INFORMATION-THEORETIC QUERY COMPLEXITY FOR LLM ENSEMBLE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting an LLM ensemble is fundamentally about complementarity: a model is valuable when it succeeds where others fail. We study how many evaluations are needed to identify the best fixed-size ensemble and show that comparison difficulty is determined entirely by disagreement. Tasks that both ensembles solve or both miss are uninformative; only outcomes on which exactly one succeeds matter. For binary feedback indicating whether each model succeeds, this yields the exact information rate governing reliable comparison and, when all models are evaluated on the same sampled tasks, the exact asymptotic number of sampled tasks required to identify the best ensemble. The same principle extends to model-ranking feedback. Building on this, we develop a greedy procedure that is asymptotically optimal in the number of accepted tasks for identifying the model that best solves tasks missed by the current ensemble. Yet, this greedy decision can require arbitrarily more sampled tasks under full-vector feedback than identifying the best ensemble directly, even when greedy ultimately selects it. Direct identification faces a separate computational limitation: even when the required number of sampled tasks is polynomially bounded, computing the exact best ensemble in time polynomial in that bound and the number of models on every instance would imply $\mathrm{NP} \subseteq \mathrm{RP}$. Experiments on published LLM evaluations and a fresh multilingual benchmark are consistent with these predictions. Even with the same overall success gap, ensembles can require substantially different numbers of sampled tasks to distinguish depending on how their successes overlap. Exploiting this complementarity can outperform selection by individual accuracy. Compared with the closest existing method, ours requires less evidence to confirm its choices while achieving similar average ensemble quality. Finally, realistic benchmarks can enable improved ensemble selection while containing too few tasks to confirm the choices with high confidence.

1 INTRODUCTION

In multilingual question answering, a model that performs well in English may be weaker in Thai or Greek, while another shows the opposite pattern. More generally, models with similar aggregate scores can succeed on different questions, domains, or reasoning problems. Modern evaluations reflect this heterogeneity: XQuAD and TyDi-QA span languages with different model strengths (Artetxe et al., 2020; Clark et al., 2020), while multi-task and routing benchmarks reveal complementary performance across broader task collections (Cobbe et al., 2021; Hu et al., 2024). When several models can be selected, it therefore matters not only how well each performs individually, but also whether it succeeds where others fail. Discovering such complementarity requires evaluating models on shared tasks, which can be expensive. The question is thus not only *which models to select*, but also *how much evaluation is needed to identify them.*

We study this problem through query complexity. Tasks are drawn from an unknown distribution, each model succeeds or fails on each task, and an ensemble succeeds if at least one member does. We seek the best ensemble of a prescribed size from evaluation feedback. Our main setting observes all models on each sampled task, revealing which models solve or miss the same tasks; we also consider ranking-based feedback. Our central question is: *how many evaluations are needed to identify the best ensemble with high confidence, and what determines this query complexity?*

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

Our contributions. We first characterize the information required to distinguish ensembles (Section 3). Under binary feedback indicating whether each model succeeds, only disagreements matter: tasks that both ensembles solve or both miss provide no information about which is better. We derive the exact information rate governing reliable comparison, determined by how often the ensembles disagree and which succeeds when they do, rather than merely by their overall success gap. We also extend this principle to ranking-based evaluation. Under full-vector feedback, where every model is evaluated on each sampled task, the same quantities yield the exact asymptotic characteristic time in sampled tasks for identifying the best ensemble and a stopping rule that attains it.

Building on this characterization, we develop a fixed-confidence greedy procedure (Section 4) that identifies, at the asymptotically optimal rate in accepted tasks, the model most likely to solve tasks missed by the current ensemble. Repeating this procedure constructs the greedy ensemble with high confidence and guarantees at least a $(1 - 1/e)$ fraction of the optimal ensemble's coverage.

We then establish two limits to efficient ensemble selection (Section 5). Identifying the next greedy model can require arbitrarily more sampled tasks under full-vector feedback than identifying the best ensemble, even when greedy ultimately selects it. Direct identification faces a different limitation: even when the required number of sampled tasks is polynomially bounded, computing the exact optimum in time polynomial in that requirement and the number of models on every instance would imply $\mathrm{NP} \subseteq \mathrm{RP}$. Thus, greedy can waste arbitrarily many samples on intermediate decisions, while statistically efficient direct identification need not be computationally efficient.

Finally, we test these predictions on published LLM evaluation matrices and a fresh multilingual benchmark (Section 6). On the published matrices, ensembles with the same overall success gap can require over four times as many evaluations to distinguish depending on how their successes overlap. Across both published and fresh evaluations, complementarity-aware selection can outperform individual-accuracy ranking, with the fresh-benchmark gain concentrated at $k = 3$. Compared with the closest method that evaluates candidates on missed tasks (Neoh et al., 2026), ours attains similar average ensemble quality but requires less evidence to confirm its choices at the specified confidence level. Finally, realistic benchmarks can improve ensemble selection while containing too few tasks for high-confidence confirmation of the choices. Together, these results show that *where* models succeed matters both for ensemble selection and its evaluation cost.

Related Work. We focus on the closest work and defer broader discussion to Appendix A.

Statistical identification. Our analysis builds on sequential experimental design and fixed-confidence identification, where instance-dependent information rates govern the observations needed to distinguish alternatives (Chernoff, 1959; Kaufmann et al., 2016; Naghshvar & Javidi, 2013; Jourdan et al., 2022; Jourdan & Degenne, 2023; Russo et al., 2025), and on combinatorial pure exploration for structured subsets (Chen et al., 2014; 2017; Du et al., 2021; Ben-David & Sabato, 2022; Wang & Zhu, 2022). Unlike these settings, where subset value is built from individual arm rewards, our non-additive coverage objective depends on overlap in the tasks models solve or miss, yielding instance-dependent complexity determined by disagreement.

Committee selection. Our full-information objective is maximum coverage, for which greedy achieves the classical $(1 - 1/e)$ approximation (Nemhauser et al., 1978; Feige, 1998); forward selection also appears in classical ensemble selection (Caruana et al., 2004). Query-efficient maximum coverage has been studied under incomplete information (Imber et al., 2025; Lindeboom et al., 2026; Zimet et al., 2026), but for a fixed electorate rather than tasks sampled from an unknown distribution. Closest to our setting, Neoh et al. (2026) study the same best-member coverage objective and greedily evaluate candidates on tasks missed by the current ensemble. We instead characterize the exact information for comparing ensembles and the asymptotic sample complexity of direct identification under full-vector feedback, and show that greedy identification can require arbitrarily more sampled tasks while exact direct identification faces a separate computational limitation. Our ranking-based setting is also related to pairwise preferences and Condorcet-style committee selection (Chiang et al., 2024; Ameli et al., 2025; Elkind et al., 2011; Charikar et al., 2025; Song et al., 2026), which primarily rank individual systems or select committees from fixed preference profiles.

LLM ensemble selection and evaluation. Routing and cascading choose a model per input rather than a fixed ensemble (Chen et al., 2024; Hu et al., 2024; Ong et al., 2025). Other work selects LLM ensembles using focal diversity (Tekin et al., 2024) or mutual information between predictions and the target label (Turkmen et al., 2026). Our objective is instead best-member coverage, with infor-

mation theory characterizing the evaluations needed to identify its optimum rather than defining the objective. Active evaluation ranks individual systems by allocating task-model evaluations (Lanctot et al., 2026), whereas we identify a fixed-size ensemble.

2 PRELIMINARIES

Ensembles. For $z \in \mathbb{N}$, write $[z] := \{1, \dots, z\}$, and for a finite set A let $\Delta(A)$ denote the probability simplex over A. Let $C = [m]$ be a set of m models. An *ensemble*, or *committee*, is a subset $S \subseteq C$. For a target size $k \in \{0, \dots, m\}$, let $\mathcal{S}_k := \{S \subseteq C : |S| = k\}$ and $M := |\mathcal{S}_k| = \binom{m}{k}$. Tasks are drawn independently from an unknown distribution D over a task space $\mathcal{D}$. For distributions P, Q on a finite space, write $D_{\mathrm{KL}}(P\|Q) := \sum_x P(x) \log(P(x)/Q(x))$ for their Kullback-Leibler (KL) divergence, and for $p, q \in [0, 1]$ write $\mathrm{kl}(p, q) := p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$ for the Bernoulli KL divergence. For a discrete random variable Y, let $H(Y) := -\sum_y \Pr[Y = y] \log_2 \Pr[Y = y]$ denote its Shannon entropy and $h_2(p) := -p \log_2 p - (1 - p) \log_2(1 - p)$ the binary entropy function. Logarithms are natural unless explicitly written $\log_2$, so KL divergences are measured in nats, whereas entropy and mutual information are measured in bits.

Binary feedback. Each task $d \in \mathcal{D}$ induces a binary outcome for every model, $u(d, c) \in \{0, 1\}$, where $u(d, c) = 1$ means that model c succeeds on task d. The utility of a committee $S \subseteq C$ is whether at least one of its members succeeds, i.e., $U(d, S) := \mathbf{1}\{\exists c \in S : u(d, c) = 1\}$, equivalently $U(d, S) = \max_{c \in S} u(d, c)$ for $S \neq \varnothing$, with $U(d, \varnothing) = 0$. The population coverage of S is $v(S) := \Pr_{d \sim D}[U(d, S) = 1]$, and the optimal size-$k$ value is $\mathrm{OPT}_k := \max_{S \in \mathcal{S}_k} v(S)$. It is convenient to represent a task by the joint success vector $X(d) := (u(d, 1), \dots, u(d, m)) \in \{0, 1\}^m$, and to write P for the distribution of $X(d)$ when $d \sim D$. For any distribution R on $\{0, 1\}^m$, define $v_R(S) := \Pr_{X \sim R}[\exists c \in S : X_c = 1]$. Thus, $v = v_P$. When the maximizer is unique, write $S^*(R) := \arg\max_{S \in \mathcal{S}_k} v_R(S)$. Our main instance-dependent results use *full-vector feedback*: on each fresh task $d \sim D$, the learner observes the entire success vector $X(d)$. Thus, one full-vector observation reveals how all models behave on the same task, including their dependence and disagreement patterns. The sample-complexity results in Section 3 count sampled tasks under full-vector feedback, not individual model-task evaluations.

Adaptive queries. For the general lower bound, let $\mathcal{F} \subseteq \mathcal{S}_k$ be a finite family of possible optimal committees, and let the unknown target $T \in \mathcal{F}$ identify the true optimum, with a fixed rule for breaking ties. Unless stated otherwise, T is drawn from a prior $\mu \in \Delta(\mathcal{F})$. The learner knows $\mathcal{F}$, the query space $\mathcal{Q}$, and the response model, but not T. For each target $t \in \mathcal{F}$ and query $q \in \mathcal{Q}$, let P_t^q denote the distribution of the response. In the main binary setting, full-vector feedback draws a fresh task $d \sim D$ and returns $X(d) \in \{0, 1\}^m$. The learner seeks to identify T by choosing queries and using their responses to guide subsequent choices. An *N-query adaptive strategy* chooses each query based on the previous queries and responses, possibly using internal randomization, and outputs an estimate $\widehat{T}$ after N queries. A *sequential strategy* may instead stop at a data-dependent stopping time. The formal model of the strategy and its sequence of queries and responses (the transcript) is given in Appendix A.6.

Information per query. For a prior $\mu \in \Delta(\mathcal{F})$ and query $q \in \mathcal{Q}$, the information provided by the response Z about the target T is their mutual information, $\mathrm{I}_\mu(T; Z \mid Q = q) := \sum_{t \in \mathcal{F}} \sum_{z \in \mathcal{Z}_q} \mu(t) P_t^q(z) \log_2 \frac{P_t^q(z)}{\sum_{t' \in \mathcal{F}} \mu(t') P_{t'}^q(z)}$. Since an adaptive strategy changes the posterior over targets, we define the *one-query information capacity* $I_{\max} := \sup_{\nu \in \Delta(\mathcal{F})} \sup_{q \in \mathcal{Q}} \mathrm{I}_\nu(T; Z \mid Q = q)$. Thus, $I_{\max}$ bounds the mutual information provided by any allowed query under any posterior encountered during the interaction.

3 INFORMATION-THEORETIC LIMITS OF ENSEMBLE IDENTIFICATION

This section characterizes the information required to identify the best ensemble. We first derive a general lower bound for adaptive identification across feedback models (Section 3.1). Under binary feedback, we show that pairwise comparison difficulty is determined entirely by tasks on which exactly one ensemble succeeds (Section 3.2). Finally, under full-vector feedback, these pairwise quantities yield the exact asymptotic sample complexity of identifying the best ensemble (Section 3.3).

3.1 A GENERAL QUERY-COMPLEXITY LOWER BOUND

We begin with a lower bound that applies to any adaptive feedback model in Section 2. The basic idea is that identifying the optimal ensemble requires enough information to distinguish it from other possible targets, while each query reveals at most $I_{\max}$ bits. The following result makes this tradeoff precise. To state it, we say that an N-query adaptive strategy succeeds with error at most $\varepsilon \in [0, 1]$ if its output $\widehat{T}$ satisfies $\Pr[\widehat{T} \neq T] \leq \varepsilon$. Recall that $\mathcal{F}$ is the set of possible target ensembles. For $|\mathcal{F}| \geq 2$, we consider $\varepsilon \leq 1 - 1/|\mathcal{F}|$, since uniform random guessing already achieves this error; for $|\mathcal{F}| = 1$, the target is known and the bounds are trivial.

Theorem 3.1. *Let T be a random optimal committee taking values in $\mathcal{F} \subseteq \mathcal{S}_k$ with prior μ, and suppose an N-query adaptive strategy identifies T with error at most $\varepsilon \in [0, 1 - 1/|\mathcal{F}|]$, where N counts queries in the feedback model under consideration. Then, $NI_{\max} \geq H_\mu(T) - h_2(\varepsilon) - \varepsilon \log_2(|\mathcal{F}| - 1)$, where $H_\mu(T)$ is the entropy of T and $h_2(\varepsilon)$ is the binary entropy. Hence, if $I_{\max} > 0$, then $N \geq [H_\mu(T) - h_2(\varepsilon) - \varepsilon \log_2(M - 1)]/I_{\max}$. If T is uniform on $\mathcal{F}$, this simplifies to $N \geq [(1 - \varepsilon) \log_2 |\mathcal{F}| - 1]/I_{\max}$.*

Proof sketch. Each query contributes at most $I_{\max}$ bits, so the transcript contains at most $NI_{\max}$ bits. Fano's inequality (Cover & Thomas, 2006) then lower-bounds the information needed to identify one of $|\mathcal{F}|$ targets with error at most ε. See Appendix B.1 for the full proof.

For the full family $\mathcal{F} = \mathcal{S}_k$ of size-k ensembles drawn from m models under a uniform prior, there are $\binom{m}{k}$ possible targets. Hence, for $1 \leq k \leq m/2$ and fixed $\varepsilon < 1$, Theorem 3.1 gives $N = \Omega(k \log(m/k)/I_{\max})$. Thus, the number of candidate ensembles enters only logarithmically, while the query complexity is inversely proportional to the information available from each query. This result concerns exact identification but also yields lower bounds for approximate ensemble selection; see Appendix B.1 for the formal statement and proof.

3.2 EXACT COMPARISON COMPLEXITY UNDER BINARY FEEDBACK

The general lower bound above applies across feedback models but takes the worst case over a family of possible instances: $I_{\max}$ summarizes the most informative query across possible targets without revealing what makes a particular pair of ensembles easy or difficult to distinguish. We now use the coverage structure to obtain an exact, instance-dependent comparison rate under binary feedback. For two ensembles, comparison depends on the two exclusive outcomes, where exactly one ensemble succeeds. The relevant KL projection is captured by $J(x, y) := x \log \frac{2x}{x+y} + y \log \frac{2y}{x+y}$ for $x > y$, and $J(x, y) := 0$ otherwise, with $0 \log 0 = 0$ and $J(0, 0) = 0$. The four-cell projection in Appendix B.2 shows that $J(x, y)$ is the KL distance to reversing the ordering of two exclusive-outcome probabilities: the closest alternative preserves the probabilities of the two agreement outcomes and makes the two exclusive outcomes equally likely.

We apply the four-cell projection to two fixed ensembles $A, B \in \mathcal{S}_k$. For $i, j \in \{0, 1\}$, let $E_{ij}^{A,B}$ be the set of success vectors on which A has utility i and B has utility j: $E_{ij}^{A,B} := \{\mathbf{x} \in \{0, 1\}^m : \mathbf{1}\{\exists c \in A : x_c = 1\} = i, \mathbf{1}\{\exists c \in B : x_c = 1\} = j\}$. For any success-vector distribution Q, let $p_{ij}^Q(A, B) := Q(E_{ij}^{A,B})$ be the probability of this eveny. Thus, $p_{10}^Q(A, B)$ and $p_{01}^Q(A, B)$ are the probabilities that only A and only B succeeds, respectively. Since $v_Q(A) = p_{10}^Q + p_{11}^Q$ and $v_Q(B) = p_{01}^Q + p_{11}^Q$, we have $v_Q(A) - v_Q(B) = p_{10}^Q - p_{01}^Q$. Hence, only these exclusive outcomes determine which ensemble has larger coverage. We define the *exclusive coverage information* as $I_{\text{cov}}^Q(A, B) := J(p_{10}^Q(A, B), p_{01}^Q(A, B))$; for the true distribution P, this is $I_{\text{cov}}^P(A, B)$.

The following result shows that this quantity is exactly the KL distance to reversing the ordering of the ensembles and determines the corresponding sequential lower bound.

Theorem 3.2. *Fix distinct $A, B \in \mathcal{S}_k$ with $v_P(A) > v_P(B)$. Then, $\inf_{Q:, v_Q(A) \leq v_Q(B)} D_{\text{KL}}(P|Q) = I_{\text{cov}}^P(A, B)$, and the infimum is attained by the distribution Q^* induced by the four-cell projection of Appendix B.2. Moreover, consider a sequential test that observes fresh i.i.d. success vectors, stops after τ sampled tasks under full-vector feedback, and decides whether A has larger coverage than B. If $\Pr_P[\text{declare } A \preceq B] \leq \beta$ and $\Pr_Q[\text{declare } A \succ B] \leq \alpha$ for every distribution Q on $\{0, 1\}^m$ satisfying $v_Q(A) \leq v_Q(B)$, where $\alpha, \beta \in (0, 1)$ and $\alpha + \beta < 1$, then $\mathbb{E}_P[\tau] \geq \text{kl}(1 - \beta, \alpha)/I_{\text{cov}}^P(A, B)$.*

Proof sketch. The KL chain rule reduces the closest order-reversing alternative to the four-cell projection of Appendix B.2; a standard change-of-measure argument then gives the sequential lower bound. See Appendix B.3 for the full proof.

The expression for $I^P_{\mathrm{cov}}(A, B)$ also clarifies what makes two ensembles difficult to distinguish. Let $s := p^P_{10}(A, B) + p^P_{01}(A, B)$ be the probability that exactly one ensemble succeeds, and let $\Delta := p^P_{10}(A, B) - p^P_{01}(A, B) = v_P(A) - v_P(B)$ be their coverage gap. For small Δ, a Taylor expansion gives $I^P_{\mathrm{cov}}(A, B) = \Delta^2/(2s) + \mathcal{O}(\Delta^4/s^3)$. Thus, when s is fixed, smaller coverage gaps are harder to distinguish. In fact, even for the same coverage gap, ensembles can have different comparison costs: increasing s makes the exclusive successes more evenly split between the two ensembles and reduces the information available for determining which is better. By contrast, the events $E^{A,B}_{00}$ and $E^{A,B}_{11}$, where both ensembles fail or both succeed, do not affect the comparison rate.

Extension to ordinal feedback. The same exclusive-outcome principle extends to a weighted ordinal setting, where each task ranks the models and an ensemble is evaluated against a sampled challenger. Comparing two ensembles again reduces to a four-outcome problem, and the exact rate depends only on outcomes on which exactly one ensemble succeeds. Appendix B.5 defines the ordinal objective and gives the formal result and proof.

3.3 Exact asymptotic sample complexity under full-vector feedback

We now move from comparing two ensembles to identifying the best size-k ensemble. Under full-vector feedback, each sampled task reveals the complete success vector $X(d) \in \{0, 1\}^m$, informing all candidate ensembles simultaneously. The pairwise rates from Section 3.2 then determine global identification difficulty through the hardest comparison between the optimal ensemble and any rival.

Assume the true success-vector distribution P has a unique optimal ensemble $S^*(P)$. Define the *full-vector information gap* $\Gamma(P) := \min_{S \in \mathcal{S}_k : S \neq S^*(P)} I^P_{\mathrm{cov}}(S^*(P), S)$ and the corresponding *characteristic time* $T_{\mathrm{full}}(P) := 1/\Gamma(P)$. Thus, $\Gamma(P)$ is the information available in one full-vector observation for distinguishing the optimum from its hardest-to-distinguish rival.

We next define a stopping rule based on the empirical counterparts of these quantities. Let $X_1, X_2, \ldots$ be i.i.d. success vectors drawn from P and $\widehat{P}_n$ their empirical distribution. After n observations, let $\widehat{S}_n \in \arg\max_{S \in \mathcal{S}_k} v_{\widehat{P}_n}(S)$, with deterministic tie-breaking, and define $\widehat{\Gamma}_n := \min_{S \in \mathcal{S}_k : S \neq \widehat{S}_n} I^{\widehat{P}_n}_{\mathrm{cov}}(\widehat{S}_n, S)$. The statistic $n\widehat{\Gamma}_n$ is the generalized likelihood ratio against alternatives under which some rival has at least as large coverage as $\widehat{S}_n$. We thus stop once this evidence exceeds a time-uniform confidence threshold. For a target error probability $\delta \in (0, 1)$, let $\beta(n, \delta) := \log(\pi^2 n^2 (\binom{m}{k} - 1)(n + 1)^4/(6\delta))$ and define $\tau_\delta := \inf\{n \geq 1 : n\widehat{\Gamma}_n \geq \beta(n, \delta)\}$, returning $\widehat{S}_{\tau_\delta}$. This is a Chernoff-style stopping rule (Chernoff, 1959) specialized to coverage.

We call an exact-identification rule *δ-correct* if it returns the optimal ensemble with probability at least $1 - \delta$. Here, stopping times count sampled tasks, equivalently full-vector observations, rather than individual model-task evaluations. The following result shows that the rule above is δ-correct and asymptotically optimal, with $\Gamma(P)$ also determining the lower bound for any δ-correct rule.

Theorem 3.3. *Assume P has a unique optimal size-k ensemble $S^*(P)$ and $\Gamma(P) > 0$. Then, for every $\delta \in (0, 1/2)$, every δ-correct exact-identification rule under full-vector feedback satisfies $\mathbb{E}_P[\tau] \geq \mathrm{kl}(1 - \delta, \delta)/\Gamma(P)$. Moreover, the rule above is δ-correct and satisfies $\lim_{\delta \downarrow 0} \mathbb{E}_P[\tau_\delta]/\log(1/\delta) = T_{\mathrm{full}}(P) = 1/\Gamma(P)$.*

Proof sketch. Each rival S induces an alternative instance, arbitrarily close to the projection in Theorem 3.2, under which S is uniquely optimal, yielding the lower bound. For the upper bound, the stopping rule controls the probability of returning an incorrect ensemble and $\widehat{\Gamma}_n \to \Gamma(P)$, so it stops asymptotically after $\log(1/\delta)/\Gamma(P)$ sampled tasks. See Appendix B.4 for the full proof.

Computing the stopping rule requires optimizing over candidate ensembles. The resulting computational difficulty is studied in Theorem 5.2.

Algorithm 1 KL-Rescue-Test(S, A, δ)

Require: Current ensemble S, active set $A \subseteq C \setminus S$, confidence $\delta \in (0,1)$
if $|A| = 1$ **then return** the unique element of A
$n \leftarrow 0$
repeat
  $n \leftarrow n+1$; draw $d_n \sim D_S$ and observe $(u(d_n, c))_{c \in A}$
  update $\widehat{q}_n^S(c)$ for all $c \in A$ and set $\widehat{a}_n \in \arg\max_{c \in A} \widehat{q}_n^S(c)$ with deterministic tie-breaking
until $\min_{b \in A \setminus \{\widehat{a}_n\}} n\widehat{I}_n^S(\widehat{a}_n, b) \geq \beta_{|A|}(n, \delta)$
return $\widehat{a}_n$

4 Statistically optimal greedy ensemble construction

Section 3 characterized the information required to distinguish competing ensembles. We now use this structure for greedy ensemble construction on tasks missed by an ensemble. Fix a current ensemble $S \subseteq C$ with miss rate $\rho(S) := \Pr_{d \sim D}[U(d, S) = 0]$. For $c \in C \setminus S$, define the *rescue rate* as the probability of success on tasks missed by S: $q(c \mid S) := \Pr_{d \sim D}[u(d, c) = 1 \mid U(d, S) = 0]$ when $\rho(S) > 0$, and $q(c \mid S) := 0$ otherwise. As $v(S \cup c) - v(S) = \rho(S) q(c \mid S)$, a greedy step is equivalent to identifying the candidate with the largest rescue rate on tasks missed by S.

This approach is also used by Adaptive-Fail-Greedy (AFG) of Neoh et al. (2026), which evaluates candidates on tasks missed by the current ensemble and eliminates them using rescue-rate confidence bounds. Our procedure instead uses likelihood-ratio evidence. For two candidates, classical sequential testing yields error-controlled decisions with leading-order complexity determined by KL information and a matching lower bound (Wald, 1947; Siegmund, 1985); see Appendices C.1 and C.2. In our setting, each missed task reveals all remaining candidates' outcomes, allowing pairwise evidence for a greedy decision to accumulate simultaneously.

To exploit this shared feedback, let $A \subseteq C \setminus S$ be the active candidates with $|A| = r \geq 2$, and let D_S be the distribution of $d \sim D$ conditioned on $U(d, S) = 0$. For $a, b \in A$, define their *exclusive rescue probabilities* $p_{10}^S(a, b) := \Pr_{d \sim D_S}[u(d, a) = 1, u(d, b) = 0]$ and $p_{01}^S(a, b) := \Pr_{d \sim D_S}[u(d, a) = 0, u(d, b) = 1]$. Since $q(a \mid S) - q(b \mid S) = p_{10}^S(a, b) - p_{01}^S(a, b)$, only tasks missed by S on which exactly one candidate succeeds determine which has the larger rescue rate. Using the function J from Section 3.2, define the *exclusive rescue information* $I_S(a, b) := J(p_{10}^S(a, b), p_{01}^S(a, b))$. By the four-cell projection in Appendix B.2, this is the KL distance to the closest alternative conditional distribution on tasks missed by S for which b has at least as large a rescue rate as a. It measures the information in one such task for distinguishing them.

We estimate this information from the tasks missed by S and retained for candidate comparison, which we call *accepted tasks*. After n i.i.d. tasks $d_1, \ldots, d_n \sim D_S$, let $\widehat{q}_n^S(c) := n^{-1} \sum_{i=1}^n u(d_i, c)$ be the empirical rescue rate of candidate c. For each pair $a, b \in A$, let $\widehat{p}_{10,n}^S(a, b)$ and $\widehat{p}_{01,n}^S(a, b)$ be the empirical frequencies of the two exclusive rescue outcomes, and define $\widehat{I}_n^S(a, b) := J(\widehat{p}_{10,n}^S(a, b), \widehat{p}_{01,n}^S(a, b))$. Thus, $n\widehat{I}_n^S(a, b)$ measures the accumulated evidence that a has a larger rescue rate than b. Algorithm 1 uses these statistics simultaneously: after each accepted task, it takes the candidate $\widehat{a}_n$ with the largest empirical rescue rate as the current winner and stops only when the evidence in its favor against every rival exceeds a time-uniform boundary. For target error probability $\delta \in (0, 1)$, we use $\beta_r(n, \delta) := \log(\pi^2 r(r-1)n^2/(6\delta)) + 4\log(n+1)$.

The boundary controls the probability of incorrectly selecting an inferior candidate at any time. We call a rescue selector δ-*correct* if, whenever there is a unique candidate with the largest rescue rate, it returns that candidate with probability at least $1 - \delta$. To characterize the number of accepted tasks required for such a decision, let $a^* := \arg\max_{c \in A} q(c \mid S)$ and define $\Delta_{\mathrm{KL}}(S, A) := \min_{b \in A \setminus \{a^*\}} I_S(a^*, b)$, the information for distinguishing the best rescuer from its hardest-to-distinguish rival. The following result shows that $\Delta_{\mathrm{KL}}(S, A)$ determines the optimal first-order complexity of the greedy step: Algorithm 1 is δ-correct and achieves the information-theoretic optimum in the expected number of accepted tasks to first order.

Theorem 4.1. *Fix S with $\rho(S) = \Pr_{d \sim D}[U(d, S) = 0] > 0$ and $A \subseteq C \setminus S$ with $|A| \geq 2$. Suppose that the unique maximizer $a^* := \arg\max_{c \in A} q(c \mid S)$ satisfies $\Delta_{\mathrm{KL}}(S, A) > 0$. Then, for any $\delta \in (0, 1)$, Algorithm 1 returns a^* with probability at least $1 - \delta$. Moreover, its number*

τ_S of accepted tasks satisfies $\mathbb{E}[\tau_S] \le (1+o(1))\log(1/\delta)/\Delta_{\mathrm{KL}}(S,A)$ as $\delta \downarrow 0$ with S, A, D fixed. Conversely, for any $\delta \in (0,1/2)$, any δ-correct rescue selector using the same vector feedback on accepted tasks satisfies $\mathbb{E}[\tau] \ge \mathrm{kl}(1-\delta,\delta)/\Delta_{\mathrm{KL}}(S,A) = (1-o(1))\log(1/\delta)/\Delta_{\mathrm{KL}}(S,A)$.

Proof sketch. The stopping threshold controls the probability of incorrectly selecting an inferior candidate, giving δ-correctness. Convergence of the empirical exclusive-rescue information yields the matching upper rate, while a change-of-measure argument against the hardest-to-distinguish rival gives the lower bound. See Appendix C.3 for the full proof.

Theorem 4.1 measures the cost of a greedy step in accepted tasks, i.e., tasks missed by the current ensemble S. Under rejection sampling from D, only a fraction $\rho(S)$ of raw tasks are accepted, giving expected raw-task cost $\mathbb{E}[\tau_S]/\rho(S)$. Thus, $\Delta_{\mathrm{KL}}(S,A)$ determines the information required once a task missed by S is observed, while $\rho(S)$ determines the cost of obtaining such observations.

We now extend the single-step procedure to construct a size-k ensemble. Starting from the empty ensemble, each greedy step runs Algorithm 1 on the remaining candidates using tasks missed by the current ensemble, adds the selected candidate, and continues until k models have been chosen. If the current ensemble has zero miss probability, the algorithm terminates early because its coverage is already one. Each of the at most k rescue-selection calls uses confidence level δ/k, giving an overall error probability of at most δ. See Algorithm 2 in Appendix C.4 for pseudocode.

To state Algorithm 2's guarantee, we compare its output with the population greedy procedure with access to the true rescue rates. Let $S_0^{\mathrm{gr}} := \varnothing$, $a_i^* := \arg\max_{c \notin S_{i-1}^{\mathrm{gr}}} q(c \mid S_{i-1}^{\mathrm{gr}})$ and $S_i^{\mathrm{gr}} := S_{i-1}^{\mathrm{gr}} \cup \{a_i^*\}$ for $i \in \{1,\dots,k\}$. Then, $S_0^{\mathrm{gr}}, \dots, S_k^{\mathrm{gr}}$ is the *population greedy path*, which adds the model with the largest true rescue rate at each step. Define $\Delta_i := \Delta_{\mathrm{KL}}(S_{i-1}^{\mathrm{gr}}, C \setminus S_{i-1}^{\mathrm{gr}})$, the information gap determining the cost of the i-th population greedy choice.

The following result shows that Algorithm 2 follows this population greedy path with high probability. Consequently, its output inherits the standard approximation guarantee for greedy maximum coverage, while the cost of each step is determined by the corresponding information gap Δ_i.

Proposition 4.1. *Assume that for each $i \le k$ the population greedy choice a_i^* is unique, $\rho(S_{i-1}^{\mathrm{gr}}) > 0$, and $\Delta_i > 0$. Then, for every $\delta \in (0,1)$, Algorithm 2 returns S_k^{gr} with probability at least $1-\delta$, and on that event $v(S_k) \ge \big(1-(1-\frac{1}{k})^k\big)\mathrm{OPT}_k \ge (1-1/e)\mathrm{OPT}_k$. Conditional on the previous steps being correct, if τ_i denotes the number of accepted tasks used by the i-th call, then $\mathbb{E}[\tau_i] \le (1+o(1))\log(k/\delta)/\Delta_i$ as $\delta \downarrow 0$ with m, k, D fixed.*

Proof sketch. Each call uses confidence δ/k, so Theorem 4.1 and a union bound imply that all k choices agree with the population greedy path with probability at least $1-\delta$. On this event, the classical greedy guarantee for monotone submodular coverage applies, and the per-step cost follows from Theorem 4.1. See Appendix C.5 for the full proof.

Each accepted task updates all active candidates simultaneously, so the cost of a greedy step is not the sum of separate pairwise comparison costs. Instead, it is determined by $\Delta_{\mathrm{KL}}(S,A)$, the difficulty of distinguishing the best rescuer from its hardest-to-distinguish rival.

5 Limits of statistically optimal ensemble selection

The preceding sections give asymptotically optimal rates for global exact ensemble identification (Theorem 3.3) and single-step greedy identification (Theorem 4.1). Here, we show that these notions of efficiency do not coincide: greedy identification can require arbitrarily more sampled tasks under full-vector feedback even when population greedy returns the global optimum, while exact global identification faces a separate computational limitation.

The local greedy decision can be much harder than global identification: greedy may need to resolve nearly tied marginal contributions that are irrelevant to identifying the optimal ensemble. This gap can be arbitrarily large, with the information for the next greedy choice vanishing while the optimal ensemble remains separated from every rival by information bounded away from zero, even when population greedy returns the unique global optimum.

Theorem 5.1. *There exists a family of success-vector distributions $(P_\varepsilon)_{\varepsilon \in (0,1/20)}$ over three models with target size $k=2$ such that, as $\varepsilon \downarrow 0$, Algorithm 2 using the true rescue rates returns the unique*

optimal ensemble, $\Gamma(P_\varepsilon)$ remains bounded away from zero, and the first-step greedy information gap satisfies $\Delta_{\mathrm{KL}}(\varnothing, C) = \frac{5}{3}\varepsilon^2 + \mathcal{O}(\varepsilon^3)$. Thus, for any fixed $\delta \in (0, 1/2)$, identifying the first greedy choice with error at most δ requires $\Omega(\varepsilon^{-2})$ more sampled tasks under full-vector feedback than globally identifying the optimal ensemble to leading order.

Proof sketch. The construction keeps the exclusive-coverage information separating the optimal ensemble from every rival bounded away from zero while making the first two greedy candidates increasingly hard to distinguish. Thus, $\Delta_{\mathrm{KL}}(\varnothing, C) = \frac{5}{3}\varepsilon^2 + \mathcal{O}(\varepsilon^3)$ while $\Gamma(P_\varepsilon)$ remains bounded away from zero, yielding the $\Omega(\varepsilon^{-2})$ separation. Direct calculation also shows that Algorithm 2 selects the unique optimum using the true rescue rates. See Appendix D.1 for the full proof.

The separation arises because greedy requires high-confidence intermediate decisions that global identification need not resolve. It can occur when the best greedy candidate is nearly tied with its closest rival, making $\Delta_{\mathrm{KL}}(S, A)$ small while the global information gap remains bounded away from zero; a near-tie between suboptimal candidates is insufficient. Thus, statistically optimal greedy steps need not yield statistically optimal global identification.

Statistical optimality also need not imply computational efficiency. The stopping rule in Theorem 3.3 repeatedly solves $\widehat{S}_n \in \arg\max_{S \in \mathcal{S}_k} v_{\widehat{P}_n}(S)$, a Max-$k$-Cover problem. More generally, exact identification requires recovering the optimal ensemble. We show this is computationally hard even when $1/\Gamma(P)$ is polynomially bounded, so the hardness is not statistical.

Theorem 5.2. *If a randomized algorithm, given sample access to any distribution P on $\{0,1\}^m$ with a unique optimal size-k ensemble, returns $S^*(P)$ with probability at least $3/4$ in time polynomial in m and $1/\Gamma(P)$, then $\mathrm{NP} \subseteq \mathrm{RP}$.*

The proof is by reduction from Max-k-Cover (Feige, 1998); see Appendix D.2. Theorem 5.2 concerns exact identification, while polynomial-time approximation remains possible. By Proposition 4.1, greedy returns a $(1 - 1/e)$-approximation with high probability in polynomial time, while $(1 - 1/e + \eta)$-approximation is NP-hard for any fixed $\eta > 0$ (Feige, 1998). Together with Theorem 5.1, this shows two limitations: greedy can waste arbitrarily more samples on intermediate decisions, while exact global identification is computationally intractable. This leaves open the statistical efficiency achievable in polynomial time. The stopping rule of Theorem 3.3 also requires optimizing over rival ensembles to verify the value of $\widehat{\Gamma}_n$; see Appendix D.2.

6 Experiments

We test whether the theoretical information quantities predict ensemble comparison and identification costs, whether using complementarity on missed tasks improves selection with a finite budget, and whether realistically sized benchmarks contain enough tasks for fixed-confidence decisions.

Benchmarks. We use two experimental tracks. *Track A* uses published per-instance evaluation matrices, allowing us to compute the relevant information quantities exactly and simulate i.i.d. sampling from the empirical populations. MMLU-PromptEval (Maia Polo et al., 2024) captures variation across models and prompt configurations, GSM8K (Cobbe et al., 2021) mathematical reasoning, and RouterBench (Hu et al., 2024) heterogeneous performance across evaluation groups; a LiveBench-derived matrix (White et al., 2025) provides weighted ordinal feedback. *Track B* is a fresh evaluation of 16 open models under a common inference protocol on multilingual XQuAD (Artetxe et al., 2020), with a TyDi-QA GoldP subset (Clark et al., 2020) as a robustness check on natively authored multilingual questions. The pool includes multilingual models and models developed for particular languages or language families. Full details are in Appendix E.

Baselines. For budgeted selection, we compare Algorithm 2 with Top-k, which selects by individual empirical accuracy (de-duplicated when configurations share a base model); Adaptive-Fail-Greedy (AFG) (Neoh et al., 2026), the closest method, which likewise evaluates candidates on tasks missed by the current ensemble but uses rescue-rate confidence bounds; and exhaustive-elicitation ERM, a budgeted non-greedy baseline. We also use full-information greedy and the exact test-split optimum OPT_k as references, and focal-diversity pruning (Tekin et al., 2024) in the offline experiments. Appendix E.6 gives full definitions.

Disagreement predicts evaluation cost. Controlled comparisons from MMLU-PROMPTEVAL support Theorem 3.2: changing the probabilities of agreement outcomes while holding those of exclusive outcomes fixed leaves comparison essentially unchanged, whereas at a fixed coverage gap, varying exclusive disagreement changes the information-theoretic difficulty by a factor of 4.20, tracked by the exclusive-outcome test but not success-rate-only bounds. Thus, comparison cost depends on disagreement, not the coverage gap alone. Across 249 settings from MMLU-PROMPTEVAL, ROUTERBENCH, and GSM8K, global identification cost follows the predicted inverse dependence on $\Gamma(P)$ despite finite-confidence lower-order costs. The LiveBench-derived experiment shows the analogous dependence on ordinal exclusive disagreement (Appendices F.1–F.4).

Complementarity improves finite-budget ensemble selection. On the fresh multilingual XQUAD evaluation, selection on missed tasks improves held-out coverage over TOP-k and exhaustive-elicitation ERM at $k = 3$ for sufficiently large budgets, but not meaningfully at $k = 5$ (Appendix G.1). Relative to de-duplicated TOP-k, this approach consistently selects less accurate models with less overlap in the tasks they solve, as expected under best-member coverage (Appendix G.4). This complementarity is not simply language specialization: the optimal size-3 committee combines two multilingual generalists with Typhoon-1.5-8b (Pipatanakul et al., 2023) and improves over the best individual model in all eight languages; no intended specialist is best on its target language, and most of Typhoon's unique successes occur outside Thai (Appendix G.2).

Our confidence criterion requires less evidence than that of AFG. Our exclusive-KL procedure and AFG obtain essentially identical average held-out selection quality on the fresh and published benchmarks, but they do not always select the same committee and differences occur in both directions. In a controlled comparison with the same observations on missed tasks, differing only in the confidence criterion, our method requires less evidence and more often confirms its choices at the specified confidence level. Thus, its advantage is that it needs fewer observations and more often meets the confidence criterion within the available budget (Appendices F.5 and G.4).

Robustness across benchmarks and baselines. The additional controls are consistent with this interpretation. On MMLU-PROMPTEVAL, de-duplicating prompt configurations of the same base model removes most of the apparent advantage over naive TOP-k, making the remaining gain more informative about complementarity rather than configuration redundancy. Focal-diversity pruning shows no systematic improvement over full-information greedy, while the TYDI-QA subset shows the same pattern of complementarity on natively authored multilingual questions but is too small to justify a separate performance claim (Appendices G.4 and G.3).

Useful selection can precede high-confidence confirmation. Finally, useful empirical selection can require substantially less data than fixed-confidence identification. On XQUAD, Algorithm 1 never meets its confidence criterion before exhausting the available training tasks despite the held-out improvement at $k = 3$; on the published matrices, at $\delta = 0.01$, 83 of 124 real greedy decisions (67%) exhaust their benchmark before meeting the confidence criterion, and 226 sampled pools have predicted global stopping times beyond the 4×10^5-task experimental horizon. Thus, a benchmark can contain enough information to improve the selected ensemble while still containing too little evidence to confirm the corresponding decision at the prescribed confidence level (Appendix F.6).

7 CONCLUSIONS AND FUTURE WORK

We characterized the statistical difficulty of LLM ensemble selection through disagreement: under binary and weighted ordinal feedback, only outcomes on which competing ensembles differ provide information for comparison. Under full-vector feedback, these quantities yield the exact asymptotic characteristic time in sampled tasks for exact identification. Separately, our likelihood tests on missed tasks attain the optimal rate in accepted tasks for a greedy step. However, we show that a greedy choice can be arbitrarily harder to identify than the optimal ensemble, while exact global identification faces a separate computational limitation. Experiments are consistent with these results: disagreement predicts evaluation cost, complementarity-aware selection can outperform individual-accuracy selection, and realistic benchmarks can allow useful selection without enough data to confirm the choices with high confidence. Future directions include polynomial-time statistical rates, benchmark design for informative disagreement, heterogeneous evaluation costs, and extending ordinal feedback to partial rankings or noisy pairwise preferences.

AI USE STATEMENT

Generative AI tools were used to assist with writing and presentation, identifying potentially relevant related work, and generating and editing experimental code. The authors consulted the original sources and independently verified all citations and characterizations of prior work. All AI-assisted text and code were reviewed and revised by the authors, and all reported experimental results were independently checked. Generative AI was not used to derive the theoretical results or proofs. These uses are distinct from the language models evaluated as part of the experiments, which constitute the objects of study and are described in the experimental methodology. All analyses, interpretations, and scientific conclusions are the authors' own, and the authors take full responsibility for the content of the paper. The authors independently verified all mathematical statements and proofs.

ETHICS STATEMENT

This work uses publicly available benchmark datasets and pretrained language models and does not involve human participants, user studies, or the collection of private user data. The reported results characterize model behavior only under the fixed prompting, quantization, decoding, and scoring protocols described in the paper; they should not be interpreted as general assessments of the evaluated models, languages, or language communities. Ensemble selection can inherit biases, errors, and uneven subgroup performance from both the candidate pool and the evaluation data. In particular, optimizing aggregate coverage can obscure weaknesses on individual languages or task groups. We therefore report language-specific analyses and recommend stratified evaluation, application-specific safety assessment, and monitoring of subgroup performance before deployment. Moreover, best-member coverage measures whether at least one selected model produces a correct answer, but does not by itself specify how a deployed system should identify or combine the correct output. Practical use therefore requires an appropriate verifier, router, or aggregation mechanism. Any use or redistribution of the datasets and model checkpoints should follow their applicable licenses and terms.

REPRODUCIBILITY STATEMENT

The assumptions, theorem statements, and complete proofs for the theoretical results are provided in the main text and Appendices A–D. Appendix E specifies the datasets and filtering procedures, candidate pools, prompts, model checkpoints, quantization and decoding settings, scoring rule, task construction, train–test splits, random seeds, query-accounting conventions, baselines, metrics, and statistical methodology. Appendices F and G provide the detailed experimental protocols, controls, implementation checks, and additional results, while Appendix H reports the hardware, software, inference configuration, computational cost, and resource requirements.

We retain the complete experimental configuration, including model repository identifiers and revision hashes, prompts, question and paragraph identifiers, data splits, random seeds, evaluation-budget grids, and accounting conventions. The retained materials also include the generated outcome matrices and raw generation logs, scoring and selection code, step-by-step selection logs, intermediate results, machine-readable tables, and figure-generation scripts. All stochastic components are reproducibly seeded, and methods in matched comparisons receive the same realized task sequences. These materials permit reproduction of the reported computational analyses without rerunning the fresh model evaluations; reproducing the inference stage additionally requires the pinned model checkpoints and fixed inference configuration documented in Appendix H.

REFERENCES

Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W Mahoney. A statistical framework for ranking LLM-based chatbots. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4623–4637, 2020.

Noa Ben-David and Sivan Sabato. A fast algorithm for pac combinatorial pure exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6064–6071, 2022.

F Thomas Bruss and James B Robertson. Wald's lemma for sums of order statistics of i.i.d. random variables. *Advances in applied probability*, 23(3):612–623, 1991.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004.

Moses Charikar, Alexandra Lassota, Prasanna Ramakrishnan, Adrian Vetta, and Kangning Wang. Six candidates suffice to win a voter majority. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1590–1601, 2025.

Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 482–534, 2017.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems 27*, 2014.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, 2nd edition, 2006.

Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*, 2024.

Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Springer Science & Business Media, 2009.

Yihan Du, Yuko Kuroki, and Wei Chen. Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7262–7270, 2021.

Edith Elkind, Jérôme Lang, and Abdallah Saffidine. Choosing collectively optimal sets of alternatives based on the condorcet criterion. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 186–191, 2011.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. Airavata: Introducing hindi instruction-tuned LLM. *arXiv preprint arXiv:2401.15006*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. `https://zenodo.org/records/12608602`, 2024.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor

Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on Apple silicon. https://github.com/ml-explore/mlx, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Aviram Imber, Jonas Israel, Markus Brill, and Benny Kimelfeld. Approval-based committee voting under incomplete information. *Artificial Intelligence*, 347:104381, 2025.

Jacques Janssen and Raimondo Manca. Renewal theory. *Applied Semi-Markov Processes*, pp. 45–104, 2006.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Marc Jourdan and Rémy Degenne. Non-asymptotic analysis of a ucb-based top two algorithm. *Advances in Neural Information Processing Systems*, 36:68980–69020, 2023.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 35:26791–26803, 2022.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Marc Lanctot, Kate Larson, Michael Kaisers, Quentin Berthet, Ian Gemp, Manfred Diaz, Roberto-Rafael Maura-Rivero, Yoram Bachrach, Anna Koop, and Doina Precup. Soft condorcet optimization for ranking of general agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1253–1262, 2025.

Marc Lanctot, Kate Larson, Ian Gemp, and Michael Kaisers. Active evaluation of general agents: Problem definition and comparison of baseline algorithms. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 753–762, 2026.

Feline Lindeboom, Martijn Brehm, Davide Grossi, and Pradeep K. Murukannaiah. Diverse committees with incomplete or inaccurate approval ballots. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 3704–3712, 2026.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems 37*, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

Ketan Mulmuley, Umesh V. Vazirani, and Vijay V. Vazirani. Matching is as easy as matrix inversion. *Combinatorica*, 7(1):105–113, 1987.

Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.

Shintaro Nakamura and Masashi Sugiyama. Fixed-budget real-valued combinatorial pure exploration of multi-armed bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 1225–1233. PMLR, 2024a.

Shintaro Nakamura and Masashi Sugiyama. Thompson sampling for real-valued combinatorial pure exploration of multi-armed bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14414–14421, 2024b.

Shintaro Nakamura and Masashi Sugiyama. A fast algorithm for the real-valued combinatorial pure exploration of the multi-armed bandit. *Neural computation*, 37(2):294–310, 2025.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Tzeh Yuan Neoh, Nicholas Teh, Je Qin Chooi, Paul W. Goldberg, and Milind Tambe. Efficient ensemble selection from binary and pairwise feedback. *arXiv preprint arXiv:2605.09588*, 2026.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 294–304, 2024.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. Vikhr: The family of open-source instruction-tuned large language models for russian. *arXiv preprint arXiv:2405.13929*, 2026.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M. Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, volume 2025, pp. 34433–34448, 2025.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. Typhoon: Thai large language models. *arXiv preprint arXiv:2312.13951*, 2023.

Alessio Russo, Yichen Song, and Aldo Pacchiano. Pure exploration with feedback graphs. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 2025.

David Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer, New York, 1985.

Haoyu Song, Thành Nguyen, and Young-San Lin. A few good choices. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 4861–4874. SIAM, 2026.

Selim Furkan Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11951–11966, 2024.

Yigit Turkmen, Baturalp Buyukates, and Melih Bastopcu. Don't always pick the highest-performing model: An information theoretic view of LLM ensemble selection. In *ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems*, 2026.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. Meltemi: The first open large language model for greek. *arXiv preprint arXiv:2407.20743*, 2024.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3):326–339, 1948.

Siwei Wang and Jun Zhu. Thompson sampling for (combinatorial) pure exploration. In *International Conference on Machine Learning*, pp. 23470–23483. PMLR, 2022.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh Agrawal, Sandeep Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yanpeng Li, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2025.

Sumeth Yuenyong, Kobkrit Viriyayudhakorn, Apivadee Piyatumrong, and Jillaphat Jaroenkantasima. OpenThaiGPT 1.5: A thai-centric open source large language model. In *International Conference on Applications of Natural Language to Information Systems*, pp. 210–219. Springer, 2025.

Itay Asher Zimet, Shiri Alouf-Heffetz, and Nimrod Talmon. Query-based committee selection. In Matteo Baldoni, Franziska Klügl, Andrei Olaru, Alexandru Sorici, and Adina Magda Florea (eds.), *Multi-Agent Systems*, pp. 124–141, Cham, 2026. Springer Nature Switzerland.

A Additional related work

Our analysis is related to several lines of work. We discuss sequential experimental design and fixed-confidence learning in Appendix A.1, and contrast our coverage objective with the additive objectives of combinatorial pure exploration in Appendix A.2. Appendix A.3 reviews coverage-based and query-efficient committee selection, including the closest setting of Neoh et al. (2026). Appendix A.4 then discusses pairwise preferences and Condorcet-style committee selection. Finally, Appendix A.5 considers the broader literature on LLM ensemble selection, routing, and evaluation.

A.1 Sequential experimental design and fixed-confidence learning

Our analysis of the number of queries required is closely related to sequential experimental design and fixed-confidence identification. Classical sequential testing asks how quickly one can distinguish competing hypotheses while controlling the probability of error (Wald, 1945; 1947; Wald &

Wolfowitz, 1948). Chernoff's sequential design of experiments extends this view to settings in which the learner can choose which experiment to perform, so the information supplied by each possible observation becomes part of the statistical problem itself (Chernoff, 1959). Related ideas appear in active hypothesis testing and controlled sensing, where sensing actions are chosen to acquire information about an unknown hypothesis (Naghshvar & Javidi, 2013; Nitinawarat et al., 2013).

Fixed-confidence best-arm identification similarly studies how statistical difficulty varies across instances. Information-theoretic change-of-measure arguments give lower bounds that depend on the alternatives that must be distinguished (Kaufmann et al., 2016), while Garivier & Kaufmann (2016) characterize the optimal asymptotic complexity for one-parameter bandits and introduce the asymptotically optimal Track-and-Stop strategy. Subsequent work has developed alternative asymptotically efficient sampling rules, notably the Top-Two family, and strengthened their guarantees beyond the original asymptotic analyses (Jourdan et al., 2022; Jourdan & Degenne, 2023). More recent extensions show that the same information-theoretic approach also applies to fixed-confidence identification with more complex feedback and decision rules (Russo et al., 2025). Our analysis builds on this broad principle: a decision is difficult when the true instance is hard to distinguish from an alternative that changes the correct answer. We combine this principle with classical sequential likelihood-ratio methods (Wald, 1947; Siegmund, 1985).

The coverage structure of ensemble selection allows us to make these information-theoretic quantities explicit. Classical and modern work on fixed-confidence identification characterizes the difficulty of an instance by optimizing over experiments, sampling allocations, and alternative instances, and develops procedures that attain or approach the resulting limits for each instance (Chernoff, 1959; Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Naghshvar & Javidi, 2013; Nitinawarat et al., 2013; Jourdan et al., 2022; Jourdan & Degenne, 2023; Russo et al., 2025). Our coverage setting has additional structure that lets us find the relevant alternative instance in *closed form*. For two competing ensembles, the closest alternative that reverses their ordering is determined entirely by the tasks on which exactly one ensemble succeeds (Theorem 3.2). Hence, tasks that both ensembles solve or both miss carry no information about their ordering, while the frequency and direction of their disagreements determine how many sampled tasks are needed. Under full-vector feedback, this pairwise characterization also gives the exact asymptotic characteristic time, measured in sampled tasks, for identifying the best ensemble among all candidates, together with a stopping rule that attains it (Theorem 3.3). Thus, whereas the general fixed-confidence literature provides information-theoretic characterizations and increasingly general procedures for attaining them, our results identify the specific task outcomes that determine the difficulty of ensemble selection and quantify their contribution to the required number of evaluations.

A.2 Combinatorial pure exploration

Combinatorial pure exploration (CPE) extends fixed-confidence identification from selecting a single arm to identifying the best member of a structured family of subsets. Chen et al. (2014) introduced a general CPE framework covering decision classes such as fixed-size subsets, matchings, spanning trees, and paths, and developed algorithms for classes with access to an offline maximization procedure. Chen et al. (2017) subsequently studied the Best-Set problem and gave a lower bound for each instance together with a sampling algorithm matching it up to a logarithmic factor in the number of feasible sets. Subsequent work has broadened the feedback and computational models. For example, Du et al. (2021) study full-bandit and partial linear feedback, Ben-David & Sabato (2022) develop a computationally efficient PAC procedure, and Wang & Zhu (2022) obtain asymptotically optimal complexity results for Thompson sampling in combinatorial pure exploration. More recent work continues to study computationally efficient and instance-dependent CPE under real-valued feedback, including exponentially large sets of possible actions (Nakamura & Sugiyama, 2024a;b; 2025).

These works are closely related to ours because they aim to identify a subset of candidates rather than an individual candidate. However, the statistical structure of the objective is fundamentally different. In standard CPE, each arm has an unknown reward mean and the value of a feasible set is the sum of those means. Estimating the individual arm means therefore provides the basic information needed to compare feasible sets. In contrast, our ensemble objective is a coverage probability: a committee succeeds whenever at least one of its members succeeds. Its value depends not only on individual model accuracies, but on which tasks they solve or miss together. Consequently, two

model pools with identical individual accuracies can have different optimal ensembles because their models complement one another in different ways.

This difference also changes what determines instance difficulty. In standard CPE, the alternatives to the optimal set are described by the arm means and the resulting differences in set values. In contrast, individual model accuracies are insufficient to characterize the difficulty of ensemble coverage. We show that the information for distinguishing two ensembles comes from tasks solved by exactly one of them (Section 3); tasks that both solve or both miss provide no information about their ordering. Under full-vector feedback, these disagreement probabilities also give the exact asymptotic characteristic time in sampled tasks for identifying the best ensemble among all candidates. Thus, while CPE provides the natural framework for fixed-confidence identification among a collection of feasible subsets, our coverage objective leads to a different notion of instance difficulty, determined by complementarity and disagreement between models rather than by their individual reward means alone.

A.3 Coverage-based and query-efficient committee selection

When all task-model outcomes are observed, selecting a size-k ensemble under our binary objective is a maximum-coverage problem. The classical greedy algorithm achieves the $(1-1/e)$ approximation guarantee for monotone submodular coverage, and this factor is tight under standard complexity assumptions (Nemhauser et al., 1978; Feige, 1998). Classical ensemble-selection methods also use forward selection over a fully evaluated model library; e.g., Caruana et al. (2004) build ensembles by repeatedly adding models to a validation-set ensemble. Their method does not use the same objective: it selects with replacement and combines predictions rather than taking the best member, but its forward selection step provides a natural reference for the fully observed coverage problem.

The more relevant literature considers what happens when the information needed to evaluate a committee is only partially available. In approval-based committee voting, knowing every voter's approvals gives the same maximum-coverage objective, and recent work studies how committee selection changes when ballots are incomplete or inaccurate. Imber et al. (2025) study the computational consequences of incomplete approval information, including whether committees could win or must win given the available information. Lindeboom et al. (2026) consider the maximum-coverage formulation directly and study both non-adaptive and adaptive querying of incomplete approval ballots, obtaining query-complexity bounds and a greedy algorithm. Zimet et al. (2026) study a different query model in which structured questions are used to refine voters' preferences under a limited query budget. These works show that incomplete information can make deciding what to query an essential part of committee selection.

Our setting differs from these fixed-election models in a fundamental way. The electorate is not a fixed set of voters whose ballots can be queried or completed. Instead, tasks are drawn from an unknown task distribution, and model evaluations reveal samples from that distribution. The statistical question is therefore not only how to complete missing information in a known election, but how many evaluations are required to identify the best ensemble for the underlying task population.

The closest prior work with tasks drawn from a distribution is that by Neoh et al. (2026), who study ensemble selection from sampled tasks under binary and pairwise feedback. Their binary objective is the same best-member coverage objective used here, and their query model allows sampled tasks to be stored and queried adaptively. They develop baselines that query all outcomes and a greedy algorithm, Adaptive-Fail-Greedy (AFG), which concentrates candidate evaluations on tasks missed by the current committee. Their analysis gives worst-case query lower bounds, an instance-dependent upper bound for the candidate-evaluation cost of greedy selection, and a matching gap-dependent lower bound up to logarithmic factors.

Our analysis differs from this prior work in what it identifies as the statistical difficulty of a selection decision. The guarantees of Neoh et al. (2026) depend on differences between candidates' probabilities of solving a task missed by the current ensemble. Their upper and lower bounds scale with the inverse square of these differences, up to logarithmic factors. In contrast, we characterize the exact information rate for comparing two ensembles: the closest alternative that reverses their ordering changes only the probabilities of the two outcomes on which exactly one ensemble succeeds (Theorem 3.2). The resulting KL rate depends not only on the coverage gap but also on how frequently the ensembles disagree. Thus, two comparisons with the same difference in overall coverage can

require different numbers of evaluations. Our experiments make this difference concrete. We construct comparisons in which two ensembles have the same difference in overall success rates but differ in how often one succeeds when the other fails. Our exact rate predicts that the number of sampled tasks needed to distinguish them varies by a factor of 4.2, whereas a conventional bound that uses only their overall success rates changes by just 1.04. This illustrates why the coverage gap alone cannot capture the evaluation cost identified by our analysis.

The scope of the problem is also different. Neoh et al. (2026) develop a greedy procedure with an approximation guarantee and characterize the cost of the individual selection steps, whereas we also study exact identification of the globally optimal ensemble. Under full-vector feedback, where one sampled task reveals the outcomes of all models on that task, we show that the pairwise disagreement rates combine into an exact asymptotic characteristic time for global identification and give a stopping rule that attains it (Theorem 3.3). This makes it possible to characterize the cost of exact global selection without treating the problem as a sequence of greedy decisions.

This leads to two further differences. First, we show that making every greedy decision statistically optimal does not make the overall selection procedure statistically optimal: there are instances where identifying the next greedy model requires arbitrarily more sampled tasks under full-vector feedback than identifying the optimal ensemble directly (Theorem 5.1). Second, our computational result concerns exact global identification: even when the full-vector characteristic time is polynomially bounded, computing the exact optimum in time polynomial in that bound on every instance would imply $\mathrm{NP} \subseteq \mathrm{RP}$ (Theorem 5.2). These results show limitations that are not described by the query-efficient greedy guarantees of prior work.

In the binary-feedback model of Neoh et al. (2026), each query evaluates one model on one sampled task and reveals only whether that model succeeds or fails. However, a sampled task can be stored and later reused: the learner may evaluate additional models on the same task, choosing them adaptively based on the outcomes observed so far. Thus, the learner can observe all model outcomes on a stored task, but this requires one query per model. In contrast, our full-vector feedback treats the outcomes of all models on a sampled task as a single statistical observation. Our evaluation count still includes each observed model-task outcome separately, so we do not claim a reduction in the underlying number of model evaluations.

A.4 Pairwise preferences and committee selection

Our ordinal feedback setting is also related to work on learning from pairwise preferences. Pairwise comparisons are widely used to evaluate and rank AI systems, including preference-based LLM evaluation and arena-style benchmarks (Chiang et al., 2024; Ameli et al., 2025). Recent work has also studied ranking general AI agents from task-level comparisons, including Soft Condorcet Optimization, which formulates agent ranking directly from pairwise evaluation data (Lanctot et al., 2025). These methods focus on estimating or optimizing a ranking of individual systems. In contrast, our ordinal analysis asks which *set* of models should be selected when the value of a committee depends on whether one of its members ranks highly on a task.

The closest analysis of committee selection is again Neoh et al. (2026), who study ensemble selection from pairwise feedback. For a randomly drawn task, they compare a committee with an individual model outside it by asking whether the committee's best-performing member is preferred to that model. They call a committee ϑ-winning, for a threshold $\vartheta \in [0, 1]$, if this happens with probability at least ϑ for *every* model outside the committee. Thus, ϑ specifies the minimum win probability that the committee must achieve against its strongest outside competitor. Because maximizing this worst-case win probability is not a submodular problem, they introduce a weighted ordinal-coverage objective in its place. This objective is submodular and can be optimized by querying tasks missed by the current committee.

Our ordinal result addresses a different question. Instead of using weighted ordinal coverage as an alternative objective to obtain a query-efficient approximation guarantee, we study how many observations are required to distinguish two committees under that objective. We show that, as in binary coverage, only outcomes on which exactly one committee succeeds against the sampled comparison model provide information for distinguishing them. This gives an exact information rate for the comparison.

Our ordinal setting is also related to the social-choice literature on Condorcet-style committee selection. Elkind et al. (2011) introduced Condorcet winning sets, which require a committee to defeat every outside candidate by majority preference. Subsequent work has substantially improved the guarantees on the existence and size of such sets, including recent constant-size bounds (Charikar et al., 2025; Song et al., 2026). These results are useful background for the preference objective used to select committees, but they study fully observed preferences and existence or optimization of Condorcet-style committees rather than the statistical cost of learning from sampled task comparisons. We therefore use this literature as conceptual background rather than as a direct baseline for our query-complexity results.

More broadly, Lanctot et al. (2026) study active evaluation, where the evaluator adaptively chooses which agents to evaluate on which tasks in order to recover an accurate ranking with fewer observations. Our problem also concerns the cost of allocating evaluations across models and tasks, but the target is different: rather than ranking individual models, we seek a fixed-size ensemble whose value depends on complementarity among its members.

A.5 LLM ENSEMBLES, ROUTING AND EVALUATION

A growing literature exploits differences among LLMs to improve quality or reduce inference cost. One approach is routing or cascading, where different models are invoked for different inputs. Chen et al. (2024) learn cost-aware cascades that can escalate a query across LLMs, while Hu et al. (2024) introduce ROUTERBENCH to benchmark multi-LLM routing and release per-instance outcomes from multiple models for training and evaluating routers. More recent learned routers similarly use query-level information to choose among models with different quality and cost profiles (Ong et al., 2025). These methods use differences among models at inference time: the system decides which model to use for a particular query.

Our problem is different. We do not learn a rule that maps each incoming query to a model. Instead, before deployment, we seek a fixed-size ensemble whose members collectively succeed on as much of the task distribution as possible, and ask how many evaluations are needed to identify that ensemble with high confidence. This distinction matters because routing can exploit features of the current input, whereas our selection problem must discover *complementarity* across the task population: a model can be valuable even if it is rarely the strongest individually, provided that it succeeds on tasks the other selected models miss. Our query-complexity results characterize the evaluation cost of discovering which models complement one another in this way.

A second line of work selects or combines multiple LLMs rather than routing each query to a single model. Tekin et al. (2024) propose LLM-TOPLA, which uses a focal-diversity criterion to select sub-ensembles and then learns how to combine their outputs. This work is particularly close to ours in its emphasis on error diversity: its focal diversity measure is computed from cases on which a designated model fails, so it rewards models with different error patterns. However, the statistical question is different. LLM-TOPLA uses diversity as a criterion for constructing an ensemble from a pool of evaluated models and then focuses on combining its outputs. In contrast, our coverage objective directly values whether at least one selected model succeeds, and our theory asks how many task evaluations are required to determine which ensemble maximizes this objective. Moreover, we show that the complementarity relevant to this identification problem can be expressed exactly in information-theoretic terms: when two ensembles are compared, only tasks solved by exactly one of them provide information about their ordering (Section 3).

Turkmen et al. (2026) share the motivation that selecting the individually strongest LLMs can produce an ensemble whose members tend to make the same errors because model errors are correlated. However, the role of information theory and the selection objective are different. They select a fixed-size ensemble by maximizing the mutual information between the true label and the selected models' predictions, and use this criterion to capture redundancy among model outputs and improve the prediction produced by the ensemble. In contrast, our objective is best-member coverage: an ensemble is valuable when at least one of its members succeeds on a task, and information theory measures the number of evaluations needed to identify the best such ensemble rather than defining which ensemble is best. This leads to a different notion of complementarity. Whereas Turkmen et al. (2026) measure redundancy through dependence among model predictions, our analysis shows that the information for comparing two ensembles comes from tasks solved by exactly one of them (Sec-

tion 3). We use this structure to obtain exact fixed-confidence comparison and global-identification rates, rather than using an information measure to approximate the ensemble objective.

Finally, our work is related to methods that reduce the cost of evaluating LLMs themselves. Maia Polo et al. (2024) study efficient multi-prompt evaluation: rather than exhaustively evaluating every model across a large collection of prompt variants, they estimate each model's performance distribution under a limited evaluation budget. Their goal is therefore to estimate model performance efficiently, whereas ours is to identify a *set* of models with the highest coverage and characterize the evaluations required to do so. Nevertheless, released per-instance evaluation matrices such as ROUTERBENCH (Hu et al., 2024) and MMLU-PROMPTEVAL (Maia Polo et al., 2024) are especially useful for our setting because they retain which models succeed on the same tasks. This information about outcomes on the same tasks makes it possible to observe complementarity and disagreement rather than only aggregate model scores, and we use these matrices to test the predictions of our theory in Section 6.

A.6 ADAPTIVE STRATEGIES AND PROBABILITIES OF QUERY-RESPONSE HISTORIES

We now define the adaptive-strategy model used in the information-theoretic arguments. For each query $q \in \mathcal{Q}$, let $\mathcal{Z}_q$ denote its set of possible responses, and recall that P_t^q is the response distribution under target $t \in \mathcal{F}$. An *N-query adaptive strategy* is a tuple $\mathcal{A} = (\psi_1, \dots, \psi_N, \varphi)$ of query rules and a final decision rule. Let ξ be an internal randomization seed independent of T. The first query is $Q_1 = \psi_1(\xi)$. For $n \geq 2$, after observing $\mathcal{H}_{n-1} := (Q_1, Z_1, \dots, Q_{n-1}, Z_{n-1})$, the learner chooses $Q_n = \psi_n(\mathcal{H}_{n-1}, \xi)$ and observes $Z_n \sim P_T^{Q_n}$ conditionally independently of the past given (T, Q_n). After N queries, it outputs $\widehat{T} = \varphi(\mathcal{H}_N, \xi)$. A *sequential strategy* instead uses a stopping time τ adapted to the observed history and returns a decision based on the observations collected before stopping and its internal randomization.

For likelihood-ratio and absolute-continuity arguments, we express the probability of the resulting query-response history as a product. Conditional on ξ, the query sequence is determined by the observed history, and the probability of a finite query-response history is the product of the corresponding response probabilities. In particular, the factor contributed by the seed is identical under every instance. We state this product formula below.

Lemma A.1. *Fix a randomized adaptive strategy and an instance θ. For a finite query-response history $h = (a_1, z_1, \dots, a_n, z_n)$, let $\Xi_h := \{\xi : \psi_j(h_{j-1}, \xi) = a_j \text{ for every } j \leq n\}$ be the set of random seeds consistent with the queries in h. Then, for every measurable set Ξ of seeds,*

$$\Pr_\theta[\mathcal{H}_n = h, \xi \in \Xi] = \Pr[\xi \in \Xi_h \cap \Xi] \prod_{j=1}^{n} P_\theta^{a_j}(z_j).$$

In particular, the seed factor is independent of θ. Hence, for two instances θ, θ', the laws of $(\mathcal{H}_n, \xi)$ are mutually absolutely continuous whenever $P_\theta^a \ll P_{\theta'}^a$ and $P_{\theta'}^a \ll P_\theta^a$ for every query a that can occur.

Proof. Condition on the random seed ξ. Along a fixed history h, the query sequence is then deterministic, and the event that the strategy chooses a_j after history h_{j-1} for every $j \leq n$ is precisely $\{\xi \in \Xi_h\}$. At each step, conditional on the target instance, seed, preceding history, and selected query, the response has the corresponding distribution $P_\theta^{a_j}$. Multiplying these conditional probabilities gives

$$\Pr_\theta[\mathcal{H}_n = h \mid \xi] = \mathbf{1}\{\xi \in \Xi_h\} \prod_{j=1}^{n} P_\theta^{a_j}(z_j).$$

Integrating over $\xi \in \Xi$ gives the claimed product formula. The final statement follows because the seed factor is common to all instances. □

B Omitted proofs for Section 3

B.1 Proof of Theorem 3.1

Theorem 3.1. *Let T be a random optimal committee taking values in $\mathcal{F} \subseteq \mathcal{S}_k$ with prior μ, and suppose an N-query adaptive strategy identifies T with error at most $\varepsilon \in [0, 1 - 1/|\mathcal{F}|]$, where N counts queries in the feedback model under consideration. Then, $NI_{\max} \geq H_\mu(T) - h_2(\varepsilon) - \varepsilon \log_2(|\mathcal{F}| - 1)$, where $H_\mu(T)$ is the entropy of T and $h_2(\varepsilon)$ is the binary entropy. Hence, if $I_{\max} > 0$, then $N \geq [H_\mu(T) - h_2(\varepsilon) - \varepsilon \log_2(M - 1)]/I_{\max}$. If T is uniform on $\mathcal{F}$, this simplifies to $N \geq [(1 - \varepsilon) \log_2 |\mathcal{F}| - 1]/I_{\max}$.*

Proof. Let ξ be the learner's internal random seed, independent of T. Since the final decision is based on $(\mathcal{H}_N, \xi)$ and $\mathrm{I}_\mu(T; \xi) = 0$, the chain rule gives

$$\mathrm{I}_\mu(T; \mathcal{H}_N, \xi) = \mathrm{I}_\mu(T; \mathcal{H}_N \mid \xi).$$

Conditioned on ξ, each query Q_n is a deterministic function of the preceding history $(\mathcal{H}_{n-1}, \xi)$, so $\mathrm{I}_\mu(T; Q_n \mid \mathcal{H}_{n-1}, \xi) = 0$. Applying the chain rule over the N queries therefore yields

$$\mathrm{I}_\mu(T; \mathcal{H}_N \mid \xi) = \sum_{n=1}^{N} \mathrm{I}_\mu(T; Q_n, Z_n \mid \mathcal{H}_{n-1}, \xi) = \sum_{n=1}^{N} \mathrm{I}_\mu(T; Z_n \mid \mathcal{H}_{n-1}, Q_n, \xi).$$

Conditioned on a history $\mathcal{H}_{n-1} = h$, query $Q_n = q$, and seed $\xi = \xi_0$, the conditional distribution of T is some posterior $\nu_{h,\xi_0} \in \Delta(\mathcal{F})$, while the response distributions are $(P_t^q)_{t \in \mathcal{F}}$. By the definition of $I_{\max}$,

$$\mathrm{I}_\mu(T; Z_n \mid \mathcal{H}_{n-1} = h, Q_n = q, \xi = \xi_0) = \mathrm{I}_{\nu_{h,\xi_0}}(T; Z \mid Q = q) \leq I_{\max}.$$

Averaging over the history, query, and seed and summing over n gives

$$\mathrm{I}_\mu(T; \mathcal{H}_N, \xi) = \mathrm{I}_\mu(T; \mathcal{H}_N \mid \xi) \leq NI_{\max}.$$

It remains to obtain a lower bound on the information required for correct identification. Since $\widehat{T} = \varphi(\mathcal{H}_N, \xi)$ is a function of $(\mathcal{H}_N, \xi)$, Fano's inequality (Cover & Thomas, 2006) gives

$$H_\mu(T \mid \mathcal{H}_N, \xi) \leq h_2(P_e) + P_e \log_2(|\mathcal{F}| - 1),$$

where $P_e := \Pr[\widehat{T} \neq T]$. The function $f(p) := h_2(p) + p \log_2(|\mathcal{F}| - 1)$ is non-decreasing on $[0, 1 - 1/|\mathcal{F}|]$, so $P_e \leq \varepsilon \leq 1 - 1/|\mathcal{F}|$ yields

$$H_\mu(T \mid \mathcal{H}_N, \xi) \leq h_2(\varepsilon) + \varepsilon \log_2(|\mathcal{F}| - 1).$$

Hence,

$$\mathrm{I}_\mu(T; \mathcal{H}_N, \xi) = H_\mu(T) - H_\mu(T \mid \mathcal{H}_N, \xi) \geq H_\mu(T) - h_2(\varepsilon) - \varepsilon \log_2(|\mathcal{F}| - 1).$$

Combining the two bounds on $\mathrm{I}_\mu(T; \mathcal{H}_N, \xi)$ proves the first bound of Theorem 3.1. Dividing by $I_{\max}$ when it is positive gives the second bound; the uniform case follows from $H_\mu(T) = \log_2 |\mathcal{F}|$ and $h_2(\varepsilon) \leq 1$. □

The preceding result concerns exact identification, but also yields lower bounds for approximate ensemble selection. For an approximation tolerance $\varepsilon > 0$, if every incorrect ensemble is more than 2ε worse than the optimum, then an ε-optimal output necessarily identifies the optimum. Formally, a finite family of instances indexed by $t \in \mathcal{F} \subseteq \mathcal{S}_k$ is *γ-separated* if, under instance t, committee t is uniquely optimal and $v_t(t) - v_t(S) > \gamma$ for every $S \in \mathcal{S}_k$ with $S \neq t$, where $v_t(S)$ denotes the probability that committee S succeeds on a task drawn under instance t.

Corollary B.1 (Lower bound for approximate ensemble selection)**.** *Let $\varepsilon > 0$ and $\delta \in (0, 1)$. Consider a 2ε-separated family indexed by $\mathcal{F} \subseteq \mathcal{S}_k$ with a uniform target. Then, any adaptive algorithm that returns an ε-optimal ensemble with probability at least $1 - \delta$ satisfies $NI_{\max} \geq (1 - \delta) \log_2 |\mathcal{F}| - 1$, where N counts adaptive queries in the feedback model under consideration. Consequently, when $I_{\max} > 0$, $N \geq [(1 - \delta) \log_2 |\mathcal{F}| - 1]/I_{\max}$.*

Proof. Under instance t, every $S \neq t$ has $v_t(S) < v_t(t) - 2\varepsilon$, so any S with $v_t(S) \geq v_t(t) - \varepsilon$ equals t. Hence, the algorithm identifies T with probability at least $1 - \delta$, and Theorem 3.1 applies.

B.2 FOUR-CELL EXCLUSIVE PROJECTION

Lemma B.1. *Let $p = (p_{00}, p_{10}, p_{01}, p_{11})$ be a probability vector with $p_{10} > p_{01}$, and let $\mathcal{Q}_0 := \{q \in \Delta(\{0,1\}^2) : q_{10} \le q_{01}\}$. Then, $\min_{q\in\mathcal{Q}_0} D_{\mathrm{KL}}(p\|q) = J(p_{10}, p_{01})$, with the unique minimizer $q^*_{00} = p_{00}$, $q^*_{11} = p_{11}$, and $q^*_{10} = q^*_{01} = (p_{10} + p_{01})/2$.*

Proof. Write $a := p_{10} + p_{01}$ and, for feasible q, $s := q_{10} + q_{01}$. Decompose the divergence as

$$D_{\mathrm{KL}}(p\|q) = \Big[p_{00} \log \tfrac{p_{00}}{q_{00}} + p_{11} \log \tfrac{p_{11}}{q_{11}}\Big] + \Big[p_{10} \log \tfrac{p_{10}}{q_{10}} + p_{01} \log \tfrac{p_{01}}{q_{01}}\Big].$$

By the log-sum inequality the first bracket is at least $(1-a)\log\frac{1-a}{1-s}$, with equality if and only if (q_{00}, q_{11}) is proportional to (p_{00}, p_{11}). For the second bracket, fix s and minimize over $q_{10}+q_{01} = s$ with $q_{10} \le q_{01}$. The objective is convex on this segment and its unconstrained minimizer is $q_{10} = sp_{10}/a > s/2$, which is infeasible because $p_{10} > p_{01}$; hence the constrained minimum is at the boundary $q_{10} = q_{01} = s/2$ and equals $p_{10}\log\frac{2p_{10}}{s} + p_{01}\log\frac{2p_{01}}{s}$. Therefore

$$D_{\mathrm{KL}}(p\|q) \;\ge\; (1-a)\log\tfrac{1-a}{1-s} + a\log\tfrac{a}{s} + J(p_{10}, p_{01}) = \mathrm{kl}(a, s) + J(p_{10}, p_{01}) \;\ge\; J(p_{10}, p_{01}),$$

where $\mathrm{kl}(a, s) := a\log\frac{a}{s} + (1-a)\log\frac{1-a}{1-s} \ge 0$ equals zero only at $s = a$. Equality throughout forces $s = a$, $q_{10} = q_{01} = a/2$, and $(q_{00}, q_{11}) = (p_{00}, p_{11})$, which is exactly q^*; and q^* attains the value $J(p_{10}, p_{01})$ by direct substitution. □

B.3 PROOF OF THEOREM 3.2

The sequential part of Theorem 3.2 uses a change-of-measure argument for stopped i.i.d. observations. We first state the form needed here, which relates the probability of any stopping-time event under two observation distributions to the expected number of observations. The same lemma is used later in the proofs of Theorems B.1 and 3.3.

Lemma B.2. *Let P and Q be two distributions on a finite observation space, and let $Y_1, Y_2, \ldots$ be i.i.d. observations drawn from either P or Q. Let ξ be an internal randomization seed, independent of the observations and with the same distribution under P and Q, and let τ be a stopping time determined by the observations so far and ξ. Let $\mathcal{F}_0 := \sigma(\xi)$ and $\mathcal{F}_n := \sigma(\xi, Y_1, \ldots, Y_n)$ denote the information available after n observations, and let*

$$\mathcal{F}_\tau := \{E : E \cap \{\tau \le n\} \in \mathcal{F}_n \text{ for every } n \ge 0\} \tag{1}$$

denote the information available when the procedure stops. Assume $P(\tau < \infty) = 1$, $0 < D_{\mathrm{KL}}(P\|Q) < \infty$ and $\mathbb{E}_1|\ell_1| < \infty$, where $\ell_i := \log\frac{dP}{dQ}(Y_i)$. If $\mathbb{E}_P[\tau] < \infty$, then for any event $E \in \mathcal{F}_\tau$,

$$\mathrm{kl}(P(E), Q(E \cap \{\tau < \infty\})) \le \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P\|Q).$$

Consequently, if $Q(E) \le P(E)$, then

$$\mathrm{kl}(P(E), Q(E)) \le \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P\|Q).$$

Proof. Let $L_n := \sum_{i=1}^n \ell_i$. Since $\mathbb{E}_1|\ell_1| < \infty$, $\mathbb{E}_P[\ell_1] = D_{\mathrm{KL}}(P\|Q)$ and $\mathbb{E}_P[\tau] < \infty$, Wald's identity (Bruss & Robertson, 1991; Janssen & Manca, 2006) gives

$$\mathbb{E}_P[L_\tau] = \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P\|Q). \tag{2}$$

Let μ_P denote the distribution of the stopped observation history, including the internal randomization, under P. Since $P(\tau < \infty) = 1$, this is a probability measure. Under Q, let μ_Q instead assign to an event A the probability $Q(A \cap \{\tau < \infty\})$, so that μ_Q is a sub-probability measure. Because $P \ll Q$ and the randomization seed has the same distribution under both distributions, $\mu_P \ll \mu_Q$ and $d\mu_P/d\mu_Q = e^{L_\tau}$ on $\{\tau < \infty\}$. Hence,

$$\mathbb{E}_P[L_\tau] = \int \log \frac{d\mu_P}{d\mu_Q}\, d\mu_P.$$

For an event E determined at stopping, write $p := P(E)$, $q := Q(E \cap \{\tau < \infty\})$, and $a := Q(\tau < \infty) \le 1$. Applying the log-sum inequality to the partition $\{E, E^c\}$ gives

$$\mathbb{E}_P[L_\tau] \ge p \log \frac{p}{q} + (1-p) \log \frac{1-p}{a-q} \ge p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = \mathrm{kl}(p, q),$$

where the second inequality holds because $a \le 1$ yields $a - q \le 1 - q$. Combining this with Wald's identity in equation 2 proves

$$\mathrm{kl}(P(E), Q(E \cap \{\tau < \infty\})) \le \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P \| Q).$$

If $Q(E) \le P(E) = p$, then $q \le Q(E) \le p$. Since $q' \mapsto \mathrm{kl}(p, q')$ is non-increasing on $[0, p]$,

$$\mathrm{kl}(p, Q(E)) \le \mathrm{kl}(p, q) \le \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P \| Q),$$

which proves the second claim. □

We now prove Theorem 3.2. The first part extends the four-cell projection of Lemma B.1 to the full success-vector distribution, identifying the closest alternative Q^* that reverses the ordering of A and B. The sequential lower bound then follows by applying Lemma B.2 to this alternative.

Theorem 3.2. *Fix distinct $A, B \in \mathcal{S}_k$ with $v_P(A) > v_P(B)$. Then, $\inf_{Q:, v_Q(A) \le v_Q(B)} D_{\mathrm{KL}}(P|Q) = I^P_{\mathrm{cov}}(A, B)$, and the infimum is attained by the distribution Q^* induced by the four-cell projection of Appendix B.2. Moreover, consider a sequential test that observes fresh i.i.d. success vectors, stops after τ sampled tasks under full-vector feedback, and decides whether A has larger coverage than B. If $\Pr_P[\text{declare } A \preceq B] \le \beta$ and $\Pr_Q[\text{declare } A \succ B] \le \alpha$ for every distribution Q on $\{0,1\}^m$ satisfying $v_Q(A) \le v_Q(B)$, where $\alpha, \beta \in (0, 1)$ and $\alpha + \beta < 1$, then $\mathbb{E}_P[\tau] \ge \mathrm{kl}(1-\beta, \alpha)/I^P_{\mathrm{cov}}(A, B)$.*

Proof. For a distribution Q on $\{0,1\}^m$, write $p^P = (p^P_{00}, p^P_{10}, p^P_{01}, p^P_{11})$ and $p^Q = (p^Q_{00}, p^Q_{10}, p^Q_{01}, p^Q_{11})$ for the probabilities of the four events $\{E^{A,B}_{ij}\}_{i,j}$. The chain rule for KL over the partition $\{E^{A,B}_{ij}\}_{i,j}$ then gives

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(p^P \| p^Q) + \sum_{i,j} p^P_{ij} D_{\mathrm{KL}}\Big(P(\cdot \mid E^{A,B}_{ij}) \,\|\, Q(\cdot \mid E^{A,B}_{ij})\Big) \ge D_{\mathrm{KL}}(p^P \| p^Q).$$

Equality holds when Q preserves the conditional distribution of P within each event. Moreover, $v_Q(A) - v_Q(B) = p^Q_{10}(A, B) - p^Q_{01}(A, B)$, so $v_Q(A) \le v_Q(B)$ is equivalent to $p^Q_{10}(A, B) \le p^Q_{01}(A, B)$. Similarly, $v_P(A) > v_P(B)$ implies $p^P_{10}(A, B) > p^P_{01}(A, B)$.

The minimization over Q therefore reduces exactly to the four-cell projection in Lemma B.1. Its minimizer preserves the probabilities of $E^{A,B}_{00}$ and $E^{A,B}_{11}$ and equalizes those of $E^{A,B}_{10}$ and $E^{A,B}_{01}$ at $(p^P_{10}(A, B) + p^P_{01}(A, B))/2$. Extending this projection to a distribution Q^* that preserves the conditional distribution of P within each event attains the lower bound above. Hence,

$$\inf_{Q:\, v_Q(A) \le v_Q(B)} D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P\|Q^*) = J\big(p^P_{10}(A, B), p^P_{01}(A, B)\big) = I^P_{\mathrm{cov}}(A, B).$$

Since $p^P_{10}(A, B) > 0$, the resulting divergence is finite.

For the sequential bound, the projection equalizes the two exclusive probabilities, so $v_{Q^*}(A) = v_{Q^*}(B)$. The assumed error guarantees therefore imply $\Pr_{Q^*}[\text{declare } A \succ B] \le \alpha$ and $\Pr_P[\text{declare } A \succ B] \ge 1 - \beta$. If $\mathbb{E}_P[\tau] = \infty$, the claimed bound is immediate. Otherwise, let $E = \{\text{declare } A \succ B\}$. Since $\Pr_{Q^*}(E) \le \alpha < 1 - \beta \le \Pr_P(E)$, Lemma B.2 gives

$$\mathrm{kl}(\Pr_P(E), \Pr_{Q^*}(E)) \le \mathbb{E}_P[\tau] D_{\mathrm{KL}}(P\|Q^*).$$

The Bernoulli KL divergence is non-decreasing in its first argument and non-increasing in its second whenever the first exceeds the second. Hence the left-hand side is at least $\mathrm{kl}(1 - \beta, \alpha)$. Using $D_{\mathrm{KL}}(P\|Q^*) = I^P_{\mathrm{cov}}(A, B)$ yields

$$\mathbb{E}_P[\tau] \ge \frac{\mathrm{kl}(1-\beta, \alpha)}{I^P_{\mathrm{cov}}(A, B)}.$$

The stopped-event formulation does not require $\tau < \infty$ under Q^*. □

B.4 Proof of Theorem 3.3

Theorem 3.3. *Assume P has a unique optimal size-k ensemble $S^*(P)$ and $\Gamma(P) > 0$. Then, for every $\delta \in (0, 1/2)$, every δ-correct exact-identification rule under full-vector feedback satisfies $\mathbb{E}_P[\tau] \geq \mathrm{kl}(1-\delta, \delta)/\Gamma(P)$. Moreover, the rule above is δ-correct and satisfies $\lim_{\delta \downarrow 0} \mathbb{E}_P[\tau_\delta]/\log(1/\delta) = T_{\mathrm{full}}(P) = 1/\Gamma(P)$.*

Proof. **Lower bound.** Fix $S \neq S^* := S^*(P)$, and let Q^* be the projection from Theorem 3.2 applied to (S^*, S). Thus, $v_{Q^*}(S^*) = v_{Q^*}(S)$ and $D_{\mathrm{KL}}(P\|Q^*) = I^P_{\mathrm{cov}}(S^*, S)$. For $\eta > 0$, define $Q_\eta := (1-\eta)Q^* + \eta R_S$, where R_S is the point mass on the vector $\mathbf{x}^S$ with $x^S_c = \mathbf{1}\{c \in S \setminus S^*\}$. The set $S \setminus S^*$ is non-empty because $S \neq S^*$ and $|S| = |S^*| = k$, and therefore $U(\mathbf{x}^S, S) = 1$ while $U(\mathbf{x}^S, S^*) = 0$. It follows that $v_{Q_\eta}(S) - v_{Q_\eta}(S^*) = \eta > 0$. Moreover, since $Q_\eta \geq (1-\eta)Q^*$ and Q^* assigns positive probability to every outcome with positive probability under P,

$$D_{\mathrm{KL}}(P\|Q_\eta) \leq D_{\mathrm{KL}}(P\|Q^*) + \log \tfrac{1}{1-\eta}.$$

The distribution Q_η need not have a *unique* optimal ensemble, whereas δ-correctness is assumed only on instances with a unique optimum. We therefore make one more small change to the distribution. The condition $v_Q(S) > v_Q(S^*)$ defines an open subset of the simplex. For any two distinct ensembles S_1, S_2, the function $Q \mapsto v_Q(S_1) - v_Q(S_2)$ is linear and not identically zero: it is positive at the point mass on $\mathbf{1}\{c \in S_1 \setminus S_2\}$. Hence, the distributions under which two ensembles have equal value form a finite union of proper hyperplanes and are of Lebesgue measure zero. Furthermore, $D_{\mathrm{KL}}(P\|\cdot)$ is continuous on the neighbourhood $\{Q : Q \geq \frac{1}{2}(1-\eta)Q^*\}$ of Q_η, throughout which $P \ll Q$. Consequently, for every sufficiently small $\zeta > 0$, there exists a distribution $Q_{\eta,\zeta}$ satisfying

$$\|Q_{\eta,\zeta} - Q_\eta\|_{\mathrm{TV}} < \zeta, \qquad v_{Q_{\eta,\zeta}}(S) > v_{Q_{\eta,\zeta}}(S^*), \qquad S^*(Q_{\eta,\zeta}) \text{ unique and } \neq S^*,$$

and $D_{\mathrm{KL}}(P\|Q_{\eta,\zeta}) \leq D_{\mathrm{KL}}(P\|Q_\eta) + o_\zeta(1)$. Indeed, the first two requirements define an open set containing Q_η, the third excludes a set of Lebesgue measure zero whose complement is dense, and the fourth follows from continuity on the neighbourhood above.

Let $(\tau, \widehat{S})$ be a δ-correct rule and set $E := \{\widehat{S} = S^*\}$. Then, $P(E) \geq 1-\delta$, while $Q_{\eta,\zeta}(E) \leq \delta$ because S^* is not optimal under $Q_{\eta,\zeta}$. If $\mathbb{E}_P[\tau] = \infty$, there is nothing to prove, so assume it is finite. Since $\delta < 1/2$, $Q_{\eta,\zeta}(E) \leq \delta < 1-\delta \leq P(E)$. Recall that $\mathcal{F}_\tau$ is the information available when the procedure stops (as defined in equation 1). Thus, Lemma B.2 applies to the $\mathcal{F}_\tau$-measurable event E and gives

$$\mathbb{E}_P[\tau]\, D_{\mathrm{KL}}\big(P\|Q_{\eta,\zeta}\big) \geq \mathrm{kl}(1-\delta, \delta), \quad \text{i.e.,} \quad \mathbb{E}_P[\tau] \geq \frac{\mathrm{kl}(1-\delta, \delta)}{D_{\mathrm{KL}}(P\|Q_{\eta,\zeta})}.$$

Letting first $\zeta \downarrow 0$ and then $\eta \downarrow 0$ yields $\mathbb{E}_P[\tau] \geq \mathrm{kl}(1-\delta, \delta)/I^P_{\mathrm{cov}}(S^*, S)$. Finally, minimizing the denominator over $S \neq S^*$ gives $\mathbb{E}_P[\tau] \geq \mathrm{kl}(1-\delta, \delta)/\Gamma(P)$.

δ-correctness. Suppose the rule stops at time n and returns $S \neq S^*$. Since S^* is one of the ensembles compared in the definition of $\widehat{\Gamma}_n$, we have $nI^{\widehat{P}_n}_{\mathrm{cov}}(S, S^*) \geq \beta(n, \delta)$. Fix such an ensemble $S \neq S^*$ and define $Y_i := (U(X_i, S), U(X_i, S^*)) \in \{0,1\}^2$. The empirical distribution $\widehat{p}_n$ of $Y_1, \ldots, Y_n$ takes at most $(n+1)^4$ values. By Lemma B.1, the event above is equivalent to

$$n \inf_{q \in \mathcal{Q}_0} D_{\mathrm{KL}}(\widehat{p}_n\|q) \geq \beta(n, \delta),$$

where $\mathcal{Q}_0$ is the set of four-cell distributions under which S is no better than S^*. The true distribution p of Y_i belongs to $\mathcal{Q}_0$ because $v_P(S) < v_P(S^*)$. By the method of types (Csiszár & Körner, 2011), $\Pr_p[\widehat{p}_n = r] \leq e^{-nD_{\mathrm{KL}}(r\|p)}$, and since $p \in \mathcal{Q}_0$, $D_{\mathrm{KL}}(r\|p) \geq \inf_{q \in \mathcal{Q}_0} D_{\mathrm{KL}}(r\|q)$. Therefore,

$$\Pr_P\Big[nI^{\widehat{P}_n}_{\mathrm{cov}}(S, S^*) \geq \beta(n, \delta)\Big] \leq (n+1)^4 e^{-\beta(n,\delta)}.$$

A union bound over the $M-1$ ensembles $S \neq S^*$ and over all $n \geq 1$, together with $\sum_{n \geq 1} n^{-2} = \pi^2/6$, gives $\Pr_P[\widehat{S}_{\tau_\delta} \neq S^*] \leq \delta$.

Expected stopping time. Fix $\varepsilon \in (0, \Gamma(P))$. Each cell probability $p^R_{ij}(A, B)$ and each value $v_R(S)$ is a linear function of R with coefficients in $\{0, 1\}$ and is therefore 1-Lipschitz in $\|\cdot\|_1$. Since

J is uniformly continuous on the compact set $\{(x, y) : x, y \geq 0,\ x + y \leq 1\}$ and $\mathcal{S}_k$ is finite, there exists $\zeta_1(\varepsilon) > 0$ such that $|I^R_{\mathrm{cov}}(A, B) - I^P_{\mathrm{cov}}(A, B)| \leq \varepsilon$ for every pair (A, B) whenever $\|R - P\|_1 \leq \zeta_1(\varepsilon)$. Since $S^*(P)$ is the unique maximizer of v_P, there also exists $\zeta_2 > 0$ such that $\widehat{S}_n = S^*$ whenever $\|\widehat{P}_n - P\|_1 \leq \zeta_2$. Hence, with $\zeta := \min(\zeta_1(\varepsilon), \zeta_2)$, $\|\widehat{P}_n - P\|_1 \leq \zeta$ implies both $\widehat{S}_n = S^*$ and $\widehat{\Gamma}_n \geq \Gamma(P) - \varepsilon$.

Define $n_0(\delta) := \min\{n : n(\Gamma(P) - \varepsilon) \geq \beta(n, \delta)\}$. Since $\beta(n, \delta) = \log(1/\delta) + \mathcal{O}(\log n)$, with constants depending only on m and k, we have

$$n_0(\delta) = (1 + o(1))\frac{\log(1/\delta)}{\Gamma(P) - \varepsilon} \qquad \text{as } \delta \downarrow 0.$$

Moreover, the derivative of $n \mapsto n(\Gamma(P) - \varepsilon) - \beta(n, \delta)$ is at least $\Gamma(P) - \varepsilon - 6/n$. Thus, once δ is small enough that $n_0(\delta) \geq 12/(\Gamma(P) - \varepsilon)$, this function is non-decreasing beyond $n_0(\delta)$, so the inequality defining this time continues to hold. Therefore, for every $n \geq n_0(\delta)$,

$$\{\tau_\delta > n\} \subseteq \{n\widehat{\Gamma}_n < \beta(n, \delta)\} \subseteq \{\widehat{\Gamma}_n < \Gamma(P) - \varepsilon\} \subseteq \{\|\widehat{P}_n - P\|_1 > \zeta\}.$$

Sanov's inequality (Dembo & Zeitouni, 2009) on the alphabet $\{0, 1\}^m$ gives $\Pr_P[\|\widehat{P}_n - P\|_1 > \zeta] \leq (n + 1)^{2^m} e^{-nc(\zeta)}$, where $c(\zeta) = \zeta^2/2 > 0$ by Pinsker's inequality (Csiszár & Körner, 2011). Hence,

$$\mathbb{E}_P[\tau_\delta] \leq n_0(\delta) + \sum_{n \geq n_0(\delta)} (n + 1)^{2^m} e^{-nc(\zeta)} \leq n_0(\delta) + K(\zeta, m),$$

where $K(\zeta, m) < \infty$ is independent of δ. Dividing by $\log(1/\delta)$, letting $\delta \downarrow 0$, and then letting $\varepsilon \downarrow 0$ gives

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_P[\tau_\delta]}{\log(1/\delta)} \leq \frac{1}{\Gamma(P)}.$$

The matching lower bound follows by applying the lower bound proved above to this δ-correct rule and using $\mathrm{kl}(1 - \delta, \delta) \sim \log(1/\delta)$. Therefore,

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_P[\tau_\delta]}{\log(1/\delta)} = \frac{1}{\Gamma(P)} = T_{\mathrm{full}}(P).$$

□

B.5 Exact comparison complexity under ordinal feedback

We now define the weighted ordinal setting and show that the same exclusive outcomes as under binary feedback determine ensemble comparison. Each task induces a strict ranking π of the models in C, with P_D denoting the resulting distribution over rankings. Let $\mathrm{rank}_\pi(c)$ be the position of model c, with rank 1 best, and for a challenger $c \in C$ define $P_\pi(c) := \{c' \in C : \mathrm{rank}_\pi(c') \leq \mathrm{rank}_\pi(c)\}$, the models ranked at least as highly as c. Given a distribution $\lambda \in \Delta(C)$ over challengers, define the *weighted ordinal coverage* of an ensemble S as $\Phi_\lambda(S) := \sum_{c \in C} \lambda_c g_c(S)$, where $g_c(S) := \Pr_{\pi \sim P_D}[S \cap P_\pi(c) \neq \varnothing]$.

Equivalently, draw $\pi \sim P_D$ and an independent challenger $c \sim \lambda$, and define $G_S(\pi, c) := \mathbf{1}\{S \cap P_\pi(c) \neq \varnothing\}$. Then $\Phi_\lambda(S) = \Pr[G_S(\pi, c) = 1]$: the ensemble succeeds when one of its members ranks at least as highly as the sampled challenger. Comparing two ensembles under weighted ordinal coverage therefore reduces to comparing two binary success indicators, yielding the same four-outcome structure as in Section 3.2.

To characterize the information required to distinguish two ensembles, we represent their ordinal outcomes using the same four-cell structure as in the binary setting. Fix two ensembles $A, B \subseteq C$. For $i, j \in \{0, 1\}$, let $p^\lambda_{ij}(A, B) := \Pr[G_A(\pi, c) = i,\ G_B(\pi, c) = j]$ denote the probability of each joint outcome, and let $P^\lambda_{A,B}$ denote the resulting distribution of (G_A, G_B) on $\{0, 1\}^2$. In particular, $p^\lambda_{10}(A, B)$ is the probability that A succeeds against the sampled challenger while B does not, whereas $p^\lambda_{01}(A, B)$ is the probability of the reverse outcome. Since $\Phi_\lambda(A) - \Phi_\lambda(B) = p^\lambda_{10}(A, B) - p^\lambda_{01}(A, B)$, only these two exclusive outcomes determine which ensemble has larger weighted ordinal coverage. Analogously to the binary case, we therefore define the *ordinal exclusive information* as $\mathcal{I}^\lambda_{\mathrm{ord}}(A, B) := J(p^\lambda_{10}(A, B), p^\lambda_{01}(A, B))$.

This four-cell representation allows us to apply Lemma B.1 to ordinal comparison. The following result shows that $\mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B)$ exactly characterizes the information required to reverse the ordering of two ensembles under weighted ordinal coverage and translates this information into a lower bound on the number of observations required for sequential comparison. To state the latter, for a target error probability $\delta \in (0, 1/2)$, we call a sequential ordinal comparison test δ-*valid* if, for i.i.d. observations from any distribution Q on $\{0, 1\}^2$, it declares A better with probability at most δ when $Q(1, 0) \le Q(0, 1)$ and with probability at least $1 - \delta$ when $Q(1, 0) > Q(0, 1)$.

Theorem B.1. *Fix $A, B \subseteq C$ with $\Phi_\lambda(A) > \Phi_\lambda(B)$. Then $\inf_{Q:\, Q(1,0)\le Q(0,1)} D_{\mathrm{KL}}(P^{\lambda}_{A,B} \| Q) = \mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B)$. Moreover, every δ-valid sequential ordinal comparison test with stopping time τ, where τ counts i.i.d. draws of (π, c) and hence induced binary comparison observations, satisfies $\mathbb{E}_{P^{\lambda}_{A,B}}[\tau] \ge \mathrm{kl}(1 - \delta, \delta)/\mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B)$.*

Proof. The first claim follows directly from Lemma B.1 applied to $P^{\lambda}_{A,B}$. Indeed, its condition $p^{\lambda}_{10}(A, B) > p^{\lambda}_{01}(A, B)$ is equivalent to $\Phi_\lambda(A) > \Phi_\lambda(B)$. Hence,

$$\inf_{Q:\, Q(1,0)\le Q(0,1)} D_{\mathrm{KL}}(P^{\lambda}_{A,B} \| Q) = \mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B).$$

For the sequential bound, let Q^* be the minimizing distribution from Lemma B.1, so that $Q^*(1, 0) = Q^*(0, 1)$. Let $\widehat{h} = 1$ denote the final decision that A is better than B. By δ-validity, $\Pr_{P^{\lambda}_{A,B}}[\widehat{h} = 1] \ge 1 - \delta$, while $\Pr_{Q^*}[\widehat{h} = 1] \le \delta$. If $\mathbb{E}_{P^{\lambda}_{A,B}}[\tau] = \infty$, there is nothing to prove. Otherwise, since $\delta < 1/2$,

$$\Pr_{Q^*}[\widehat{h} = 1] \le \delta < 1 - \delta \le \Pr_{P^{\lambda}_{A,B}}[\widehat{h} = 1].$$

Applying Lemma B.2 to the $\mathcal{F}_\tau$-measurable event $\{\widehat{h} = 1\}$ therefore gives

$$\mathrm{kl}(1 - \delta, \delta) \le \mathrm{kl}\Big(\Pr_{P^{\lambda}_{A,B}}[\widehat{h} = 1], \Pr_{Q^*}[\widehat{h} = 1]\Big) \le \mathbb{E}_{P^{\lambda}_{A,B}}[\tau]\, D_{\mathrm{KL}}(P^{\lambda}_{A,B} \| Q^*).$$

Since the first part gives $D_{\mathrm{KL}}(P^{\lambda}_{A,B} \| Q^*) = \mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B)$, rearranging yields

$$\mathbb{E}_{P^{\lambda}_{A,B}}[\tau] \ge \frac{\mathrm{kl}(1 - \delta, \delta)}{\mathcal{I}^{\lambda}_{\mathrm{ord}}(A, B)},$$

as claimed. □

Thus, binary and weighted ordinal feedback share the same pairwise information structure. In both cases, the comparison rate depends only on the probabilities of the two outcomes on which exactly one ensemble succeeds. Outcomes on which both ensembles succeed or neither succeeds do not affect the information required to distinguish them.

Connection to greedy selection on missed tasks. The same focus on tasks missed by the current ensemble appears in the greedy procedure of Section 4. For a current ensemble S and candidates a, b, take $A = S \cup a$ and $B = S \cup b$. A challenger already covered by S is covered by both enlarged ensembles and cannot distinguish them; only challenger instances missed by the current ensemble on which exactly one of a and b succeeds are informative. This parallels the exclusive rescue probabilities $p^{S}_{10}(a, b)$ and $p^{S}_{01}(a, b)$ under binary feedback on tasks missed by the current ensemble.

C Omitted proofs for Section 4

C.1 Sequential probability ratio tests for pairwise comparison

We begin with the basic decision underlying a greedy step: comparing two candidate models on tasks missed by the current ensemble. Fix a current ensemble $S \subseteq C$ with miss rate $\rho(S) := \Pr_{d\sim D}[U(d, S) = 0]$. For $c \in C \setminus S$, define the *rescue rate*, or success rate on tasks missed by the current ensemble, $q(c \mid S) := \Pr_{d\sim D}[u(d, c) = 1 \mid U(d, S) = 0]$ when $\rho(S) > 0$, and $q(c \mid S) := 0$ otherwise. Since $v(S \cup \{c\}) - v(S) = \rho(S)q(c \mid S)$, maximizing marginal gain at a fixed greedy

step is exactly equivalent to finding the candidate with the largest rescue rate on tasks missed by the current ensemble.

This approach of conditioning on missed tasks is also used by Neoh et al. (2026), whose ADAPTIVE-FAIL-GREEDY (AFG) procedure evaluates active candidates on tasks missed by the current ensemble and eliminates them using confidence bounds on their rescue rates. Here, we study these decisions through likelihood ratios. We begin with the *Sequential Probability Ratio Test* (SPRT), Wald's sequential test between two simple hypotheses, which accumulates log-likelihood ratios until the evidence crosses a prescribed threshold (Wald, 1947; Siegmund, 1985). We next analyze this pairwise test; the simultaneous multi-candidate procedure is developed in Section 4.

To formalize the pairwise problem, fix two candidates $a, b \in C \setminus S$ and let $Y_1, Y_2, \ldots$ be i.i.d. observations from tasks missed by the current ensemble; for example, $Y_i = (u(d_i, a), u(d_i, b))$, where d_i is drawn from D conditioned on $U(d_i, S) = 0$. We consider two simple hypotheses: H_1, under which a has the larger rescue rate, and H_0, under which b is at least as good. Let P_1 and P_0 denote the distributions of an observation Y_i under H_1 and H_0, respectively. We also write $\Pr_j$ and $\mathbb{E}_j$ for probabilities and expectations under P_j. The composite problem in which the relevant rescue rates are unknown is handled by the generalized statistic in Section 4. Here, the simple-hypothesis setting lets us examine the basic likelihood-ratio comparison.

The goal of the SPRT is to reach a decision while controlling the probability of choosing the wrong hypothesis. There are two possible errors, depending on which hypothesis is true. The *type-I error* is the probability of incorrectly declaring H_1 when H_0 is true, $\Pr_0[\text{declare } H_1]$, while the *type-II error* is the probability of incorrectly declaring H_0 when H_1 is true, $\Pr_1[\text{declare } H_0]$. Let $\alpha, \beta \in (0, 1)$ denote the target levels for these two errors, respectively.

The SPRT determines when there is sufficient evidence by accumulating log-likelihood ratios. For each observation Y_i, define $\ell_i := \log \frac{dP_1}{dP_0}(Y_i)$ and $L_n := \sum_{i=1}^{n} \ell_i$. Given thresholds $A > 0 > B$, the test stops at $\tau := \inf\{n \geq 1 : L_n \geq A \text{ or } L_n \leq B\}$, declaring H_1 if $L_\tau \geq A$ and H_0 if $L_\tau \leq B$. Let $\kappa_1 := D_{\mathrm{KL}}(P_1 \| P_0)$ and $\kappa_0 := D_{\mathrm{KL}}(P_0 \| P_1)$ denote the KL information per observation.

The thresholds trade off error control and stopping time. We use $A = \log(1/\alpha)$ and $B = -\log(1/\beta)$, which directly guarantee the target type-I and type-II error levels. The following result bounds both these errors and the expected stopping time.

Theorem C.1. *Assume P_0 and P_1 are mutually absolutely continuous with $\kappa_0, \kappa_1 > 0$, and that $|\ell_i| \leq K$ almost surely under both hypotheses. For $\alpha, \beta \in (0, 1)$ with $\alpha + \beta < 1$, set $A = \log(1/\alpha)$ and $B = -\log(1/\beta)$. Then, the SPRT has type-I error at most α and type-II error at most β. Here τ counts the i.i.d. observations Y_i, and $\mathbb{E}_1[\tau] \leq [\log(1/\alpha) + K]/\kappa_1$ and $\mathbb{E}_0[\tau] \leq [\log(1/\beta) + K]/\kappa_0$.*

Proof. Let $\mathcal{E}_1$ and $\mathcal{E}_0$ denote the events that the test declares H_1 and H_0, respectively. We first note that $\tau < \infty$ almost surely under both hypotheses. Indeed, $\mathbb{E}_1[\ell_i] = D_{\mathrm{KL}}(P_1 \| P_0) = \kappa_1 > 0$, whereas $\mathbb{E}_0[\ell_i] = -D_{\mathrm{KL}}(P_0 \| P_1) = -\kappa_0 < 0$. Since the increments are bounded, the strong law of large numbers gives $L_n/n \to \kappa_1$ almost surely under P_1 and $L_n/n \to -\kappa_0$ almost surely under P_0. Hence, L_n eventually crosses one of the finite boundaries under either hypothesis.

We first bound the type-I error. Under P_0, the likelihood-ratio process $(e^{L_n})_{n \geq 0}$ is a nonnegative martingale with $\mathbb{E}_0[e^{L_n}] = 1$. For $n \geq 1$, let $\tau_n := \tau \wedge n$. Since τ_n is bounded, the optional stopping theorem gives $\mathbb{E}_0[e^{L_{\tau_n}}] = 1$. On $\mathcal{E}_1 \cap \{\tau \leq n\}$, we have $L_{\tau_n} = L_\tau \geq A$, and therefore $e^{L_{\tau_n}} \geq e^A$. It follows that

$$1 = \mathbb{E}_0[e^{L_{\tau_n}}] \geq \mathbb{E}_0\big[e^{L_{\tau_n}} \mathbf{1}_{\mathcal{E}_1 \cap \{\tau \leq n\}}\big] \geq e^A \Pr_0(\mathcal{E}_1 \cap \{\tau \leq n\}).$$

Since $\tau < \infty$ almost surely under P_0, $\mathcal{E}_1 \cap \{\tau \leq n\} \uparrow \mathcal{E}_1$. Letting $n \to \infty$ therefore gives $\Pr_0(\mathcal{E}_1) \leq e^{-A} = \alpha$.

The type-II bound follows symmetrically. Under P_1, $(e^{-L_n})_{n \geq 0}$ is a nonnegative martingale of mean one, and on $\mathcal{E}_0 \cap \{\tau \leq n\}$ we have $-L_{\tau_n} \geq -B$. Hence, $\Pr_1(\mathcal{E}_0) \leq e^B = \beta$.

We next bound the expected stopping time under P_1. Since τ_n is bounded and the increments are i.i.d. and integrable, Wald's identity (Bruss & Robertson, 1991; Janssen & Manca, 2006) gives

$$\mathbb{E}_1[L_{\tau_n}] = \mathbb{E}_1[\ell_1]\, \mathbb{E}_1[\tau_n] = \kappa_1 \mathbb{E}_1[\tau_n].$$

Moreover, $L_{\tau_n} \le A + K$ almost surely. If $\tau > n$, then $L_{\tau_n} = L_n < A$ because neither boundary has yet been crossed. If $\tau \le n$, then either the test exits through the lower boundary, in which case $L_\tau \le B < A + K$, or it exits through the upper boundary. In the latter case $L_{\tau-1} < A$ and $|\ell_\tau| \le K$, so $L_\tau = L_{\tau-1} + \ell_\tau < A + K$. Thus,

$$\kappa_1 \mathbb{E}_1[\tau_n] = \mathbb{E}_1[L_{\tau_n}] \le A + K.$$

Since $\tau_n \uparrow \tau$, monotone convergence yields

$$\mathbb{E}_1[\tau] \le \frac{A + K}{\kappa_1} = \frac{\log(1/\alpha) + K}{\kappa_1}.$$

The argument under P_0 is symmetric. Applying Wald's identity to $-L_{\tau_n}$, whose increments have mean κ_0 under P_0, and using $-L_{\tau_n} \le |B| + K$ gives

$$\mathbb{E}_0[\tau] \le \frac{|B| + K}{\kappa_0} = \frac{\log(1/\beta) + K}{\kappa_0}.$$

□

Classical Wald thresholds. The conservative thresholds in Theorem C.1 give direct control of the individual type-I and type-II errors. For completeness, we also state the corresponding guarantees for the narrower classical Wald thresholds (Wald, 1945; Wald & Wolfowitz, 1948). Because the log-likelihood ratio can pass beyond a boundary, these thresholds do not in general guarantee the individual nominal error levels α and β. Instead, the following bounds hold for the actual error probabilities.

Proposition C.1 (Classical Wald thresholds)**.** *Under the assumptions of Theorem C.1, set* $A = \log \frac{1-\beta}{\alpha}$ *and* $B = \log \frac{\beta}{1-\alpha}$. *If* $\alpha' := \Pr_0[\text{declare } H_1]$ *and* $\beta' := \Pr_1[\text{declare } H_0]$ *are the actual error probabilities, then* $\alpha'/(1-\beta') \le \alpha/(1-\beta)$ *and* $\beta'/(1-\alpha') \le \beta/(1-\alpha)$*, and hence* $\alpha' + \beta' \le \alpha + \beta$*. Moreover, with* τ *measured in the same i.i.d. observations* Y_i *as in Theorem C.1,* $\mathbb{E}_1[\tau] \le [\log((1-\beta)/\alpha) + K]/\kappa_1$ *and* $\mathbb{E}_0[\tau] \le [\log((1-\alpha)/\beta) + K]/\kappa_0$.

Proof. Let $\mathcal{E}_1$ and $\mathcal{E}_0$ denote the events that the test declares H_1 and H_0, respectively, and write $\alpha' := \Pr_0(\mathcal{E}_1)$ and $\beta' := \Pr_1(\mathcal{E}_0)$. As in the proof of Theorem C.1, $\tau < \infty$ almost surely under both hypotheses.

Since $\tau < \infty$ almost surely under both hypotheses and P_0 and P_1 are mutually absolutely continuous, the likelihood ratio of the stopped observations under P_1 relative to P_0 is e^{L_τ}. Therefore, for any event E determined by the observations up to τ,

$$\Pr_0(E) = \mathbb{E}_1\left[e^{-L_\tau} \mathbf{1}_E\right]. \tag{3}$$

Applying this identity to $E = \mathcal{E}_1$ gives

$$\alpha' = \Pr_0(\mathcal{E}_1) = \mathbb{E}_1\left[e^{-L_\tau} \mathbf{1}_{\mathcal{E}_1}\right] \le e^{-A} \Pr_1(\mathcal{E}_1) = e^{-A}(1 - \beta'),$$

where the inequality follows because $L_\tau \ge A$ on $\mathcal{E}_1$. Since $e^{-A} = \alpha/(1-\beta)$, this gives

$$\frac{\alpha'}{1-\beta'} \le \frac{\alpha}{1-\beta}. \tag{4}$$

Similarly, applying the reverse identity in equation 3 to $\mathcal{E}_0$ and using $L_\tau \le B$ on this event gives

$$\beta' = \Pr_1(\mathcal{E}_0) = \mathbb{E}_0\left[e^{L_\tau} \mathbf{1}_{\mathcal{E}_0}\right] \le e^{B} \Pr_0(\mathcal{E}_0) = e^{B}(1 - \alpha').$$

Since $e^B = \beta/(1-\alpha)$, we have

$$\frac{\beta'}{1-\alpha'} \le \frac{\beta}{1-\alpha}. \tag{5}$$

Equivalently, $\alpha'(1-\beta) \le \alpha(1-\beta')$ and $\beta'(1-\alpha) \le \beta(1-\alpha')$. Adding the inequalities in equations 4–5 and cancelling the common terms $\alpha'\beta$ and $\alpha\beta'$ yields $\alpha' + \beta' \le \alpha + \beta$.

It remains to bound the stopping times. The argument is the same as in Theorem C.1. For $\tau_n := \tau \wedge n$, Wald's identity (Bruss & Robertson, 1991; Janssen & Manca, 2006) gives $\mathbb{E}_1[L_{\tau_n}] = \kappa_1 \mathbb{E}_1[\tau_n]$,

while bounded increments imply $L_{\tau_n} \le A + K$. Hence $\kappa_1 \mathbb{E}_1[\tau_n] \le A + K$, and monotone convergence gives

$$\mathbb{E}_1[\tau] \le \frac{A+K}{\kappa_1} = \frac{\log((1-\beta)/\alpha) + K}{\kappa_1}.$$

Applying the symmetric argument to $-L_{\tau_n}$ under P_0 gives

$$\mathbb{E}_0[\tau] \le \frac{|B|+K}{\kappa_0} = \frac{\log((1-\alpha)/\beta) + K}{\kappa_0}.$$

□

The classical Wald boundaries lie closer to zero than the conservative boundaries, so a test using them can stop sooner. More precisely,

$$A_{\mathrm{Wald}} = \log(1/\alpha) + \log(1-\beta) < \log(1/\alpha), \quad |B_{\mathrm{Wald}}| = \log\Big(\frac{1-\alpha}{\beta}\Big) < \log(1/\beta).$$

Thus, the interval between the two thresholds is narrower, so the test typically reaches a decision after accumulating less evidence, which may reduce the expected stopping time. However, when the log-likelihood ratio passes beyond a boundary, the individual guarantees $\alpha' \le \alpha$ and $\beta' \le \beta$ need not hold[1] and Proposition C.1 gives the guarantees that remain valid. The conservative thresholds are therefore used in the main text because they provide direct control of the two error probabilities.

C.2 Optimality of pairwise sequential testing

The preceding subsection gave an upper bound on the number of observations required by the SPRT. We now show that its leading dependence on the error levels and KL information cannot be improved by any sequential test between the same two simple hypotheses. Recall that α and β denote the type-I and type-II errors, respectively, and that $\kappa_1 = D_{\mathrm{KL}}(P_1\|P_0)$ and $\kappa_0 = D_{\mathrm{KL}}(P_0\|P_1)$ measure the KL information per observation in the two directions.

Theorem C.2 (Lower bound for pairwise sequential testing)**.** *Consider any sequential test between the simple hypotheses H_0 and H_1 based on i.i.d. observations with distributions P_0 and P_1, respectively. Let τ be its stopping time, measured in the number of i.i.d. observations, and suppose its type-I and type-II errors are at most α and β, where $\alpha, \beta \in (0,1)$ and $\alpha + \beta < 1$. If $0 < \kappa_1 := D_{\mathrm{KL}}(P_1\|P_0) < \infty$, then $\mathbb{E}_1[\tau] \ge \mathrm{kl}(1-\beta, \alpha)/\kappa_1$. Symmetrically, if $0 < \kappa_0 := D_{\mathrm{KL}}(P_0\|P_1) < \infty$, then $\mathbb{E}_0[\tau] \ge \mathrm{kl}(1-\alpha, \beta)/\kappa_0$.*

Proof. It suffices to prove the bound under H_1; the bound under H_0 follows by interchanging the two hypotheses. If $\mathbb{E}_1[\tau] = \infty$, there is nothing to prove, so assume $\mathbb{E}_1[\tau] < \infty$. Let $\ell_i := \log \frac{dP_1}{dP_0}(Y_i)$ and $L_n := \sum_{i=1}^n \ell_i$. Since $\mathbb{E}_1|\ell_1| < \infty$ and $\mathbb{E}_1[\ell_1] = D_{\mathrm{KL}}(P_1\|P_0) = \kappa_1$, Wald's identity (Bruss & Robertson, 1991; Janssen & Manca, 2006) gives

$$\mathbb{E}_1[L_\tau] = \kappa_1 \mathbb{E}_1[\tau].$$

We next relate the accumulated log-likelihood ratio to the final decision. Let ξ denote the test's internal randomization seed. Define $\mathcal{F}_0 := \sigma(\xi)$ and $\mathcal{F}_n := \sigma(\xi, Y_1, \dots, Y_n)$ for $n \ge 1$. Thus, $\mathcal{F}_n$ represents the information available to the test after n observations. The information available when the test stops is described by the stopped σ-algebra

$$\mathcal{F}_\tau := \{E : E \cap \{\tau \le n\} \in \mathcal{F}_n \text{ for every } n \ge 0\}.$$

In particular, the final decision is determined by $\mathcal{F}_\tau$.

Let μ_1 be the restriction of P_1 to $\mathcal{F}_\tau$, which is a probability measure because $\tau < \infty$ P_1-almost surely, and define the sub-probability measure μ_0 on $\mathcal{F}_\tau$ by $\mu_0(E) := P_0(E \cap \{\tau < \infty\})$. Since $P_1 \ll P_0$ for each observation and the randomization seed has the same distribution under both hypotheses, $\mu_1 \ll \mu_0$, with $\frac{d\mu_1}{d\mu_0} = e^{L_\tau}$ on $\{\tau < \infty\}$. Hence,

$$\mathbb{E}_1[L_\tau] = \int \log \frac{d\mu_1}{d\mu_0}\, d\mu_1.$$

[1] For instance, with $P_1(1) = 0.9$, $P_0(1) = 0.1$, $\alpha = 0.05$, and $\beta = 0.55$, one has $A = \log 9 = |\ell_i|$ and $|B| < \log 9$, so $\tau \equiv 1$ and the actual type-I error is $0.1 > \alpha$. The guarantee $\alpha' \le \alpha/(1-\beta) = 1/9$ does hold.

The use of the sub-probability measure μ_0 is important here: we do not assume that $\tau < \infty$ under P_0.

Let E be the event that the test declares H_1, and write $p := \mu_1(E) = P_1(E)$, $q := \mu_0(E) = P_0(E \cap \{\tau < \infty\})$, and $a := \mu_0(\Omega) = P_0(\tau < \infty)$. Then $\mu_0(E^c) = a - q$. Applying the log-sum inequality to the partition $\{E, E^c\}$ gives

$$\mathbb{E}_1[L_\tau] \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{a-q} \geq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = \mathrm{kl}(p, q),$$

where the second inequality follows from $a \leq 1$, and hence $a - q \leq 1 - q$. Combining this inequality with Wald's identity yields

$$\kappa_1 \mathbb{E}_1[\tau] \geq \mathrm{kl}(p, q).$$

By the assumed error guarantees, $P_1(E) \geq 1 - \beta$ and $P_0(E) \leq \alpha$. Moreover, $q \leq P_0(E) \leq \alpha < 1 - \beta \leq P_1(E) = p$, where the strict inequality follows from $\alpha + \beta < 1$. Since $\mathrm{kl}(p, q)$ is non-decreasing in its first argument when $p \geq q$ and non-increasing in its second argument when $q \leq p$,

$$\mathrm{kl}(p, q) \geq \mathrm{kl}\big(p, P_0(E)\big) \geq \mathrm{kl}(1 - \beta, \alpha).$$

Therefore,

$$\mathbb{E}_1[\tau] \geq \frac{\mathrm{kl}(1 - \beta, \alpha)}{\kappa_1}.$$

Interchanging P_0 and P_1 gives, under the corresponding assumptions,

$$\mathbb{E}_0[\tau] \geq \frac{\mathrm{kl}(1 - \alpha, \beta)}{\kappa_0},$$

which proves the result. □

Comparing this lower bound with Theorem C.1 shows that the SPRT is optimal to first order for pairwise comparison: its expected stopping time matches the information-theoretic lower bound up to the bounded overshoot term K and lower-order differences in the error terms. This optimality applies to the comparison of two simple hypotheses. It does not imply that a sequence of such pairwise decisions is globally optimal for ensemble identification.

C.3 Proof of Theorem 4.1

Theorem 4.1. *Fix S with $\rho(S) = \Pr_{d\sim D}[U(d, S) = 0] > 0$ and $A \subseteq C \setminus S$ with $|A| \geq 2$. Suppose that the unique maximizer $a^* := \arg\max_{c \in A} q(c \mid S)$ satisfies $\Delta_{\mathrm{KL}}(S, A) > 0$. Then, for any $\delta \in (0, 1)$, Algorithm 1 returns a^* with probability at least $1 - \delta$. Moreover, its number τ_S of accepted tasks satisfies $\mathbb{E}[\tau_S] \leq (1 + o(1)) \log(1/\delta)/\Delta_{\mathrm{KL}}(S, A)$ as $\delta \downarrow 0$ with S, A, D fixed. Conversely, for any $\delta \in (0, 1/2)$, any δ-correct rescue selector using the same vector feedback on accepted tasks satisfies $\mathbb{E}[\tau] \geq \mathrm{kl}(1 - \delta, \delta)/\Delta_{\mathrm{KL}}(S, A) = (1 - o(1)) \log(1/\delta)/\Delta_{\mathrm{KL}}(S, A)$.*

Proof. **Validity of the stopping threshold.** Fix an ordered pair (a, b) satisfying $q(a \mid S) \leq q(b \mid S)$ and a sample size $n \geq 1$. The empirical distribution of $(u(d_i, a), u(d_i, b)) \in \{0, 1\}^2$ can take at most $(n+1)^4$ values. By Lemma B.1, the event $n\widehat{I}_n^S(a, b) \geq x$ means that this empirical distribution has KL distance at least x/n from the set of four-outcome distributions satisfying $q_{10} \leq q_{01}$. The true distribution belongs to this set because $q(a \mid S) \leq q(b \mid S)$. The method of types (Csiszár & Körner, 2011) therefore gives

$$\Pr\Big[n\widehat{I}_n^S(a, b) \geq x\Big] \leq (n+1)^4 e^{-x}.$$

Taking $x = \beta_r(n, \delta)$ and applying a union bound over the at most $r(r-1)$ ordered pairs and all $n \geq 1$, together with $\sum_{n \geq 1} n^{-2} = \pi^2/6$, yields

$$\Pr\Big[\exists n \geq 1,\ \exists a, b \in A : q(a \mid S) \leq q(b \mid S),\ n\widehat{I}_n^S(a, b) \geq \beta_r(n, \delta)\Big] \leq \delta. \tag{6}$$

Correctness. Denote the event in equation 6 by $\mathcal{E}$. On its complement $\mathcal{E}^c$, no ordered pair (a, b) with $q(a \mid S) \leq q(b \mid S)$ ever meets the stopping criterion in favor of a. Suppose the algorithm

nevertheless stops and returns some $a \neq a^*$. Since a^* is the unique maximizer, $q(a \mid S) < q(a^* \mid S)$. However, the stopping condition requires $\tau_S \widehat{I}^S_{\tau_S}(a, a^*) \geq \beta_{|A|}(\tau_S, \delta)$, contradicting the definition of $\mathcal{E}^c$. Therefore, the algorithm returns a^* on $\mathcal{E}^c$, and hence does so with probability at least $1 - \delta$.

Stopping time. Fix $\varepsilon \in (0, \Delta_{\mathrm{KL}}(S, A))$ and set $d := \Delta_{\mathrm{KL}}(S, A) - \varepsilon$. By uniform continuity of J and uniqueness of a^*, there exists $\zeta > 0$ such that whenever the empirical distribution $\widehat{P}^S_n$ of the outcome vectors from accepted tasks satisfies $\|\widehat{P}^S_n - P_S\|_1 \leq \zeta$, we have both $\widehat{a}_n = a^*$ and $\min_{b \neq a^*} \widehat{I}^S_n(a^*, b) \geq d$. Let $n_0(\delta) := \min\{n \geq 1 : nd \geq \beta_{|A|}(n, \delta)\}$. Since $\beta_{|A|}(n, \delta) = \log(1/\delta) + \mathcal{O}(\log n)$,

$$n_0(\delta) = (1 + o(1)) \frac{\log(1/\delta)}{d} \qquad \text{as } \delta \downarrow 0.$$

It remains to verify that once this threshold is reached, the inequality continues to hold. For real $x > 0$, it holds that

$$\frac{\partial}{\partial x} \beta_{|A|}(x, \delta) = \frac{2}{x} + \frac{4}{x+1} \leq \frac{6}{x},$$

and thus

$$\frac{d}{dx} \big[xd - \beta_{|A|}(x, \delta) \big] \geq d - \frac{6}{x}.$$

The latter is nonnegative for $x \geq 6/d$. Since $n_0(\delta) \to \infty$ as $\delta \downarrow 0$, for all sufficiently small δ we have $n_0(\delta) \geq 6/d$, and therefore $nd \geq \beta_{|A|}(n, \delta)$ for every $n \geq n_0(\delta)$.

Consequently, for every $n \geq n_0(\delta)$, we have

$$\{\tau_S > n\} \subseteq \{\|\widehat{P}^S_n - P_S\|_1 > \zeta\}.$$

Sanov's inequality (Dembo & Zeitouni, 2009) on the finite alphabet $\{0, 1\}^A$ gives

$$\Pr\Big[\|\widehat{P}^S_n - P_S\|_1 > \zeta\Big] \leq (n+1)^{2^{|A|}} e^{-n\zeta^2/2},$$

where the exponent follows from Pinsker's inequality (Csiszár & Körner, 2011). The resulting tail sum is finite and independent of δ, so $\mathbb{E}[\tau_S] \leq n_0(\delta) + \mathcal{O}(1)$. Dividing by $\log(1/\delta)$, letting $\delta \downarrow 0$, and then letting $\varepsilon \downarrow 0$ yields

$$\mathbb{E}[\tau_S] \leq (1 + o(1)) \frac{\log(1/\delta)}{\Delta_{\mathrm{KL}}(S, A)}.$$

Lower bound. Let P_S denote the distribution of $(u(d, c))_{c \in A}$ conditional on S missing the task, and fix $b \neq a^*$. Since a^* is the unique maximizer, $x := p^S_{10}(a^*, b) > y := p^S_{01}(a^*, b)$. By Lemma B.1 and the argument conditioning on each of the four outcomes in the proof of Theorem 3.2, there is a distribution Q^* that preserves the conditional distributions within the four outcomes and replaces the two exclusive probabilities by $(x + y)/2$. It satisfies

$$D_{\mathrm{KL}}(P_S \| Q^*) = J(x, y) = I_S(a^*, b), \qquad q_{Q^*}(a^* \mid S) = q_{Q^*}(b \mid S).$$

Let $\mathcal{M} := \arg\max_{c \in A} q_{Q^*}(c \mid S)$ be the set of maximizers under Q^*. Then, $\mathcal{M} \backslash \{a^*\} \neq \varnothing$. Indeed, this is immediate if $a^* \notin \mathcal{M}$; if $a^* \in \mathcal{M}$, then $b \in \mathcal{M}$ as well because $q_{Q^*}(a^* \mid S) = q_{Q^*}(b \mid S)$. Fix $b^* \in \mathcal{M} \backslash \{a^*\}$ and choose weights $w \in \Delta(A)$ with pairwise distinct entries and strict maximum at b^*. For each $c \in A$, let $e_c \in \{0, 1\}^A$ be the indicator vector of $\{c\}$ and let δ_{e_c} denote the point mass at e_c. Define

$$Q_\eta := (1 - \eta) Q^* + \eta \sum_{c \in A} w_c \delta_{e_c}.$$

Since under δ_{e_c} only candidate c succeeds, the rescue rates under Q_η are exactly

$$q_{Q_\eta}(c \mid S) = (1 - \eta) q_{Q^*}(c \mid S) + \eta w_c.$$

Candidates outside $\mathcal{M}$ lie below the maximum under Q^* by a fixed positive amount. Hence, for every η smaller than this gap divided by $1 + \max_c w_c$, the maximizer under Q_η remains in $\mathcal{M}$. Within $\mathcal{M}$, the values under Q^* coincide, so the distinct weights w_c make b^* the unique maximizer. Thus, for sufficiently small η, the unique best rescuer under Q_η is $b^* \neq a^*$.

Algorithm 2 KL-greedy on missed tasks (k, δ)

Require: Target size k, confidence level $\delta \in (0, 1)$
$S_0 \leftarrow \varnothing$
for $i = 1, \dots, k$ **do**
  if $\rho(S_{i-1}) = \Pr_{d\sim D}[U(d, S_{i-1}) = 0] = 0$ **then**
    return S_{i-1}
  $a_i \leftarrow$ KL-RESCUE-TEST$(S_{i-1}, C \setminus S_{i-1}, \delta/k)$
  $S_i \leftarrow S_{i-1} \cup \{a_i\}$
return S_k

Every distribution on $\{0,1\}^A$ can be realized as the conditional distribution of outcomes on tasks missed by S: take a task distribution D' under which every member of S misses every task, so $\rho_{D'}(S) = 1$, and let the A-coordinates have the prescribed distribution. Moreover, $P_S \ll Q_\eta$ and

$$D_{\mathrm{KL}}(P_S \| Q_\eta) \le D_{\mathrm{KL}}(P_S \| Q^*) + \log \frac{1}{1-\eta}.$$

A δ-correct selector returns a^* with probability at least $1-\delta$ under P_S. Under Q_η, the unique best rescuer is $b^* \neq a^*$, so the same selector returns a^* with probability at most δ. If $\mathbb{E}_{P_S}[\tau] = \infty$, there is nothing to prove; otherwise, since $\delta < 1/2$, the event $E := \{\text{return } a^*\}$ satisfies

$$\Pr_{Q_\eta}(E) \le \delta < 1 - \delta \le \Pr_{P_S}(E).$$

Applying Lemma B.2 with $(P_1, P_0) = (P_S, Q_\eta)$ and using the same monotonicity of the Bernoulli KL divergence kl as in the proof of Theorem 3.2 gives

$$\mathrm{kl}(1-\delta, \delta) \le \mathrm{kl}\big(\Pr_{P_S}(E), \Pr_{Q_\eta}(E)\big) \le \mathbb{E}_{P_S}[\tau] D_{\mathrm{KL}}(P_S \| Q_\eta).$$

Letting $\eta \downarrow 0$ therefore yields

$$\mathbb{E}_{P_S}[\tau] \ge \frac{\mathrm{kl}(1-\delta, \delta)}{I_S(a^*, b)}.$$

Finally, taking the smallest denominator over $b \neq a^*$ gives

$$\mathbb{E}_{P_S}[\tau] \ge \frac{\mathrm{kl}(1-\delta, \delta)}{\Delta_{\mathrm{KL}}(S, A)},$$

which proves the lower bound. □

C.4 KL-GREEDY ON MISSED TASKS

For completeness, Algorithm 2 gives the pseudocode for the size-k greedy construction described in Section 4. Starting from the empty ensemble, it repeatedly applies Algorithm 1 to the remaining candidates using tasks missed by the current ensemble, with confidence level δ/k at each step.

C.5 PROOF OF PROPOSITION 4.1

Proposition 4.1. *Assume that for each $i \le k$ the population greedy choice a_i^* is unique, $\rho(S_{i-1}^{\mathrm{gr}}) > 0$, and $\Delta_i > 0$. Then, for every $\delta \in (0,1)$, Algorithm 2 returns S_k^{gr} with probability at least $1-\delta$, and on that event $v(S_k) \ge \big(1 - (1 - \frac{1}{k})^k\big)\mathrm{OPT}_k \ge (1 - 1/e)\mathrm{OPT}_k$. Conditional on the previous steps being correct, if τ_i denotes the number of accepted tasks used by the i-th call, then $\mathbb{E}[\tau_i] \le (1 + o(1)) \log(k/\delta)/\Delta_i$ as $\delta \downarrow 0$ with m, k, D fixed.*

Proof. For each $i = 1, \dots, k$, condition on the first $i-1$ calls being correct. The ensemble entering the i-th call is then S_{i-1}^{gr}, so by Theorem 4.1, Algorithm 1 returns the population greedy choice a_i^* with probability at least $1 - \delta/k$. Hence, the probability that any of the k calls makes an incorrect choice is at most δ by a union bound. With probability at least $1-\delta$, therefore, Algorithm 2 follows the population greedy path exactly and returns S_k^{gr}.

On this event, each selected model maximizes the marginal coverage gain at the corresponding greedy step, since maximizing the rescue rate is equivalent to maximizing marginal gain. Because v is a monotone submodular coverage function, the classical greedy guarantee of Nemhauser et al. (1978) gives

$$v(S_k) \geq \left(1 - \left(1 - \frac{1}{k}\right)^k\right) \mathrm{OPT}_k \geq (1 - 1/e)\mathrm{OPT}_k.$$

Finally, conditional on the preceding $i-1$ calls being correct, the i-th call is applied to S_{i-1}^{gr} with confidence level δ/k. Applying the stopping-time upper bound in Theorem 4.1 therefore gives

$$\mathbb{E}[\tau_i] \leq (1 + o(1))\frac{\log(k/\delta)}{\Delta_i}$$

as $\delta \downarrow 0$ with m, k, D fixed, completing the proof. □

D Omitted proofs for Section 5

D.1 Proof of Theorem 5.1

Theorem 5.1. *There exists a family of success-vector distributions $(P_\varepsilon)_{\varepsilon\in(0,1/20)}$ over three models with target size $k = 2$ such that, as $\varepsilon \downarrow 0$, Algorithm 2 using the true rescue rates returns the unique optimal ensemble, $\Gamma(P_\varepsilon)$ remains bounded away from zero, and the first-step greedy information gap satisfies $\Delta_{\mathrm{KL}}(\varnothing, C) = \frac{5}{3}\varepsilon^2 + \mathcal{O}(\varepsilon^3)$. Thus, for any fixed $\delta \in (0, 1/2)$, identifying the first greedy choice with error at most δ requires $\Omega(\varepsilon^{-2})$ more sampled tasks under full-vector feedback than globally identifying the optimal ensemble to leading order.*

Proof. We begin with the explicit construction. Let $C = \{a, b, c\}$, let $k = 2$, and fix $\varepsilon \in (0, 1/20)$. Define P_ε on $\{0,1\}^3$, with coordinates ordered as (X_a, X_b, X_c), by

x	$(1,0,0)$	$(0,1,0)$	$(0,0,1)$	$(0,1,1)$	$(1,1,0)$	$(0,0,0)$
$P_\varepsilon(x)$	$\frac{3}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{1}{10} - \varepsilon$	$\frac{1}{5}$	$\frac{7}{20} + \varepsilon$

and set $P_\varepsilon(x) = 0$ for all other x. These masses are nonnegative for $\varepsilon < 1/10$ and sum to $\frac{3}{20} + \frac{1}{20} + \frac{3}{20} + \frac{2}{20} - \varepsilon + \frac{4}{20} + \frac{7}{20} + \varepsilon = 1$.

We first characterize the globally optimal size-2 ensemble. The singleton coverages are $v(\{a\}) = \frac{3}{20} + \frac{1}{5} = \frac{7}{20}$, $v(\{b\}) = \frac{1}{20} + (\frac{1}{10} - \varepsilon) + \frac{1}{5} = \frac{7}{20} - \varepsilon$, and $v(\{c\}) = \frac{3}{20} + (\frac{1}{10} - \varepsilon) = \frac{1}{4} - \varepsilon$. For pairs, an ensemble misses a task exactly on the outcomes where all its members have coordinate zero. Hence,

$$v(\{a, c\}) = 1 - P_\varepsilon(0,1,0) - P_\varepsilon(0,0,0) = \frac{3}{5} - \varepsilon,$$

whereas

$$v(\{a, b\}) = v(\{b, c\}) = 1 - \frac{3}{20} - \left(\frac{7}{20} + \varepsilon\right) = \frac{1}{2} - \varepsilon,$$

using $P_\varepsilon(0,0,1) = P_\varepsilon(1,0,0) = 3/20$ for the two rivals, respectively. Thus, $\{a, c\}$ is the unique optimal size-2 ensemble, with coverage gap exactly $1/10$ against either rival.

We next compute its full-vector information gap. For $A = \{a, c\}$ and $B = \{a, b\}$, the exclusive event on which A succeeds and B misses the task is the outcome $(0, 0, 1)$, of probability $3/20$, while the reverse exclusive event is $(0, 1, 0)$, of mass $1/20$. For the other rival $B = \{b, c\}$, the corresponding exclusive events are $(1, 0, 0)$ and $(0, 1, 0)$, again with masses $3/20$ and $1/20$. Therefore,

$$I_{\mathrm{cov}}^{P_\varepsilon}(\{a, c\}, B) = J\left(\frac{3}{20}, \frac{1}{20}\right)$$

for both rivals, and hence

$$\Gamma(P_\varepsilon) = J\left(\frac{3}{20}, \frac{1}{20}\right) > 0.$$

In particular, the global information gap is independent of ε. By Theorem 3.3, the globally optimal identification rate is therefore

$$\mathbb{E}[\tau_{\mathrm{global}}] = (1 + o(1))\frac{\log(1/\delta)}{J(3/20, 1/20)} \qquad \text{as } \delta \downarrow 0,$$

whose leading constant does not depend on ε.

We now turn to the first greedy step. Since $S = \varnothing$, $\rho(\varnothing) = 1$ and $D_\varnothing = D$, the rescue rates coincide with the singleton coverages computed above. Thus, a is the unique first greedy choice for every $\varepsilon > 0$. Its exclusive rescue probabilities against b are

$$p_{10}^{\varnothing}(a, b) = \Pr[X_a = 1, X_b = 0] = P_\varepsilon(1, 0, 0) = \frac{3}{20},$$

and

$$p_{01}^{\varnothing}(a, b) = P_\varepsilon(0, 1, 0) + P_\varepsilon(0, 1, 1) = \frac{3}{20} - \varepsilon.$$

Against c, they are $p_{10}^{\varnothing}(a, c) = P_\varepsilon(1, 0, 0) + P_\varepsilon(1, 1, 0) = 7/20$ and $p_{01}^{\varnothing}(a, c) = P_\varepsilon(0, 0, 1) + P_\varepsilon(0, 1, 1) = 1/4 - \varepsilon$. Consequently,

$$I_\varnothing(a, c) = J\left(\frac{7}{20}, \frac{1}{4} - \varepsilon\right) \geq J\left(\frac{7}{20}, \frac{1}{4}\right) > 8 \cdot 10^{-3},$$

whereas

$$I_\varnothing(a, b) = J\left(\frac{3}{20}, \frac{3}{20} - \varepsilon\right) \leq J\left(\frac{3}{20}, \frac{1}{10}\right) < 5.1 \cdot 10^{-3}$$

for $\varepsilon \in (0, 1/20)$, where the inequalities follow from monotonicity of $J(x, \cdot)$. Hence, throughout this range,

$$\Delta_{\mathrm{KL}}(\varnothing, C) = J\left(\frac{3}{20}, \frac{3}{20} - \varepsilon\right).$$

To obtain its asymptotic behavior, write $x = 3/20$, $u = \varepsilon/2$, and $\mu = x - u$. Then

$$J(x, x - \varepsilon) = (\mu + u)\log\left(1 + \frac{u}{\mu}\right) + (\mu - u)\log\left(1 - \frac{u}{\mu}\right).$$

Expanding the logarithms gives

$$\mu \log\left(1 - \frac{u^2}{\mu^2}\right) + 2u\left(\frac{u}{\mu} + \frac{u^3}{3\mu^3} + \cdots\right) = \frac{u^2}{\mu} + O\left(\frac{u^4}{\mu^3}\right),$$

and therefore

$$\Delta_{\mathrm{KL}}(\varnothing, C) = J\left(\frac{3}{20}, \frac{3}{20} - \varepsilon\right) = \frac{5}{3}\varepsilon^2 + \mathcal{O}(\varepsilon^3).$$

Since the first greedy choice is unique and $\Delta_{\mathrm{KL}}(\varnothing, C) > 0$, the lower bound in Theorem 4.1 implies that, for every $\delta \in (0, 1/2)$, any δ-correct procedure identifying this choice from full-vector feedback satisfies

$$\mathbb{E}[\tau_{\mathrm{greedy}}] \geq \frac{\mathrm{kl}(1 - \delta, \delta)}{\Delta_{\mathrm{KL}}(\varnothing, C)}.$$

Comparing this lower bound with the global identification rate above, and using $\mathrm{kl}(1 - \delta, \delta)/\log(1/\delta) \to 1$ as $\delta \downarrow 0$, shows that for each fixed ε the ratio of the smallest expected sample size of any δ-correct first-step greedy selector to the expected sample size of the globally optimal rule is asymptotically at least

$$\frac{\Gamma(P_\varepsilon)}{\Delta_{\mathrm{KL}}(\varnothing, C)} = \frac{J(3/20, 1/20)}{J(3/20, 3/20 - \varepsilon)} = \frac{3}{5}J\left(\frac{3}{20}, \frac{1}{20}\right)\varepsilon^{-2}(1 + \mathcal{O}(\varepsilon)).$$

Thus, this ratio is $\Omega(\varepsilon^{-2})$ and diverges as $\varepsilon \downarrow 0$.

It remains to verify that this difference in sample requirements is not caused by greedy selecting a suboptimal ensemble. After the first choice $S = \{a\}$, the ensemble misses a task on the event $\{X_a = 0\}$, which has probability $13/20$. On this event, b rescues the ensemble with unnormalized probability $P_\varepsilon(0, 1, 0) + P_\varepsilon(0, 1, 1) = 3/20 - \varepsilon$, whereas c rescues it with unnormalized probability $P_\varepsilon(0, 0, 1) + P_\varepsilon(0, 1, 1) = 1/4 - \varepsilon$. Hence, c has the higher success rate on tasks missed by S and is the second greedy choice. Therefore, Algorithm 2, when each step uses the true rescue rates, follows $\varnothing \to \{a\} \to \{a, c\}$ and returns the unique globally optimal ensemble. □

D.2 Proof of Theorem 5.2

Here, we prove Theorem 5.2. Recall that RP is the class of decision problems admitting a randomized polynomial-time algorithm that never accepts a NO instance and accepts every YES instance with probability at least $1/2$ (equivalently, any fixed constant bounded away from zero).

Theorem 5.2. *If a randomized algorithm, given sample access to any distribution P on $\{0,1\}^m$ with a unique optimal size-k ensemble, returns $S^*(P)$ with probability at least $3/4$ in time polynomial in m and $1/\Gamma(P)$, then* $\mathrm{NP} \subseteq \mathrm{RP}$.

Proof. Suppose there exists a randomized algorithm ALG that, given sample access to any distribution P on $\{0,1\}^m$ with a unique optimal size-k ensemble, returns $S^*(P)$ with probability at least $3/4$ and runs in time polynomial in m and $1/\Gamma(P)$.

We reduce from Max-k-Cover, whose decision version is NP-hard (Feige, 1998). Let $([n], F_1, \ldots, F_m, k, v)$ be an instance asking whether some k sets $F_c \subseteq [n]$ cover at least vn elements. Draw $w_1, \ldots, w_m$ independently and uniformly from $[2m]$, and let $w'_c := 2m + 1 - w_c$, which are again i.i.d. uniform on $[2m]$. We use the Isolation Lemma (Mulmuley et al., 1987): for any family of subsets of an m-element ground set, assigning independent uniform weights from $[N]$ produces a unique minimum-weight member with probability at least $1 - m/N$. Thus, with $N = 2m$, the probability of a non-unique minimum is at most $m/2m = 1/2$.

Let $W := \sum_c w_c \le 2m^2$, fix $\theta := 1/(2n+2)$, and define

$$P' := (1-\theta)\, P_0 + \theta \sum_{c \in C} \frac{w_c}{W}\, \delta_{e_c},$$

where P_0 is the distribution of $\big(\mathbf{1}\{i \in F_c\}\big)_{c \in [m]}$ for i uniform on $[n]$, and $e_c \in \{0,1\}^m$ is the indicator vector of $\{c\}$. Sampling from P' takes $\mathcal{O}(m)$ time given the instance. Since the outcome e_c is covered by S exactly when $c \in S$,

$$v_{P'}(S) = (1-\theta)\, \frac{|\bigcup_{c \in S} F_c|}{n} + \frac{\theta}{W} \sum_{c \in S} w_c.$$

Distinct values of the first term differ by at least $(1-\theta)/n > \theta$, while the second term lies in $[0, \theta]$. Hence, every maximizer of $v_{P'}$ maximizes $|\bigcup_{c \in S} F_c|$ and, among those maximizers, maximizes $\sum_{c \in S} w_c$. Because every committee has cardinality k,

$$\sum_{c \in S} w'_c = k(2m+1) - \sum_{c \in S} w_c \qquad \text{for every } S \in \mathcal{S}_k.$$

Thus, maximizing $\sum_{c \in S} w_c$ over the family of optimal covers is equivalent to minimizing $\sum_{c \in S} w'_c$ over that family. By the Isolation Lemma (Mulmuley et al., 1987) specified above, this minimizer is unique with probability at least $1/2$ over the weights. Call the corresponding committee S^*. Then $S^* = S^*(P')$ is an optimal k-cover.

We next lower bound the coverage gap. Write $C(S) := |\bigcup_{c \in S} F_c|$ and $W(S) := \sum_{c \in S} w_c$, so that

$$v_{P'}(S) = (1-\theta)\frac{C(S)}{n} + \frac{\theta}{W} W(S).$$

On the event that the weighted optimum is unique, S^* maximizes C and, among its maximizers, uniquely maximizes W. Fix $S \neq S^*$ and consider two cases. If $C(S) = C(S^*)$, then uniqueness of the weighted optimum and the integer weights give $W(S^*) - W(S) \ge 1$, and hence

$$v_{P'}(S^*) - v_{P'}(S) = \frac{\theta}{W}\big(W(S^*) - W(S)\big) \ge \frac{\theta}{W}.$$

If instead $C(S) \le C(S^*) - 1$, then, since $0 \le W(S), W(S^*) \le W$,

$$v_{P'}(S^*) - v_{P'}(S) = (1-\theta)\frac{C(S^*) - C(S)}{n} + \frac{\theta}{W}\big(W(S^*) - W(S)\big) \ge \frac{1-\theta}{n} - \theta = \frac{1}{2n} > \theta \ge \frac{\theta}{W},$$

where the equality $(1-\theta)/n - \theta = 1/(2n)$ follows from $\theta = 1/(2n+2)$. Thus, in both cases,

$$v_{P'}(S^*) - v_{P'}(S) \ge \gamma := \frac{\theta}{W} \ge \frac{1}{2m^2(2n+2)},$$

since $W \le 2m^2$.

We now translate this coverage gap into a lower bound on the information gap. For $x > y$,

$$J(x,y) \ge \frac{(x-y)^2}{2(x+y)} \ge \frac{(x-y)^2}{2}.$$

The first inequality follows from $J(x,y) = (x+y)\mathrm{kl}\big(\frac{x}{x+y}, \frac{1}{2}\big)$ and Pinsker's inequality (Csiszár & Körner, 2011), while the second uses $x + y \le 1$. For the exclusive outcomes of S^* and S, $p_{10} + p_{01} \le 1$ because the two events are disjoint, and

$$p_{10} - p_{01} = v_{P'}(S^*) - v_{P'}(S) \ge \gamma.$$

Therefore,

$$\Gamma(P') \ge \frac{\gamma^2}{2} \ge \frac{1}{\mathrm{poly}(m,n)}.$$

Finally, run ALG on P' with a fixed running-time limit. Let p be a fixed polynomial bounding the running time of ALG on uniquely optimal instances as a function of m and $1/\Gamma(P)$, and let $\widetilde{p}$ be a polynomial with nonnegative coefficients that dominates p on $\mathbb{R}_+^2$. The nonnegative coefficients ensure that $\widetilde{p}$ is non-decreasing in each argument, allowing an upper bound on $1/\Gamma(P')$ to give a valid time budget. On the event that the weighted optimum is unique, $\Gamma(P') \ge \gamma^2/2$, so

$$\frac{1}{\Gamma(P')} \le \frac{2}{\gamma^2}.$$

Since $\gamma = \theta/W$ is computable from the sampled weights,

$$\mathrm{time}(\mathsf{ALG}) \le \widetilde{p}\big(m, 2/\gamma^2\big),$$

which is polynomial in the instance size. Run ALG for at most this many steps and stop the run otherwise. Because the budget $\widetilde{p}(m, 2/\gamma^2)$ depends only on the sampled weights, the time limit keeps the reduction polynomial-time for *every* choice of weights, whether or not the weighted optimum is unique.

When the weighted optimum is not unique, we make no assumption about ALG. It may return anything or reach the time limit. Its correctness and running-time guarantees are used only when the weighted optimum is unique, so P' has a unique optimum. With probability at least $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$, the weighted optimum is unique and ALG returns an optimal k-cover within the time bound. Accept if the returned committee covers at least vn elements, which is checkable in polynomial time. A NO instance is therefore never accepted, while a YES instance is accepted with probability at least $3/8$; repetition amplifies this probability to $2/3$. Hence, the decision problem is in RP. □

There is also a computational issue specific to implementing the stopping rule of Theorem 3.3. Even after computing the empirical maximizer $\widehat{S}_n$, evaluating $\widehat{\Gamma}_n = \min_{S \ne \widehat{S}_n} I_{\mathrm{cov}}^{\widehat{P}_n}(\widehat{S}_n, S)$ requires minimizing over all rival size-k ensembles. Restricting this minimum to a subset of rivals overestimates $\widehat{\Gamma}_n$ and can therefore cause the rule to stop too early. An implementation that adds rival constraints as needed would require a procedure, called a separation oracle, that, for a threshold Λ, either finds a rival S with $I_{\mathrm{cov}}^{\widehat{P}_n}(\widehat{S}_n, S) < \Lambda$ or verifies that no such rival exists. With such an oracle, the stopping rule can be implemented in polynomial time when oracle calls count as single operations. Finding a good ensemble does not resolve the difficult step of verifying that no rival lies sufficiently close to the empirical winner.

E Experimental Setup

This appendix gives the experimental setup underlying the results in Section 6. Appendix E.1 describes the two sampling protocols and explains why they are kept separate. Appendix E.2 describes the benchmarks, and Appendix E.3 specifies the candidate model pools used in each experiment. Appendix E.4 gives the protocol for our fresh multilingual evaluation, while Appendix E.5 defines how model evaluations are counted. Finally, Appendices E.6, E.7, and E.8 describe the baselines, reported metrics, and statistical methodology, respectively.

E.1 SAMPLING PROTOCOLS

The experiments serve two different purposes and therefore use two sampling protocols. Track A uses published evaluation matrices to test the information-rate predictions of the theory under the i.i.d. task model assumed in our main results. Track B uses a fresh evaluation of open models on a finite benchmark to study ensemble selection and confidence guarantees under practical model-evaluation costs. We describe the two protocols before explaining why their results must be interpreted separately.

Track A: validating the information rates. Track A starts from a published binary evaluation matrix $U \in \{0,1\}^{m\times n}$, whose rows correspond to models and columns to tasks. We treat the uniform distribution over its n columns as the success-vector distribution P introduced in Section 2. Because the entire matrix is observed, the population quantities studied in Section 3 can be computed directly rather than estimated from the simulated samples. In particular, for a pair of ensembles A, B, $p_{10}^P(A,B)$ and $p_{01}^P(A,B)$ are the probabilities that exactly one of the two succeeds, and $I_{\mathrm{cov}}^P(A,B) = J(p_{10}^P(A,B), p_{01}^P(A,B))$ is their exclusive-coverage information, as defined in Theorem 3.2. For global identification, $\Gamma(P)$ is the minimum such information separating the optimal ensemble $S^*(P)$ from any rival, as defined in Theorem 3.3.

We then draw task columns independently with replacement from this empirical population. Thus, each sampled column is a fresh i.i.d. observation from P, even if the same column index is drawn more than once. This protocol matches the sampling model of Theorems 3.2 and 3.3 and lets us compare their predicted information rates with observed stopping times. The primary cost in Track A is therefore the number of sampled tasks.

Track B: fresh finite-benchmark evaluation. Track B uses a fresh evaluation that we run ourselves to study how the selection procedures behave when the evaluator has a fixed collection of tasks and pays for individual model-task evaluations. For each seed, we randomly shuffle the benchmark and process its tasks without replacement. Each task is therefore encountered at most once, and each model-task outcome is charged at most once. The primary cost is the number of model evaluations rather than the number of sampled tasks. If a fixed-confidence stopping condition has not been reached when all available tasks have been used, we mark the decision as unconfirmed at the specified confidence level. This is the protocol used in Track B for practical selection and for testing whether the confidence criterion can be met.

Why the protocols are kept separate. The distinction matters because Track A is intended to test the information-theoretic rates derived for an i.i.d. task sequence, whereas Track B measures what can be selected or confirmed at a specified confidence level from a finite benchmark. Sampling without replacement does not invalidate the finite-sample confidence bounds used by our procedures: the relevant concentration bounds remain conservative under without-replacement sampling (Hoeffding, 1963). However, the lower bounds and asymptotic rates in Theorems 3.2, 3.3 and B.1 are statements about stopped i.i.d. experiments and therefore are not claimed for the Track-B protocol.

A finite task sequence also cannot be used to test the asymptotic characteristic time in Theorem 3.3: its stopping time is bounded by the benchmark size, whereas the theorem studies the growth of the required number of observations as $\delta \downarrow 0$. Further, sampling without replacement becomes increasingly easier than i.i.d. sampling as a substantial fraction of the benchmark is consumed because of the finite-population correction. For example, with $N = 4{,}400$ available tasks, an i.i.d. stopping time of $4{,}000$ corresponds to only about $2{,}095$ without-replacement draws under the variance-matching approximation $k_{\mathrm{wor}} = k_{\mathrm{iid}}N/(N + k_{\mathrm{iid}})$, a reduction of about 48%. Using Track B to validate the i.i.d. rates would thus understate their sample requirements. For this reason, all rate-validation results use Track A, while Track B is reserved for finite- benchmark selection, evaluation cost, and whether the confidence criterion can be met. No rate or characteristic-time claim combines the two protocols.

E.2 BENCHMARKS

We use four published evaluation datasets in Track A and two multilingual question-answering benchmarks in Track B. For each benchmark, we describe the observations used in our experiments, any filtering or preprocessing, and the role the benchmark plays in the evaluation.

XQuAD (Artetxe et al., 2020) (fresh multilingual evaluation). XQuAD contains 240 paragraphs and 1,190 questions from the SQuAD development set, professionally translated into 12 languages. The parallel translations are particularly useful for studying complementarity: the same question and gold answer appear across languages, so differences in model success can be examined while holding the underlying question fixed. In our experiments, a task is a question-language pair, and a model succeeds when its extracted answer matches a gold answer according to the scoring rule described later in Appendix E.4. This gives the binary feedback model of Section 2, with an ensemble succeeding whenever at least one of its members answers correctly.

We use eight languages for which the candidate pool contains models specialized for the corresponding language or language family: Thai, Hindi, Greek, Russian, Chinese, Vietnamese, German, and Spanish. The candidate pool is described in Appendix E.3. We sample 300 questions at the paragraph level using seed 20260823, yielding 63 paragraphs and 2,400 question-language tasks. Because XQuAD consists of translated Wikipedia questions, we additionally use TyDi-QA to test whether the observed complementarity persists on natively authored questions.

TyDi-QA GoldP (Clark et al., 2020) (native-question robustness check). TyDi-QA contains information-seeking questions written by native speakers who did not see the answer passage when formulating the question. Unlike XQuAD, the questions are not translations of a common English source. We use 200 Russian and 200 Arabic questions, for a total of 400 tasks. Russian provides a direct comparison between translated and natively authored questions for a language represented in both benchmarks, while Arabic extends the check to a language outside the eight-language XQuAD evaluation. We use this subset to examine whether the complementarity observed on XQuAD persists under this change in question construction and language coverage. Its small size is insufficient for a separate selection experiment, so we use it only as a robustness check.

MMLU-PromptEval (Maia Polo et al., 2024) (large Track-A population). MMLU-PromptEval provides per-question binary correctness for 15 base models evaluated with 100 prompt templates on the 14,042 MMLU test questions (Hendrycks et al., 2021). We therefore obtain a complete $1{,}500 \times 14{,}042$ binary matrix in which each candidate is a base-model and prompt-template configuration and each task is an MMLU question. Its large number of tasks and candidates makes it our primary Track-A population, including the controlled disagreement experiments in Appendices F.1 and F.2 and the characteristic-time experiments in Appendix F.3. Because each base model appears under multiple prompt templates, the candidate pool can contain highly similar configurations of the same underlying model; Appendix E.6 describes how we account for this when defining the accuracy-based baselines. The released matrix uses fixed evaluation software and a fixed number of examples in each prompt and does not uniquely identify every model revision.

RouterBench (Hu et al., 2024) (exact global enumeration). RouterBench provides per-instance outcomes for 11 open- and closed-source models across 80 evaluation groups and 28,590 items. We retain the 61 groups with documented correctness criteria, comprising the 57 MMLU subjects and four additional benchmarks covering commonsense reasoning, science question answering, pronoun resolution, and code generation. This leaves 27,248 tasks whose outcomes map directly to binary correctness. The remaining 19 groups contain 1,342 items and consist primarily of free-form generation and Chinese-language-knowledge evaluations. Although their released scores take values in $\{0, 1\}$, no corresponding scoring rule is documented. Moreover, 13 of these groups contain fewer than 35 items, so observing only binary-valued scores provides limited evidence that the underlying evaluation itself represents binary correctness. We therefore exclude all groups whose outcomes cannot be unambiguously interpreted as $u(d, c) \in \{0, 1\}$, removing 4.7% of the benchmark. The resulting 11-model pool is particularly useful for global identification because every size-k ensemble can be enumerated. We can therefore compute the optimal ensemble $S^*(P)$ and the global information gap $\Gamma(P)$ exactly, without approximate combinatorial optimization. This makes RouterBench useful for the characteristic-time experiments in Appendix F.3.

GSM8K (Cobbe et al., 2021) (distinct-model population). For GSM8K, we use per-problem 5-shot exact-match results from the lm-evaluation-harness (Gao et al., 2024) for 73 distinct open models on the 1,319 test problems. We align outcomes across models using a hash of the question text rather than relying on row position. Unlike MMLU-PromptEval, this pool contains distinct models rather than prompt variants of a smaller set of base models, so accuracy-based Top-k is a substantially stronger reference. The benchmark is too short for the controlled Track-A experiments

that require long task sequences, but under the i.i.d. sampling protocol of Appendix E.1 it can still be used in the characteristic-time experiment of Appendix F.3.

LiveBench (White et al., 2025) (weighted ordinal evaluation). We use LiveBench programmatic-grader scores to construct the ordinal evaluation population for Theorem B.1. After removing models scored on fewer than 80% of the questions and questions without scores for every retained model, we obtain a complete score matrix with 215 questions and 37 models. We exclude task types with only two distinct score levels because they already reduce to binary feedback.

For any question i and challenger model c', we convert the scores into the binary outcome $\widetilde{U}_i(c, c') := \mathbf{1}\{\mathrm{rank}_i(c) \le \mathrm{rank}_i(c')\}$, indicating whether model c ranks at least as highly as challenger c'. The resulting matrix has $215 \times 37 = 7{,}955$ columns indexed by question-challenger pairs. This is precisely the binary representation of weighted ordinal coverage used in Theorem B.1, so the same comparison procedure can be applied to these columns. We use competition ranks, treating tied scores as weak wins. This differs from the strict-ranking assumption in the theorem and is the main approximation in the empirical construction. Further, a column now represents a question-challenger pair instead of a single model evaluation, so this dataset is used to study comparison rates in task units and not model-evaluation cost. See Appendix F.4 for the corresponding results.

E.3 Candidate pools

This subsection describes the candidate models used in the fresh multilingual evaluation and the published evaluation matrices. The fresh pool was designed to combine broadly multilingual models with models expected to have stronger performance in particular languages, allowing different models to succeed on different tasks. The offline pools are fixed by the models available in their respective released evaluation matrices.

Fresh multilingual pool ($m = 16$). We evaluate 16 open models under a common inference configuration. All models use 4-bit MLX quantization with group size 64 and fixed model versions. The candidate pool and inference configuration were fixed before computing any ensemble-selection results. The pool combines multilingual generalists from the Qwen2 and Qwen2.5 families (Yang et al., 2024; 2025), Llama 3.1 and 3.2 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), and Aya (Dang et al., 2024), with models developed for particular languages or language families. These include OpenThaiGPT and Typhoon for Thai (Yuenyong et al., 2025; Pipatanakul et al., 2023), OpenHathi for Hindi (Gala et al., 2024), Meltemi for Greek (Voukoutis et al., 2024), Vikhr for Russian (Nikolich et al., 2026), Yi-6B-Chat for Chinese (Young et al., 2025), and SeaLLM for Vietnamese and other Southeast Asian languages (Nguyen et al., 2024). The complete 16-model pool is: Qwen2.5-7B/14B/3B/1.5B-Instruct, Qwen2-7B-Instruct, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, aya-expanse-8b, OpenThaiGPT-1.0.0-7b-chat, Typhoon-1.5-8b-instruct, OpenHathi-7B-Hi-v0.1-Base, Meltemi-7B-Instruct-v1, Vikhr-7B-instruct-0.2, Yi-6B-Chat, and SeaLLM-7B-v2.

The models included for their expected language focus do not all exhibit language-specific advantages under our common prompting and inference protocol. Appendix G.2 examines their measured language-specific strengths and their contribution to complementarity. OpenThaiGPT, Meltemi, and Yi-6B-Chat exhibit clear relative advantages on Thai, Greek, and Chinese, respectively, whereas OpenHathi and Vikhr do not show the corresponding advantage on Hindi and Russian. Moreover, none is the most accurate candidate on its target language. We therefore use intended language focus only to diversify the candidate pool. The selection procedures themselves rely solely on observed task-level outcomes and make no specialization assumptions.

Published evaluation pools. For Track A and the offline selection experiments, the candidate pools are inherited from the corresponding published evaluation matrices. MMLU-PromptEval (Maia Polo et al., 2024) contains 15 base models each paired with 100 prompt templates, giving $1{,}500$ candidate configurations; the base models span the CodeLlama, FLAN-T5, FLAN-UL2, Gemma, Merlinite, Llama 3, Mistral, Mixtral, and Falcon families. RouterBench (Hu et al., 2024) contains 11 open and proprietary models, including models from the Claude, GPT, Llama 2, Mistral, Mixtral, WizardLM, Code Llama, and Yi families. The GSM8K pool (Gao et al., 2024; Cobbe et al., 2021) contains 73 distinct open models spanning the Llama, Mistral/Mixtral, Qwen, Yi, Gemma, Phi, Falcon, DeepSeek, DBRX, Command-R, SOLAR, CodeLlama, StableLM, MPT, OLMo, Pythia, and OPT families.

E.4 FRESH EVALUATION PROTOCOL

This appendix describes the protocol used for the fresh multilingual evaluation, from task construction and model inference to scoring, data splitting, and budgeted ensemble selection. The candidate pool and benchmark sources are described in Appendices E.3 and E.2, respectively; Appendix E.5 defines the evaluation-cost accounting, Appendix E.6 defines the baselines, and Appendix H gives additional implementation and reproducibility details.

Task construction. We use the validation split of XQuAD in eight languages: Thai, Hindi, Greek, Russian, Chinese, Vietnamese, German, and Spanish. Because XQuAD is parallel, the same $1{,}190$ questions occur in every language. We verify this alignment before constructing the evaluation set. Question selection is performed once at the paragraph level and then used across all eight languages. Using fixed seed 20260823, we randomly permute the paragraphs and traverse them in that order, adding all questions from a paragraph whenever doing so does not exceed the target of 300 questions. This produces 63 paragraphs and exactly 300 questions. Each selected question is then included in all eight languages, giving $300 \times 8 = 2{,}400$ tasks. A task is a (question, language) pair containing the corresponding passage, question, and gold answer strings. Task order is not fixed at this stage; each selection run independently randomizes the training task sequence as described below.

Prompt. Every candidate receives the same two-shot completion prompt. The two demonstrations are fixed and written in English, while the passage and question of the target task are inserted in their original XQuAD language. The exact template is reproduced below, where `<PASSAGE>` denotes the passage of the target task and `<QUESTION>` denotes its question:

```
Answer the question using the shortest exact span copied from the passage.

Passage: The Amazon rainforest covers most of the Amazon basin of South
America. This basin encompasses 7,000,000 square kilometres.
Question: How large is the Amazon basin?
Answer: 7,000,000 square kilometres

Passage: Nikola Tesla was a Serbian-American inventor born in 1856 in the
village of Smiljan. He is best known for his work on alternating current.
Question: In which year was Tesla born?
Answer: 1856

Passage: <PASSAGE>
Question: <QUESTION>
Answer:
```

The two demonstration passages are single lines in the actual prompt and are wrapped above only to fit the page width. The final line ends immediately after `Answer:`, with no trailing space, so that the model continues the answer directly. The passage and question are inserted verbatim in the target language, while the instruction and demonstrations remain in English. Thus, across languages, only the target task itself varies.

We use the same completion format for every candidate, with no chat template, system message, or model-specific prompt wrapper. Three of the sixteen candidates do not provide a chat template. In particular, OpenHathi is a base model, and Vikhr and OpenThaiGPT provide none. Therefore, applying chat templates where available and using raw completions for the remaining candidates would make the input format depend on model identity. Using a single completion format avoids this source of variation, at the cost of not using each model's preferred prompting format. This choice is appropriate here because the experiment compares the candidates with one another under a common evaluation protocol.

Inference. All sixteen candidates are served under the same inference configuration. We use the pinned 4-bit MLX checkpoints listed in Appendix E.3, with quantization group size 64; no candidate uses a different precision, group size, or inference procedure. Each model is loaded with the tokenizer shipped with its checkpoint. Prompts are tokenized once, sorted by token length, and generated in batches of 16. The sorting reduces padding during batched generation but does not change the generated text because each prompt is decoded independently. Decoding is greedy with a maximum of 24 new tokens. The experiment configuration file specifies temperature 0.0, and the generation call specifies no additional sampling parameters. In particular, no top-p, top-k, repetition penalty, stop sequence, or minimum-generation-length constraint is set. Generation therefore uses deterministic argmax decoding and stops at the model's end-of-sequence token or at the 24-token

limit, whichever comes first. No input is truncated because all prompts fit within every candidate's context window.

Scoring. Each generation is converted to a binary success indicator using a scoring rule fixed before the full evaluation matrix was generated. First, we retain only the text before the first newline. We then apply the same text normalization to this first line and to every gold answer: Unicode NFKC normalization, lower-casing, removal of leading and trailing whitespace, replacement of consecutive whitespace characters with a single space, and removal of leading and trailing punctuation. The punctuation set contains standard Latin punctuation together with the corresponding fullwidth and CJK punctuation marks, including ideographic commas and periods, CJK quotation marks, and fullwidth parentheses, colons, semicolons, question marks, and exclamation marks.

An empty normalized prediction scores 0. Otherwise, the task scores 1 if the normalized prediction equals or contains at least one normalized gold answer, and 0 otherwise. The containment test is directional: a prediction may contain additional text around the gold answer, but a prediction that is only a proper substring of the gold answer is not accepted. XQuAD provides a single gold answer string per question, but the scorer accepts a list of gold answers and returns a match if any one satisfies this criterion, so the same rule applies without modification to tasks with multiple acceptable answers. For example, in the two-shot completion prompt described above, `7,000,000 square kilometres.` normalizes to `7,000,000 square kilometres` and thus matches that gold answer. The prediction `The year 1856.` also scores 1 for gold answer `1856`, as the normalized prediction contains the normalized gold answer. In the reverse direction, `square kilometres` scores 0 for gold answer `7,000,000 square kilometres`. Since only the first generated line is scored, a generation beginning with a newline has an empty first line and scores 0 regardless of subsequent text.

The above scoring rule was fixed before the full evaluation matrix was generated and was not modified afterwards.

Training and test splits. We split the $16 \times 2{,}400$ outcome matrix at the paragraph level rather than the question level. All questions derived from the same passage, together with their versions in all eight languages, are assigned to the same side of the split. Thus, no passage appears in both training and test, while the parallel structure of XQuAD preserves exact language balance. For each split s, we randomly permute the 63 paragraphs using seed $1000 + s$ and assign 47 paragraphs to training and the remaining 16 to test. Because paragraphs contain different numbers of questions, the number of tasks varies across splits. Over the ten executed splits, the training sets contain between $1{,}656$ and $1{,}848$ tasks and the held-out sets between 552 and 744, with averages of $1{,}770$ and 630, respectively.

The final analysis uses 10 independent splits and 20 algorithm seeds per split, for each ensemble size $k \in \{3, 5\}$ and each requested model-evaluation budget $Q \in \{10^3, 3 \times 10^3, 10^4, 3 \times 10^4, 10^5, 3 \times 10^5\}$. Thus, an experimental configuration is specified by the ensemble size k, the split, the algorithm seed, the budget Q, and the selection method. The same program logs the evaluation-budget breakdown for each method, the termination status of every greedy step, and the committee-composition statistics used in the subsequent analyses. Within each split, every method receives the same training and held-out matrices. The held-out tasks are used only to evaluate the coverage of an ensemble after it has been selected. They are never accessed during selection, used to determine a stopping rule, or used to choose any experimental parameter.

The randomized training task sequences are generated reproducibly from the corresponding experimental configuration. For comparisons between our method and AFG by Neoh et al. (2026), the two procedures are paired more directly: for each $(k, \text{split}, \text{seed}, Q)$, we generate a single random permutation of the training tasks and pass that same permutation to both methods. They therefore encounter exactly the same sequence of tasks, ensuring that differences between them cannot be attributed to the realized task order.

Implementation of the selection procedure. We run Algorithm 2 with confidence level $\delta = 0.1$ and ensemble sizes $k \in \{3, 5\}$. The confidence level is divided evenly across the k greedy steps, so each step applies Algorithm 1 at level δ/k. For each run, the training tasks are uniformly randomly permuted and partitioned into k contiguous blocks of equal size, one for each greedy step. Each step draws only from its assigned block. This prevents an early step from exhausting the finite training task sequence and, because the blocks are disjoint, ensures that the tasks used at a later step were

not used to select the earlier committee members. The requested evaluation budget Q is divided in the same way, giving each step a budget of Q/k model evaluations.

At a given step, the active set initially contains every candidate not already in the current committee. Tasks are processed from the step's block in order and without replacement. For each task, the current committee is evaluated first, in a fixed order: evaluation stops as soon as a committee member succeeds, and only the members actually evaluated are charged. If the committee covers the task, the procedure moves to the next task without evaluating the active candidates. If every committee member answers incorrectly, the task is accepted for candidate comparison and the active candidates are evaluated on it. Their empirical rescue counts and pairwise joint counts are then updated, and the count of accepted tasks increases by one.

For computational efficiency, the stopping condition is checked at predetermined sample counts rather than after every accepted task. The first checks occur when the number of accepted tasks reaches $8, 9, 10, \ldots$, after which the spacing grows by a factor of 1.02. Because the boundary in Theorem 4.1 is time-uniform, checking it only at a predetermined subset of times preserves its validity. At each check, the candidate with the largest empirical rescue rate is taken as the current leader and every other surviving candidate is compared with it. A candidate shown to be inferior at the specified confidence level is eliminated and is not evaluated on subsequent accepted tasks. The step terminates with a selection that meets the confidence criterion as soon as only one candidate remains, and that candidate is added to the committee.

Because the experiment uses a finite task sequence and a finite evaluation budget, a step can also terminate before meeting the confidence criterion. This occurs if its assigned task block is exhausted or if the next task cannot be processed without exceeding the remaining per-step budget. In either case, the procedure returns the candidate with the largest empirical rescue rate among the surviving candidates, without confirming the winner at the specified confidence level. We explicitly label each step as *certified*, *stream exhausted*, or *budget exhausted*, the implementation's labels for meeting the confidence criterion, exhausting the task sequence, or exhausting the evaluation budget, respectively. A run is marked as censored if at least one of its k steps terminates without meeting the confidence criterion. This distinction is important: the fixed-confidence guarantee of Theorem 4.1 applies when a step meets that criterion, whereas a censored step returns an empirical selection without this guarantee; its held-out coverage can still be evaluated.

There is one additional fallback case. If a step observes no accepted tasks, for example because the current committee covers every task in its assigned block, no empirical rescue rates can be compared. The implementation then returns the first active candidate under the fixed candidate ordering and labels the step as *no accepted failures*, indicating that no task was accepted for candidate comparison. This branch is not encountered in the reported experiments: every greedy step observes at least one task missed by the current committee.

For the comparison with AFG by Neoh et al. (2026), we use the same selection loop on missed tasks, including the same realized task sequence, block partition, committee evaluation that stops at the first correct answer, checking schedule, confidence allocation, and budget accounting. The methods differ only in the statistical confidence criterion used to eliminate candidates and determine when a selection meets the specified confidence level. This matched implementation measures the effect of the confidence criterion while holding the surrounding greedy-selection procedure fixed.

Budgets and cost accounting. The budget Q is measured in model evaluations, where one evaluation means evaluating one candidate on one task. Drawing or selecting a task itself incurs no cost. Because tasks are processed without replacement, a given candidate-task pair is evaluated at most once within a run. We distinguish evaluations of the current committee, which are used to determine whether the committee misses a task, from evaluations of active candidates on those missed tasks. Both are charged to the same total budget. Appendix E.5 gives the full accounting convention, which is applied consistently across methods.

We consider the six budgets $Q \in \{10^3, 3 \times 10^3, 10^4, 3 \times 10^4, 10^5, 3 \times 10^5\}$. Each is a hard upper bound rather than a target expenditure. For the methods that select on missed tasks, the budget is divided evenly across the k greedy steps. Before processing a new task, a step determines whether the evaluations required to process that task fit within its remaining budget. These include the committee evaluations needed to determine whether the committee misses the task and, if so, one evaluation for each candidate that remains active. If the required evaluations would exceed the remaining budget,

the task is not processed and the step terminates as budget exhausted. Consequently, no reported run spends more than its requested budget. Conversely, a method that confirms a selection at the specified confidence level or exhausts its assigned task block earlier is charged only for the evaluations it actually performs. The same accounting applies to runs that do not meet the confidence criterion: they are charged for work performed, not for the full budget available.

For the budgeted baselines, the same budget is translated into model evaluations as follows; Appendix E.6 gives their complete definitions. Top-k and exhaustive-elicitation ERM evaluate all $m = 16$ candidates on a without-replacement prefix of $T = \min\{\lfloor Q/m \rfloor, n_{\text{train}}\}$ training tasks. Top-k returns the k candidates with the highest empirical individual accuracies, whereas exhaustive-elicitation ERM enumerates all $\binom{16}{k}$ committees on the same observations and returns the exact empirical coverage maximizer. All methods operate on the same training matrix and are evaluated on the same held-out matrix, and no budgeted baseline receives task-level information unavailable to the methods that select on missed tasks. Full-information greedy and OPT_k are reported only as reference quantities rather than budget-matched competitors. Full-information greedy uses the entire training split and is charged m evaluations per training task, while OPT_k is the exact size-k coverage maximizer on the held-out split, computed by enumeration and reported without an evaluation budget.

Randomness and determinism. The protocol contains randomness in task construction, train-test splitting, and the task sequences used by the selection methods; model generation itself is deterministic. The initial paragraph sample is generated using the fixed seed 20260823, and split s uses seed $1000 + s$ to partition the selected paragraphs into training and test. The per-run randomization determines the order of the training task sequence and, for the budgeted baselines, the prefix of training tasks that is observed.

For each experimental configuration, the per-run random seed is derived deterministically from the candidate pool, run role, ensemble size k, split index, algorithm-seed index, and budget Q. We map these values to an integer seed using the BLAKE2B cryptographic hash function, making task permutations reproducible across executions and independent of process-specific randomization. Comparisons between our method and AFG use stronger pairing: for each configuration, a single task permutation is generated and passed explicitly to both methods. Thus, individual runs are reproducible, and the two methods that select on missed tasks use exactly the same realized task sequence.

Pool validation before the full evaluation. Before generating the full evaluation matrix, we used a small fixed subsample to verify that the candidate pool contained enough variation for complementarity to be meaningfully studied. This check was performed before any ensemble-selection results were computed, so that a pool consisting primarily of near-duplicate models, or one with essentially no room for ensemble improvement, would not simply produce an uninformative selection experiment. The diagnostic used the first 12 of the 300 selected questions in each of the eight languages, giving a 16×96 outcome matrix.

Five acceptance criteria, together with their justification, were fixed before this diagnostic matrix was generated. Every candidate was required to have accuracy at least 0.05; total pool coverage had to be at most 0.99; the coverage gain of the best size-3 ensemble over the best singleton had to be at least 0.04; the median pairwise exclusive mass had to be at least 0.10; and at most 25% of candidate pairs could have Jaccard overlap above 0.9. These conditions respectively rule out unusable candidates, a pool that already solves essentially every task, negligible room for improvement through ensemble selection, insufficient disagreement between candidates, and excessive redundancy. The observed values were 0.219, 0.958, 0.052, 0.302, and 0.008, respectively, so all five criteria were satisfied.

After this check, the candidate pool, prompt, decoding configuration, and scoring rule were frozen before any ensemble-selection result was computed. The criteria and their calibration were documented in the experiment configuration file. We describe them as fixed in advance rather than preregistered, since they were not filed with an external registry.

Fixed parameters. The main parameters fixed for the fresh evaluation are the confidence level $\delta = 0.1$, ensemble sizes $k \in \{3, 5\}$, the six budgets above, the two-shot completion prompt, the 24-token generation limit, batch size 16, the $75/25$ paragraph-level split, and the 1.02 growth factor

used for the schedule for checking the confidence criterion. None of these quantities was selected using held-out tasks or changed after the ensemble-selection results were observed.

The five pool-validation thresholds were calibrated on the three published evaluation matrices in Appendix E.2, where complementarity could be measured before the fresh evaluation, and were not derived from the fresh XQuAD selection results. Their calibration and rationale are documented in the experiment configuration file. Batch size 16 was chosen from a throughput scan on the evaluation hardware and affects computational throughput, but not the task presented to a model or its output under deterministic decoding.

E.5 Query accounting

We measure experimental cost in model evaluations, using the same distinction between committee-evaluation and candidate-evaluation queries as Neoh et al. (2026). The following rules are applied consistently to all methods in Track B.

1. One *model evaluation* consists of observing $u(d, c)$ for one candidate c on one task d. Drawing or selecting a task itself incurs no cost.
2. Tasks are processed without replacement within each run. Consequently, a given candidate-task pair is evaluated at most once within that run.
3. Evaluating whether the current committee S covers a task uses a fixed order and stops at the first correct answer. Committee members are evaluated sequentially until one succeeds or every member has been evaluated. Thus, establishing $U(d, S) = 1$ may cost fewer than $|S|$ evaluations, whereas establishing that the committee misses the task, $U(d, S) = 0$, always costs exactly $|S|$.
4. Rejection sampling is charged in full. In particular, committee evaluations spent on tasks that are rejected because the current committee covers them remain part of the evaluation cost.
5. Once a task missed by the current committee is accepted, evaluating the active candidate set A costs $|A|$ model evaluations. This matches the feedback used by Algorithm 1, which observes $(u(d, c))_{c\in A}$ on an accepted task rather than evaluating every candidate in C.
6. Following the decomposition of query cost in Neoh et al. (2026), we report committee-evaluation and candidate-evaluation costs separately. We denote by Q_{eval} the evaluations used to determine whether sampled tasks are missed by the current committee, and by Q_{cand} the evaluations of active candidates on accepted tasks. The total cost is $Q = Q_{\mathrm{eval}} + Q_{\mathrm{cand}}$. This distinction is useful because the two components arise from different parts of selection on missed tasks: Q_{eval} is the cost of finding such tasks, whereas Q_{cand} is the cost of distinguishing among possible additions once such tasks have been found.
7. The requested budget Q is a hard upper bound. Before processing a new task, a greedy step checks whether the evaluations required to process that task can fit within its remaining budget. The cost reserved for the task consists of the committee evaluations needed to determine whether the committee misses the task and, if so, one evaluation for every candidate that remains active. If this worst-case cost would exceed the remaining budget, the task is not processed and the step terminates. Consequently, no reported run exceeds its requested budget.
8. A greedy step that terminates without meeting its statistical confidence criterion is marked as censored. We distinguish whether this occurs because the step exhausts its assigned task block or because its evaluation budget is exhausted, and retain this termination status in the reported results.

The finite training task sequence is randomly permuted and partitioned into k disjoint contiguous blocks, one for each greedy step. Each step uses only its assigned block. This prevents an early step from consuming the entire finite benchmark and leaving no tasks for subsequent selections. More importantly, the disjoint blocks preserve the conditional sampling structure required by the greedy analysis: conditional on the committee selected at the previous steps, the tasks used at the next step have not themselves been used to construct that committee. This is the finite-benchmark implementation of the fresh-sample requirement underlying Proposition 4.1. Reusing the same tasks across greedy steps would not have this property, because the committee conditioning the later rescue test would already depend on those tasks.

E.6 BASELINES

We compare against baselines that separate three possible sources of improvement: using complementarity rather than individual accuracy, allocating evaluations adaptively rather than observing a complete evaluation matrix, and changing the statistical confidence criterion used within greedy selection on missed tasks. All budgeted methods are charged according to Appendix E.5; the full-information methods below are reported only as references.

Top-k by individual accuracy. This baseline estimates each candidate's individual accuracy from the same sample of training tasks and returns the k candidates with the largest estimates. Under budget Q, it evaluates all m candidates on $T = \min\{\lfloor Q/m \rfloor, n_{\text{train}}\}$ tasks drawn without replacement from the training task sequence, for a total cost of mT model evaluations. Top-k ignores overlap between candidates' successes: two models are preferred whenever their individual accuracies are larger, even if they solve almost exactly the same tasks. It therefore represents the natural selection rule induced by an individual-model leaderboard and serves as the main accuracy-only baseline.

De-duplicated Top-k. On pools containing multiple configurations of the same base model, naive Top-k can select several nearly redundant configurations of that model. We therefore also report a de-duplicated version that traverses candidates in decreasing empirical individual accuracy and skips a candidate if its base model is already represented in the committee. It uses exactly the same task-level evaluations as Top-k and additionally uses only the base-model identity associated with each configuration, which we treat as free metadata. This is our primary accuracy-only baseline for pools with multiple configurations of each base model, such as MMLU-PROMPTEVAL. On the fresh multilingual pool, where each candidate corresponds to a distinct model, Top-k and de-duplicated Top-k coincide. Appendix G.4 reports both versions where the distinction matters.

Sampled ERM (Neoh et al., 2026). This baseline approximates the committee with the highest empirical coverage by sampling N distinct size-k committees uniformly from the $\binom{m}{k}$ possibilities, evaluating their coverage on a training prefix, and returning the best. The number of committees and prefix length are chosen subject to the model-evaluation budget. Candidate-task outcomes shared across committees are evaluated once and reused, so the budget counts distinct model evaluations. Sampled ERM is non-greedy and searches over complete committees, but only a subset of all possible committees.

Exhaustive-elicitation ERM (Neoh et al., 2026). This baseline evaluates all m candidates on $T = \min\{\lfloor Q/m \rfloor, n_{\text{train}}\}$ training tasks, observing the complete $m \times T$ outcome matrix, then returns the size-k committee with maximum empirical coverage over all $\binom{m}{k}$ committees. Optimization after observing the matrix is not charged because the budget counts model evaluations rather than computation. Thus, it selects the best committee from a completely observed training prefix under the evaluation budget, but unlike methods that select on missed tasks, allocates evaluations uniformly across candidates and tasks before knowing which outcomes are relevant.

ADAPTIVE-FAIL-GREEDY (AFG) (Neoh et al., 2026). This is the closest prior method and the most direct comparison with our procedure. Like Algorithm 2, AFG constructs the committee greedily and evaluates possible additions primarily on tasks missed by the current committee. Conditional on the committee missing a task, a candidate's rescue rate is its probability of solving that task, so maximizing rescue rate is equivalent to maximizing the candidate's marginal coverage gain. AFG maintains empirical rescue rates for the active candidates and eliminates candidates once its confidence bounds establish that they are inferior to the current empirical leader.

For the experiments, we use AFG in exactly the same selection loop on missed tasks as our method: the two procedures receive the same task sequence, use the same block partition across greedy steps, the same committee evaluation that stops at the first correct answer, the same checking schedule, the same budget accounting, and the same confidence allocation δ/k. Their difference is therefore the statistical confidence criterion used to compare active candidates. AFG uses the Hoeffding-based confidence construction of Neoh et al. (2026), whereas Algorithm 1 uses the paired exclusive-outcome KL confidence criterion developed in this paper. We run the comparison as exact identification, corresponding to the $\eta \to 0$ limit of AFG's η-optimal elimination rule. This matched implementation is designed to measure the effect of the confidence criterion rather than differences in the surrounding greedy procedure.

Random committee. We draw a size-k subset uniformly from the candidate pool without observing any task outcomes. Its evaluation cost is therefore zero. This baseline is included only as a basic reference for how much the data-dependent methods gain from using task outcomes.

Full-information greedy. This reference observes the complete training outcome matrix and applies the standard forward greedy algorithm for maximum coverage. Starting from the empty committee, each step adds the candidate with the largest empirical marginal coverage gain until k candidates have been selected. Observing the complete matrix costs $m\, n_{\mathrm{train}}$ model evaluations, so the method is reported as a reference using all training outcomes rather than as a budget-matched competitor.

This construction is related to classical ensemble selection (Caruana et al., 2004), but it should not be interpreted as an implementation of the method of Caruana et al. (2004). Their procedure selects models with replacement and combines their predictions by averaging, whereas our objective is best-member coverage. Under a fully observed binary coverage objective, the corresponding forward-selection step is simply the standard greedy maximum-coverage rule used here; we claim no novelty for this full-information reference.

OPT_k **on the test split.** We compute the exact size-k committee with maximum coverage on the held-out split by enumerating all $\binom{m}{k}$ committees using bitset operations, with the implementation checked against direct brute-force enumeration on small instances. Because this procedure uses the held-out outcomes themselves to choose the committee, it is an oracle reference and is never treated as a selection method or budgeted competitor. It indicates the maximum held-out coverage attainable by any size-k committee in the candidate pool.

Focal-diversity pruning (Tekin et al., 2024). For the offline experiments of Appendix G.4, we additionally compare with the focal-diversity pruning criterion of Tekin et al. (2024), adapted to the best-member coverage objective studied here. We use only their sub-ensemble pruning idea. Their complete LLM-TOPLA pipeline then trains a learn-to-ensemble combiner, whose prediction rule differs from our objective $U(d,S) = \max_{c\in S} u(d,c)$. We therefore refer to this baseline as *focal-diversity pruning*, rather than as a reproduction of the complete LLM-TOPLA procedure.

We first define the two quantities used by our adaptation: empirical best-member coverage and focal diversity. Let $\mathcal{D}_{\mathrm{tr}}$ denote the complete training block. The *empirical best-member coverage* of a committee S is $\widehat{v}_{\mathrm{tr}}(S) := |\mathcal{D}_{\mathrm{tr}}|^{-1} \sum_{d\in\mathcal{D}_{\mathrm{tr}}} U(d,S)$, the empirical counterpart of the population coverage $v(S)$ defined in Section 2.

Focal diversity is computed from the pattern of incorrect answers across committee members. For a committee S with $|S| = k \geq 2$, let $\phi(d,S) := |\{c \in S : u(d,c) = 0\}|$ be the number of committee members that answer task d incorrectly. Each member $c \in S$ serves in turn as the *focal* member, and we restrict attention to the set of tasks it answers incorrectly $F_c := \{d \in \mathcal{D}_{\mathrm{tr}} : u(d,c) = 0\}$. When $F_c \neq \varnothing$, define $p_j^{(c)} := |\{d \in F_c : \phi(d,S) = j\}|/|F_c|$ for $j \in \{1,\dots,k\}$. Thus, $p_j^{(c)}$ is the conditional probability that exactly j members of S answer a training task incorrectly, given that the focal member c also answers incorrectly. Following the focal-diversity construction of Tekin et al. (2024), define $P_1^{(c)} := \sum_{j=1}^{k} (j/k) p_j^{(c)}$ and $P_2^{(c)} := \sum_{j=1}^{k} [j(j-1)/(k(k-1))] p_j^{(c)}$. These are respectively the probabilities that one and two uniformly chosen committee members answer incorrectly on a task drawn from F_c. The focal negative-correlation score of c within the committee S is $\rho_{\mathrm{focal}}(c;S) := 1 - P_2^{(c)}/P_1^{(c)}$. The *focal diversity* of committee S is then $\lambda_{\mathrm{focal}}(S) := k^{-1} \sum_{c\in S} \rho_{\mathrm{focal}}(c;S)$. If a focal member answers every task in the training block correctly, its contribution is defined to be zero; for $k < 2$, we set $\lambda_{\mathrm{focal}}(S) = 0$.

Our adapted pruning objective combines focal diversity with the best-member coverage objective studied in this paper. We use equal weights and select $\widehat{S} \in \arg\max_{S\in\mathcal{C}_k} \{\frac{1}{2}\lambda_{\mathrm{focal}}(S) + \frac{1}{2}\widehat{v}_{\mathrm{tr}}(S)\}$, where $\mathcal{C}_k$ is the set of candidate size-k committees considered by the search. Equal weighting is an untuned design choice of our adaptation, rather than a parameter choice inherited from Tekin et al. (2024); Appendix G.4 verifies that the empirical comparison is insensitive to this choice.

The focal-diversity statistic above follows the construction of Tekin et al. (2024), while the performance term, weighting, and search procedure are our adaptation to the present setting. LLM-TOPLA combines focal diversity with validation accuracy and uses weights $(0.6, 0.4)$ for its MMLU and GSM8K experiments, whereas we replace validation accuracy with empirical best-member coverage and weight the two terms equally. LLM-TOPLA also searches over sub-ensembles using a genetic

algorithm, while our experiment fixes the committee size at k. When the candidate pool contains more than 24 models, we retain the 24 candidates with the highest individual accuracy on $\mathcal{D}_{\mathrm{tr}}$. If the resulting number of size-k committees is at most 4,000, we evaluate all of them; otherwise, we evaluate 4,000 uniformly sampled size-k committees using the fixed experimental random number generator. In either case, we also include the full-information greedy committee in $\mathcal{C}_k$. The search space therefore always contains the greedy solution, although the adapted focal-diversity objective may assign a higher score to another committee.

Proposed method and experimental variants. The proposed statistical procedure is Algorithm 1, used within the greedy construction of Algorithm 2. In the experimental implementation, a candidate is removed from the active set once the confidence criterion establishes that it is inferior to the current empirical leader. We refer to this implementation as KL-greedy on missed tasks with elimination. This elimination is an implementation optimization rather than a separate statistical procedure: the simultaneous confidence guarantee underlying Theorem 4.1 holds over all candidate pairs and sample sizes, so a candidate shown to be inferior at the specified confidence level need not be evaluated on subsequent accepted tasks. Elimination therefore preserves the guarantee of Theorem 4.1 while reducing candidate-evaluation cost from evaluating the full active set until stopping to $\sum_{n \le \tau_S} |A_n|$, where A_n denotes the candidates still active after n accepted tasks.

The paired-Bernstein, paired-Hoeffding, and marginal-KL variants in Table 1 are experimental controls used to separate the effects of the observed feedback and the statistical confidence criterion. They are not additional methods proposed by this paper or baselines taken from prior work. Appendix F.5 defines these variants and the component changed by each one.

E.7 Metrics

Our primary performance metric is held-out ensemble coverage, $v_{\mathrm{test}}(S)$, defined as the fraction of held-out tasks solved by at least one member of the selected ensemble S. This is the empirical counterpart of the coverage objective $v(S)$ defined in Section 2. The exact held-out optimum OPT_k therefore provides an oracle upper bound for comparing ensembles of size k.

Our primary cost metric is the total number Q of model evaluations, where one model evaluated on one task counts as one evaluation. As described in Appendix E.5, we decompose this cost as $Q = Q_{\mathrm{eval}} + Q_{\mathrm{cand}}$. Here, Q_{eval} counts evaluations of the current committee used to find tasks it misses, whereas Q_{cand} counts evaluations of active candidates on those accepted tasks. We report both components because they correspond to different stages of selection on missed tasks, but use Q for comparisons across methods because it captures the total evaluation cost.

We additionally measure T_{raw}, the number of tasks processed from the training task sequence, and T_{acc}, the number of those tasks that the current committee misses and that are therefore accepted for candidate comparison. These task counts are not interchangeable with model evaluations: one raw task can require several committee evaluations, and one accepted task can additionally require evaluations of multiple active candidates. We report T_{acc} in particular because accepted tasks missed by the current committee are the statistical observations used by Algorithm 1 and the unit in which the sample complexity of Theorem 4.1 is expressed.

When assessing whether the confidence criterion can be met, a run is *censored* if at least one greedy step terminates without meeting that criterion because its assigned task block or evaluation budget is exhausted. We report the censoring rate as the fraction of runs satisfying this condition. Thus, censoring measures whether the finite benchmark permits a selection at the specified confidence level and is distinct from the held-out quality of the ensemble returned when that level is not reached.

The remaining metrics are used in specific analyses. For two candidates or ensembles, the pairwise exclusive mass is $s = p_{10} + p_{01}$, the probability that exactly one succeeds; this is the disagreement quantity studied in Theorem 3.2. Jaccard overlap is the intersection-over-union of the two candidates' solved-task sets and is used to quantify redundancy in the candidate pool. For global identification, $\Gamma(P)$ is the information gap defined in §3.3, and $1/\Gamma(P)$ is the corresponding first-order characteristic time from Theorem 3.3.

E.8 Statistical methodology

The fresh XQuAD analysis uses 10 independent paragraph-level splits, 20 algorithm seeds per split, 6 evaluation budgets, and both $k \in \{3, 5\}$. Comparisons are matched within each experimental configuration: methods are evaluated on the same training and held-out split, and matched comparisons between methods that select on missed tasks use the same realized task permutation as described in Appendix E.4. We therefore analyze differences in held-out coverage between methods as paired observations rather than as independent samples.

For each comparison, we report the mean paired difference in held-out coverage with a 95% confidence interval based on variation across the executed runs. As a robustness check, a $20{,}000$-resample percentile bootstrap gives nearly identical intervals and unchanged conclusions. We therefore report effect sizes and confidence intervals rather than p-values.

Method comparisons are reported separately at each evaluation budget. This is important because the relative performance of the methods changes with the available budget: selection on missed tasks can underperform simpler baselines when very few evaluations are available and outperform them once enough data have been observed to identify complementary candidates. Averaging across the budget grid would combine these distinct settings and obscure the dependence on evaluation cost. When we summarize performance over a range of budgets, we do so only where that range corresponds to a common pattern of empirical performance and state the range explicitly.

Runs that exhaust their assigned task sequence or evaluation budget before meeting the confidence criterion remain part of the empirical analysis and are explicitly marked as censored; they are neither extended beyond the available data nor discarded. Their returned ensembles are still evaluated on the held-out split, while their lack of a confidence guarantee is reported separately. Finally, held-out tasks are used only for evaluating already selected ensembles. No hyperparameter, stopping threshold, evaluation budget, or other experimental choice is selected using held-out performance.

F Track A: Statistical validation of the information rates and confidence criteria

Track A tests the statistical predictions underlying our selection procedure. We first ask whether the information quantities identified by the theory predict the empirical difficulty of comparison and identification on real LLM evaluation data. Because the complete outcome matrices are available, these experiments can treat each matrix as a known empirical population, compute the relevant information quantities exactly, and sample tasks i.i.d. from that population as described in Appendix E.1. We then turn from the information rates themselves to the sequential stopping criterion, using controlled experiments to examine how the observed feedback, confidence bound, and comparison scope affect the amount of evidence required to reach a decision at the specified confidence level. Track A therefore studies the statistical difficulty and evidence required for identification at that confidence level, while Track B asks the complementary practical question of which ensembles are selected under finite model-evaluation budgets.

Appendix F.1 tests the invariance prediction of Theorem 3.2: once the probabilities of the two exclusive outcomes are fixed, varying the probability that both ensembles succeed should not change the intrinsic difficulty of comparing them. Appendix F.2 tests the complementary prediction by holding the coverage gap fixed while varying the amount of exclusive disagreement, which the same theorem predicts should change comparison difficulty. Appendix F.3 moves from pairwise comparison to global identification and tests whether the characteristic time of Theorem 3.3 predicts identification cost across candidate pools and confidence levels. Appendix F.4 examines the weighted ordinal extension of Theorem B.1 using the LiveBench-derived evaluation matrix.

Appendix F.5 then examines the sequential stopping criterion separately from the selection objective. Holding the evidence from tasks missed by the current ensemble fixed, it varies the measurement, confidence bound, and comparison scope to determine which components explain why the required evidence differs between our exclusive-KL procedure and AFG (Neoh et al., 2026). Appendix F.6 asks whether the fixed-confidence decisions studied in these controlled experiments are actually reachable on finite benchmark populations, connecting the statistical rates and evidence requirements to the informative observations available in real evaluation matrices. Finally, Appendix F.7

reports implementation checks that verify the numerical information quantities, evaluation accounting, and sequential stopping behavior used in these experiments.

F.1 HOLDING THE EXCLUSIVE OUTCOMES FIXED

We first test the invariance prediction of Theorem 3.2. For two ensembles A and B, recall that p_{ij} denotes the probability of the joint outcome $(U(d, A), U(d, B)) = (i, j)$. Thus, p_{10} is the probability that A succeeds and B fails, p_{01} is the probability that B succeeds and A fails, p_{11} is the probability that both succeed, and p_{00} is the probability that both fail. The two exclusive outcomes determine both the coverage gap $\Delta = v(A) - v(B) = p_{10} - p_{01}$ and the exclusive mass $s = p_{10} + p_{01}$, the probability that exactly one ensemble succeeds. Moreover, Theorem 3.2 shows that the information determining comparison difficulty is $I^P_{\mathrm{cov}}(A, B) = J(p_{10}, p_{01})$ and therefore does not depend on the agreement probabilities p_{00} and p_{11}.

This experiment tests that prediction directly by fixing p_{10} and p_{01} and varying p_{11}. The coverage gap Δ, the exclusive mass s, and the information $I^P_{\mathrm{cov}}(A, B)$ are consequently unchanged throughout the range tested. If the theory identifies the relevant information correctly, changing how often both ensembles succeed should not materially affect a comparison procedure that exploits their shared task-level outcomes.

Construction. We begin with a 40-candidate subsample of MMLU-PROMPTEVAL and search over candidate ensemble pairs of size 2 and size 3, both ensembles in a pair always having the same size. We keep the pair that maximizes the smallest of its four joint-outcome cell counts, so that every cell contains a large number of real tasks. The selected pair consists of two size-2 ensembles with empirical cell counts 3,277, 2,383, 2,208 and 6,174 for $(00, 01, 10, 11)$. This confirms that the experimental populations can be constructed: these counts exceed the number required by every controlled population used below, so each prescribed four-cell profile is attainable from real evaluation outcomes rather than being an arbitrary choice of numbers. The ensemble size itself carries no significance for what follows, since Theorem 3.2 concerns the four joint-outcome cells of two fixed ensembles and not their size.

For each condition we then construct a population of $N = 4{,}500$ joint-outcome labels with exactly the prescribed cell proportions, and the sequential procedures draw from it i.i.d. with replacement. Here we fix $(p_{10}, p_{01}) = (0.30, 0.12)$ exactly while varying $p_{11} \in \{0.02, 0.05, 0.10, 0.16, 0.24, 0.32, 0.40, 0.48, 0.55\}$; the remaining probability mass is assigned to p_{00}. Hence, the coverage gap is fixed at $\Delta = 0.18$, the exclusive mass at $s = 0.42$, and the exclusive-coverage information at $J(0.30, 0.12)$. Only the agreement probabilities change across the values tested, through the probability p_{11} that both ensembles succeed. Both procedures compared here depend on a task only through the joint coverage outcome $(U(d, A), U(d, B))$, so task identities carry no further information for this experiment.

Comparison procedures. We compare two sequential procedures at confidence level $\delta = 0.05$ over 80 independent seeds. The first is the paired exclusive-KL test, which uses the joint outcomes of the two ensembles and stops when the empirical exclusive-outcome evidence crosses a confidence boundary that is valid simultaneously at all sample sizes. The statistic is therefore determined by the empirical analogues of p_{10} and p_{01}. The second procedure deliberately discards this pairing. It evaluates the two ensembles on disjoint halves of the sampled task sequence, estimates their marginal coverages separately, and compares empirical-Bernstein confidence intervals. This represents the common practice of evaluating two systems on separate benchmark samples and comparing their estimated success rates. In particular, it does not use the exclusive-outcome information I_{cov}.

The experiment therefore does not simply plot a procedure against a quantity it explicitly optimizes. The experiment varies p_{11}, which Theorem 3.2 predicts should be irrelevant to the information rate, and the unpaired comparator never computes I_{cov}. The empirical prediction is that the paired exclusive test should remain stable as p_{11} changes, whereas a procedure based on separate marginal estimates may respond to the changing marginal variances.

Results. As shown in Figure 1(b), the paired exclusive-KL test consumes between 1,108 and 1,191 tasks across the entire range tested, a ratio of only 1.075 between its largest and smallest mean stopping times. Its mean standard error across the values tested is approximately 32 tasks. Thus, changing p_{11} from 0.02 to 0.55, a 27-fold change, produces only a 7.5% variation in stopping time.

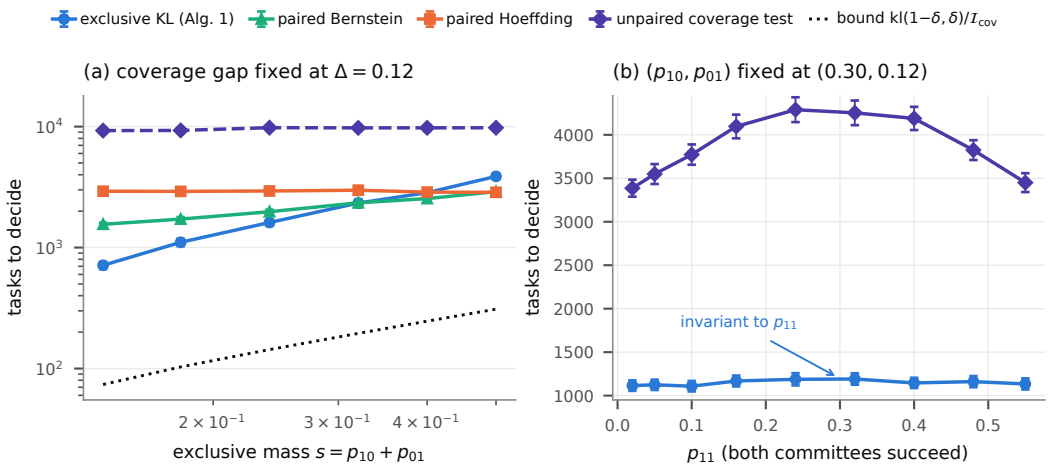


Figure 1: **Controlled tests of exclusive disagreement.** The vertical axis shows the number of sampled tasks consumed before the stopping criterion is met; for the unpaired comparator, this is the total across both halves of the task sequence. Error bars show 95% confidence intervals. **(a)** Holding the coverage gap $\Delta = 0.12$ fixed, we vary the exclusive mass $s = p_{10} + p_{01}$. The dotted curve is the information-theoretic lower bound from Theorem 3.2. Results use 40 seeds at $\delta = 10^{-2}$. **(b)** Holding $(p_{10}, p_{01}) = (0.30, 0.12)$ fixed, we vary the agreement probabilities p_{11} and p_{00}. Results use 80 seeds at $\delta = 0.05$.

The unpaired comparison behaves differently. Its mean stopping time ranges from 3,386 to 4,289 tasks, a $1.267\times$ spread, and peaks at $p_{11} = 0.24$. At this point $v(A) = 0.54$, close to the value $1/2$ at which a Bernoulli coverage estimate has its largest variance. The procedure is therefore responding to changes in the marginal coverage variance even though the exclusive probabilities that determine which ensemble is better have not changed.

Conclusion. Therefore, holding the exclusive outcomes fixed makes the exclusive-KL comparison nearly invariant to large changes in the probability that both ensembles succeed. An unpaired comparison based on separate coverage estimates instead changes by about 27% as the agreement mass varies. This is consistent with the prediction of Theorem 3.2 that agreement outcomes do not determine the intrinsic difficulty of comparing two ensembles; their apparent influence arises when the comparison procedure estimates the two marginal coverages separately rather than exploiting their shared task-level outcomes.

F.2 Holding the coverage gap fixed and varying exclusive disagreement

Next, we test the complementary prediction proved in Theorem 3.2. Recall that $\Delta = p_{10} - p_{01} = v(A) - v(B)$ is the coverage gap between the two ensembles, while $s = p_{10} + p_{01}$ is their *exclusive mass*, i.e., the probability that exactly one ensemble succeeds. The first experiment held both quantities fixed and varied only the agreement probabilities. Here, we instead hold Δ fixed and vary s.

This distinction matters because the coverage gap alone does not determine the information available for comparison. For small Δ relative to s, the exclusive-coverage information satisfies $I^P_{\mathrm{cov}}(A, B) = \Delta^2/(2s) + \mathcal{O}(\Delta^4/s^3)$, as discussed after Theorem 3.2. Thus, at a fixed coverage gap, increasing s decreases the information available per task and should increase the number of tasks required to determine which ensemble has larger coverage. The experiment tests this prediction while keeping the difference Δ between the two marginal coverage values unchanged.

Construction. Using the same ensemble pair with sufficient joint-outcome counts as in Appendix F.1, we fix the coverage gap at $\Delta = 0.12$ and vary the exclusive mass over $s \in \{0.14, 0.18, 0.24, 0.32, 0.40, 0.50\}$. For each value of s, the two exclusive probabilities are therefore $p_{10} = (s + \Delta)/2$ and $p_{01} = (s - \Delta)/2$. We hold $p_{11} = 0.30$ fixed and assign the remaining

mass to p_{00}, again constructing a population of $N = 4{,}500$ joint-outcome labels and drawing from it i.i.d. with replacement. The selected pair's cell counts exceed the demand of every profile in this experiment, so each is attainable from real evaluation outcomes. Across the values tested, the two ensembles consequently have the same coverage difference but different probabilities of disagreeing.

Comparison procedures. We compare four sequential tests for determining the sign of $\Delta = v(A) - v(B)$. All tests are run at confidence level δ and use the same time-uniform confidence allocation. Define

$$L(n, \delta) := \log\big(\pi^2 n^2/(6\delta)\big), \qquad \beta(n, \delta) := L(n, \delta) + 4\log(n + 1). \tag{7}$$

The choice of L allocates error probability $\exp(-L(n, \delta)) = 6\delta/(\pi^2 n^2)$ to sample size n, which sums to δ over $n \geq 1$ due to $\sum_{n\geq 1} 6/(\pi^2 n^2) = 1$. The resulting boundaries therefore hold simultaneously over all sample sizes.

For computational efficiency, we evaluate the stopping conditions only on the common multiplicative grid $\mathcal{G} = \{n_1 = 8,\ n_{j+1} = \max(n_j + 1, \lfloor 1.02 n_j \rfloor + 1)\}$. Since the confidence boundaries are valid for every n, checking them only on $\mathcal{G}$ does not affect validity and can only delay stopping. In the reported experiments, this discretization increases stopping times by at most 2.1%.

Exclusive-KL. After n paired observations, let $\widehat{p}_{10,n}$ and $\widehat{p}_{01,n}$ denote the empirical frequencies of the two exclusive outcomes: A succeeds and B fails, and B succeeds and A fails, respectively. The test stops at the first $n \in \mathcal{G}$ such that

$$nJ(\widehat{p}_{10,n}, \widehat{p}_{01,n}) \geq \beta(n, \delta) \qquad \text{or} \qquad nJ(\widehat{p}_{01,n}, \widehat{p}_{10,n}) \geq \beta(n, \delta), \tag{8}$$

returning A in the first case and B in the second. Because $J(x, y) = 0$ when $x \leq y$, the two conditions cannot hold simultaneously. This is the two-ensemble version of the stopping criterion used in Algorithm 1. Since only one pair of ensembles is compared here, the pairwise multiplicity term of $\beta_r(n, \delta)$ in the general boundary is set to one.

Paired Hoeffding. For each sampled task, define the paired coverage difference $Z_t = U(d_t, A) - U(d_t, B) \in \{-1, 0, 1\}$ and its empirical mean $\widehat{\Delta}_n = n^{-1} \sum_{t\leq n} Z_t$. The test stops at the first $n \in \mathcal{G}$ for which $|\widehat{\Delta}_n| > r_n^{\mathrm{H}}$, where

$$r_n^{\mathrm{H}} := \sqrt{2L(n, \delta)/n}, \tag{9}$$

and returns the ensemble indicated by the sign of $\widehat{\Delta}_n$. This is the Hoeffding radius for a variable with range 2. It depends only on the range $[-1, 1]$ and therefore does not adapt to how frequently the two ensembles disagree.

Paired Bernstein. This test uses the same paired differences Z_t and empirical mean $\widehat{\Delta}_n$, but adapts to their observed variance. Let

$$V_n := \frac{1}{n-1} \sum_{t\leq n} \big(Z_t - \widehat{\Delta}_n\big)^2 = \frac{n}{n-1} \left[\widehat{p}_{10,n} + \widehat{p}_{01,n} - \widehat{\Delta}_n^2\right]. \tag{10}$$

The second equality follows because $Z_t^2 = 1$ exactly on the two exclusive outcomes and is zero otherwise. With $L_{\mathrm{MP}}(n, \delta) := L(n, \delta) + \log 2$, define

$$r_n^{\mathrm{B}} := \sqrt{\frac{2V_n L_{\mathrm{MP}}(n, \delta)}{n}} + \frac{14 L_{\mathrm{MP}}(n, \delta)}{3(n-1)}. \tag{11}$$

The test stops at the first $n \in \mathcal{G}$ such that $|\widehat{\Delta}_n| > r_n^{\mathrm{B}}$ and again returns the ensemble indicated by the sign of $\widehat{\Delta}_n$. The radius in equation 11 is obtained from the empirical-Bernstein bound of Maurer & Pontil (2009). Its one-sided form for variables in $[0, 1]$ uses the unbiased sample variance, a $\log(2/\delta)$ confidence term, and an additive constant 7. Applying the bound to $(Z_t + 1)/2$ and rescaling back to $Z_t \in [-1, 1]$ leaves the variance term in the form above and doubles the additive constant from 7 to 14. Unlike paired Hoeffding, this radius can therefore respond to changes in the exclusive mass through the empirical variance of Z_t.

Unpaired. The final comparator discards task-level pairing by estimating the two ensemble coverages from separate task samples. For each comparison, A and B receive disjoint samples of equal size. After each has been evaluated on n tasks, let $\widehat{v}_{A,n}$ and $\widehat{v}_{B,n}$ denote their empirical coverages, and define the unbiased Bernoulli sample variances $V_{A,n} = \frac{n}{n-1}\widehat{v}_{A,n}(1 - \widehat{v}_{A,n})$ and

$V_{B,n} = \frac{n}{n-1}\widehat{v}_{B,n}(1-\widehat{v}_{B,n})$. With $L_{\mathrm{U}}(n,\delta) := L(n,\delta) + \log 4$, the corresponding empirical-Bernstein radius for A is $r_{A,n} := \sqrt{\frac{2V_{A,n}L_{\mathrm{U}}(n,\delta)}{n}} + \frac{7L_{\mathrm{U}}(n,\delta)}{3(n-1)}$, with $r_{B,n}$ defined analogously. The $\log 4$ term allocates the confidence budget across the two marginal bounds. The test stops when the confidence bounds separate: it returns A if $\widehat{v}_{A,n} - r_{A,n} > \widehat{v}_{B,n} + r_{B,n}$ and returns B under the symmetric condition. Because the two ensembles have each been evaluated on n separately sampled tasks, the reported stopping time is $2n$. This comparator represents evaluating the two ensembles separately rather than exploiting their outcomes on shared tasks; it therefore discards the task-level disagreement information available to the three paired procedures.

Confidence allocation. The factor n^{-2} in $L(n,\delta)$ makes the fixed-time error probabilities summable over all sample sizes: $\sum_{n\geq 1}\exp(-L(n,\delta)) = \delta$. This provides the time-uniform guarantee used by the sequential tests without an additional construction that groups sample sizes or combines bounds or tests. For the exclusive-KL and paired-Hoeffding tests, under either fixed true ordering only a crossing in the wrong direction constitutes an error. It is therefore sufficient to control the corresponding one-sided event at total level δ. Paired Bernstein uses the same one-sided allocation; the additional $\log 2$ in $L_{\mathrm{MP}}(n,\delta)$ comes from the confidence-level convention in the empirical-Bernstein bound of Maurer & Pontil (2009).

The unpaired test instead constructs separate confidence bounds for the two marginal coverages. We allocate error probability $\delta/2$ to each marginal bound and combine the two guarantees by a union bound. Together with the $\log 2$ convention of the empirical-Bernstein inequality, this yields the $\log 4$ term in $L_{\mathrm{U}}(n,\delta)$. Under these allocations, all four comparison procedures have error probability at most δ.

The exclusive-KL stopping criterion uses a more conservative boundary than the two paired controls: $\beta(n,\delta)$ contains the additional $4\log(n+1)$ term needed by the four-outcome confidence bound underlying Algorithm 1. The Hoeffding and Bernstein controls do not incur this term. For example, at $n = 1{,}493$ and $\delta = 10^{-2}$, $4\log(n+1) = 29.2$, compared with a total boundary $\beta(n,\delta) = 49.0$. Thus, any empirical advantage of exclusive-KL in this experiment is obtained despite its larger stopping threshold.

The four tests use progressively different aspects of the available feedback. Exclusive-KL retains the two exclusive-outcome frequencies separately and is the only procedure that explicitly computes I_{cov}. Paired Bernstein uses the variance of the paired differences, while paired Hoeffding uses only their fixed range. The unpaired comparator additionally discards shared-task pairing. We use 40 independent seeds for each condition.

Results. The table below reports the results at $\delta = 10^{-2}$.

s	I_{cov}	bound	exclusive KL	paired Bernstein	paired Hoeffding	unpaired
0.14	0.0610	74	714 ± 26	$1{,}557 \pm 25$	$2{,}923 \pm 52$	$9{,}265 \pm 228$
0.18	0.0437	103	$1{,}106 \pm 40$	$1{,}719 \pm 35$	$2{,}914 \pm 63$	$9{,}296 \pm 228$
0.24	0.0314	143	$1{,}610 \pm 50$	$1{,}978 \pm 40$	$2{,}937 \pm 66$	$9{,}797 \pm 223$
0.32	0.0231	195	$2{,}334 \pm 91$	$2{,}344 \pm 70$	$2{,}979 \pm 93$	$9{,}759 \pm 264$
0.40	0.0183	246	$2{,}844 \pm 108$	$2{,}540 \pm 81$	$2{,}874 \pm 104$	$9{,}767 \pm 273$
0.50	0.0145	310	$3{,}877 \pm 111$	$2{,}900 \pm 75$	$2{,}862 \pm 104$	$9{,}781 \pm 249$
spread across s		4.20×	5.43×	1.86×	1.04×	1.06×

Our exclusive-KL test provides the largest savings when the exclusive mass is small. At $s = 0.14$, it stops after 714 tasks on average, compared with $1{,}557$ for paired Bernstein, $2{,}923$ for paired Hoeffding, and $9{,}265$ for the unpaired comparison. Thus, exclusive-KL uses $2.2\times$ fewer tasks than the strongest paired control, $4.1\times$ fewer than paired Hoeffding, and $13.0\times$ fewer than the unpaired comparison. It remains the fastest procedure through $s = 0.24$. These are the cases in which the information identified by Theorem 3.2 is largest: when disagreement is relatively rare but favors one ensemble consistently, the exclusive outcomes provide particularly strong evidence about which ensemble has larger coverage.

The advantage changes systematically as the exclusive mass increases. Across the range from $s = 0.14$ to $s = 0.50$, the information-theoretic lower bound increases by $4.20\times$, and the stopping time of exclusive-KL increases by $5.43\times$ in the same direction. Paired Bernstein responds more

moderately, with a $1.86\times$ spread, because its radius adapts to the variance of the paired difference. Paired Hoeffding and the unpaired comparison are nearly insensitive to these changes, varying by only $1.04\times$ and $1.06\times$, respectively. Consequently, exclusive-KL is approximately tied with paired Bernstein at $s = 0.32$ and is slower than the paired controls at the two largest values of s. The experiment therefore does not show that exclusive-KL uniformly dominates generic concentration bounds. Rather, it shows that its cost adapts strongly to the information structure identified by the theory, yielding substantial savings precisely on instances with favorable exclusive disagreement.

The Bernstein comparison helps separate this information effect from variance adaptation alone. Theorem 3.2 does not predict the paired-Bernstein stopping curve: Bernstein uses the same paired observations but summarizes their information through the variance of $U(d, A) - U(d, B)$, whereas exclusive-KL retains the two exclusive frequencies separately. Its intermediate response to s shows that variance adaptation captures part, but not all, of the changing structure. Across every tested value of s and all three confidence levels, no measured mean stopping time falls below the information-theoretic lower bound. Figure 1(a) shows results across the complete range.

Conclusion. Accordingly, the coverage gap alone does not determine either comparison difficulty or the benefit of exploiting exclusive outcomes. At fixed Δ, exclusive-KL is more than twice as sample-efficient as the strongest paired control at low exclusive mass and substantially more efficient than range-only or unpaired comparisons. As s increases, its advantage disappears at exactly those values where I_{cov} predicts that exclusive-outcome evidence becomes less informative. Together with Appendix F.1, these experiments provide evidence for both implications of Theorem 3.2: agreement outcomes have little effect once the exclusive outcomes are fixed, while changing exclusive disagreement can substantially alter both the intrinsic comparison difficulty and the value of using the information identified by the theory.

F.3 Characteristic time under full-vector feedback

We now test the global prediction of Theorem 3.3: under full-vector feedback, the difficulty of identifying the optimal ensemble is determined by the information gap $\Gamma(P)$, with characteristic time $1/\Gamma(P)$.

Experimental design. We construct instances from MMLU-PromptEval, RouterBench, and GSM8K. For MMLU-PromptEval and GSM8K, we use candidate-pool sizes $m \in \{6, 8, 12, 16, 22\}$; for RouterBench, whose candidate pool is smaller, we use $m \in \{6, 8, 9, 11\}$. We consider ensemble sizes $k \in \{1, 2, 3, 4\}$ whenever feasible. Because the full population matrix is available, $\Gamma(P)$ and the optimal ensemble are computed exactly for each instance before the sequential simulation.

The instances are grouped by predicted difficulty so that the predicted stopping time n_0 spans from 626 to 3.99×10^5 tasks, a factor of 637. We evaluate $\delta \in \{10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}\}$ with 12 independent seeds for each (instance, δ) cell and impose a compute cap of 4×10^5 sampled tasks per run. As throughout Track A, tasks are drawn i.i.d. with replacement from the corresponding empirical population. The resulting task sequence is therefore not limited by the number of columns in the original matrix. This is why the 1,319-task GSM8K matrix can be used here even though it is too short for the finite without-replacement task sequences used in Track B.

Stopping rule and comparator. The disagreement-aware procedure is the full-vector identification rule associated with Theorem 3.3. After n sampled tasks, we compute the empirical coverage $\widehat{v}_n(S)$ of every $S \in \mathcal{S}_k$, let $\widehat{S}_n$ be the empirical maximizer, and compare it with every competing ensemble using their empirical exclusive-outcome frequencies. The procedure stops once the resulting empirical information against every alternative crosses the confidence boundary valid simultaneously for all alternatives and sample sizes. Thus, the stopping decision is computed entirely from the observations seen so far. In particular, the population quantity $\Gamma(P)$ is not supplied to the procedure: because the complete evaluation matrix is available in Track A, we compute $\Gamma(P)$ separately from the population and use it only to test the predicted relationship between instance difficulty and stopping time.

As a control, we run a uniform full-vector Hoeffding procedure on exactly the same observation sequence. After n tasks, it likewise computes $\widehat{v}_n(S)$ for every $S \in \mathcal{S}_k$ and identifies the empirical leader $\widehat{S}_n$. It then compares the leader's empirical coverage with that of every alternative using the

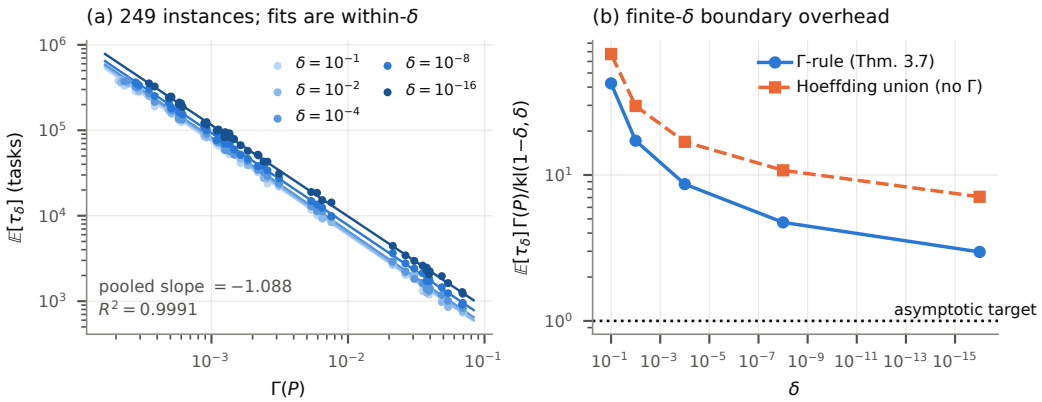


Figure 2: **Characteristic time under full-vector feedback.** Results for the 249 reachable $(\text{instance}, \delta)$ cells, with 12 seeds per cell. **(a)** Mean stopping time versus the population information gap $\Gamma(P)$; lines show least-squares fits within each confidence level. **(b)** The same stopping times normalized by the first-order scale $\mathrm{kl}(1-\delta, \delta)/\Gamma(P)$. The full-vector Hoeffding comparator uses marginal coverage confidence bounds rather than exclusive-outcome information.

time-uniform Hoeffding radius r_n^{H} defined in Appendix F.2, with the confidence allocation adjusted simultaneously over the $M = \binom{m}{k}$ candidate ensembles. The procedure stops only when the lower confidence bound for $\widehat{S}_n$ exceeds the upper confidence bound for every $S \neq \widehat{S}_n$; otherwise, sampling continues. The resulting rule is δ-correct and, unlike the disagreement-aware procedure, uses only the marginal empirical coverages of the ensembles when deciding whether to stop.

The comparison therefore holds the feedback and sampling policy fixed. Both procedures observe the complete success vector on every sampled task, both evaluate the same M ensembles, and both must establish at the specified confidence level that the same empirical leader is better than all alternatives. They differ only in how the observed full-vector feedback is converted into evidence for stopping: the proposed procedure uses the pairwise exclusive-outcome information that determines $\Gamma(P)$, whereas the Hoeffding control uses marginal coverage differences to establish separation at the specified confidence level, using the range-based radius of Appendix F.2.

Reachable instances. The compute cap is applied after the instances are constructed, rather than used to select them. Of the sampled instances, 226 have predicted stopping times above 4×10^5 tasks, including instances with $\Gamma(P)$ as small as 1.6×10^{-8}. We mark these instances as unreachable under the experimental compute budget rather than removing them from the construction; their distribution is summarized in Figure 4(a). The scaling analysis below uses the 249 $(\text{instance}, \delta)$ cells that are reachable under this cap. It is therefore necessarily concentrated on the easier part of the constructed instance family.

Results. As shown in Figure 2(a), the central prediction is strongly reflected in the data. Within each fixed confidence level, mean stopping time follows the predicted inverse dependence on $\Gamma(P)$ closely. Regressing $\log \mathbb{E}[\tau_\delta] - \log \mathrm{kl}(1-\delta, \delta)$ on $\log \Gamma(P)$ separately at each δ gives slopes between -1.067 and -1.100. The coefficient of determination is $R^2 \geq 0.9991$, meaning that more than 99.9% of the variation in log stopping time within each confidence level is explained by the fitted dependence on $\log \Gamma(P)$; the residual standard deviation is only 0.047 to 0.060 in log space. Pooling the cells with confidence-level fixed effects gives a slope of -1.088 (95% CI $[-1.092, -1.084]$), $R^2 = 0.9991$, and residual standard deviation 0.058 over the 249 reachable cells. Thus, across three benchmarks, four ensemble sizes, and candidate families ranging from $M = 6$ to $M = 1{,}540$, the information gap $\Gamma(P)$ explains nearly all of the within-confidence variation in identification cost.

The fitted exponent is close to, but detectably steeper than, the -1 first-order exponent in Theorem 3.3. We do not interpret this difference as a contradiction of the theorem, whose statement concerns the limit $\delta \to 0$ with the instance fixed. At finite confidence, the stopping boundary also

contains terms that grow with the sample size and the number of candidate ensembles, so exact proportionality to $1/\Gamma(P)$ is not expected at the confidence levels used here.

As shown in Figure 2(b), the remaining multiplicative constant depends much more strongly on the confidence level than on the benchmark or ensemble size. After normalizing by $\mathrm{kl}(1-\delta,\delta)/\Gamma(P)$, its median decreases from 42.4 at $\delta = 0.1$ to 2.97 at $\delta = 10^{-16}$, a factor of 14.3, and δ accounts for 97% of its variance. Its dependence on confidence is close to $1/\log(1/\delta)$, with fitted exponent -0.953 and $R^2 = 0.985$. After controlling for δ, benchmark identity changes the median normalized constant by only 0.2%, while pool size and ensemble size produce larger but still secondary variation of 17% and 23%, respectively. These finite-confidence constants are therefore far from the asymptotic constant at practical values of δ, even though the scaling across instances with $\Gamma(P)$ is already very precise.

The disagreement-aware stopping rule also substantially outperforms the Hoeffding comparator. The comparator is slower in every one of the 249 reachable cells, with a median paired slowdown of $2.35\times$ on the normalized scale. It also does not meet the confidence criterion before the compute cap in 49% of runs, compared with 2.4% for the disagreement-aware rule. Among runs that terminate within the cap, neither procedure makes an observed identification error in any of the 249 cells. This comparison is not needed to establish the scaling with $\Gamma(P)$, but shows that exploiting the information gap also yields a substantial finite-confidence advantage over a generic coverage-based confidence rule.

Conclusion. Across the reachable instances, $\Gamma(P)$ is thus an accurate predictor of global identification cost: within a fixed confidence level, variation in $1/\Gamma(P)$ explains essentially all of the observed variation in stopping time. The experiment does not establish the exact asymptotic constant in Theorem 3.3. Indeed, the normalized constant remains large at practical confidence levels and approaches the first-order asymptotic behavior only as δ becomes very small. The data provide strong evidence for the instance-dependent prediction of the theorem: ensembles with smaller $\Gamma(P)$ require proportionally more evidence to identify, across substantially different model pools and ensemble sizes.

F.4 Ordinal feedback

We finally examine the weighted ordinal extension in Theorem B.1. In this setting, a task ranks the candidate models rather than assigning each a binary success indicator, and an ensemble receives credit according to the best-ranked member it contains. Theorem B.1 shows that, once a challenger is sampled, comparing two ensembles under this objective reduces to an ordinary binary coverage comparison. We test this reduction on ordinal LLM evaluations and ask whether the binary comparison procedure, together with the same disagreement-based controls used above, behaves as predicted without any testing procedure specific to ordinal feedback.

Ordinal evaluation matrix. We construct the experiment from the programmatic-grader scores released with LiveBench (White et al., 2025). We exclude task types with only two distinct score levels, since these already provide effectively binary feedback, and retain models scored on at least 80% of the questions. We then remove questions without scores for every retained model. The resulting dense matrix contains 215 questions evaluated by 37 models.

For each question i, we convert the grader scores to competition ranks. For a candidate c and challenger x, define $\widetilde{U}[c,(i,x)] = \mathbf{1}\{\mathrm{rank}_i(c) \le \mathrm{rank}_i(x)\}$. Thus, $\widetilde{U}[c,(i,x)] = 1$ exactly when c ranks at least as highly as x on question i. Pairing each of the 215 questions with each of the 37 possible challengers yields $215 \times 37 = 7{,}955$ derived binary tasks. Once the challenger is included in the task, an ensemble succeeds whenever at least one of its members ranks at least as highly as that challenger. The ordinal comparison can therefore be expressed through the same four binary joint outcomes used earlier.

In particular, for two ensembles A and B, let $p^{\lambda}_{10}(A,B)$ be the probability that A succeeds against the sampled challenger while B does not, and let $p^{\lambda}_{01}(A,B)$ denote the reverse outcome. As shown in Theorem B.1, their weighted ordinal coverage difference is $\Phi_\lambda(A) - \Phi_\lambda(B) = p^{\lambda}_{10}(A,B) - p^{\lambda}_{01}(A,B)$, and the corresponding ordinal exclusive information is $\mathcal{I}^{\lambda}_{\mathrm{ord}}(A,B) = J(p^{\lambda}_{10}(A,B), p^{\lambda}_{01}(A,B))$. Thus, after introducing the sampled challenger, $\mathcal{I}^{\lambda}_{\mathrm{ord}}$ is the ordinal analogue of the exclusive information used in the binary experiments, which quantifies the information

available for distinguishing A from B from the outcomes on which exactly one ensemble succeeds. As such, larger values of $\mathcal{I}^{\lambda}_{\mathrm{ord}}$ correspond to easier comparisons.

Pair selection and protocol. We apply the binary sequential comparison procedure of Appendices F.1 and F.2 directly to this derived matrix, with no ordinal-specific statistic or stopping rule. We evaluate 60 pairs of size-3 ensembles at $\delta = 0.05$, using 40 independent task sequences for each pair. These pairs are deliberately selected to span the observed range of $\mathcal{I}^{\lambda}_{\mathrm{ord}}$, rather than sampled uniformly: we retain at most 12 pairs per half-decade of $\log_{10} \mathcal{I}^{\lambda}_{\mathrm{ord}}$. The resulting set spans $\mathcal{I}^{\lambda}_{\mathrm{ord}}$ from 0.0012 to 0.161, corresponding to finite-confidence predicted stopping times from 224 to 5.6×10^4 derived tasks.

To determine whether the observed behavior is specific to the exclusive-outcome statistic, we compare exclusive KL with two controls from Appendix F.2. Paired Hoeffding uses the same paired observations but a range-only radius, $\sqrt{2L(n,\delta)/n}$, and therefore cannot adapt to the amount of exclusive disagreement. Paired empirical Bernstein uses the same observations while adapting to the variance of their paired difference. All three procedures use the same task sequences, confidence allocation, multiplicative checking grid, and underlying implementation; they differ only in the information extracted from the paired outcomes. No additional model inference is required: all comparisons are performed on the fixed derived matrix.

Ties. Ties are common in the retained LiveBench scores. Across the 37 models, a question has a median of only three distinct score levels; all 215 retained questions contain at least one tie and, averaged over questions, 47.6% of ordered pairs of distinct models are tied. We use competition ranks and allow a win to include a tie with the challenger (a weak win). This convention agrees with the definition of $P_\pi(c)$ in Section 2, which contains all models ranked at least as highly as c.

The presence of ties does not invalidate the binary reduction for this chosen objective. The four-cell argument in Theorem B.1 depends only on the joint binary law of (G_A, G_B), so it continues to apply once weak wins define the binary outcomes. The convention does, however, determine the ordinal objective being measured. This choice is empirically consequential: replacing weak wins with strict wins changes $\mathcal{I}^{\lambda}_{\mathrm{ord}}$ by a median factor of 1.32 (IQR $[1.04, 2.08]$), and reverses the ordering of two of the 60 ensemble pairs.

Results. Across the 60 ensemble pairs, three procedures, and 40 task sequences per pair, all 7,200 runs reach a decision at the specified confidence level before the 2×10^6-task compute cap, so no run is censored by the cap. We also observe no incorrect comparisons. To compare costs across procedures, for each ensemble pair we first take the median, over its 40 matched task sequences, of the ratio between a control's stopping time and that of exclusive KL. This gives one cost ratio for each of the 60 ensemble pairs. The geometric mean of these pair-level ratios is $4.72\times$ for paired Hoeffding (log-scale t interval $[4.19, 5.31]$), which is slower than exclusive KL on all 60 pairs. For paired Bernstein, the corresponding factor is $1.35\times$ ($[1.21, 1.52]$), with Bernstein slower on 41 of the 60 pairs. Because the ensemble pairs were deliberately selected to span $\mathcal{I}^{\lambda}_{\mathrm{ord}}$, these intervals summarize variation across the evaluated instances. In particular, they are not population intervals for arbitrary LiveBench ensemble pairs.

These aggregate comparisons show that the exclusive-outcome procedure is substantially cheaper on the selected instances, but they do not by themselves determine the contribution of exclusive disagreement. In the naturally occurring pairs, the coverage gap Δ and exclusive mass s are strongly correlated. We therefore specify, before inspecting stopping times, a matched-gap analysis: two ensemble comparisons i and j are matched when $|\Delta_i - \Delta_j| / \{\frac{1}{2}(\Delta_i + \Delta_j)\} \le 0.05$. This criterion produces 56 matched pairs-of-pairs, all of which are included in the analysis. Within each match, the coverage gaps are nearly equal, allowing us to examine how differences in exclusive disagreement are reflected in the comparison cost.

As shown in Figure 3, exclusive KL responds strongly to differences in ordinal exclusive information among comparisons with nearly equal coverage gaps. Its median cost ratio is 1.21, with rank correlation $+0.85$ between the stopping-time ratio and the corresponding information ratio. Paired Hoeffding is nearly insensitive, with median ratio 0.99 and rank correlation -0.05, while paired Bernstein lies between the two, with median ratio 1.07 and rank correlation $+0.74$, consistent with variance adaptation recovering part, but not all, of the structure used by exclusive KL. Among the seven matches with the largest differences in exclusive mass, the exclusive mass differs by a median factor of 1.80; exclusive KL requires $2.24\times$ as many derived tasks for the higher-mass comparison,

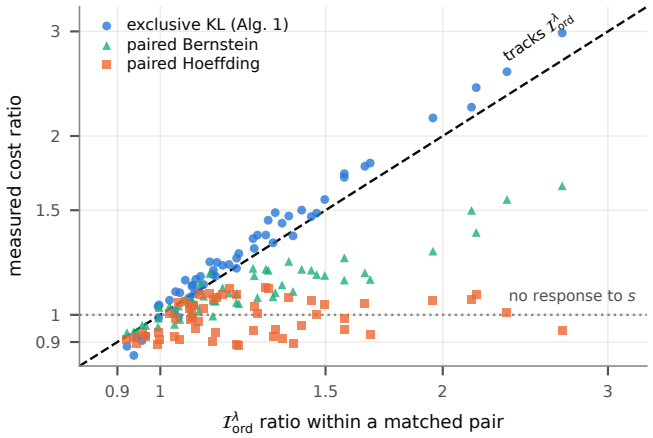


Figure 3: **Matched-gap comparison under ordinal feedback.** Each point represents one of the 56 matched pairs-of-pairs whose ordinal coverage gaps differ by at most 5%. The horizontal axis is the ratio of their exact ordinal exclusive information $\mathcal{I}^{\lambda}_{\mathrm{ord}}$, and the vertical axis is the corresponding ratio of mean stopping times for each procedure, averaged over 40 matched task sequences at $\delta = 0.05$. All matches are shown; no intervals are drawn.

close to the $2.15\times$ ratio implied by the exact ordinal information, whereas paired Hoeffding changes by only $1.01\times$.

As a separate implementation check, the stopping time of exclusive KL divided by the finite-δ fixed point of its own boundary has median 1.007 and IQR $[0.985, 1.026]$. This agreement is not independent evidence for the information rate: the observed and predicted stopping times use the same boundary, evaluated at empirical and population quantities, respectively. The main evidence here instead comes from the matched control comparison: the three procedures receive identical observations, while Hoeffding and Bernstein never compute $\mathcal{I}^{\lambda}_{\mathrm{ord}}$.

The control comparison therefore shows that, on the ordinal-derived tasks, procedures using different aspects of the same paired feedback respond differently to exclusive disagreement. This conclusion should be interpreted in light of two features of the experimental design. First, the sampling unit (i, x) is a question-challenger pair rather than an original benchmark question. Reading one cell of $\widetilde{U}$ therefore does not correspond to one model evaluation, so the stopping times in this experiment are measured in derived tasks and should not be interpreted as model-evaluation costs. Second, the 60 ensemble pairs were deliberately selected to span $\mathcal{I}^{\lambda}_{\mathrm{ord}}$. The resulting range of instance difficulties is therefore by design. The matched-gap comparison addresses a different question within this selected set: it compares naturally occurring pairs whose coverage gaps differ by at most 5%. Unlike the controlled construction in Appendix F.2, it does not hold Δ exactly fixed.

The experiment also tests the instance dependence identified by Theorem B.1, rather than how closely its lower bound is attained at the specified confidence level. For every evaluated pair, the measured stopping time exceeds $\mathrm{kl}(1-\delta, \delta)/\mathcal{I}^{\lambda}_{\mathrm{ord}}$ by more than an order of magnitude. We therefore do not interpret the experiment as evidence that the lower bound is attained at the confidence level considered here. Instead, the relevant empirical finding is that comparison cost varies with the ordinal exclusive information in the manner revealed by the control comparison.

Conclusion. The LiveBench experiment provides evidence for two conclusions. First, after pairing a question with a sampled challenger, weighted ordinal ensemble comparison can be handled by the same binary coverage procedure used earlier, without an ordinal-specific testing procedure. Second, among naturally occurring comparisons with nearly equal coverage gaps, the cost of exclusive KL varies systematically with ordinal exclusive disagreement, while the range-only paired-Hoeffding rule is nearly insensitive and the variance-adaptive Bernstein rule exhibits an intermediate response. This is the qualitative ordering predicted by Theorem B.1, now observed on ordinal rather than constructed binary comparisons.

Table 1: **Measurement × bound × scope grid at $\delta = 0.05$.** Results over 20 held-out greedy steps from the 41-step reachable subset and 24 seeds per step (480 runs per configuration). "Tasks" and "evaluations" are the mean numbers of accepted tasks and model evaluations before termination, respectively. "vs Alg. 1" is the median matched-instance ratio of task counts to Algorithm 1, and "censored" is the fraction of runs that exhaust the available benchmark before meeting the confidence criterion. The six pairwise configurations are summarized by their ranges in the final row.

measurement	bound	scope	tasks	vs Alg. 1	evaluations	censored
paired	KL	simultaneous	1,142	1.00×	12,716	0.0%
marginal	KL	simultaneous	1,558	1.30×	16,367	15.0%
paired	Hoeffding	simultaneous	1,718	1.32×	17,805	25.0%
marginal	Hoeffding	simultaneous	1,718	1.32×	17,805	25.0%
paired	Bernstein	simultaneous	1,522	1.36×	20,449	25.0%
marginal	Bernstein	simultaneous	2,594	2.54×	32,949	25.0%
all six pairwise configurations			4,765–5,533	3.79–5.20×	21,240–27,315	78–99%

F.5 The measurement × bound × scope grid

This experiment examines separately the two components combined in Theorem 4.1: the statistic used to determine which candidate is better at the specified confidence level and the reuse of each accepted task across the active candidates. Here, accepted tasks are tasks missed by the current ensemble and retained for candidate comparison. We vary three design choices. The *measurement* is either marginal, using the two candidates' empirical rescue rates separately, or paired, using their two exclusive outcomes. The *bound* is Hoeffding, empirical Bernstein (Maurer & Pontil, 2009), or KL. The *scope* is either pairwise, where the current leading candidate is compared with one challenger at a time, or simultaneous, where every active candidate is updated from the same accepted task. This gives twelve configurations.

All configurations are run on the same populations of tasks missed by the current ensemble, active sets, confidence levels, and checking schedule. For each instance and seed, they also receive exactly the same shuffled benchmark task sequence. The experiment uses 20 held-out greedy steps from the 41-step reachable subset characterized in Appendix F.6, drawn from MMLU-PromptEval, GSM8K, and RouterBench, with 24 seeds per step and $\delta \in \{0.1, 0.05, 0.01\}$. In the pairwise configurations, challengers are ordered by their *true* rescue rates, giving pairwise testing the most favorable possible ordering rather than constructing a weak tournament baseline.

The measurement axis specifies which information is used from the same observations. A marginal configuration retains the two empirical rescue rates but discards their joint outcome, whereas a paired configuration retains the two exclusive cells. With Hoeffding, this distinction has no numerical effect: the resulting range-based threshold for the difference is identical under the two representations. With empirical Bernstein or KL, the joint outcome structure can affect the boundary through the variance or likelihood, respectively. To ensure that the comparison is not confounded by differences in the sampled task sequences, each instance and seed deterministically defines a single benchmark permutation, which is then shared by all twelve configurations. Thus, for a given instance and seed, every configuration receives exactly the same sequence of tasks, and the complete grid is reproducible across independent runs. Table 1 reports the results at $\delta = 0.05$.

Two of the twelve configurations correspond to the procedures compared in this paper: (marginal, Hoeffding, simultaneous) implements AFG (Neoh et al., 2026), while (paired, KL, simultaneous) implements Algorithm 1. The remaining ten configurations are controls constructed to distinguish the effects of measurement, confidence bound, and comparison scope. Their confidence guarantees follow from the concentration arguments given below, but we do not propose these controls as standalone selection algorithms. In particular, combining a measurement with a given confidence bound in this grid should not be interpreted as attributing that configuration to the work from which the bound is drawn.

What the grid separates. The grid reveals that comparison scope is the largest source of cost in this experiment. Among the twelve configurations at their stated thresholds, Algorithm 1 has both the lowest mean evaluation cost and the lowest observed censoring rate, where a run is censored if the

available benchmark task sequence is exhausted before the confidence criterion is met. Replacing simultaneous reuse with pairwise comparison increases the number of accepted tasks to 3.79–$5.20\times$ that of the reference configuration and leaves 78–99% of runs without a confirmed decision when the benchmark is exhausted (Table 1). This is a large effect, but it is not an advantage of our method over AFG: both published methods already reuse each accepted task across all active candidates. Rather, the scope comparison quantifies the value of this simultaneous reuse relative to testing challengers one at a time.

The two procedures, compared directly. The grid contains a direct head-to-head comparison between Algorithm 1 and AFG (Neoh et al., 2026). They correspond to the (paired, KL, simultaneous) and (marginal, Hoeffding, simultaneous) cells, respectively. These two cells use the same populations of accepted tasks, task sequences, active sets, confidence level, and checking schedule, and both reuse each accepted task across all active candidates. The comparison therefore holds the feedback on accepted tasks and comparison scope fixed while changing the confidence criterion: Algorithm 1 uses the paired exclusive outcomes with a KL confidence bound, whereas AFG uses marginal rescue rates with a Hoeffding confidence bound.

At $\delta = 0.05$, Algorithm 1 uses a mean of $1{,}142$ accepted tasks and $12{,}716$ model evaluations, compared with $1{,}718$ tasks and $17{,}805$ evaluations for AFG. The median matched-run ratio in accepted tasks is $1.32\times$ in favor of our procedure. The difference in reachability is larger: none of the 480 runs of our reference configuration is censored, whereas AFG does not meet its confidence criterion before exhausting the benchmark in 25% of its runs (Table 1).

These two comparisons should be interpreted differently. The task and evaluation ratios are finite-sample quantities under the confidence boundaries specified for the two procedures, and should therefore not be interpreted as intrinsic constants of the underlying statistics. The censoring comparison answers a complementary question: under the same finite benchmark and the stated confidence guarantees, does the procedure reach a decision at the specified confidence level before the available task sequence is exhausted? Under this protocol, Algorithm 1 reaches such a decision in all 480 observed runs, whereas AFG leaves 25% without such a decision. We therefore report both the evidence required for a confirmed decision and how often such a decision is reached rather than reducing the comparison to a single ratio.

Measurement and bound interact. The simultaneous-scope rows show that the effects of measurement and confidence bound cannot be interpreted as independent additive contributions. Under Hoeffding, changing from marginal rescue rates to paired exclusive outcomes changes nothing: both rows use $1{,}718$ accepted tasks and $17{,}805$ model evaluations on average, with 25% censoring. This is expected because the Hoeffding confidence criterion uses only the fixed range of the paired difference and therefore cannot exploit how probability mass is distributed between the two exclusive outcomes.

With a likelihood-based confidence criterion, the representation matters. Under KL, changing from marginal to paired measurement reduces the mean number of accepted tasks from $1{,}558$ to $1{,}142$ and observed censoring from 15% to 0%. Empirical Bernstein lies between these cases because its boundary can adapt to the variance of the paired difference. Thus, the exclusive-outcome representation is useful only when the accompanying confidence criterion can exploit the additional joint information. The grid should therefore be read as a study of how these design choices interact, not as a decomposition into independent "measurement" and "bound" effects.

Confidence accounting. Every configuration in the grid is δ-correct. The confidence allocation differs across configurations, however, because the number and form of the deviation events that must be controlled depend on the measurement, bound, and comparison scope. We first describe the allocation across candidate pairs and checking times, and then specify how each confidence bound uses the resulting per-comparison budget. For a comparison involving P ordered candidate pairs, define $L(n, \delta) := \log(P/\delta) + \log(\pi^2/6) + 2\log n$. For simultaneous comparisons, $P = r(r-1)$ for r active candidates. For pairwise comparisons, the overall confidence budget is first divided among the $r-1$ challengers, and the two comparison directions give $P = 2$ within each challenger comparison. Since $e^{-L(n,\delta)} = 6\delta/(P\pi^2 n^2)$ and $\sum_{n\geq 1} 6/(\pi^2 n^2) = 1$, each ordered pair receives total error probability δ/P across all checking times. Summing over the P ordered pairs therefore controls the overall error probability by δ.

The cells differ in how this per-comparison budget enters their confidence bounds. The paired Hoeffding and empirical-Bernstein bounds are the same time-uniform controls defined in Appendix F.2; here we describe how their confidence budgets, together with those of the KL controls, are allocated across the twelve grid configurations. For paired measurement, the statistic is the empirical difference of the two exclusive-outcome frequencies, so each of the three bounds acts directly on a single one-sided comparison event. Marginal Hoeffding can be written in the same way: because both candidates are evaluated on the same accepted tasks, the difference of their rescue indicators is a range-2 random variable, and the implemented threshold $2\sqrt{L/(2n)} = \sqrt{2L/n}$ is a single Hoeffding bound on that difference. Marginal KL and marginal empirical Bernstein instead bound the two candidate rescue rates separately, so each marginal bound receives half of the per-comparison budget.

For empirical Bernstein, we use the one-sided bound of Maurer & Pontil (2009), as defined for the paired comparison in Appendix F.2, including its unbiased sample variance and $\log(2/d)$ confidence term. In the paired cell, rescaling the bound from range 1 to the range-2 paired difference leaves the variance contribution in the corresponding scale and doubles the additive constant from 7 to 14. The KL configuration uses the exclusive-outcome likelihood statistic and includes an additional $4\log(n+1)$ term to account for the possible empirical four-outcome frequency patterns. Consequently, the twelve configurations need not have identical numerical thresholds: each uses the confidence boundary justified for its statistic. Notably, the paired-KL reference remains the least costly simultaneous configuration despite this additional term.

Task counts and model-evaluation counts also measure different costs. A pairwise configuration evaluates only the two candidates involved in the current comparison, whereas a simultaneous configuration evaluates the active candidate set on every accepted task. Pairwise testing can therefore consume many more accepted tasks without increasing model evaluations by the same factor. This is why the pairwise task ratios in Table 1 reach 3.79–$5.20\times$ even though the corresponding evaluation ratios are much smaller. We report both quantities rather than treating either as a complete measure of cost.

Conclusion of the experiment. The grid separates two effects that should not be conflated. First, simultaneous reuse of each accepted task across all active candidates is the dominant source of savings relative to pairwise testing in this experiment: pairwise controls require 3.79–$5.20\times$ as many accepted tasks as the reference configuration and leave 78–99% of decisions unconfirmed within the available benchmark. This reuse is shared by our procedure and AFG, so it is not an advantage over prior work.

Second, once comparison scope and feedback on accepted tasks are matched, the two confidence criteria differ. At $\delta = 0.05$, AFG requires a median $1.32\times$ as many accepted tasks as Algorithm 1 under their stated confidence boundaries. Moreover, 25% of the AFG runs exhaust the benchmark without meeting the confidence criterion, whereas none of the 480 observed runs of our procedure does. These results concern the cost and reachability of a decision *confirmed at the specified confidence level*, not the quality of the ensemble ultimately selected: Appendices G.1 and G.4 show that the two procedures attain essentially the same held-out selection quality.

F.6 Reaching a decision within a finite benchmark

The information quantities studied above determine not only how the cost of a fixed-confidence decision varies across instances, but also whether that decision can be reached within a benchmark of realistic length. We examine this question for confirmation of greedy choices, global ensemble identification, and the fresh multilingual experiment. Across all three settings, a substantial fraction of the decisions require more evidence than the available benchmark provides (Figure 4).

Confirming greedy choices at a fixed confidence level. We begin with all 124 real greedy steps drawn from the three offline matrices. At $\delta = 0.01$, Algorithm 1 reaches its confidence threshold before the benchmark is exhausted on only 41 steps. On the remaining 83, or 67%, the benchmark ends before enough evidence has accumulated to confirm the greedy choice. Thus, for two thirds of these real greedy decisions, the amount of evidence required from tasks missed by the current ensemble exceeds what is available in the benchmark from which the decision is drawn.

This result also explains the instance selection used in Appendix F.5. The purpose of that experiment is to compare the costs of alternative confidence criteria on instances for which the reference proce-

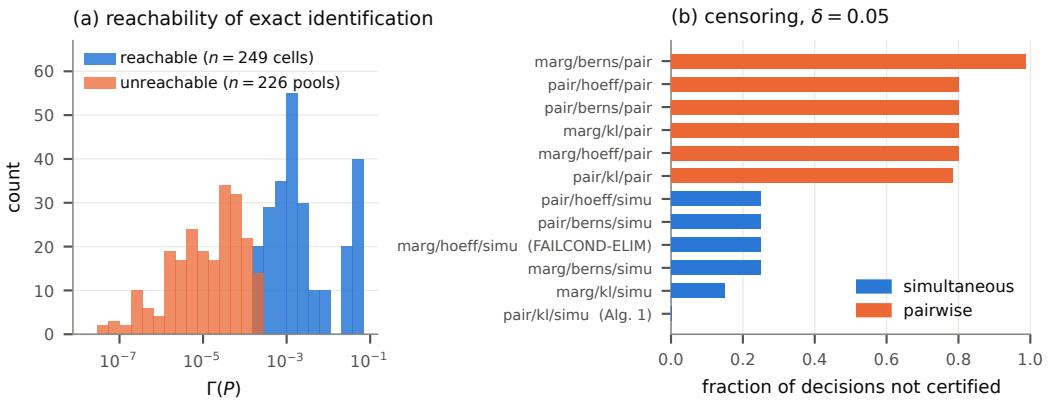


Figure 4: **Fixed-confidence identification can exceed the available benchmark. (a)** Distribution of the global information gap $\Gamma(P)$ for the pools in Appendix F.3, separated by whether the predicted stopping time falls within the 4×10^5-task compute cap. **(b)** Fraction of greedy decisions in the twelve configurations of Appendix F.5 that exhaust the benchmark without meeting the confidence criterion at $\delta = 0.05$, shown separately for simultaneous and pairwise scopes.

dure can actually reach a fixed-confidence decision. The grid therefore uses 20 held-out steps from the 41 steps on which Algorithm 1 confirms the decision within the benchmark at $\delta = 0.01$, and excludes the other 83 steps from the grid. The 67% figure here is measured before that restriction, on all 124 steps. By contrast, the censoring rates in Table 1 are measured on the 20 held-out steps used by the grid, all of which belong to the 41-step reachable subset. We did not evaluate all twelve grid configurations on the original set of 124 steps, so these two sets of censoring rates should not be compared directly.

The restriction used by the grid is important precisely because it makes that experiment a conservative test of whether the confidence criterion can be met. Even after restricting attention to decisions on which our reference confidence criterion can succeed within the benchmark at the stricter level $\delta = 0.01$, alternative confidence criteria still remain unmet on a substantial fraction of runs at $\delta = 0.05$ (Table 1 and Figure 4(b)). The differences in reaching a decision observed in the grid therefore arise within a set of decisions already known to meet the reference procedure's confidence criterion.

Global exact identification. The same phenomenon appears at the level of the final ensemble. Recall that $\Gamma(P)$ measures the information separating the optimal ensemble from its hardest-to-distinguish rival, with characteristic time $1/\Gamma(P)$. Among the pools considered in the global-identification experiment, 226 have finite-confidence predicted stopping times beyond the 4×10^5-task compute cap. As Figure 4(a) shows, these instances extend to global information gaps as small as $\Gamma(P) = 1.6 \times 10^{-8}$. For such small information gaps, distinguishing the optimum from its closest rival requires correspondingly large amounts of evidence. The inability to reach a decision on these finite benchmarks thus reflects the same instance-dependent difficulty quantified by $\Gamma(P)$.

Conclusion of the experiment. The local and global analyses show the same practical consequence of the information-theoretic rates. When the relevant information gap is small, the competing decisions are difficult to distinguish, so fixed-confidence identification requires correspondingly more observations. For many of the real instances studied here, that requirement exceeds the number of tasks available in the benchmark. At the local level, Algorithm 1 does not reach a fixed-confidence decision within the benchmark for 83 of the 124 offline greedy steps at $\delta = 0.01$. At the global level, 226 sampled pools have finite-confidence predicted stopping times beyond the 4×10^5-task experimental horizon.

Exhausting the available benchmark before crossing the confidence boundary does not indicate a problem with the stopping rule. It means that the benchmark has not supplied the evidence required for fixed-confidence confirmation. Nor does it imply that the available data cannot provide a basis

for a useful empirical choice: selection quality and fixed-confidence confirmation can have different data requirements. Track B examines this distinction directly under finite sequences of benchmark tasks.

F.7 Implementation checks

The experiments above test predictions of the theory. We separately perform three implementation checks whose purpose is narrower: to verify quantities, accounting rules, and stopping behavior against calculations that can be checked independently. These checks validate the experimental implementation; they do not provide additional evidence for the predicted information rates.

Closed-form information quantities. We first verify the numerical implementation of J, the exclusive-outcome KL quantity used throughout the paper to measure the information available for a pairwise comparison. We use the explicit family P_ε from Theorem 5.1, evaluated at $\varepsilon \in \{0.04, 0.02, 0.01, 0.005, 0.002, 0.001\}$. Recall that $\Gamma(P_\varepsilon)$ is the global information gap separating the optimal ensemble from its hardest-to-distinguish rival, whereas $\Delta_{\mathrm{KL}}(\varnothing, C)$ is the local information gap for identifying the best candidate at the first greedy step. The construction in Theorem 5.1 keeps $\Gamma(P_\varepsilon)$ bounded away from zero while making $\Delta_{\mathrm{KL}}(\varnothing, C)$ vanish as $\varepsilon \to 0$.

At all six values, the implementation returns $\Gamma(P_\varepsilon) = J(3/20, 1/20) = 0.026162$, independently of ε. The local information gap also approaches its predicted asymptotic form: $\Delta_{\mathrm{KL}}(\varnothing, C)/(\frac{5}{3}\varepsilon^2)$ decreases from 1.158 at $\varepsilon = 0.04$ to 1.003 at $\varepsilon = 0.001$. Consequently, the ratio of the global to the local information gap, $\Gamma(P_\varepsilon)/\Delta_{\mathrm{KL}}(\varnothing, C)$, grows from 8.5 to $15{,}645$ over the same range. Finally, the implementation reproduces the population greedy path $\varnothing \to \{a\} \to \{a, c\}$ and returns the globally optimal ensemble at every tested value of ε. These calculations verify that the code reproduces the quantities and greedy choices used in the theoretical separation rather than constituting a separate empirical test of that separation.

Evaluation accounting. We next check the shared short-circuit accounting used for pairwise ensemble comparison. On a 500-task instance with two size-3 committees, we independently count, task by task, the distinct models that must be evaluated under the short-circuit rule and compare this count with the vectorized accounting routine. The two counts agree exactly on every task.

This check concerns only the shared short-circuit accounting. Track B uses the broader decomposition into committee-evaluation cost Q_{eval} and candidate-evaluation cost Q_{cand}, which is checked separately by the experimental driver. The driver enforces the requested evaluation budget before admitting additional work and flags any run that exceeds it. A separate partition check verifies that each greedy step receives its assigned block of the sequence of tasks, that no task is charged to more than one step, and that all methods receive the same total requested budget.

Sequential stopping rule. Finally, we verify that the implemented pairwise rule behaves consistently with its prescribed stopping boundary. We test 273 pairs of size-3 ensembles: 100 from RouterBench, 104 from GSM8K, and 69 from MMLU-PromptEval. Each pair is evaluated over 40 independent sequences of tasks at $\delta = 0.05$, giving $10{,}920$ runs in total. No incorrect comparison is observed.

For each pair, we also compare the observed stopping time with the finite-confidence fixed point $n_0(I_{\mathrm{cov}}, \delta)$ obtained by evaluating the same stopping boundary at the population value of I_{cov}. The ratio of observed to predicted stopping time has median 1.009 and IQR $[0.988, 1.032]$, while every pair-level mean ratio lies in $[0.898, 1.081]$. We also find no material systematic dependence of this ratio on I_{cov}, the exclusive mass s, the coverage gap Δ, or the predicted stopping time: the absolute rank correlation with each quantity is below 0.13. A systematic dependence on any of these instance characteristics would indicate that the empirical stopping behavior departs from what the implemented boundary predicts; we observe no such pattern.

The stopping-time agreement checks consistency of the implementation; it does not independently test the information rate. The procedure stops when its empirical exclusive-outcome statistic crosses the prescribed boundary, while $n_0(I_{\mathrm{cov}}, \delta)$ is obtained from that same boundary after replacing the empirical statistic by its population counterpart. Agreement near one therefore checks the implementation of the sequential updates, boundary evaluation, and stopping-time accounting. Likewise, observing no errors in $10{,}920$ runs is consistent with the δ-correctness guarantee but does not establish it; the guarantee itself is provided by Theorem 3.2. The independent empirical evidence that

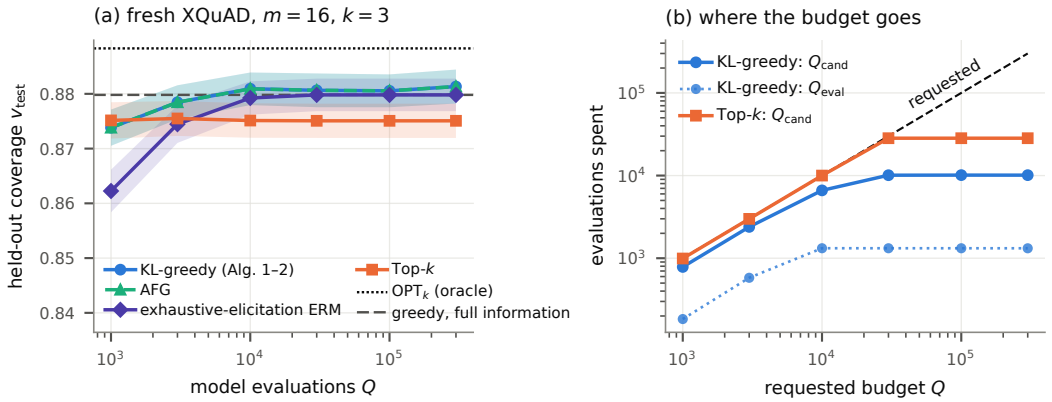


Figure 5: **Fresh multilingual XQuAD selection at $k = 3$.** Results over 10 paragraph-level splits and 20 algorithm seeds per budget; bands show 95% confidence intervals. **(a)** Held-out coverage of the selected committee versus total model evaluations. Horizontal lines show the full-information references. **(b)** Actual evaluation expenditure versus requested budget, decomposed into candidate evaluations Q_{cand} and committee evaluations Q_{eval} used to identify tasks missed by the committee.

exclusive disagreement determines comparison difficulty comes instead from the controlled experiments in Appendices F.1 and F.2.

G Track B: Practical Selection

Track B asks whether selection on tasks missed by the current ensemble produces better ensembles when model evaluations are limited and selection must be performed from a finite training set. Unlike Track A, which examines the statistical difficulty of identification and the efficiency of the corresponding confidence criteria, the experiments here evaluate the quality of the committees selected under practical evaluation budgets. We measure held-out coverage, account for the model evaluations required to construct each committee, and examine the structure of the resulting selections across newly generated and published evaluation matrices.

Appendix G.1 presents the main fresh evaluation on multilingual XQuAD, comparing held-out ensemble quality across evaluation budgets and examining how the available training data constrain the selection procedure. Appendix G.2 analyzes the language-specific structure of the same candidate pool, asking whether its complementarity can be reduced to a simple assignment of languages to specialist models or instead reflects task-level differences among both specialist and multilingual candidates. Appendix G.3 tests whether the complementarity observed on translated XQuAD questions persists on TyDi-QA questions originally written in their respective languages. Finally, Appendix G.4 evaluates the same selection principles on the published MMLU-PromptEval, GSM8K, and RouterBench matrices, including the effects of candidate-pool construction and committee redundancy.

G.1 Fresh XQuAD selection experiment

Outcome matrix and available complementarity. The fresh evaluation produces a $16 \times 2{,}400$ binary outcome matrix with mean cell accuracy 0.5295. The strongest individual candidate covers 0.7671 of the tasks, whereas the union of all 16 candidates covers 0.9354. Exact enumeration gives optimal coverage 0.8825 for a size-3 ensemble and 0.9054 for a size-5 ensemble, so at $k = 3$ there is 0.1154 additional coverage available beyond the best individual model. The median pairwise exclusive mass is 0.301, and no candidate pair has Jaccard overlap above 0.9, confirming that the pool contains substantial variation in which tasks different models solve.

The optimal size-3 ensemble is {Qwen2.5-14B-Instruct, aya-expanse-8b, Typhoon-1.5-8B}. Its advantage is not confined to a single language: relative to the best individual candidate for each lan-

guage, optimal size-3 coverage increases from 0.710 to 0.820 on Hindi, from 0.787 to 0.850 on Greek, and from 0.810 to 0.893 on German, with an improvement on all eight languages. Appendix G.2 examines this language-specific structure in detail and shows that it cannot be reduced to assigning each language to an obvious specialist.

Selection quality at $k = 3$. Pooling the four budgets from 10^4 to 3×10^5 model evaluations, selection on missed tasks improves held-out coverage over Top-k by $+0.00578$ (95% CI $[+0.00529, +0.00627]$; bootstrap $[+0.00529, +0.00627]$; $n = 800$ paired runs), with a positive difference in 8 of the 10 splits. The advantage is smaller but remains positive against exhaustive-elicitation ERM, the strongest budgeted non-greedy comparator, at $+0.00120$ ($[+0.00081, +0.00159]$), and is $+0.0078$ against sampled-ERM. Figure 5(a) shows the full budget curve.

The advantage does not extend uniformly to the smallest budgets. At 10^3 evaluations, the difference from Top-k is -0.00133 and its interval contains zero. Selection on missed tasks therefore becomes useful only after enough training tasks have been observed, and averaging performance over the entire budget grid would obscure this transition. This is why we report differences separately by budget; Figure 6(a) shows the full budget dependence at both committee sizes.

The same pattern persists when the amount of training data is varied while keeping the paragraph grouping fixed. With train/test fractions of $60/40$, $75/25$, and $90/10$, the margin over Top-k at $k = 3$ is respectively $+0.0054$, $+0.0058$, and $+0.0070$, with all three intervals excluding zero. Thus, within the range tested, additional training data strengthens rather than eliminates the $k = 3$ advantage.

Selection quality at $k = 5$. At $k = 5$, we find no practically meaningful improvement over either baseline. Pooling the four budgets from 10^4 through 3×10^5, the difference from Top-k is -0.00079 (95% CI $[-0.00118, -0.00041]$; bootstrap $[-0.00118, -0.00041]$; $n = 800$), with a positive difference in 5 of the 10 splits. The corresponding difference from exhaustive-elicitation ERM is -0.00104 (95% CI $[-0.00142, -0.00065]$). Although both intervals exclude zero, the effects are small in magnitude: each is below 0.0011 in absolute coverage and less than one fifth of the $k = 3$ improvement over Top-k. Accordingly, we interpret the $k = 5$ results as showing no practically meaningful improvement, rather than as evidence that selection on missed tasks performs substantially worse. Figure 6(a) shows this distinction across the full budget range: the $k = 3$ difference becomes positive as the evaluation budget increases, whereas the $k = 5$ difference remains close to zero throughout.

The per-step diagnostics suggest that the later greedy decisions may have too little data on tasks missed by the current ensemble to identify complementary candidates reliably, although the experiment does not establish this explanation. At $k = 5$, the five greedy steps receive on average 207, 64, 40, 34, and 30 accepted tasks, respectively. The candidates selected at these steps agree with those selected using the corresponding full training blocks in 54%, 80%, 40%, 35%, and 36% of runs. For comparison, the three steps at $k = 3$ agree in 64%, 88%, and 43% of runs. Thus, the later $k = 5$ decisions are based on progressively fewer accepted tasks and often differ from the choices obtained using the full training block.

Increasing the amount of training data, however, does not resolve the $k = 5$ result. With train/test fractions of $60/40$, $75/25$, and $90/10$, the differences from Top-k are -0.0016, -0.0008, and -0.0007, respectively; only the $90/10$ interval contains zero. The per-step diagnostics are therefore consistent with limited data on tasks missed by the current ensemble contributing to the weaker $k = 5$ performance, but the training-fraction experiment does not establish data availability as its cause.

Selection agreement and fixed-confidence decisions. The exclusive-KL confidence criterion of Algorithm 1 and the marginal Hoeffding confidence criterion used by AFG (Neoh et al., 2026) produce essentially identical selected ensembles on this benchmark. Across the 800 matched runs at budgets of 10^4 and above, their held-out coverages differ by $+0.00001$ (95% CI $[-0.00001, +0.00002]$), and they select the same committee in 99.9% of matched runs and in every run of 9 of the 10 splits. Their held-out coverage curves are consequently indistinguishable in Figure 5(a).

This agreement concerns ensemble quality, not the statistical efficiency of the two confidence criteria. The fresh XQUAD task sequences are too short for either procedure to reach a fixed-confidence

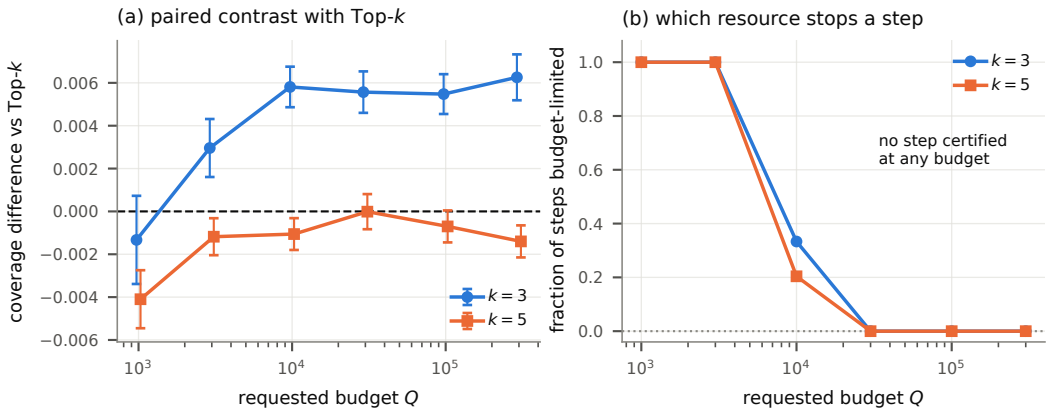


Figure 6: **Selection quality and limits of finite task sets on fresh XQuAD. (a)** Paired difference in held-out coverage between selection on missed tasks and Top-k versus requested evaluation budget, for $k \in \{3, 5\}$. Points are means over 200 matched runs and bars show 95% confidence intervals. **(b)** Fraction of greedy steps terminating because their evaluation budget is exhausted, shown by budget and committee size. The two methods that select using missed tasks are pooled because their termination fractions are essentially identical.

decision, so this experiment cannot reveal the difference between them. We make that comparison in Appendix F.5, where both methods receive the same simultaneous feedback on tasks missed by the current ensemble and differ only in the confidence criterion applied to it. In that controlled experiment, AFG requires a median $1.32\times$ as many accepted tasks as our exclusive-KL procedure and leaves 25% of decisions unconfirmed at the stated confidence level within the available benchmark, whereas our method confirms the decision in all 480 runs. Thus, the two methods achieve essentially identical ensemble quality on fresh XQuAD, but the controlled experiment shows a clear advantage for our method in the amount of evidence from missed tasks required for a fixed-confidence decision and in the probability of reaching that decision within a finite benchmark.

The reason this advantage is not visible on XQuAD is that neither confidence criterion can be met within the available training tasks. Across both committee sizes, no greedy step terminates with a fixed-confidence decision. Instead, each step returns the empirical rescue-rate maximizer after exhausting either its evaluation budget or its assigned block of training tasks. At the two smallest budgets, every step is budget-limited. At 10^4 evaluations, the budget-limited fraction falls to 33% at $k = 3$ and 20% at $k = 5$, and from 3×10^4 onward every step is limited by the available training tasks. Figure 6(b) shows this transition. Pooled across the design, 2,800 of the 7,200 steps at $k = 3$ and 4,408 of the 12,000 steps at $k = 5$ terminate because of the budget, with the remaining steps exhausting their training block.

The requested budget therefore ceases to determine actual expenditure once the number of available training tasks becomes the limiting factor. From 3×10^4 requested evaluations onward, the procedures that select using missed tasks spend on average only 11,435 evaluations at $k = 3$, ranging from 10,643 to 12,375, even when the requested budget is 3×10^5; Figure 5(b) shows the resulting saturation in actual expenditure. The fresh XQuAD experiment therefore establishes the selection behavior of the two procedures with finite data and shows why their difference in statistical efficiency cannot be observed here. The controlled experiment of Appendix F.5 supplies that complementary comparison.

Sensitivity to the split scheme. Our reported results use paragraph-level train/test splits, so all questions derived from the same passage remain on the same side. Under the weaker question-level alternative, approximately 45 paragraphs appear in both training and test sets on average, producing about 16,200 train-test task pairs that share a passage.

We rerun the same selection design under both splitting schemes, using 10 splits, 20 algorithm seeds, and all six budgets. At $k = 3$ under budgets of 10^4 and above, the margin over Top-k is

larger under question-level splitting at every budget, by between $+0.0005$ and $+0.0016$ coverage. At $k = 5$, question-level splitting likewise shifts the result upward; e.g., the difference moves from -0.0008 to $+0.0002$ at 3×10^4 evaluations and from -0.0002 to $+0.0004$ at 3×10^5. Thus, in the budget range on which our positive $k = 3$ claim is based, question-level splitting makes the results more favorable to selection on missed tasks at both committee sizes. We therefore report the more conservative paragraph-level results. At the two smallest budgets the direction reverses, reinforcing the need to compare methods separately at each budget rather than averaging over the grid.

Committee composition. At $k = 3$ and the largest budget, our exclusive-KL procedure and AFG select the same committee in all 200 matched runs, so a single summary describes both. Their selected ensembles have mean member accuracy 0.72 and mean pairwise Jaccard overlap 0.72, compared with 0.74 and 0.79 for Top-k. Selection on missed tasks therefore gives up some individual accuracy in exchange for less redundant success patterns. The difference is smaller than on the published evaluation matrices, consistent with the absence of near-duplicate candidates in the fresh pool.

Conclusion. To summarize, the fresh XQuAD experiment separates useful finite-data selection from fixed-confidence confirmation. At $k = 3$, once the evaluation budget is large enough to collect sufficient data on missed tasks, selecting models by how well they solve tasks missed by the current ensemble improves held-out coverage over ranking models by individual accuracy, by about 0.58 percentage points at budgets of 10^4 and above. At smaller budgets this advantage is absent, and at $k = 5$ we find no practically meaningful improvement. The exclusive-KL and AFG confidence criteria produce essentially identical ensembles here because neither reaches a fixed-confidence decision before the training tasks are exhausted. The benefit observed at $k = 3$ therefore comes from using missed tasks to select complementary candidates rather than from the choice of confidence criterion. Together with the results on reaching a decision in Appendix F.6, this shows that a benchmark can contain enough information to improve an empirical committee choice without containing enough evidence to confirm that choice at the prescribed confidence level. The statistical cost of the two confidence criteria themselves is compared separately in Appendix F.5.

G.2 Language-specific strengths do not make selection trivial

The fresh candidate pool combines broadly multilingual models with models developed for particular languages or language families (Appendix E.3). This raises a natural question about the source of complementarity in the experiment. If each language had an obvious specialist, committee selection could largely reduce to assigning each language to its most accurate model. In that case, the observed complementarity would mainly reflect how the pool was constructed rather than meaningful task-level differences among candidates. We test this possibility directly using the measured XQuAD outcomes, considering both the language-focused candidates and the multilingual generalists.

Measuring language specialization. The eight languages differ substantially in difficulty for this candidate pool. Mean accuracy across the 16 models ranges from 0.366 on Hindi to 0.684 on Chinese. Raw accuracy on an intended language therefore reflects both the model's relative strength on that language and the language's overall difficulty for the pool. To separate these effects, we define a model's *language lift* as its accuracy on a language minus the mean accuracy of all 16 candidates on that language. A positive lift indicates that the model performs better than the pool average on that language.

We report two additional quantities in Table 2. The first is the model's rank among all 16 candidates on its intended language. The second is its *complementarity share*, defined as the fraction of the model's correct answers that are missed by aya-expanse-8b, the most accurate individual candidate in the pool. This quantity is particularly relevant to ensemble selection. A candidate can improve a committee despite modest standalone accuracy if it succeeds on tasks that a stronger model already in the committee misses.

Intended specialization only partly predicts measured performance. Five of the seven language-focused candidates have positive lift on their intended language. The lift ranges from $+0.221$ for Typhoon on Thai to $+0.092$ for Meltemi on Greek. The correspondence is not universal, however. Vikhr is 0.092 below the pool mean on Russian, while OpenHathi is 0.249 below the pool mean on Hindi. Intended language focus therefore does not by itself determine how a model performs under the common prompt, decoding, and scoring protocol used here. The selection procedures do not

Table 2: **Language specialization in the fresh XQuAD pool.** Accuracy, rank, and language lift are measured on each candidate's intended language. "Compl. share" is the fraction of its correct answers missed by aya-expanse-8b.

model	intended language	accuracy	rank	lift	compl. share
Typhoon-1.5-8b	Thai	0.707	4/16	+0.221	0.115
SeaLLM-7B-v2	Vietnamese	0.697	6/16	+0.138	0.074
Yi-6B-Chat	Chinese	0.793	5/16	+0.109	0.061
OpenThaiGPT-1.0.0-7b	Thai	0.590	7/16	+0.104	0.134
Meltemi-7B-Instruct	Greek	0.480	6/16	+0.092	0.064
Vikhr-7B-instruct	Russian	0.383	12/16	−0.092	0.039
OpenHathi-7B	Hindi	0.117	13/16	−0.249	0.025

Table 3: **Cross-language performance of the multilingual generalists.** Candidates are ordered by overall accuracy on fresh XQuAD. "Langs. above mean" counts languages on which a candidate exceeds the pool mean; min and max lift summarize its eight language lifts. Complementarity share is relative to aya-expanse-8b.

model	accuracy	langs. above mean	min lift	max lift	compl. share
aya-expanse-8b	0.767	8/8	+0.058	+0.399	—
Qwen2.5-14B-Instruct	0.756	8/8	+0.135	+0.301	0.121
Qwen2.5-7B-Instruct	0.716	8/8	+0.105	+0.268	0.099
Llama-3.1-8B-Instruct	0.687	8/8	+0.071	+0.301	0.102
Qwen2-7B-Instruct	0.628	7/8	−0.028	+0.221	0.099
Llama-3.2-3B-Instruct	0.604	8/8	+0.004	+0.214	0.109
Qwen2.5-1.5B-Instruct	0.560	5/8	−0.091	+0.141	0.103
Qwen2.5-3B-Instruct	0.544	5/8	−0.065	+0.109	0.088
Mistral-7B-Instruct-v0.3	0.491	2/8	−0.095	+0.049	0.064

use these intended specializations; they operate only on observed correct and incorrect answers on individual tasks.

Multilingual generalists differ in performance. Table 3 applies the same measurements to the nine candidates included as multilingual generalists. Five perform well across languages under this measure: aya-expanse-8b, Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Llama-3.2-3B-Instruct exceed the pool mean on all eight languages. Even within this group, the strength of that advantage varies substantially. Their minimum language lift ranges from $+0.135$ for Qwen2.5-14B-Instruct to only $+0.004$ for Llama-3.2-3B-Instruct.

Broad performance is not automatic for models included in this group. Qwen2-7B-Instruct exceeds the pool mean on seven of the eight languages, the two smallest Qwen2.5 models do so on five, and Mistral-7B-Instruct-v0.3 on only two. Mistral's language lift, for example, ranges from -0.095 on Russian to $+0.049$ on Chinese. Inclusion as a multilingual generalist and performance above the pool average across languages are therefore distinct properties here. The same distinction appeared for the intended specialists in Table 2. In either case, these labels are not available to the selection procedures.

Language leaders do not follow intended specialization. A more direct test of the explanation based on assigning languages to specialists is whether the intended specialists lead the languages for which they were developed. We observe that none does. Qwen2.5-14B-Instruct has the highest accuracy on five languages: German, Spanish, Russian, Thai, and Chinese. aya-expanse-8b leads the remaining three: Greek, Hindi, and Vietnamese. The language-focused candidates in Table 2 rank between fourth and thirteenth among the 16 candidates on their own target languages. Assigning each language to its intended specialist would therefore not select the most accurate model on any of the eight languages.

The winning margins are generally small. On seven of the eight languages, the gap between the best and second-best candidate is at most 0.043, ranging from 0.013 on Spanish to 0.043 on German and Hindi. Greek is the exception, where aya-expanse-8b leads Llama-3.1-8B-Instruct by 0.183. Thus, leadership alternates between two candidates, neither of which dominates every language, and the

winner is usually only slightly ahead of the runner-up. The benchmark is therefore neither divided into languages each led by its intended specialist nor uniformly dominated by a single model.

Strong generalists are themselves complementary. The two language leaders have similar aggregate accuracies but substantially different task-level success patterns. aya-expanse-8b answers 0.7671 of the $2{,}400$ tasks correctly, compared with 0.7558 for Qwen2.5-14B-Instruct. However, they disagree on 465 tasks: Qwen2.5-14B-Instruct solves 219 tasks missed by aya-expanse-8b, and aya-expanse-8b solves 246 tasks missed by Qwen2.5-14B-Instruct. Together, the pair covers 0.8583 of the tasks, compared with 0.7671 for the more accurate model alone.

Some of this disagreement reflects differences across languages. Qwen2.5-14B-Instruct exceeds aya-expanse-8b by 0.243 on Thai, whereas aya-expanse-8b exceeds Qwen2.5-14B-Instruct by 0.263 on Greek and 0.127 on Hindi. The disagreement also occurs within languages, since each model solves tasks missed by the other in all eight languages.

Complementarity is therefore already substantial between the two most accurate individual candidates, rather than arising only from lower-accuracy language-focused models. Their aggregate performance differs by just 27 correct answers, but the identities of the tasks they solve differ on 465 occasions. Marginal accuracy alone does not reveal this structure. Feedback on missed tasks preserves this information by measuring whether a candidate succeeds specifically on tasks missed by the current committee. This is the task-level information used by the rescue-based procedures in Appendix G.1.

Standalone accuracy and complementary value differ. The complementarity shares in Table 2 and Table 3 show that the same distinction also appears among less accurate candidates. OpenThaiGPT is the clearest example. Its overall accuracy is only 0.258, among the lowest in the pool, yet it has the largest complementarity share of any candidate: 0.134 of the tasks it solves are missed by aya-expanse-8b. OpenHathi illustrates the opposite case. It combines low accuracy with a complementarity share of only 0.025, so few of its correct answers cover tasks missed by the most accurate individual model. Low standalone accuracy can therefore accompany either substantial or negligible complementary value. This is precisely the distinction used by the procedures that select using missed tasks in Appendix G.1.

Complementarity within the optimal committee. The exact optimal size-3 committee on the complete fresh matrix is {Qwen2.5-14B-Instruct, aya-expanse-8b, Typhoon-1.5-8b}, with coverage 0.8825. Thus, the optimum combines the two most accurate individual candidates Qwen2.5-14B-Instruct and aya-expanse-8b with one language-focused model (Typhoon-1.5-8b). To quantify what each member adds, we define its *leave-one-out contribution* as the number of tasks it solves that the other two committee members both miss. This quantity is relative to the committee; it does not imply that the remaining 13 candidates also miss those tasks. All three members contribute across several languages. aya-expanse-8b contributes 178 tasks, with its largest contributions on Greek (68) and Hindi (37). Qwen2.5-14B-Instruct contributes 108, led by Thai (28) and Russian (22). Typhoon contributes 58. Only 17 of these are Thai, while the remaining 41 are distributed across the other seven languages, including 11 German and 8 Hindi tasks. Thus, each member fills a distinct set of gaps in the other two members' coverage, and even Typhoon's contribution is mostly outside its intended language.

A different notion of uniqueness provides a complementary view of the full matrix. Only 131 of the $2{,}400$ tasks are solved by exactly one of the 16 candidates, so globally unique successes are uncommon. Useful complementarity therefore need not mean that one model is the only candidate capable of solving a task. What matters for selection is whether a candidate solves tasks missed by the particular committee assembled so far.

Conclusion. Language specialization contributes to variation in the fresh pool, but it does not make selection trivial. Neither does the presence of strong multilingual models. No intended specialist is the most accurate candidate on its target language, and two of the seven have negative lift there. Among the models included as multilingual generalists, measured breadth also varies substantially: five exceed the pool mean on all eight languages, whereas one does so on only two.

More importantly, aggregate accuracy does not determine complementary value. The two most accurate individual candidates differ by only 27 correct answers across $2{,}400$ tasks but disagree on 465, and both make substantial, differently distributed contributions to the optimal size-3 committee.

Typhoon adds another 58 leave-one-out successes, most of them outside Thai. The selection problem is therefore neither "choose one specialist per language" nor "choose the globally most accurate models." Its relevant empirical structure is the pattern of which models solve which tasks, the same structure represented by the information quantities in our theory. This conclusion applies to the candidate pool and evaluation protocol studied here; we do not generalize it to language-specialized models in general.

G.3 TyDi-QA robustness check

The fresh XQuAD experiment uses translated parallel questions, so we perform a separate robustness check using TyDi-QA questions originally written in their respective languages. The purpose of this experiment is narrower than that of Appendix G.1: we ask whether the complementarity observed in the fresh candidate pool persists on data from a different source and whether the agreement between our exclusive-KL procedure and AFG is preserved. The subset is too small to provide a second well-powered comparison of selection quality.

Outcome matrix and complementarity. The TyDi-QA subset contains 400 questions, producing a 16×400 binary outcome matrix with mean cell accuracy 0.5216. The union of the 16 candidates covers 0.8750 of the questions. The most accurate individual candidate is aya-expanse-8b at 0.7925, while the exact best size-3 committee reaches 0.8525, adding 0.0600 coverage beyond the best individual candidate. The median pairwise exclusive mass is 0.261, and no candidate pair has Jaccard overlap above 0.9. Thus, the basic structure that motivates complementarity-aware selection on XQuAD is also present on these untranslated questions: candidates differ substantially in which questions they solve, and combining models can improve coverage over the strongest individual candidate.

Agreement between our method and AFG. We apply the same deterministic selection design used for the fresh XQuAD experiment. Our exclusive-KL procedure and AFG by Neoh et al. (2026) select exactly the same committee in every one of the 1,200 matched runs at $k = 3$ and every one of the 1,200 matched runs at $k = 5$. Their held-out coverages are therefore identical in every matched run. As on XQuAD, this agreement concerns the selected ensembles rather than the statistical efficiency of the confidence criteria. Neither procedure reaches a fixed-confidence decision on this shorter benchmark. The controlled comparison of confidence criteria in Appendix F.5 is therefore the relevant experiment for the difference between our exclusive-KL confidence criterion and the marginal-Hoeffding confidence criterion used by AFG. In the experiment discussed in Appendix F.5, our method requires fewer accepted tasks and reaches a fixed-confidence decision substantially more often.

However, the subset cannot resolve selection quality. The limiting factor is the amount of data available to each greedy decision. Each split contains only 300 training questions. At $k = 3$, partitioning the training questions across the three greedy steps leaves 100 questions per step. After restricting attention to questions missed by the current committee, the three steps receive only 100, 21, and 17 accepted questions, respectively. No greedy step reaches a fixed-confidence decision at either committee size. The held-out set is also small. Each split contains only 100 test questions, so one additional covered question changes held-out coverage by 0.01. The observed differences from Top-k are substantially smaller than this unit: selection on missed tasks is ahead by $+0.0029$ at $k = 3$ (95% CI $[+0.0022, +0.0036]$) and behind by -0.0048 at $k = 5$ ($[-0.0053, -0.0043]$), with both differences stable across budgets. Although the intervals are narrow across repeated runs, the underlying held-out measurement for any individual split changes only in increments of 0.01. We therefore do not interpret the opposite signs at $k = 3$ and $k = 5$ as evidence that selection on missed tasks helps at one committee size and hurts at the other.

Conclusion. The TyDi-QA subset provides a check of robustness to the data source rather than a second selection benchmark. On questions originally written in their respective languages, the fresh candidate pool continues to exhibit substantial complementarity: the best size-3 committee improves coverage by 0.060 over the strongest individual candidate, the median pairwise exclusive mass remains substantial, and no pair is nearly identical by the Jaccard criterion. Our method and AFG also select identical committees throughout the experiment. However, the short sequences of training tasks prevent either confidence criterion from being met, and the 100-question held-out sets are too coarse to allow a meaningful comparison of the small observed selection margins. The

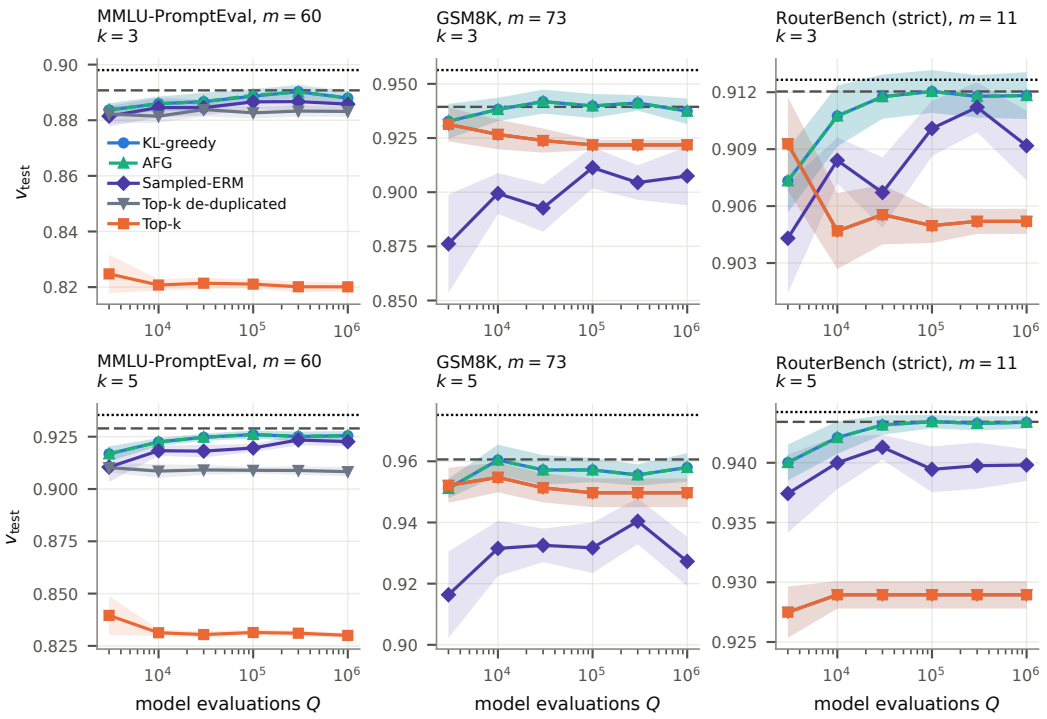


Figure 7: **Selection performance on the published evaluation matrices.** Held-out coverage versus total model evaluations on MMLU-PROMPTEVAL (left), GSM8K (center), and ROUTERBENCH (right), for $k = 3$ (top) and $k = 5$ (bottom). Curves show mean held-out coverage and bands show 95% confidence intervals. Dotted horizontal lines denote the exact test-split optimum OPT_k and dashed lines full-information greedy.

stronger selection comparison therefore remains the larger XQUAD experiment of Appendix G.1, while the evidence required for fixed-confidence decisions is evaluated separately in Appendix F.5.

G.4 OFFLINE SELECTION RESULTS

We next evaluate the selection procedures of Appendix E.6 on three published outcome matrices: MMLU-PROMPTEVAL, GSM8K, and ROUTERBENCH. Because these matrices provide complete model outcomes, the experiments require no additional model inference and allow us to test whether the selection patterns observed on the fresh benchmark persist across different candidate pools and task collections. Figure 7 reports held-out coverage as a function of total model evaluations for $k \in \{3, 5\}$ under the common accounting convention of Appendix E.5. For each matrix and committee size, we run 90 matched trials per method: five train/test splits, three seeds, and six requested budgets. The two procedures that select using missed tasks receive the same sequence of tasks by design. Table 4 summarizes the principal comparisons.

Selection quality. Selection on missed tasks consistently outperforms de-duplicated accuracy ranking. Across the three matrices and both committee sizes, our exclusive-KL procedure improves held-out coverage by $+0.0044$ to $+0.0144$, with all six 95% intervals excluding zero (Table 4). The gains are not ordered uniformly by matrix or committee size: they are largest at $k = 5$ on MMLU-PROMPTEVAL and ROUTERBENCH, and at $k = 3$ on GSM8K.

The pooled results mask a clear dependence on the evaluation budget, visible in the budget curves of Figure 7. At the smallest budget, 3×10^3 evaluations, selection on missed tasks trails de-duplicated Top-k in two of the six settings, by 0.0019 on ROUTERBENCH at $k = 3$ and 0.0010 on GSM8K at $k = 5$. From 10^4 evaluations onward, the mean difference is positive in all six settings. Restricting to the three largest requested budgets increases the mean gain in every setting, to a range of $+0.0059$

Table 4: **Offline selection results.** Paired held-out coverage differences between exclusive-KL and de-duplicated Top-k, and agreement between exclusive-KL and AFG, over 90 matched runs per setting. Intervals are reported as mean ± half-width. For exclusive-KL versus AFG, "agree" counts identical selected committees and "max" is the largest absolute run-level coverage difference.

		vs de-dup. Top-k	exclusive-KL vs AFG		
matrix	k	mean ± half-width	agree	mean	max
MMLU-PROMPTEVAL	3	$+0.0044 \pm 0.0011$	$85/90$	-3.8×10^{-5}	0.0060
MMLU-PROMPTEVAL	5	$+0.0144 \pm 0.0013$	$87/90$	-1.6×10^{-5}	0.0037
GSM8K	3	$+0.0140 \pm 0.0026$	$90/90$	0	0
GSM8K	5	$+0.0054 \pm 0.0022$	$90/90$	0	0
ROUTERBENCH	3	$+0.0051 \pm 0.0010$	$90/90$	0	0
ROUTERBENCH	5	$+0.0139 \pm 0.0006$	$90/90$	0	0

to $+0.0176$. Thus, as on the fresh benchmark, the benefit of selection on missed tasks emerges once the budget is large enough to collect informative observations on missed tasks.

The flattening of the curves at the upper end of Figure 7 reflects a different constraint: the finite number of training tasks, rather than the requested evaluation budget, becomes the limiting factor. At a requested budget of 10^6 model evaluations, the procedures that select using missed tasks consume, on average, only 19,321 evaluations on GSM8K at $k = 3$, 37,944 on ROUTERBENCH, and 148,706 on MMLU-PROMPTEVAL. References below to the "largest budget" therefore mean the largest *requested* budget, not the number of evaluations actually consumed. At that setting, the number of available training tasks is the limiting factor.

Figure 7 also places the procedures that select using missed tasks relative to the other selection baselines and the two full-information references. Selection on missed tasks outperforms Sampled-ERM in all six settings, by $+0.0023$ to $+0.0399$, with every interval excluding zero. At the largest requested budget, its selected committees are already close to the full-information references shown in the figure: held-out coverage is at most 0.0035 below greedy selection on the complete training matrix and 0.0008–0.0190 below the exact test-split optimum OPT_k. Thus, under the available sequences of tasks, selection on missed tasks achieves most of the coverage improvement attainable with full information.

Selection agreement between exclusive-KL and AFG. Our exclusive-KL procedure and AFG (Neoh et al., 2026) achieve essentially identical held-out ensemble quality, although they do not always select the same committee. Across the three offline matrices (MMLU-PROMPTEVAL, GSM8K, and ROUTERBENCH) and both committee sizes $k \in \{3, 5\}$, their largest absolute mean coverage difference is only 3.8×10^{-5}, and every confidence interval contains zero. The two procedures select the same committee in all 90 matched runs on GSM8K and ROUTERBENCH, at both committee sizes. On MMLU-PROMPTEVAL, they select the same committee in $85/90$ runs at $k = 3$ and $87/90$ runs at $k = 5$ (Table 4).

The few committee-selection disagreements on MMLU-PROMPTEVAL are concentrated at intermediate budgets. Across both committee sizes, exclusive-KL and AFG select different committees in only eight of the 180 matched runs: five at $k = 3$ and three at $k = 5$. All eight occur at requested budgets of 10^5 or 3×10^5 evaluations. At each of the three smaller budgets, the two procedures select the same committee in every matched run, and they also agree in all 15 runs at the largest requested budget, 10^6 evaluations, for each committee size.

When the selected committees differ, their held-out coverages can differ by a noticeable amount in an individual run: the largest absolute difference is 0.0060 at $k = 3$. These differences do not consistently favor either procedure. In some of the eight runs exclusive-KL selects the higher-coverage committee, whereas in others AFG does. As a result, the positive and negative run-level coverage differences largely offset one another when averaged over all 90 matched runs for a given committee size. This explains why the mean coverage difference remains close to zero despite the occasional selection of different committees. The offline results therefore indicate similar average held-out selection quality, not identical run-by-run behavior.

The termination results further clarify this pattern. On MMLU-PROMPTEVAL and GSM8K, no greedy step reaches a fixed-confidence decision at any budget. Each step instead ends when either its budget or its finite sequence of training tasks is exhausted, after which both procedures choose the candidate with the highest observed rescue rate. Differences can therefore arise only through the candidates they eliminate along the way. ROUTERBENCH is the only matrix on which the confidence criteria are met: our procedure confirms the choice on 22.2% of steps at $k = 3$ and 11.6% at $k = 5$, compared with 17.8% and 10.0% for AFG, respectively. Despite this difference in how often the criteria are met, the two procedures select the same committee in all 90 matched runs on ROUTERBENCH.

This agreement concerns selection quality, not the amount of evidence required for a fixed-confidence decision. The controlled experiment of Appendix F.5 holds the feedback on tasks missed by the current ensemble fixed and compares only the confidence criteria themselves. There, exclusive-KL requires fewer accepted tasks than AFG and reaches a fixed-confidence decision substantially more reliably before the available benchmark is exhausted. The two experiments therefore answer complementary questions: Figure 7 and Table 4 compare the quality of the ensembles selected by the two procedures, whereas Appendix F.5 compares the statistical evidence required to confirm their decisions.

Effect of prompt-expanded candidate pools. The comparison with Top-k requires particular care on MMLU-PROMPTEVAL, where the candidate pool contains multiple prompt configurations of the same underlying model. A naive accuracy ranking can therefore fill a committee with highly redundant variants of a single base model. On the 60-candidate subsample, naive Top-3 selects three configurations of the same base model and achieves coverage 0.822. Restricting Top-3 to at most one configuration per base model raises coverage to 0.883, compared with an optimum of 0.895. The same pattern appears at $k = 5$, where naive and de-duplicated Top-5 achieve coverages of 0.831 and 0.907, respectively.

This redundancy explains most of the apparently large advantage of selection on missed tasks over naive Top-k. On the prompt-expanded MMLU-PROMPTEVAL pool, the margin over naive Top-k is $+0.066$ at $k = 3$ and $+0.091$ at $k = 5$. Simply preventing multiple configurations of the same base model from entering the committee removes 93% and 84% of these margins, respectively. This issue is specific to the configuration-expanded pool: GSM8K and ROUTERBENCH contain no configuration variants, so naive and de-duplicated Top-k coincide on those matrices.

We obtain the same conclusion from a complementary analysis that removes prompt expansion altogether. Restricting MMLU-PROMPTEVAL to its 15 base models under a single fixed prompt template reduces the Top-3 gap to the optimum to only 0.004–0.012, with a median of 0.008 across all 100 templates in the release. The much larger gap observed for naive Top-k on the prompt-expanded pool is therefore primarily a consequence of allowing several configurations of the same base model to occupy the committee, rather than a general weakness of accuracy ranking. For this reason, we use de-duplicated Top-k as the primary accuracy-based baseline whenever the candidate pool contains configuration variants.

Committee composition. Having removed the duplication effect from the baseline, we next ask what accounts for the remaining advantage over de-duplicated Top-k. At the largest requested budget, our exclusive-KL procedure and AFG select the same committee in every matched run on all three matrices and at both committee sizes. Their committee-composition statistics are therefore identical at this budget, and we report them jointly. To make these statistics comparable across splits, we evaluate the members of each selected committee on the complete pooled outcome matrix rather than on the particular training or test split that produced the selection.

At $k = 3$, the committees selected by these two procedures contain members with mean individual accuracies 0.642, 0.738, and 0.695 on MMLU-PROMPTEVAL, GSM8K, and ROUTERBENCH, respectively. The corresponding values for de-duplicated Top-k are 0.727, 0.780, and 0.722. The committees selected using missed tasks are also less redundant: their mean pairwise Jaccard overlaps are 0.577, 0.663, and 0.617, compared with 0.741, 0.755, and 0.688 for de-duplicated Top-k.

The same pattern holds at $k = 5$. Mean individual accuracies are 0.624, 0.700, and 0.592 for the committees selected using missed tasks, compared with 0.693, 0.745, and 0.686 for de-duplicated Top-k. Their mean pairwise Jaccard overlaps are likewise lower: 0.590, 0.631, and 0.469, compared with 0.724, 0.724, and 0.667. Thus, across the three matrices and both committee sizes, selecting

candidates based on missed tasks produces committees whose members have lower average standalone accuracy but also substantially less overlap in the tasks they solve. This is precisely the tradeoff encouraged by best-member coverage: a candidate can improve the committee even with lower individual accuracy if it succeeds on tasks that the current members miss.

The distinction is especially pronounced if selection on missed tasks is compared with *naive* Top-k on MMLU-PROMPTEVAL. At $k = 3$, for example, mean pairwise Jaccard overlap is 0.577 for the committee selected using missed tasks but 0.959 for naive Top-k, which selects multiple configurations of the same base model. This much larger contrast primarily reflects the configuration-duplication effect established above, rather than the residual advantage of complementarity-aware selection after duplication is controlled. The comparison with de-duplicated Top-k is therefore the more informative one for understanding what selection on missed tasks contributes beyond simply avoiding redundant model variants.

Comparison with focal-diversity pruning. We also compare with the focal-diversity pruning criterion of Tekin et al. (2024), adapted to best-member coverage as described in Appendix E.6. Focal-diversity pruning and full-information greedy observe the same complete training block on each of five deterministic $75/25$ splits, select a committee, and are evaluated on the same held-out block. The comparison is therefore paired by split and measures selection quality under full information, independently of the query-accounting convention used elsewhere in this appendix. We use the same candidate pools and outcome matrices as in the preceding experiments, including the strict ROUTERBENCH matrix.

Across the three matrices and the two committee sizes $k \in \{3, 5\}$, focal-diversity pruning does not improve on full-information greedy in five of the six settings when coverage is averaged over splits. On MMLU-PROMPTEVAL, it selects exactly the greedy committee on all five splits at both committee sizes, giving a difference of zero. On GSM8K, its mean held-out coverage is lower by 0.0042 at $k = 3$ and 0.0006 at $k = 5$, and on ROUTERBENCH it is lower by 0.0058 at $k = 3$. The only positive average difference occurs on ROUTERBENCH at $k = 5$, where focal-diversity pruning is ahead by 0.0005.

The split-level results likewise provide no evidence of a systematic advantage. The positive mean on ROUTERBENCH at $k = 5$ is generated by a single split, while the other four select the same committee as full-information greedy. Across all 30 split-level comparisons, focal-diversity pruning achieves higher held-out coverage in four, lower coverage in eight, and identical coverage in eighteen. We therefore find no consistent improvement from the focal-diversity criterion over ordinary full-information greedy for the best-member coverage objective studied here. We do not make the stronger claim that focal-diversity pruning can never improve on greedy: individual splits provide counterexamples. With only five splits per setting, we report these results as paired point comparisons rather than attaching interval estimates.

Sensitivity to the pruning weight. The equal weighting used by our focal-diversity adaptation in Appendix E.6 is an untuned design choice, so we additionally test whether the comparison with full-information greedy depends on this choice. Holding the candidate committees, train–test splits, and all other experimental settings fixed, we replace the focal-diversity weight $1/2$ by $w \in \{0, 0.1, \ldots, 1\}$ and score each candidate committee according to $w\lambda_{\text{focal}}(S) + (1 - w)\widehat{v}_{\text{tr}}(S)$. Thus, $w = 0$ uses empirical best-member coverage alone, $w = 1$ uses focal diversity alone, and $w = 0.5$ is the weighting used in our primary comparison. Table 5 reports the resulting held-out comparison with full-information greedy.

The conclusion is stable throughout the sweep. At no value of w does focal-diversity pruning achieve higher held-out coverage than full-information greedy in more than 5 of the 30 comparisons, and its mean paired difference from greedy remains negative at every weight. In particular, $w = 0.5$ and $w = 0.6$ select exactly the same committee in all 30 comparisons. The latter comparison is useful because Tekin et al. (2024) use $w = 0.6$ for their MMLU and GSM8K experiments, although their second term is the validation accuracy of a learned ensemble combiner rather than the best-member coverage used here. The absence of a systematic improvement over full-information greedy is therefore not a consequence of the equal weighting chosen for our adaptation.

Conclusion. The published evaluation matrices reproduce the main selection pattern observed on the fresh benchmark. Across MMLU-PROMPTEVAL, GSM8K, and ROUTERBENCH, and at both $k = 3$ and $k = 5$, selecting complementary candidates based on missed tasks improves held-out

Table 5: **Sensitivity of focal-diversity pruning to the diversity weight.** "Ahead," "behind," and "tied" count held-out comparisons with full-information greedy across the 30 matrix, committee-size, and split settings; "mean difference" is the corresponding mean paired coverage difference. The primary baseline uses $w = 0.5$.

w	ahead	behind	tied	mean difference
0.0	0	1	29	−0.000404
0.1	3	6	21	−0.000600
0.2	4	8	18	−0.001097
0.3	4	8	18	−0.001195
0.4	4	8	18	−0.001606
0.5	**4**	**8**	**18**	**−0.001694**
0.6	4	8	18	−0.001694
0.7	4	8	18	−0.001694
0.8	4	8	18	−0.001694
0.9	5	8	17	−0.001593
1.0	5	8	17	−0.001593

coverage over a properly de-duplicated accuracy ranking. The pooled gains range from $+0.0044$ to $+0.0144$, with all six 95% intervals excluding zero. As on the fresh benchmark, however, the gain is budget-dependent: it emerges once enough observations on tasks missed by the current ensemble are available and is absent at the smallest budget in two of the six settings.

Our exclusive-KL procedure and AFG achieve essentially the same average held-out selection quality, despite occasional committee disagreements on MMLU-PROMPTEVAL. This similarity in selection quality should not be confused with equivalence of the confidence criteria. In the controlled experiment of Appendix F.5, where both procedures receive the same feedback on tasks missed by the current ensemble, exclusive-KL requires fewer accepted tasks and reaches a fixed-confidence decision more reliably. The offline experiment and the grid of confidence criteria therefore establish complementary results. Specifically, the two procedures ultimately select ensembles of comparable quality, while our confidence criterion requires less statistical evidence to confirm those decisions.

Finally, the MMLU-PROMPTEVAL results show that candidate-pool construction can strongly affect comparisons with accuracy-based selection. Much of the apparent advantage over naive Top-k disappears once multiple prompt configurations of the same base model are prevented from occupying the committee. A smaller but consistent advantage remains against the de-duplicated baseline, and the committee-composition analysis explains its source. In particular, selection on missed tasks trades some individual model accuracy for less overlapping success patterns, thereby increasing best-member coverage.

H COMPUTATIONAL RESOURCES AND REPRODUCIBILITY

Hardware and software. All experiments were run on a single Apple M4 Max laptop with 128 GB of unified memory (macOS 26.5.2, arm64). We used no GPU cluster, external accelerator, or paid inference API. Fresh model inference used MLX (Hannun et al., 2023) through `mlx_lm` 0.19.1, with 4-bit quantized weights at group size 64, batch size 16 with length-sorted batching, greedy decoding, and a 24-token generation limit. The analysis code used Python 3.11.15, NumPy 2.2.6, SciPy, and Numba.

Cost of the fresh evaluation. The fresh XQUAD evaluation required $16 \times 2{,}400 = 38{,}400$ model evaluations, and the TYDI-QA robustness subset required a further 6,400. Sustained throughput ranged from 0.45 generations per second for the 14B candidate to 4.93 for the 1.5B candidate. Generating the XQUAD outcomes took 39,375 s (10.9 h) of wall-clock time, and the TYDI-QA subset required approximately 4,200 s (1.2 h), for approximately 12.1 h in total. Peak MLX memory allocation reached 79 GB on the longest runs. This is the maximum memory allocated by the framework during execution, including its cache, rather than the memory occupied by the model weights alone. Model weights were already present in the local Hugging Face cache, so download time is not included. We distinguish wall-clock cost from the evaluation cost studied in the paper. A *model evaluation* is one model-task evaluation and is the hardware-independent query unit used by our

theoretical and empirical accounting. Wall-clock time instead depends on the hardware, inference software, quantization, batching, and model size. We report it here for reproducibility, but do not use it to compare selection procedures.

Cost of the offline experiments. All Track-A experiments and the offline Track-B experiments operate on precomputed outcome matrices and therefore require no model inference. Together, they require only a few CPU-hours on the same machine.

Reproducibility materials. We retain the complete experimental configuration, including model repositories and revision hashes, quantization and decoding settings, prompts, question and paragraph identifiers, data splits, evaluation-budget grids, and accounting conventions. The experimental materials also include the generated outcome matrices and raw generation logs, together with the scoring and selection code, step-by-step selection logs, intermediate results, machine-readable tables, and figure-generation scripts. These materials document the configurations used for the reported experiments and allow the computational results to be reproduced without rerunning the fresh model evaluations.

Inference protocol. The fresh evaluation uses the same fixed 4-bit quantized inference configuration for all 16 candidates. Quantization and other inference choices can affect both model accuracy and the pattern of successes across tasks, so the reported coverage values characterize this evaluation configuration rather than the models' full-precision benchmark performance. All selection methods operate on the same resulting outcome matrices, with the candidate pool, inference configuration, and scoring rule held fixed.