doi.org/10.1016/S1566-2535(02)00051-9).
- [9] C. Ivaşcu, R. M. Everson, and J. E. Fieldsend (2022), “Optimising Diversity in Classifier Ensembles,” *SN Computer Science*, 3, 191, <https://doi.org/10.1007/s42979-022-01063-7>.
- [10] E. M. Kim, A. Garg, K. Peng, and N. Garg (2025), “Correlated Errors in Large Language Models,” *Proceedings of Machine Learning Research*, 267, 30038–30066, <https://proceedings.mlr.press/v267/kim25e.html>.
- [11] S. Seto et al. (2026), “Optimal Splitting of Language Models from Mixtures to Specialized Domains,” *Proceedings of Machine Learning Research*, 306, 109364–109387, <https://proceedings.mlr.press/v306/seto26a.html>.
- [12] N. Chen et al. (2026), “Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation,” *Findings of ACL 2026*, 251–306, <https://doi.org/10.18653/v1/2026.findings-acl.13>.
- [13] Y. Kim et al. (2025), “Towards a Science of Scaling Agent Systems,” <https://arxiv.org/abs/2512.08296>.