WHEN ARE SMALLER BENCHMARKS BETTER?

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated evaluations often rank systems differently from humans. We show that agreement improves substantially when the benchmark is reduced to a small subset of its existing items, without adding new data. Sieve selects as small as 20 items on a Pareto front over two objectives: per-item accuracy match with the source benchmark, and correlation with the human system ranking. We evaluate on held-out systems and held-out items, so the gain cannot be circular. The 20-item subsets agree with humans better than the full 80- to 5,000-item benchmarks they came from, closing about half the remaining gap to perfect agreement on average and all of it for the weakest judges. Across the 36-cell task-judge grid the mean gain is $+0.37$ at one-sided Wilcoxon $p = 2.7 \times 10^{-6}$. At the extreme, LLaMA-3B on MT-Bench swings from an inverted ranking, $\rho_{\mathcal{T}} = -0.49$, to perfect agreement, $\rho_{Q^\star} = +1.00$. What a judge gains, it turns out, depends on how much room it had to improve. We call this the *Headroom Law*: gain scales with headroom at slope $\alpha = 0.86$, 95% CI $[0.75, 0.93]$, with per-judge slopes in $[0.65, 0.94]$ and seven independent selectors recovering $\alpha \in [0.87, 1.01]$. So we can tell whether Sieve will help on a new task-judge pair before running the search at all: leave-one-out $r = 0.92$ on fit cells, and live prediction $r = 0.935$ on 66 cells we had never seen.

1 INTRODUCTION

Benchmarks are how the field decides which models are better, yet the test sets behind them are rarely checked against the human judgment they aim to approximate. Many items are irrelevant, redundant, or contaminated, adding noise rather than signal, and the resulting failure modes are well documented (§2). The consequence is striking: two benchmarks built to measure the same capability can crown different winners. If our instruments cannot agree, we should ask what, exactly, they are measuring.

A growing line of work makes benchmarks more efficient to run. Methods such as tinyBenchmarks (Polo et al., 2024) and metabench (Kipnis et al., 2025) select a small subset of items that reproduces the full benchmark's ranking, shrinking tens of thousands of items down to one to two hundred with little loss in fidelity. These methods, however, treat the full benchmark as ground truth. When its ranking diverges from human judgment, a faithful subset inherits that divergence intact: it does not correct the error, it *compresses* it (Table 3). Efficiency, in other words, is only as valuable as the target it preserves.

This paper asks two questions. **(Q1)** *Can a smaller test set agree with humans more than the full benchmark does?* **(Q2)** *If so, can we predict when, before selecting a single item?*

At first, Q1 sounds counterintuitive. Why would throwing away most of a benchmark make it better? Because an item only helps if it captures a distinction humans care about. Suppose humans rank three systems $A > B > C$. An item that every system gets right, or every system gets wrong, says nothing about that order, and an item that hinges on an irrelevant cue can scramble it. Two well-chosen items, on the other hand, can recover the whole ranking: one separates A from $\{B, C\}$, the other separates B from C. More items do not guarantee more signal. Past a point they mostly add noise.

This changes what compression should aim for. We do not want *lossless* compression, which reproduces the benchmark's ranking, but *human-agreement-gaining* compression, which improves on the benchmark itself. Following Kane's criterion-agreement view of measurement (§2), we measure agreement against human judgment directly. Let R_Q be the system ranking an LLM-as-Judge induces on an item set Q, and let $\mathcal{H}$ be the human ranking of the same systems. Human agreement is the

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

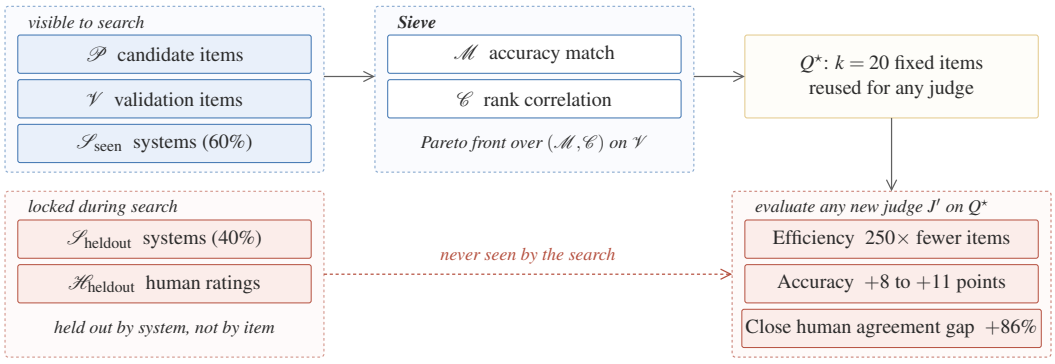


Figure 1: Sieve at a glance. The source benchmark is split once, before any search: $\mathcal{P}$ is the candidate pool, $\mathcal{V}$ the validation set on which objectives are computed, $\mathcal{S}_{\text{seen}}$ the 60% of systems visible to the search (§3.2). Sieve returns a fixed k=20 subset $Q^\star$ by Pareto search over accuracy match $\mathcal{M}$ and ranking correlation $\mathcal{C}$ on $\mathcal{V}$, and $Q^\star$ is reused across future judges. Outcomes at right are measured on the held-out 40% of systems $\mathcal{S}_{\text{heldout}}$.

Spearman correlation $\rho(R_Q, \mathcal{H})$. For a source benchmark $\mathcal{T}$ this is $\rho_{\mathcal{T}} = \rho(R_{\mathcal{T}}, \mathcal{H})$, and for a learned subset $Q^\star$ the *human-agreement gain* is

$$\text{HAG} = \rho_{Q^\star} - \rho_{\mathcal{T}}.$$

Positive HAG means the subset ranks systems more like humans than its source benchmark does.

A non-circular protocol. Gains of this kind are easy to manufacture by accident: if items are chosen using the same human labels they are scored against, improvement is guaranteed and meaningless. We therefore fix three separations before search begins. *Items* are split into a candidate pool $\mathcal{P}$, a validation set $\mathcal{V}$, and a locked test set $\mathcal{T}$, with human labels split to match. Search draws from $\mathcal{P}$, scores on $\mathcal{V}$, and never touches $\mathcal{T}$ or its labels. *Systems* are split 60/40 into seen and held-out sets, and search sees only the former. *References* are kept apart as well: search never observes the human system ranking, only item-level labels on the search split and a panel of reference judge prompts. Every gain we report is measured on items and systems the search never saw.

Sieve. Our method, Sieve, looks for a 20-item subset that scores well on two objectives computed on $\mathcal{V}$ (§3.4). *Accuracy match* ($\mathcal{M}$) keeps the subset's accuracy from underestimating the source, and *ranking correlation* ($\mathcal{C}$) preserves the system ordering induced by the reference judge panel. Starting from 60 random 20-item subsets drawn from $\mathcal{P}$, Sieve runs NSGA-II within a budget of 3,000 judge calls and returns the Pareto-front candidate closest to $(\max \mathcal{M}, \max \mathcal{C})$ (§3.4). Figure 1 summarises the pipeline.

The Headroom Law. The answer to Q2 turns out to be one number, the slope α. Call the gap between a benchmark's human agreement and perfect agreement its *headroom*, $1 - \rho_{\mathcal{T}}$. Across 36 benchmark-judge pairs, HAG can be predicted from headroom alone:

$$\text{HAG} \approx \alpha\,(1 - \rho_{\mathcal{T}}) + \beta,$$

with slope $\alpha \approx 0.86$ (bootstrap 95% CI $[0.75, 0.93]$) and intercept $\beta \approx -0.11$ for Sieve's default setting. Headroom is also a hard ceiling on the gain (Corollary 1), so α, which we call the *bound-saturation coefficient*, tells us how much of the available gain Sieve actually captures. The Law is robust: $\alpha \in [0.65, 0.94]$ across four judges and $[0.87, 1.01]$ across seven independent subset selectors (Appendix J.1). Furthermore, the Headroom Law predicts leave-one-out HAGs at Pearson $r = 0.92$ across the 36 cells (Figure 2). The practical reading is simple:

When a benchmark disagrees with humans, a 20-item subset closes most of the gap.

When it already agrees, there is little left to gain, and compression will not help (§5.3).

Evidence. We test this with up to seven judges, from a 3B open-weight model to frontier-scale GPT-5, across 11 system-level and 4 item-level benchmarks (§5.1). On held-out items and systems, 20-item subsets beat their full benchmarks by an average of $+0.45$ for LLaMA-3.2-3B and $+0.23$ and $+0.30$ for Qwen-7B and Qwen-14B. MT-Bench shows both ends of the Law: the weak judge

LLaMA-3.2-3B flips from $\rho_{\mathcal{T}} = -0.49$ to $\rho_{Q^\star} = +1.00$ with Sieve, while Luna barely moves. The selected subsets also transfer across judges, in proportion to their in-domain gain (Appendix H.2).

Contributions.

1. **A non-circular protocol.** If the search shares items, systems, or human labels with the human reference, the reference moves with the search and any subset looks good. Prior compression work (Jiang et al., 2026b, Polo et al., 2024, Kipnis et al., 2025) blocks the first two axes. We block all three, so every gain is measured against a reference the search could not shape. The protocol applies to any learned-benchmark method.
2. **The Headroom Law.** A single linear rule predicts HAG *before* any items are selected. It explains 83% of the variance ($R^2 = 0.83$) across 36 benchmark-judge pairs (falsification in §5.3).
3. **Sieve.** Where prior compression preserves the source benchmark's ranking, Sieve searches for the subset whose ranking is closer to the human ranking, under a tiny call budget, deterministically.
4. **Evidence at scale.** Over $1{,}000$ (task, judge, seed) trials show that smaller subsets outperform their full benchmarks when the judge has headroom, and stay within the noise band at the ceiling.

2 Related work

Benchmark failures and repairs. Benchmark scores can be precise yet invalid. Audits and behavioural probes (Ribeiro et al., 2020, Röttger et al., 2021, Perlitz et al., 2024, Zheng et al., 2023, Dubois et al., 2024b) trace this to annotation shortcuts (Gururangan et al., 2018, Poliak et al., 2018, McCoy et al., 2019, Belinkov & Glass, 2019), weak links to the target capability (Bowman & Dahl, 2021, Ethayarajh & Jurafsky, 2020, Raji et al., 2021), contamination (Sainz et al., 2023, Golchin & Surdeanu, 2024, Xu et al., 2024, Zhang et al., 2024), and prompt-format sensitivity (Sclar et al., 2024, Mizrahi et al., 2024, Alzahrani et al., 2024). Past fixes either *expand* a benchmark with harder items (Kiela et al., 2021) or *compress* it into IRT-anchored subsets that preserve the source ranking (Polo et al., 2024, Kipnis et al., 2025). Sieve also compresses, but it selects for agreement with humans rather than fidelity to the source, under a non-circular protocol. Its ranking objective $\mathcal{C}$ borrows from item-response theory (Birnbaum, 1968, Lord, 1980, Embretson & Reise, 2000) without fitting a latent-ability model, avoiding doubts about IRT's unidimensionality and monotonicity assumptions for AI benchmarks (Jiang et al., 2026a). It puts into practice the item-level measurement programme of Jiang et al. (2026b), grounded in Kane (2013).

LLM-as-Judge. Since LLM judges were introduced (Zheng et al., 2023, Chiang et al., 2024, Liu et al., 2023), work has made them more robust, less biased, more specialised, and pooled them into panels (Wei et al., 2025, Atil et al., 2025, Raina et al., 2024, Chen et al., 2024, Liu et al., 2024, Dubois et al., 2024a, Wang et al., 2024b, Li et al., 2024, Zhu et al., 2025, Kim et al., 2024, Wang et al., 2024a, Verga et al., 2024, Lambert et al., 2025). All of it improves the judge for a fixed test set. We do the reverse: fix the judge and improve the test set. The Headroom Law (Law 1) says when this helps: when the judge still has room to agree more with humans.

3 Method

3.1 Problem setup

Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a collection of systems and J an LLM-as-Judge. Let $\mathcal{B}$ be the original benchmark and $Q \subset \mathcal{B}$ a candidate test set of fixed size k. We use $k = 20$ throughout. The judge scores every system-item pair, then aggregates over Q to rank the systems, giving R_Q.

What we want is for this ranking to match the human ranking $\mathcal{H}$ of the same systems. We measure how closely they agree as

$$\mathrm{HA}(Q) = \rho(R_Q, \mathcal{H})\,. \tag{1}$$

Ideally, then, we would pick the subset that maximizes this agreement:

$$Q^\star = \arg\max_{\substack{Q \subseteq \mathcal{B} \\ |Q| = k}} \rho(R_Q, \mathcal{H})\,. \tag{2}$$

But $\mathcal{H}$ is not available while we select the subset, so we need a surrogate.

In practice, we solve Equation 2 through

$$Q^\star \in \mathrm{Knee}\big[\mathrm{Pareto}\left\{\big(\mathcal{M}(Q), \mathcal{C}(Q)\big) : Q \subseteq \mathcal{P},\ |Q| = k\right\}\big], \tag{3}$$

where $\mathcal{M}$ (accuracy match) and $\mathcal{C}$ (ranking correlation) are the two subset objectives defined in §3.4, and the candidate pool $\mathcal{P}$, validation set $\mathcal{V}$, and seen-system split $\mathcal{S}_{\mathrm{seen}}$ are defined in §3.2. Equation 2 is what we care about. Equation 3 is what we can actually observe. The paper succeeds to the extent that the second approximates the first on systems and items never used during selection. Here $\mathrm{Pareto}\{\cdot\}$ keeps the candidates that no other candidate strictly dominates, and $\mathrm{Knee}[\cdot]$ picks the one closest to the ideal point $(\max \mathcal{M},\ \max \mathcal{C})$. §3.4 gives the search details.

3.2 Non-circular evaluation protocol

If $\mathcal{H}$ were visible during subset selection, Equation 2 would become circular. The protocol below prevents this. It applies to any learned-benchmark method, not just Sieve, and we see it as a contribution in its own right. It keeps evaluation non-circular along three independent axes: items, systems, and reference aggregation.

Item axis. We split the source benchmark into a candidate pool $\mathcal{P}$, a validation set $\mathcal{V}$, and a locked test set $\mathcal{T}$. Candidates come from $\mathcal{P}$, objectives are computed and the search runs on $\mathcal{V}$, and $\mathcal{T}$ stays untouched until the subset is frozen and evaluation begins.

System axis. We split the systems $60/40$ into seen $\mathcal{S}_{\mathrm{seen}}$ and held-out $\mathcal{S}_{\mathrm{heldout}}$. The search sees only $\mathcal{S}_{\mathrm{seen}}$, so the human ranking on $\mathcal{S}_{\mathrm{heldout}}$ is never observed during search.

Reference-aggregation axis. We build the human reference $\mathcal{H}_{\mathrm{clean}}$ only from $\mathcal{T}$ items and $\mathcal{S}_{\mathrm{heldout}}$ systems, so no label seen during search leaks into it. With the item and system splits, this closes all three axes. From here on, $\mathcal{H}$ means $\mathcal{H}_{\mathrm{clean}}$. Appendix I.3 compares it with $\mathcal{H}_{\mathrm{leaky}}$ across a 200-cell grid.

3.3 Evaluation metrics

We evaluate the selected subset on the locked test set along three distinct dimensions.

Efficiency reproduction gap. This measures how well a smaller set of items reproduces the behaviour of the source benchmark:

$$\mathrm{Rep}(Q) = \left|\mathrm{Acc}(Q) - \mathrm{Acc}(\mathcal{T})\right|. \tag{4}$$

Smaller is better, meaning the compression is more faithful. In tables we report it as $\mathrm{Rep}{\downarrow}$.

Accuracy gain. This measures how much the learned subset raises item-level agreement between judge and human, compared with the source benchmark. Item-level accuracy is

$$\mathrm{Acc}(Q) \ = \ 1 - \frac{1}{R\,|Q|} \sum_{x \in Q} \big|J(x;p) - y(x)\big|, \tag{5}$$

where $J(x;p)$ is the judge's score for item x under prompt p, $y(x)$ is the aggregated human score, and R is the range of the rating scale. The accuracy gain over the source benchmark is

$$\Delta \ = \ \mathrm{Acc}(Q^\star) - \mathrm{Acc}(\mathcal{T}), \tag{6}$$

which is the item-level counterpart of HAG.

Human agreement is the central metric of this paper. Acc and HAG ask different questions: Acc asks whether the judge rates each item the way a human would, while HAG asks whether the judge ranks systems the way a human would.

Human agreement. This measures how closely the judge's system-level decisions match human ones. For each test set Q, the judge aggregates its item scores into a system ranking R_Q, and human agreement is

$$\mathrm{HA}(Q) = \rho(R_Q, \mathcal{H}), \tag{7}$$

where $\mathcal{H}$ is the human system ranking and ρ is Spearman rank correlation (Spearman, 1904).

Definition 1 (Human-agreement gain)**.** *For a learned subset $Q^\star$, the human-agreement gain over the default benchmark $\mathcal{T}$ is*

$$\mathrm{HAG}(Q^\star) = \rho(R_{Q^\star}, \mathcal{H}) - \rho(R_{\mathcal{T}}, \mathcal{H})\,. \tag{8}$$

A positive value means that swapping the full benchmark for the learned subset (for example, 20 items) brings the judge's system ranking closer to human judgment.

Throughout the main text, we compute HAG against the fair-test reference $\mathcal{H}_{\mathrm{clean}}$ (§3.2). The comparison with a leaky reference $\mathcal{H}_{\mathrm{leaky}}$ appears only in Appendix I.3.

We treat a negative gain that falls within finite-sample noise, $|\mathrm{HAG}| \le \varepsilon$ with $\varepsilon \le 0.07$, as statistically indistinguishable from zero. For significance, we use bootstrap confidence intervals (Efron, 1979) with paired t-tests or Wilcoxon signed-rank tests (Wilcoxon, 1945).

The key point is that HAG is never optimized directly, nor seen during search. It is a held-out result of optimizing the three observable objectives that follow.

3.4 Learning objectives

The two subset objectives $\mathcal{M}$ and $\mathcal{C}$ that enter Equation 3 are defined below. Both use per-item labels $y(x)$ via Acc on $\mathcal{V}$ ($\mathcal{S}_{\mathrm{seen}}$ only), and $\mathcal{C}$ additionally uses a panel of three reference judge prompts. Neither uses the aggregated system ranking $\mathcal{H}$.

Accuracy match ($\mathcal{M}$). The idea behind $\mathcal{M}$ is simple: the subset Q should never make the judge look worse than it really is on the validation set $\mathcal{V}$. So we penalise Q when its accuracy falls below that of $\mathcal{V}$, and leave it alone when its accuracy is higher.

To make this precise, let $J(x;p) \in [m, M]$ be the score the judge gives item x under prompt p, let $y(x) \in [m, M]$ be the aggregated human score, and let $\Delta = M - m$ be the range of the rating scale (so Δ=4 on a 1-5 scale). For any set of items A and prompt p, we measure how closely the judge tracks humans as

$$\mathrm{Acc}(A;p) = 1 - \frac{1}{\Delta|A|}\sum_{x\in A} |J(x;p) - y(x)|\,. \tag{9}$$

The judge and the human can differ by at most Δ on any item, so $\mathrm{Acc}(A;p)$ always lies between 0 and 1. When the prompt is obvious, we drop it and write $\mathrm{Acc}(A)$.

We then define

$$\mathcal{M}(Q) = 1 - \max\bigl(0,\ \mathrm{Acc}(\mathcal{V}) - \mathrm{Acc}(Q)\bigr). \tag{10}$$

Both accuracies are measured with the current prompt. If the judge does at least as well on Q as on $\mathcal{V}$, $\mathcal{M}(Q) = 1$. If it does worse, $\mathcal{M}$ drops in step with the shortfall. We always compute $\mathcal{M}$ on the validation set and never touch the locked test set.

Why one-sided. Only $\mathrm{Acc}(Q) < \mathrm{Acc}(\mathcal{V})$ is harmful: a subset that understates the judge's accuracy misleads anyone who uses it, so we penalise it. The reverse, $\mathrm{Acc}(Q) > \mathrm{Acc}(\mathcal{V})$, is expected, since Sieve picks the items with the clearest signal. A symmetric version, $1 - |\mathrm{Acc}(Q) - \mathrm{Acc}(\mathcal{V})|$, would penalise both. Under joint prompt-subset search, that rewards closing the gap the wrong way: swapping in noisier items, or choosing a prompt that makes the judge *worse* on Q. The one-sided form removes this incentive while still anchoring Q to $\mathcal{V}$.

Ranking correlation ($\mathcal{C}$). Matching average accuracy is not enough: if every judge scores the same on Q, it cannot rank them. $\mathcal{C}$ asks that Q separate judges the way $\mathcal{V}$ does. To stay backend-agnostic, our panel is the default judge under three prompts of deliberately different strength: weak, medium, and strong (Appendix D.3). Let $\mathbf{a}_Q = \bigl(\mathrm{Acc}(Q;p_r)\bigr)_{r=1}^{3}$ and $\mathbf{a}_{\mathcal{V}} = \bigl(\mathrm{Acc}(\mathcal{V};p_r)\bigr)_{r=1}^{3}$ be the panel's accuracies (Eq. 9) on Q and $\mathcal{V}$, and $\rho(\mathbf{a}_Q, \mathbf{a}_{\mathcal{V}}) \in [-1, 1]$ their Spearman correlation (0 if either vector is fully tied, Appendix B.1). Then

$$\mathcal{C}(Q) = \underbrace{\frac{1 + \rho(\mathbf{a}_Q,\, \mathbf{a}_{\mathcal{V}})}{2}}_{\text{alignment} \in [0,1]} \cdot \underbrace{\min\left(1,\ \frac{\mathrm{range}(\mathbf{a}_Q)}{\mathrm{range}(\mathbf{a}_{\mathcal{V}})}\right)}_{\text{spread ratio} \in [0,1]}, \tag{11}$$

where $\mathrm{range}(\mathbf{a}) = \max(\mathbf{a}) - \min(\mathbf{a})$.

The two factors do different jobs. *Alignment* checks that Q ranks the panel in the same order as $\mathcal{V}$. *Spread ratio* checks that Q keeps the panel as far apart as $\mathcal{V}$ does, penalising subsets that squeeze the accuracies together. Because the three prompts are built to differ in strength, almost any subset keeps their order (weak $<$ medium $<$ strong), so alignment sits at 1 and cannot tell subsets apart. The signal comes from the spread ratio, which rewards subsets that keep the prompts as far apart as $\mathcal{V}$ does. If $\mathcal{V}$ itself cannot tell the prompts apart ($\mathrm{range}(\mathbf{a}_{\mathcal{V}}) = 0$), the ratio is undefined, so we set $\mathcal{C} = 0.5$: a neutral value that leaves the search to be driven by $\mathcal{M}$ alone.

Why a prompt panel stands in for systems. At evaluation time, ρ compares rankings over the $|\mathcal{S}|$ held-out systems (§3.2). Those are off limits during search, so $\mathcal{C}$ uses the three-prompt panel instead. The rationale comes from item-response theory: if Q separates judges of different strength the way $\mathcal{V}$ does, it should also preserve the ranking of any system panel drawn from the same population. We do not derive this equivalence. The fit of $\alpha = 0.86$ across 116 cells (§5.3) is empirical support for it.

Search procedure and budget. Our default search, *SimpleSearch*, is deliberately plain. For each cell, we draw N=60 random 20-item subsets from $\mathcal{P}$, score each with the default prompt on $\mathcal{S}_{\text{seen}}$ (caching every system-item pair), keep the Pareto front (Deb et al., 2002), and return the candidate closest to $(\max \mathcal{M}, \max \mathcal{C})$. No mutation, crossover, or learned surrogate is involved. Each cell costs about 3,000 judge calls (Algorithm 1, Figure 1, Appendix D.3). Seven other selectors give $\alpha \in [0.87, 1.01]$ (Appendix J.1), so the result comes from the objectives, not the search.

4 The Headroom Law

Our central quantity is the human-agreement gain HAG (Definition 1): when $\mathrm{HAG} > 0$, the learned subset ranks systems more like humans than the default benchmark does. We call $1 - \rho(R_{\mathcal{T}}, \mathcal{H})$ the judge's *headroom*, the distance to perfect agreement. Headroom caps the gain:

Corollary 1 (Headroom / weak-judge lift)**.** *Because $\rho_{Q^\star} \leq 1$, $\mathrm{HAG} \leq 1 - \rho_{\mathcal{T}}$ on the same set of systems used to compute both quantities. The bound holds under any protocol (search-time or held-out), as long as both are read on the same systems.*

The Headroom Law says Sieve reaches a roughly fixed fraction of this cap.

Law 1 (Headroom Law)**.** *Across (task, judge) cells,* HAG *grows linearly with headroom:*

$$\mathrm{HAG} \;=\; \alpha\,\big(1 - \rho(R_{\mathcal{T}}, \mathcal{H})\big) \;+\; \beta \;+\; \eta,$$

where $\alpha \in (0, 1]$, the bound-saturation coefficient*, is the fraction of the cap in Corollary 1 that Sieve reaches, and the noise η is bounded by the finite-sample tolerance $\varepsilon(|Q^\star|, |\mathcal{S}|)$. Empirically, $\alpha \approx 0.86$ pooled, with per-judge slopes in $[0.65, 0.94]$.*

Reading the law. On average, Sieve closes 86% of the gap to perfect agreement, so the gain depends on where a benchmark starts: about 0.09 near the ceiling (headroom 0.1), 0.86 at random (headroom 1), and 1.3 when reversed (headroom 1.5). The worse the benchmark, the larger the gain. The law is falsifiable: it predicts gains on unseen cells, and those predictions can fail. With $|Q^\star| = 20$, bootstrap gives $\varepsilon \leq 0.06$, so single-cell readings inside that band count as null.

Why a smaller learned set can be better. Ranking systems takes surprisingly little information: $\log_2(|\mathcal{S}|!)$ bits, or about 9.5 bits for six systems. What matters is how sharply each item separates the systems. Under a simple item-response model, rank reliability grows as $\bar{d}_Q \sqrt{|Q|}/\sigma$, where $\bar{d}_Q$ is the average per-item discrimination and σ the scoring noise. A learned $Q^\star$ that concentrates on high-discrimination items can therefore beat a larger random draw.

Theorem 1 (Appendix C) covers the shape of this behavior: some size-m subset reproduces $\mathcal{T}$'s per-classifier accuracy within $O(\sqrt{d/m})$ error, and m depends on the classifier class's pseudo-dimension d rather than on $|\mathcal{T}|$, so one fixed small exam works across corpora ranging from a few hundred to several thousand items. It does not pin down the plateau's height, the choice $|Q|$=20, or $Q^\star$'s human-agreement gain over $\mathcal{T}$. Those are empirical laws verified in Section 5.2.

	LLaMA-3.2-3B			Qwen-2.5-7B		
Dataset	**base→ $+Q$**	**Rep.↓**	$\rho\uparrow$	**base→ $+Q$**	**Rep.↓**	$\rho\uparrow$
Amazon ESCI	72.0→78.8	6.4	0.67	60.67→74.5	13.4	0.62
WANDS	64.2→75.0	6.0	1.00	75.4→84.8	7.6	1.00
SummEval°	82.1→81.7	1.4	1.00	79.3→82.7	2.8	0.33
TruthfulQA°	71.2→73.0	7.6	0.79	71.9→77.3	7.1	0.83

Table 1: Judge accuracy, default benchmark → learned subset (prompt fixed). $\mathcal{M}$+$\mathcal{C}$ front, k=50, 3-trial mean on locked $\mathcal{T}$. Columns: base→ $+Q$ accuracy, reproduction gap Rep↓, system-rank correlation $\rho\uparrow$. ° rows carry 95% bootstrap CI and Wilcoxon p in Appendix G.1.

5 EXPERIMENTS

Our experiments ask three questions. Does a small $Q^\star$ keep the judge's item-level accuracy? Does it rank systems more like humans do? And does the gain follow the Headroom Law (Law 1)? We answer them in turn, using the metrics of Section 3.3 (Eqs. 4, 6, 7). Appendix A gives a three-step recipe (score → Pareto search → deploy) and the eleven $Q^\star$ subsets used here.

5.1 EXPERIMENTAL SETUP

Protocol. Each experiment runs on a *(task, judge) cell*: one task paired with one judge. Within a cell we compare the default benchmark $\mathcal{T}$ with the learned $Q^\star$, holding the judge, systems, and human reference fixed. Search follows Section 3.4. The optimiser itself uses no LLM: it works only from black-box judge scores under the fixed default prompt.

Datasets and tasks. We use four item-level datasets: Amazon ESCI (Reddy et al., 2022), WANDS (Chen et al., 2022), SummEval (Fabbri et al., 2021), and TruthfulQA (Lin et al., 2022). We use eleven system-level benchmarks: SummEval, LMSYS Arena, Nectar (Zhu et al., 2024), MT-Bench, AlpacaEval, UltraFeedback (Cui et al., 2023), JudgeLM, MMLU-Pro (Wang et al., 2024c), HELM GSM8K (Liang et al., 2022), PromptBench, and HumanEval (Chen et al., 2021).

Judges and human references. Five judges span the headroom axis: LLaMA-3.2-3B (Grattafiori et al., 2024), LLaMA-3.3-70B, Qwen-2.5-7B, Qwen-2.5-14B (Yang et al., 2024), and GPT-5. Luna serves as a near-ceiling reference. The human reference $\mathcal{H}$ depends on the task: expert Likert ratings, Arena Elo, or GPT-4 preference judgments.

5.2 RESULTS

Accuracy. Learned subsets raise item-level judge accuracy in 7 of 8 dataset-judge pairs, with a small reproduction gap (Table 1). The exception, LLaMA on SummEval, is flat ($82.1 \rightarrow 81.7$). Gains are largest where the judge has the most room to improve (Amazon ESCI, WANDS). $Q^\star$ also preserves *which* items the judge gets wrong, not just its average: item-level error correlation with $\mathcal{T}$ is Kendall τ=0.88 for LLaMA and 0.55 for Qwen (Appendix I.2).

Human agreement. Across 36 cells, mean HAG is $+0.37$ (median $+0.22$, one-sided Wilcoxon $p = 2.7\times10^{-6}$). On held-out systems (Table 2, right), it is $+0.45$ (LLaMA-3B), $+0.23$ (Qwen-7B), $+0.30$ (Qwen-14B), and $+0.02$ (near-ceiling Luna), about 27% below the all-systems figure. The gain shrinks but holds: weak judges gain, strong judges do not (Appendix I.3, Table 24).

On MT-Bench, LLaMA-3B ranks six chat systems in reverse human order ($\rho_{\mathcal{T}} = -0.49$). A 20-item $Q^\star$ with the same prompt makes it perfect ($\rho_{Q^\star} = +1.00$). $Q^\star$ keeps items where humans agree and weak systems fail plainly, which flips six system pairs (Appendix J.3). The new ranking also matches LMSYS Arena Elo exactly, which the search never saw.

Comparison with compression baselines. $Q^\star$ beats IRT-style compression on both human agreement and efficiency (Table 3). On our 23-cell grid, mean HAG is $+0.46$ (LLaMA-3B) and $+0.14$ (Qwen-7B), and IRT/gp-IRT beats ours on no cell. $Q^\star$ cuts tinyBenchmarkRep from 2.08 to 1.65. Under $60/40$ held-out systems, Sieve extends the lead (HAG-L = $+0.57$, HAG-Q = $+0.33$) while all four compression baselines stay near zero or negative. tinyBenchmarks aims to *preserve* a source ranking. Sieve can also *improve* an imperfect judge.

Task	**No system split** (100% systems, upper bound) LLaMA-3B	Qwen-7B	Qwen-14B	Luna	**60/40 system split** (deployment, in-samp. / held-out per judge) LLaMA-3B	Qwen-7B	Qwen-14B	Luna
SummEval	$+0.79$	$+0.11$	$+0.08$	$+0.02$	$+0.61/-0.15$	$+0.29/+0.16$	$+0.14/+0.11$	$+0.00/+0.04$
LMSYS Arena	$+0.87$	$+0.55$	$+0.45$	$+0.28$	n/a†	n/a†	n/a†	n/a†
MT-Bench	$+0.87$	$+0.15$	$+0.27$	$+0.06$	$+0.88/+1.28$	$+0.16/+0.56$	$+0.20/+0.60$	$+0.00/+0.00$
JudgeLM	$+1.13$	$+0.06$	-0.14	-0.08	$+0.85/+1.07$	$-0.03/-0.03$	$-0.03/+0.03$	$+0.01/-0.01$
PromptBench	$+0.72$	$+0.05$	$+0.05$	$+0.07$	$+0.85/+0.05$	$+0.26/-0.20$	$+0.12/-0.25$	$+0.37/-0.11$
UltraFeedback	$+0.67$	$+0.05$	$+0.34$	$+0.03$	$+0.63/+0.72$	$+0.35/+0.06$	$+0.97/+0.61$	$+0.24/-0.21$
AlpacaEval	-0.08	$+0.03$	$+0.12$	$+0.30$	$+0.08/-0.21$	$+0.16/+0.17$	$+0.21/+0.23$	$+0.46/+0.21$
HELM Lite	-0.06	-0.02	-0.01	-0.02	$+0.04/+0.03$	$-0.01/-0.04$	$+0.00/-0.01$	$-0.01/-0.02$
MMLU-Pro	$+0.46$	$+0.52$	$+0.27$	$+0.03$	$+0.86/+0.46$	$+0.56/+0.56$	$+0.40/+0.54$	$+0.04/-0.04$
Nectar	$+1.20$	$+0.60$	$+0.51$	$+0.21$	$+1.56/+1.24$	$+1.16/+1.12$	$+1.16/+1.08$	$+0.32/+0.36$
HumanEval	$+0.18$	-0.01	-0.00	-0.01	$+0.38/+0.00$	$-0.00/-0.00$	$+0.01/+0.03$	$-0.00/-0.01$
Mean	$\mathbf{+0.59}$	$\mathbf{+0.15}$	$\mathbf{+0.15}$	$\mathbf{+0.06}$	$\mathbf{+0.67/+0.45}$	$\mathbf{+0.29/+0.23}$	$\mathbf{+0.32/+0.30}$	$\mathbf{+0.14/+0.02}$

Table 2: Human-agreement gain (HAG) by task and judge. Means over 5 seeds (42–46), $|Q^\star|$=20. *Left:* all systems. *Right:* systems the search saw / never saw (60/40). The held-out number is what a practitioner gets. $|\mathrm{HAG}| \leq 0.06$ is noise. †LMSYS Arena has no split: its Elo ratings cannot be divided by system.

Method (estimator)	**No system split** *(100% systems, upper bound)* HAG-L↑	HAG-Q↑	**60/40 system split** *(deployment, held-out)* HAG-L↑	HAG-Q↑	beats ours	**MMLU compression** Rep↓	Rank τ↑
Our subset search	$\mathbf{+0.46}$	$\mathbf{+0.14}$	$\mathbf{+0.57}$	$\mathbf{+0.33}$	ref.	**1.65**	**0.864**
IRT / tinyBench, gp-IRT	$+0.044$	-0.039	$+0.14$	$+0.04$	0/23, 0/23	2.08	0.833
k-center, embed. ridge	-0.13	-0.11	-0.12	-0.15	0/23, 2/23	2.27	0.828
Random, mean	-0.14	-0.11	-0.14	-0.10	1/23, 6/23	3.31	0.774
Ours, entropy+ridge	$+0.14$	$+0.20$	-0.12	-0.31	4/23, 16/23	2.20	0.857

Table 3: Human agreement vs. compression baselines. HAG on the 23-cell open-judge grid (no-split and 60/40 held-out). MMLU compression: k=100, 395 models, gp-IRT fit-then-test. Rank τ is Kendall's.

Transfer. $Q^\star$ transfers without re-running the search. Selected with one judge, it preserves the system ranking for other LLaMA and Qwen judges (Spearman 0.83–1.00) and raises GPT-5's item-level accuracy on Amazon-ESCI by $+27.3$ and $+29.3$ (Appendix H.2).

Cost. Each cell takes 3,000 black-box judge calls, a few minutes on one consumer GPU (Appendix D.4). Deployed, $Q^\star$ has 20 items against up to $\sim$5,000 in the source, making it 250$\times$ cheaper to evaluate, at a mean deployment gain of $+0.16$ HAG.

Subset size. Any k from 10 to 100 gives 77–82% accuracy (Fig. 4, Table 13), and weak judges peak by k=5 (Fig. 3). k=20 is a compute pin within the plateau, not a threshold.

5.3 Verifying the Headroom Law

The fit. On the 36-cell open-judge grid, fitting $\mathrm{HAG} = \alpha(1 - \rho_{\mathcal{T}}) + \beta$ gives $\alpha = 0.86$ (95% CI $[0.75, 0.93]$, Fig. 2). AIC and BIC both prefer the linear form, and the $\pm 1.96\hat{\sigma}$ band covers 94% of cells (Appendix J.5). Per-judge slopes span $[0.65, 0.94]$, and a 149-cell multi-seed refit on the same three open judges gives $\alpha = 0.84$ with 95% CI $[0.75, 0.91]$ (Appendix J.4, Table 27).

α belongs to the objectives, not the optimiser. Seven different selectors give $\alpha \in [0.87, 1.01]$, $R^2 \geq 0.79$ (Table 4). A larger sweep agrees (Appendix J.1).

It predicts unseen cells. With the fit frozen, we ran Sieve for the first time on 66 held-out (judge, task, seed) cells. Predicted and observed HAG agree at $r = 0.935$, with 64/66 cells inside the band (Table 5), and all 14 near-ceiling cells stay within the noise band.

Falsification and mechanism. The Law can fail, and one cell nearly does. LLaMA-3B $\times$ MMLU-Pro (h=1.19) should gain $+0.91$, but SimpleSearch reaches only $+0.08$. The other six selectors all exceed $+0.3$ there (Appendix J.1). The Law describes the best gain across selectors, so only a cell where every selector fails would falsify it. We found none. Why $\alpha \approx 0.86$? If every item carries the same irreducible noise ν, a rate-distortion sketch gives $\alpha \approx 1 - \nu$, so $\nu \approx 0.13$ (Appendix J.6). Measuring ν independently on a new judge family would test this, but we have not yet done so.

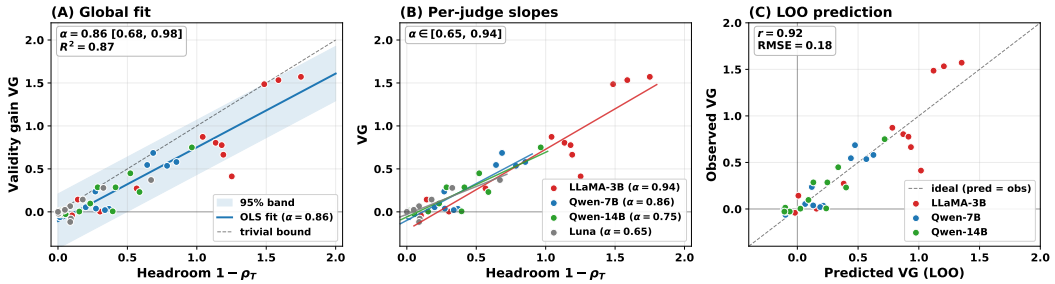


Figure 2: The Headroom Law, three views. *(A)* Global OLS fit on 36 open-judge cells (dashed line, trivial bound $\text{HAG} \leq 1 - \rho_{\mathcal{T}}$, Corollary 1. Shaded band is the 95% prediction interval). *(B)* Per-judge fits. *(C)* Leave-one-out prediction. Numeric fit constants are in §5.3.

Selector	n	α	β	R^2	mean HAG (No system split) *(100% systems, upper bound)*	mean HAG (60/40 system split) *(deployment, held-out)*
SimpleSearch (default)	22	0.868	−0.149	0.790	+0.450	+0.50
CoDeepNEAT	22	0.924	−0.171	0.873	+0.467	+0.52
MAP-Elites	22	0.946	−0.145	0.918	+0.508	+0.48
HyperNEAT	22	0.949	−0.131	0.935	+0.524	+0.59
Reg. Evolution	22	0.957	−0.153	0.921	+0.508	+0.50
WANN	22	1.008	−0.117	0.949	+0.579	+0.61
CMA-ES-embed	22	0.949	−0.215	0.907	+0.440	+0.32

Table 4: Per-selector Headroom Law fits. Left block (n=22, α, β, R^2): shared 22-cell grid at seed 42, budget 3,000, $|Q^\star|$=20, fitting $\text{HAG} = \alpha(1-\rho_{\mathcal{T}}) + \beta$. Two rightmost columns: mean HAG over LLaMA-3B and Qwen-7B on the extended sweep, under the no-split (all-systems) and 60/40 held-out protocols respectively. Cells per selector on the 60/40 sweep: SimpleSearch 61, CoDeepNEAT 61, MAP-Elites 75, HyperNEAT 17, Reg. Evolution 72, WANN 62, CMA-ES-embed 12. Full breakdown in Appendix J.1, Table 25.

Cell	Seed	$\rho_{\mathcal{T}}$	h=1−$\rho_{\mathcal{T}}$	Regime	Pred HAG	Obs HAG	in 95%
LLaMA-3B × Nectar	49	−0.71	1.71	weak	+1.36	+1.39	1/1
Qwen-7B × MT-Bench	49	+0.43	0.57	mid	+0.38	+0.57	1/1
Qwen-14B × HumanEval	49	+0.99	0.01	near-ceiling	−0.10	−0.02	1/1
LLaMA-3B × 11 tasks (n=11)	49	+0.00	1.00	weak	+0.75	+0.70	10/11
LLaMA-3B × 11 tasks (n=11)	50	+0.00	1.00	weak	+0.75	+0.76	10/11
Qwen-7B × 11 tasks (n=11)	49	+0.63	0.37	mid	+0.21	+0.22	11/11
Qwen-7B × 11 tasks (n=11)	50	+0.61	0.39	mid	+0.23	+0.25	11/11
Qwen-14B × 11 tasks (n=11)	49	+0.65	0.35	near-ceiling	+0.19	+0.24	11/11
Qwen-14B × 11 tasks (n=11)	50	+0.67	0.33	near-ceiling	+0.17	+0.20	11/11
Aggregate (n=66). $r = 0.935$, RMSE $= 0.161$, band coverage 64/66 (97.0%).							

Table 5: Predicting HAG on 66 new cells (3 judges × 11 tasks × unseen seeds 49, 50). Each cell's gain is predicted from $\rho_{\mathcal{T}}$ alone via $\text{HAG} = 0.86(1 - \rho_{\mathcal{T}}) - 0.11$, then observed by running Sieve. Rows 1–3 span weak to near-ceiling. Qwen-14B × HumanEval is a **negative control** (no gain predicted). Rows 4–9 average 11 tasks per (judge, seed). Band: ± 0.35 ($1.96\hat{\sigma}$, Appendix J.5).

6 Conclusion

A benchmark does not need to be large to agree with humans, only informative. Across 36 general-task cells, a learned 20-item subset improved human agreement over 80–5,000-item benchmarks with mean $\text{HAG} = +0.37$ ($p = 2.7 \times 10^{-6}$), and the gain is strongly predicted by headroom ($r = -0.93$), as formalised by $\text{HAG} \leq 1 - \rho_{\mathcal{T}}$. Evaluation should therefore learn what to judge: a small, deliberately selected test can agree with humans more closely than the benchmark it replaces.

7 Limitations

Sieve has limits. A subset chosen for one task must be chosen again for a new one, and a judge that already agrees with humans has nothing left to gain. Human rankings are themselves estimates, so very small gains on a single task may be noise. The gain should also be measured on systems the search never saw, since it can look very different on those it did. Finally, because many of our tests share tasks, the 0.86 figure is reliable in direction but not to two decimal places.

8 USE OF LARGE LANGUAGE MODELS

An LLM was used for drafting prose and for running experiments. All ideas, algorithms, design decisions, analyses, and interpretations are the authors' own, and the authors take full responsibility for the content of this paper.

9 ETHICS STATEMENT

Sieve reduces the compute burden of LLM-as-Judge evaluation and thereby energy use. A compact learned $Q^\star$ that ranks judges correctly on aggregate may still be less faithful on rare subgroups, so we recommend it as a triage tool paired with human oversight, not a replacement for full evaluation in high-stakes settings.

10 REPRODUCIBILITY STATEMENT

Code, hyperparameters and seeds will be released on acceptance. The subset-selection procedure is specified in Section 3.4 and Algorithm 1. The frozen-prompt protocol and the item/system 60/40 split are in Section 3.2, complete hyperparameters in Appendix D.3. The 20-item $Q^\star$ subsets for all eleven benchmarks are attached as supplementary material (`supplementary/sieve_subsets.zip`, per-benchmark JSON schema and README).

11 AUTHOR CONTRIBUTIONS

The authors carried out ideation, algorithm design, conceptualization, methodology, software, formal analysis, and investigation. The authors have read and agreed to the published version of the manuscript.

12 FUNDING

This research received no external funding.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: a regret minimization approach. *Mathematical Programming*, 186:439–478, 2021.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M. Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13787–13805, 2024. URL https://aclanthology.org/2024.acl-long.744/.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. Non-determinism of "deterministic" LLM settings. In *Proceedings of the Eval4NLP Workshop at EMNLP*, 2025.

Anthony C. Atkinson, Alexander N. Donev, and Randall D. Tobias. *Optimum Experimental Designs, with SAS*. Oxford University Press, 2007.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. URL https://aclanthology.org/Q19-1004/.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick (eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

Elizabeth L. Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World*. Worth Publishers, 2011.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4843–4855, 2021. URL https://aclanthology.org/2021.naacl-main.385/.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8301–8327, 2024. URL https://aclanthology.org/2024.emnlp-main.474/.

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

Yan Chen, Shujian Liu, Zheng Liu, Weiyi Sun, Linas Baltrunas, and Benjamin Schroeder. WANDS: Dataset for product search relevance assessment. In *European Conference on Information Retrieval (ECIR)*, 2022.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas 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 *ICML*, 2024.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024a. arXiv:2404.04475.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://arxiv.org/abs/2305.14387.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, 2000.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4846–4853, 2020. URL https://aclanthology.org/2020.emnlp-main.393/.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Adam Gaier and David Ha. Weight agnostic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.04358.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 107–112, 2018. URL https://aclanthology.org/N18-2017/.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for ai evaluation? *arXiv preprint arXiv:2607.15190*, 2026a.

Han Jiang, Susu Zhang, Dongyao Zhu, Yuzhuo Bai, Sang T. Truong, Xiaoyuan Yi, Sanmi Koyejo, Xing Xie, and Ziang Xiao. Ai evaluation should require standardized item-level data releases. *arXiv preprint arXiv:2604.03244*, 2026b.

Michael T. Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 2013.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4110–4124, 2021. URL https://aclanthology.org/2021.naacl-main.324/.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *EMNLP*, 2024.

Alex Kipnis, Konstantinos Voudouris, Luca M. Schulze Buschoff, and Eric Schulz. metabench: A sparse benchmark to measure general ability in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.12844.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of NAACL*, pp. 1755–1797, 2025.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.05470 (Auto-J).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *EMNLP*, 2023.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2403.16950.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, 1980.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3428–3448, 2019. URL https://aclanthology.org/P19-1334/.

Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, and Babak Hodjat. Evolving deep neural networks. In *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, pp. 293–312, 2019. URL https://arxiv.org/abs/1703.00548.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. URL https://aclanthology.org/2024.tacl-1.52/.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015. URL https://arxiv.org/abs/1504.04909.

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.

Aleksandar Nikolov, Mohit Singh, and Uthaipon (Tao) Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. Do these LLM benchmarks agree? fixing benchmark evaluation with BenchBench. *arXiv preprint arXiv:2407.13696*, 2024. URL https://arxiv.org/abs/2407.13696.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 180–191, 2018. URL https://aclanthology.org/S18-2023/.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *ICML*, 2024.

Chao Qian, Yang Yu, and Zhi-Hua Zhou. Subset selection by pareto optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-Judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2111.15366.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://arxiv.org/abs/1802.01548.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale esci benchmark for improving product search. In *KDD Cup*, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4902–4912, 2020. URL https://aclanthology.org/2020.acl-main.442/. Best Paper Award.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 41–58, 2021. URL https://aclanthology.org/2021.acl-long.4/.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787, 2023. URL https://aclanthology.org/2023.findings-emnlp.722/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11324.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Kenneth O. Stanley, David B. D'Ambrosio, and Jason Gauci. A hypercube-based encoding for evolving large-scale neural networks. *Artificial Life*, 15(2):185–212, 2009. URL https://direct.mit.edu/artl/article/15/2/185/2617.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv preprint arXiv:2404.18796, 2024.

Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. arXiv preprint arXiv:2408.02666, 2024a.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *International Conference on Learning Representations (ICLR)*, 2024b. arXiv:2306.05087.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024c.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of LLM-as-a-Judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. In *ICLR Workshop on Building Trust in LLMs and LLM Applications*, 2025.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024. URL https://arxiv.org/abs/2406.04244.

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*, 2024.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *Conference on Language Modeling (COLM)*, 2024.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations (ICLR, Spotlight)*, 2025. arXiv:2310.17631.

Appendix

APPENDIX CONTENTS

The appendix is organised so a reader can navigate by scientific question, from formal grounding through experimental protocol to ablations, transfer, and diagnostics. Section headings, subsection headings, and every table/figure are listed with their page.

A Notation and Formal Definitions . . . . . p. 17
Table 6: Notation used in this paper . . . . . p. 17
B Theory and Proofs . . . . . p. 18
B.1 Formal framework . . . . . p. 18
B.2 Sample efficiency: a small exam suffices p. 19
B.3 Positioning of the Q-optimizer . . . . . p. 21
C Subset Search: Algorithm and Details . . . . . p. 21
C.1 Subset-selection algorithm . . . . . p. 21
C.2 Budget $\times$ performance tradeoff curve . . . p. 22
C.3 Sieve hyperparameters . . . . . p. 22
C.4 Computational cost . . . . . p. 24
Figure 3: Budget $\times$ performance tradeoff . . p. 22
Table 7: Sieve hyperparameters . . . . . p. 23
D Experimental Setup and Reproducibility . . p. 24
E Subset-Objective Ablations . . . . . p. 25
E.1 A-only ablation: dropping $\mathcal{C}$ . . . . . p. 25
E.2 Full 7-configuration Pareto sweep on ESCI p. 26
E.3 Subset-size sweep . . . . . p. 26
E.4 Subset-size ablation . . . . . p. 26
E.5 Algorithm-agnostic verification (subset-selector swap) . . . . . p. 26
Table 8: $\mathcal{M}$+$\mathcal{C}$ vs. $\mathcal{A}$+$\mathcal{C}$ fronts . . . . . p. 25
Table 9: Sanity check without $\mathcal{C}$ . . . . . p. 25
Table 10: $\mathcal{C}$ guards $\mathcal{M}$ . . . . . p. 26
Table 11: Component analysis (agnostic panels) p. 27
Figure 4: Compression is free . . . . . p. 26
F Additional Benchmark Results . . . . . p. 28
F.1 Bootstrap CIs on subset accuracy gain . . . p. 28
F.2 7-configuration sweep on public benchmarks p. 29
Table 14: Subset accuracy CIs, public benchmarks p. 29
Table 15: 7-config sweep, public benchmarks p. 29
G Transfer, Scaling, and Sensitivity . . . . . p. 30
G.1 Scaling: Qwen-2.5-14B . . . . . p. 30
G.2 Cross-judge transfer of $Q^\star$ . . . . . p. 30
G.4 Transfer of $Q^\star$ to frontier judges . . . . . p. 30
Table 17: Scaling: Qwen-2.5-14B . . . . . p. 30
Table 18: Cross-judge & frontier transfer . . p. 30
G.8 Reproduction of LOST tables: $|V|/|Q|$ sweeps p. 30
Table 19: Component-D, $|V|$ sweep, regenerated p. 31
Table 20: Component-D, $|Q|$ sweep, regenerated p. 31
H Mechanistic and Diagnostic Analyses . . . . . p. 31
H.1 Bootstrap CIs on subset accuracy gain . . . p. 31
H.5 Error-pattern preservation . . . . . p. 31
H.6 Fair-test $\mathcal{H}$ split: per-cell breakdown (200 cells) p. 31
Table 21: Bootstrap CIs on subset accuracy Δ p. 31
Table 22: Fair-test $\mathcal{H}$ split: per-judge summary p. 32
Table 24: Fair-test $\mathcal{H}$ split: per-judge, per-dataset means $\pm$ std . . . . . p. 33
I Additional Ablations and Comparisons . . . . p. 32
I.3 Algorithm-independence of the VG Law . p. 32
J Worked Examples . . . . . p. 38
Table 28: Learned subset samples . . . . . p. 38
K Provenance and Supplementary Material . . p. 38
K.1 Provenance markers used . . . . . p. 38

Symbol	Description
(a) Primitives: what the search sees.	
Q	The size-k subset drawn from the labelled pool $\mathcal{P}$; the single search variable.
$\mathcal{P}, \mathcal{V}, \mathcal{T}$	The pool from which Q is drawn, the validation reference on which every search-time objective is scored, and the locked test read only for final reporting.
J	The base judge, treated as a black box that returns an integer score.
$\mathcal{H}$	The independent human ranking of the systems in $\mathcal{S}$, held out of the search and consulted only for human agreement.
(b) Outcome quantities: what we report.	
$\mathrm{HAG}(Q^\star)$	The human-agreement-gain, $\rho_{Q^\star} - \rho_{\mathcal{T}}$, our headline metric.
$\rho_{\mathcal{T}}$	The Spearman agreement of the default $\mathcal{T}$ with $\mathcal{H}$, our baseline.
$\rho_{Q^\star}$	The Spearman agreement of the learned pair $Q^\star$ with $\mathcal{H}$, our target.
ε	The finite-sample noise band on HAG. Bootstrap estimates put it at $\varepsilon \leq 0.07$ on our grid.
(c) Subset objectives: what shapes Q during search.	
$\mathcal{M}(Q)$	Accuracy match: how closely Q's accuracy matches $\mathrm{Acc}(\mathcal{V})$ from below.
$\mathcal{C}(Q)$	Ranking correlation: how faithfully the panel-accuracy ranking induced by Q tracks the ranking induced by $\mathcal{V}$.
$\mathcal{A}(Q)$	Adversariality: the agreement-gated mean judge error on Q, so hard items on which humans agree are rewarded.
$\mathrm{Acc}(S)$	The rescaled mean-absolute agreement between J and human labels on set S, defined in Eq. 9.
Δ	The rating range used to normalise Acc. On our 1 to 5 scale, Δ=4.
(d) System panel and per-system quantities.	
$\mathcal{S}$, $\lvert\mathcal{S}\rvert$	The set of systems ranked in a cell, and its size. $\mathcal{H}$ orders these.
S_i, s_i	System i in $\mathcal{S}$ and its aggregate judge score J averaged over the items in $S \in \{Q, \mathcal{T}\}$.
R_Q	The ranking of $\mathcal{S}$ induced by the judge under the default prompt scored on subset Q.
k	The subset budget, $k = \lvert Q^\star \rvert$. We use k=20 throughout the main grid.
ρ, τ	Spearman and Kendall rank correlations.

Table 6: Notation used in this paper, grouped by the role each symbol plays in the method: (a) primitives the search sees, (b) human-agreement-gain outcome quantities we report, (c) the three subset objectives, (d) the system panel.

A How to Sieve your own benchmark

(i) Score. Score every candidate item in your source benchmark with the default judge prompt to obtain per-system item scores. *(ii) Pareto search.* Compute the two objectives $(\mathcal{M}, \mathcal{C})$ from Equations 10–11 on candidate subsets drawn from the validation pool $\mathcal{V}$, and run any non-dominated selector for a fixed budget of $\sim$3000 judge calls (Section 3.4). *(iii) Deploy.* Take the Pareto knee $Q^\star$, freeze it, and evaluate any future judge on its 20 items only. Before running, use the Headroom Law (§5.3) to predict whether a gain is even possible from $\rho_{\mathcal{T}}$ alone: high-headroom cells ($\rho_{\mathcal{T}} < 0.6$) are the ones where a Sieved subset is likely to be worth deploying. Reference implementation and the eleven $Q^\star$ subsets used in this paper are attached as supplementary material.

B Notation and Formal Definitions

Table 6 groups the symbols used in this paper by the role they play in our method, and the three subset objectives are stated in full below.

B.1 Full objective definitions

This appendix states the three subset objectives in full. The compact form is in Section 3.4. All share one base signal: agreement between the judge's scores and human ratings on an item set $S \in \{Q, \mathcal{V}, \mathcal{T}\}$, via the rescaled mean absolute error $\mathrm{Acc}(S)$ of Eq. 9, with $J(x)$ the judge's score under the default prompt, $y(x)$ the human rating, and Δ the rating range (Δ=4 for a 1–5 scale). This scalar feeds the three subset objectives, all truncated to $[0, 1]$.

Accuracy match ($\mathcal{M}$). Q should not underestimate the accuracy on $\mathcal{V}$. $\mathcal{M}$ is the one-sided *accuracy match* between Q and $\mathcal{V}$:

$$\mathcal{M}(Q) = 1 - \max\big(0,\ \mathrm{Acc}(\mathcal{V}) - \mathrm{Acc}(Q)\big). \tag{12}$$

Both terms use the same judge and prompt. The one-sided penalty targets the pathological direction (a Q on which the judge does substantially worse than on $\mathcal{V}$) while leaving Q free to be locally easier than $\mathcal{V}$ without demoting the prompt. The symmetric $1 - |\mathrm{Acc}(Q) - \mathrm{Acc}(\mathcal{V})|$ has a joint-optimisation Goodhart mode (prompts that depress both accuracies in lockstep score 1); the one-sided form removes it.

Ranking correlation ($\mathcal{C}$). A good subset separates a panel of reference judges the way $\mathcal{V}$ does. Instantiating the base judge under $m{=}3$ reference prompts of deliberately different strength gives panel-accuracy vectors $\mathbf{a}_Q = [\mathrm{Acc}(Q; p_k)]_{k=1}^3$ and $\mathbf{a}_\mathcal{V} = [\mathrm{Acc}(\mathcal{V}; p_k)]_{k=1}^3$. For a vector $\mathbf{a} = (a_1, \ldots, a_n)$, the (average) rank of its i-th entry is

$$\mathrm{rank}(\mathbf{a})_i \ = \ \big|\{j :\ a_j < a_i\}\big| \ + \ \tfrac{1}{2}\big(1 + \big|\{j :\ a_j = a_i\}\big|\big)\,, \tag{13}$$

where $|\{j : a_j = a_i\}|$ counts i itself. Distinct entries receive positions $1, \ldots, n$ and tied entries share the mean of the positions they would otherwise occupy (e.g., ranks of $(5, 3, 8, 3)$ are $(3,\ 1.5,\ 4,\ 1.5)$). Spearman rank correlation is Pearson correlation applied to the ranks,

$$\rho(\mathbf{a}_Q, \mathbf{a}_\mathcal{V}) \ = \ \frac{\mathrm{cov}\big(\mathrm{rank}(\mathbf{a}_Q),\ \mathrm{rank}(\mathbf{a}_\mathcal{V})\big)}{\mathrm{std}\big(\mathrm{rank}(\mathbf{a}_Q)\big)\ \mathrm{std}\big(\mathrm{rank}(\mathbf{a}_\mathcal{V})\big)} \ \in \ [-1,\ 1], \tag{14}$$

with $\rho = 0$ when either denominator vanishes. $\mathcal{C}$ asks whether Q ranks the panel as $\mathcal{V}$ does *and* preserves its spread:

$$\mathcal{C}(Q) = \mathrm{clip}\Big(\tfrac{1}{2}\big(\rho(\mathbf{a}_Q, \mathbf{a}_\mathcal{V}) + 1\big) \cdot \min\Big(1,\ \tfrac{\mathrm{range}(\mathbf{a}_Q)}{\mathrm{range}(\mathbf{a}_\mathcal{V})}\Big)\ ,\ 0,\ 1\Big)\,, \tag{15}$$

with ρ Spearman's rank correlation and $\mathrm{range}(\mathbf{a}) = \max(\mathbf{a}) - \min(\mathbf{a})$; if $\mathrm{range}(\mathbf{a}_\mathcal{V}){=}0$ we set $\mathcal{C}(Q){=}0.5$. The alignment term is the IRT item-discrimination criterion (Birnbaum, 1968, Lord, 1980); on a small $m{=}3$ panel it saturates to 1 under any Q drawn from the same pool as $\mathcal{V}$, so the spread factor is what supplies the actual gradient to the optimiser.

Adversariality ($\mathcal{A}$). For diagnostic use Q should target hard, failure-revealing items: the mean per-item difficulty when annotated, else the agreement-gated mean judge error

$$\mathcal{A}(Q) = \min\Big(1,\ \tfrac{1}{\Delta} \underset{\{x \in Q:\ \mathrm{agree}(x) \geq \tau\}}{\mathrm{mean}} \big|J(x) - y(x)\big|\Big), \tag{16}$$

where the gate $\mathrm{agree}(x) \geq \tau$ keeps only items humans agree on, so Q mines genuine judge weaknesses rather than label noise.

Why Q's own accuracy is not an objective. Rewarding Q for its own accuracy would let the search shed items until only the easy ones survived. Instead Q is scored by how well it *tracks* the reference ($\mathcal{M}$, $\mathcal{C}$) and how usefully hard its retained items are ($\mathcal{A}$). All search-time signal is read from $\mathcal{V}$, and $\mathcal{T}$ is consulted only after search closes.

C Theory and Proofs

This appendix gives the formal counterpart of the informal "why it works" paragraph in Section 3.4.

C.1 Formal framework

Definition 2 (Judge game)**.** *A judge game is a tuple $\big(\mathcal{X}, \mathcal{D}, y, J, \mathcal{C}\big)$ where $\mathcal{X}$ is a space of questions, $\mathcal{D}$ a distribution over $\mathcal{X}$ (the domain), $y : \mathcal{X} \to [0, 1]$ the ground-truth grade, J a base judge (an LLM), and $\mathcal{C}$ a class of students (classifiers). The judge is configured by a fixed default instruction and an exam $Q \subseteq \mathcal{X}$. For a student $c \in \mathcal{C}$ and question x, the judge returns a correctness score $\phi_c(x) \in [0, 1]$ obtained by grading c's response to x.*

For a student c, write its population (full-exam) score and its measured score on the exam Q as

$$S(c) = \mathbb{E}_{x\sim\mathcal{D}}\big[\phi_c(x)\big],$$
$$\widehat{S}_Q(c) = \frac{1}{|Q|}\sum_{x\in Q}\phi_c(x). \tag{17}$$

The judge's own *quality* on a set S is the human agreement $\mathrm{Acc}(S)$ of Eq. (9). Tuning Q targets the instrument properties below.

Definition 3 (ϵ-faithful exam)**.** *An exam Q is* ϵ-faithful *for student class $\mathcal{C}$ (w.r.t. $\mathcal{D}$) if* $\sup_{c\in\mathcal{C}}\big|\widehat{S}_Q(c) - S(c)\big| \le \epsilon$.

Definition 4 (γ-separated students)**.** *$\mathcal{C}$ is* γ-separated *under $\mathcal{D}$ if $|S(c) - S(c')| \ge \gamma$ for every $c \ne c' \in \mathcal{C}$.*

The three tuning objectives (Eqs. 12, 15) are the empirical surrogates of these population quantities: *accuracy match* ($\mathcal{M}$) targets Definition 3. *ranking correlation* ($\mathcal{C}$) targets order preservation under Definition 4 and equals, in the Bayesian-design view, the information the exam carries about student quality (Lindley, 1956, MacKay, 1992, Chaloner & Verdinelli, 1995, Cover & Thomas, 2006). *difficulty* ($\mathcal{A}$) places items where item information peaks, i.e. at the student's ability frontier, following Item Response Theory (Birnbaum, 1968, Lord, 1980, Embretson & Reise, 2000) and the pedagogy of desirable difficulty (Vygotsky, 1978, Bjork & Bjork, 2011). Selecting Q to maximize exam information is a D-optimal experimental design (Atkinson et al., 2007).

C.2 Sample efficiency: a small exam suffices

Assumption 1 (Bounded class complexity)**.** *The score family $\mathcal{F} = \{\phi_c : c \in \mathcal{C}\}$ takes values in $[0,1]$ and has pseudo-dimension $\mathrm{Pdim}(\mathcal{F}) = d < \infty$ (or $\mathcal{C}$ is finite with $|\mathcal{C}| = K$).*

Theorem 1 (Sample efficiency / accuracy match)**.** *Under Assumption 1, let Q consist of m i.i.d. draws from $\mathcal{D}$. For any $\epsilon, \delta \in (0,1)$, Q is ϵ-faithful with probability at least $1-\delta$ provided*

$$m \;\ge\; \frac{C}{\epsilon^2}\Big(d\,\ln\tfrac{1}{\epsilon} + \ln\tfrac{1}{\delta}\Big), \tag{18}$$

for a universal constant C (for a finite class, $m \ge \frac{1}{2\epsilon^2}\ln\frac{2K}{\delta}$ suffices). In particular, m is independent of the full-corpus size $|\mathcal{T}|$. This is a worst-case, distribution-free sufficiency *guarantee: its constant is deliberately loose (it predicts m of order 10^3 for a 1 to 6 pt gap), whereas empirically a far smaller learned exam ($|Q^\star|$=30 to 50) already attains single-digit reproduction gaps (Table 1). The operative, and empirically borne-out, prediction is therefore the* corpus-size independence *of m, not its numerical value: one fixed small exam reproduces full-test accuracy across corpora ranging from a few hundred to several thousand items.*

Proof. Fix c. The variables $\phi_c(x_i) \in [0,1]$ are i.i.d. with mean $S(c)$, so Hoeffding's inequality gives $\Pr[\,|\widehat{S}_Q(c) - S(c)| > \epsilon\,] \le 2e^{-2m\epsilon^2}$. For a finite class, a union bound over the K students and setting $2Ke^{-2m\epsilon^2} \le \delta$ yields the stated m. For an infinite class, the event $\sup_c|\widehat{S}_Q(c) - S(c)| > \epsilon$ is controlled by the uniform-convergence bound for bounded function classes of pseudo-dimension d (Vapnik, 1998, Shalev-Shwartz & Ben-David, 2014), giving Eq. (18). Because the right-hand side does not depend on $|\mathcal{T}|$, a fixed-size exam represents arbitrarily large corpora. □

Since the learned $Q^\star$ is *selected* to minimise the empirical reproduction gap against the validation reference $\mathcal{V}$ (objective A), it is, in expectation, no worse than an equally sized random exam. Theorem 1 is therefore an upper bound on the exam size that suffices, and selection can only reduce it.

The next result shows this size is essentially unavoidable: no exam can be much smaller in the worst case.

Proposition 1 (Lower bound)**.** *There is a student class of pseudo-dimension d and a distribution $\mathcal{D}$ such that any ϵ-faithful exam (Definition 3) with confidence $1-\delta$ must contain $m = \Omega\big(\frac{d+\ln(1/\delta)}{\epsilon^2}\big)$ questions.*

Proof. Estimating a single Bernoulli mean to additive error ϵ with confidence $1-\delta$ requires $\Omega(\epsilon^{-2}\ln(1/\delta))$ samples (binomial tail / Le Cam two-point method). Embedding d independent such mean-estimation problems, one per coordinate of a class shattering d points, forces $\Omega(d/\epsilon^2)$ questions for simultaneous ϵ-accuracy (Vapnik, 1998, Shalev-Shwartz & Ben-David, 2014). Combining the two gives the bound. □

Remark 1 (Cost-accuracy-match frontier). Reading Theorem 1 and Proposition 1 together, the best achievable accuracy match at exam budget m scales as $\epsilon(m) = \Theta(\sqrt{d/m})$. Evaluation cost is proportional to m, so the attainable (cost, accuracy match) pairs lie on a $1/\sqrt{m}$ frontier. The $\mathcal{M}$+$\mathcal{C}$ Pareto front of Table 8 is the empirical realization of this curve.

Corollary 2 (Ranking preservation / ranking correlation)**.** *If Q is ϵ-faithful and $\mathcal{C}$ is γ-separated with $\epsilon < \gamma/2$, then the ranking of $\mathcal{C}$ by $\widehat{S}_Q$ is identical to the ranking by S. Hence any rank correlation between the two equals* 1.

Proof. Take c, c' with $S(c) \geq S(c') + \gamma$. Then $\widehat{S}_Q(c) \geq S(c) - \epsilon \geq S(c') + \gamma - \epsilon > S(c') + \epsilon \geq \widehat{S}_Q(c')$, using $\epsilon < \gamma/2$. So strict order is preserved for every pair, and the induced rankings coincide. □

Corollary 2 explains the empirically observed rank correlation of 1.0 between $Q^\star$ and the full test set: once the exam is faithful below half the student-separation margin, it reproduces the full leaderboard exactly. This links objectives A and B: reproduction implies discrimination whenever students are separated.

Information-theoretic form of ranking correlation. Our Spearman-based objective B (Eq. (15)) is a model-free surrogate for a sharper, parametric quantity. Under a two-parameter IRT model, a student of ability θ answers item x correctly with probability $P_x(\theta) = \sigma\big(a_x(\theta - b_x)\big)$, where a_x is the item's discrimination and b_x its difficulty. The exam's *test information* at ability θ is

$$\mathcal{I}_Q(\theta) = \sum_{x \in Q} a_x^2 \, P_x(\theta)\big(1 - P_x(\theta)\big), \tag{19}$$

the Fisher information about θ, with standard error $\mathrm{SE}(\hat{\theta}) = \mathcal{I}_Q(\theta)^{-1/2}$ (Birnbaum, 1968, Lord, 1980). Maximising Eq. (19) is the D-optimal exam-design objective (Atkinson et al., 2007, Chaloner & Verdinelli, 1995): each summand peaks at $\theta = b_x$ (an item is most informative at the ability frontier, objective C), and the a_x^2 weight is the discrimination of objective B. Eq. (19) may be substituted for Eq. (15) when a calibrated item bank is available.

Computing $\bar{d}_Q$ in practice (correlation with system quality). The item-response bound in the main text (rank reliability $\propto \bar{d}_Q\sqrt{|Q|}/\sigma$, Section 3.4) uses $\bar{d}_Q$ as the mean per-item discrimination, but the paper does not fit a full 2PL model. When only aggregate judge scores are available, a model-free surrogate is the correlation of an item's per-system scores with an overall system-quality signal. Let $J(s,x)$ be the judge's score on item x for system s, and let $\mu(s) = |Q|^{-1}\sum_x J(s,x)$ be system s's aggregate score (or, when available, a human-derived quality). Then per-item discrimination is

$$d_x \;=\; \mathrm{corr}_s\big(J(s,x),\ \mu(s)\big), \qquad \bar{d}_Q \;=\; \tfrac{1}{|Q|}\sum_{x \in Q} d_x, \tag{20}$$

with correlation computed across the $|\mathcal{S}|$ evaluation systems using Pearson (or Spearman on ranks). A high d_x means item x's per-system scores align with the overall system ordering, so item x contributes signal in the same direction the target ranking requires. In code, $\mu \leftarrow J.\mathrm{mean}(\mathrm{axis}=1)$ and $d_x \leftarrow \mathrm{corrcoef}(J[:,x], \mu)[0,1]$. This is the definition the paper uses conceptually. It matches Definition 4 up to the choice of aggregation and is what the $\mathcal{C}$ objective approximates on the three-prompt panel without requiring a fitted latent-ability model.

Proposition 2 (Reliability)**.** *Fix a faithful exam Q ($\sup_c |\widehat{S}_Q(c) - S(c)| \leq \epsilon$). At judge temperature* 0 *the measured score $\widehat{S}_Q(c)$ is deterministic, so its test-retest difference is* 0 *(perfect reliability) and its deviation from the population score is at most ϵ (human agreement). At temperature $\tau > 0$, if per-question grades are independent with variance at most $s^2 \leq \frac{1}{4}$, then* $\mathrm{Var}\big[\widehat{S}_Q(c)\big] \leq s^2/|Q|$.

Proof. At temperature 0 each $\phi_c(x)$ is a fixed function of x, so $\widehat{S}_Q(c)$ is constant across repetitions. Faithfulness bounds its bias by ϵ. At temperature τ, $\widehat{S}_Q(c)$ averages $|Q|$ independent bounded grades, so its variance is at most $s^2/|Q|$ by independence. □

Proposition 2 formalises the paper's test-retest reliability claim: a faithful, deterministic exam is simultaneously valid (bias $\leq \epsilon$) and perfectly reliable, and any residual stochasticity shrinks as $1/|Q|$.

Remark 2 (Label noise and the agreement gate). The results above assume access to the true grade $y(x)$. With noisy annotations $\tilde{y}(x) = y(x) + \eta(x)$, write $\alpha(x) \in [0, 1]$ for the inter-annotator agreement on x (high α implies low $\mathbb{E}|\eta(x)|$). The adversarial objective ($\mathcal{A}$) gates on $\alpha(x) \geq \tau$, restricting selection to a low-noise sub-population $\mathcal{X}_\tau = \{x : \alpha(x) \geq \tau\}$ on which the effective label noise is bounded by some $\nu(\tau)$ that is non-increasing in τ. Re-running the proof on $\mathcal{X}_\tau$ inflates every error bound only additively by $\nu(\tau)$, so a stricter gate trades coverage for a tighter bound. This is what prevents objective C from collapsing onto mislabeled items: they are screened out before they can masquerade as "hard."

Summary. Theorem 1 justifies objective A (a corpus-size-independent exam reproduces any classifier's score) and Corollary 2 justifies objective B (faithful, well-separated exams preserve the leaderboard). These are the theoretical counterparts of the locked-$\mathcal{T}$ reproduction gap and rank correlation reported in Table 8.

C.3 Theoretical positioning of the Q-optimizer

Selecting Q is a discrete optimisation problem, and our objectives place it inside two well-studied families with established approximation results, so the selector is not merely a heuristic. *Coverage and entropy* objectives (balanced, diverse selection) are monotone submodular, for which the greedy algorithm is a $(1-1/e)$-approximation and this factor is optimal in the value-oracle model (Nemhauser et al., 1978). $\mathcal{C}$ (maximising the test/Fisher information that separates classifiers) is D-optimal experimental design, and $\mathcal{M}$ (minimising the variance of the score estimate) is A-optimal design. Both admit near-optimal approximation via proportional volume sampling (Nikolov et al., 2019) and regret minimisation (Allen-Zhu et al., 2021).

Crucially, the multi-objective evolutionary selector we use (NSGA-II) is an instance of the Pareto-optimisation-for-subset-selection family, for which Pareto optimisation provably attains the optimal $(1 - 1/e)$ approximation on monotone submodular subset selection (Qian et al., 2015). Our Q-search therefore connects to an approximation result rather than relying on pure heuristic search:

Proposition 3 (Approximation result for the Q-selector)**.** *When the subset objective is monotone submodular (e.g. the coverage/entropy term), the Pareto-optimisation selector returns, in expected polynomial time, a size-k subset whose objective value is at least $(1 - 1/e)$ times the optimum (Qian et al., 2015). For the A/D-optimal (reproduction/discrimination) objectives, proportional volume sampling yields a subset within a $(1 + \epsilon)$ factor of optimal in the regime $k = \Omega(d/\epsilon)$ (Nikolov et al., 2019).*

This grounds the empirical selectors (Eqs. 12, 15): the entropy/diversity selector is the submodular instance ($(1 - 1/e)$ via greedy or Pareto optimisation), and the $\mathcal{C}$ objective is the D-optimal-design instance, connecting our method to the optimal-design and submodular-optimisation literature while keeping the judge a black box.

D Subset Search: Algorithm and Details

D.1 Subset-selection algorithm

The subset-selection procedure referenced in Section 3.4 is stated below. It is a single Pareto pass over uniformly random 20-item candidates. There is no prompt search: the default prompt is held fixed throughout.

Eq. 21 states the subset-selection problem in closed form, subject to the size budget k:

$$Q^\star = \arg \max_{Q \subseteq \mathcal{P},\, |Q|=k} \big\{\, \beta\, \mathcal{M}(Q) + \gamma\, \mathcal{C}(Q) + \delta\, \mathcal{A}(Q) \,\big\}, \tag{21}$$

Algorithm 1: Learn $Q^\star$ under the fixed default prompt. Sample random candidate subsets, score each on three label-free objectives, and return the NSGA-II knee point.

Input: judge J; default prompt; *pool* $\mathcal{P}$; *validation* $\mathcal{V}$; *locked test* $\mathcal{T}$ (disjoint from $\mathcal{P}, \mathcal{V}$); subset size k; candidate count N.
Output: $Q^\star$; reproduction gap and rank ρ on $\mathcal{T}$.

Search $(\mathcal{P}, \mathcal{V}) \to Q^\star$**:**
  Sample N candidate subsets $\{Q^{(i)}\}_{i=1}^{N}$ uniformly at random from $\mathcal{P}$, each of size k
  for $i = 1$ **to** N **do**
    Score $Q^{(i)}$ on $\mathcal{V}$ under $(\mathcal{M}, \mathcal{C}, \mathcal{A})$ [Eqs. 12, 15, 16]
  NSGA-II non-dominated sort over the N candidates — no persistent population
  return front-0 knee point $Q^\star$
Inference $(Q^\star, \mathcal{T}) \to (\text{gap}, \rho)$**:**
  SCORE: $Q^\star$ on $\mathcal{T}$
  return (gap, ρ)

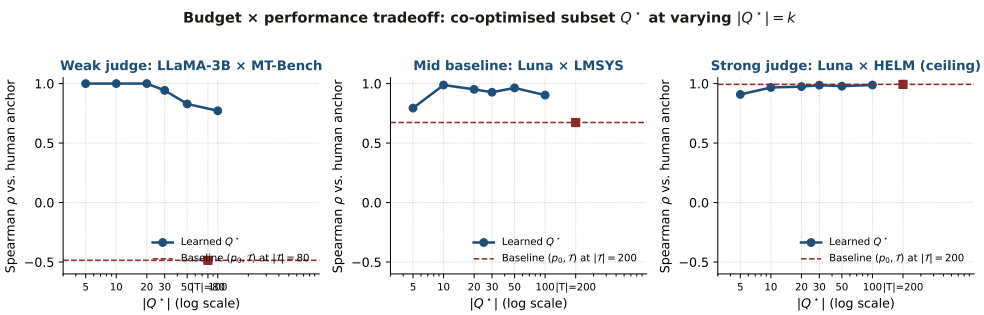


Figure 3: Budget × performance tradeoff. A learned $Q^\star$ matches or beats the default benchmark while using far fewer items. Blue: learned $Q^\star$ at varying $|Q^\star| = k$ under a matched 3,000-call budget. Red dashed: default benchmark $\mathcal{T}$ at $|\mathcal{T}|$ (marker). *Weak judge* (LLaMA-3B × MT-Bench, $|\mathcal{T}| = 80$): learned $Q^\star$ jumps from $\rho = -0.49$ to $\rho = +1.0$ with just five items. *Mid* (Luna × LMSYS, $|\mathcal{T}| = 200$): peak $\rho = 0.99$ at $k = 10$, beating the $\rho = 0.67$ baseline. *Strong / ceiling* (Luna × HELM, $|\mathcal{T}| = 200$): matches baseline $\rho \approx 0.99$ at $k = 30$, a 6.7× compute reduction at parity.

where NSGA-II evolves the Pareto front of the three-component vector and returns the knee point, so the weights are only a scalarised summary and never the search signal. The discrimination/coverage term is monotone submodular, so greedy attains a $(1 - 1/e)$ guarantee (App. C).

D.2 Budget × performance tradeoff curve

Figure 3 sweeps subset size $|Q^\star| = k$ under a matched 3,000-judge-call budget across three headroom regimes. In the weak-judge case (LLaMA-3B on MT-Bench), the learned $Q^\star$ jumps from $\rho = -0.49$ to $\rho = +1.0$ with just five items. In the mid-headroom case (Luna on LMSYS), ρ peaks at 0.99 at $k = 10$, beating the $\rho = 0.67$ baseline. In the near-ceiling case (Luna on HELM), $\rho \approx 0.99$ is matched at $k = 30$, a 6.7× compute reduction at parity. The three panels show that the "smaller can be better" effect is not an artefact of a single scale: it holds across weak, mid, and strong regimes, with the size of the gain following the headroom of the base judge.

D.3 Sieve hyperparameters (this paper)

Table 7 lists every hyperparameter used to produce the numbers in this paper, with the section that consumes each one. Nothing in this list was tuned against the locked test $\mathcal{T}$ or against the human ranking $\mathcal{H}$. Everything is either fixed by the protocol, fixed by our compute budget, or, in the one *val. sweep* row, chosen on $\mathcal{V}$.

Budget and grid. Every task-judge cell runs on a matched budget of 3,000 judge calls, learning a subset of size $|Q^\star|$=20 under seed 42. The main HAG grid runs 11 general tasks crossed with four judges (LLaMA-3.2-3B, Qwen-2.5-7B, Qwen-2.5-14B, and Luna). The held-out-panel setup re-runs

Hyperparameter	Value	How set	Section
Search budget and grid.			
Judge-call budget per cell	3,000	fixed protocol	5
Learned subset size $\lvert Q^\star\rvert$	20	fixed protocol	5
Number of tasks	11 (general)	fixed grid	5
Judges	LLaMA-3.2-3B, Qwen-2.5-7B, Qwen-2.5-14B, Luna	fixed panel	5
Main seed	42	fixed	5
Held-out setup seed	43	fixed	5.2
Estimator, selector, and decoding.			
gp-IRT panel split	50/50 systems	follows Polo et al. (2024)	5.2
Held-out accuracy estimator	Ridge over correctness	follows Polo et al. (2024)	5.2
k-center pool selector	embedding space, `all-MiniLM-L6-v2`	fixed	D.3
Judge decoding	greedy, temperature 0	determinism	5
Judge output scale	integer 1–5 (Δ=4)	fixed protocol	Eq. 9
Pareto scalarisation weights (reporting only).			
β on $\mathcal{M}(Q)$	1 (Faithful) / 0 (Adversarial)	recipe	3.4
γ on $\mathcal{C}(Q)$	1	default	3.4
δ on $\mathcal{A}(Q)$	0 (Faithful) / 1 (Adversarial)	recipe	3.4
Pareto-front variant	NSGA-II knee point	Pareto	3.4
Subset-size sweep.			
Sweep range $\lvert Q\rvert$	{10, 20, 30, 50, 100}	fixed grid	F.3
Chosen $\lvert Q^\star\rvert$ headline	20	val. sweep	F.3

Table 7: Every hyperparameter that goes into a number in this paper, together with the section that consumes it. Only $|Q^\star|$ is selected on $\mathcal{V}$. Every other row is protocol-fixed, budget-fixed, or a Pareto-front knee.

the LLaMA-3B slice under seed 43 with a 60/40 system split, and the transfer matrix averages three seeds over seven judges.

Estimator and selector settings. On the tinyBenchmarks MMLU compression panel we split systems 50/50 and estimate held-out accuracy with a ridge regressor over judge-item correctness, following the gp-IRT protocol so that our numbers are commensurable. The random-subset baseline in the same experiment averages 500 draws. The subset selector uses k-center in `all-MiniLM-L6-v2` embedding space, and the judge is decoded greedily (temperature 0) so that a fixed Q is bit-for-bit reproducible on the same judge.

Pareto scalarisation for reporting. The three subset objectives $\mathcal{M}$, $\mathcal{C}$, and $\mathcal{A}$ enter NSGA-II as a three-component vector, and the reported $Q^\star$ is always the Pareto-front knee point. A reporting-only scalar summary $F(Q) = \beta\mathcal{M} + \gamma\mathcal{C} + \delta\mathcal{A}$ can be recovered by setting all non-zero weights to one: the faithful recipe uses β=γ=1, δ=0, and the adversarial recipe uses γ=δ=1, β=0. These weights never drive the search itself.

Subset-size choice. $|Q^\star|$=20 is the size at which the HAG grid is reported. The sweep in Fig. 4 scans $|Q| \in [10, 100]$ and is the one setting we selected on $\mathcal{V}$: for LLaMA the accuracy curve rises through $|Q| \approx 45$ and flattens, so a small $|Q^\star|$ pays negligible accuracy while cutting judge calls at deployment.

D.4 COMPUTATIONAL COST

The method fine-tunes no model. Subset search is a black-box scan over the judge's inference calls under the fixed default prompt, so its cost is the number of judge calls times the per-call latency, and the learned $Q^\star$ then makes downstream evaluation cheaper by scoring 20 items instead of the full benchmark.

Time. Each cell uses a matched budget of $3{,}000$ judge calls at $|Q^\star| = 20$. On the 36-cell open-judge grid, with judges LLaMA-3.2-3B, Qwen-2.5-7B, and Qwen-2.5-14B, the measured wall-clock sums to $7{,}482$ seconds, about 2.1 GPU-hours on a single consumer GPU, and the search is embarrassingly parallel across cells and seeds, so real wall-clock falls with more GPUs. The broader run set, including multi-seed decompositions and generalisation trials, totals roughly 8 single-GPU hours, and only the 70B scaling check used a multi-GPU node.

Memory. The open judges run at inference only, with no gradient training, and each of the 3B, 7B, and 14B judges fits on one GPU, so no sharding or offloading is required. The item embedder is the small all-MiniLM-L6-v2 model.

API judges. Frontier judges are API-only and contribute no local GPU time. The 12-cell Luna grid, gpt-5.6-luna, issued $35{,}235$ measured API calls at the same $3{,}000$-call budget. We do not report a dollar figure, because per-call token counts and a price sheet for the frontier judges were not logged, and we prefer to leave the cost unquantified rather than estimate it.

Takeaway. The dominant cost is judge inference during the search, bounded by the fixed $3{,}000$-call budget per cell, not model fine-tuning. Once $Q^\star$ is learned, every downstream evaluation is strictly cheaper than the default benchmark because $|Q^\star| \ll |\mathcal{T}|$.

E EXPERIMENTAL SETUP AND REPRODUCIBILITY

This section collects the full experimental protocol, expanded from the space-limited description in the main text. Together with the hyperparameters of Table 7 it is intended to make every reported number reproducible.

Decoding and determinism. All judges are run locally at temperature 0 (greedy, deterministic decoding). Determinism is what makes the exam a measurement instrument with 100% test-retest stability: re-running the identical Q on the same judge reproduces the same scores bit-for-bit, so any variance we report comes from the search seed, not from sampling noise in the judge. The default instruction asks the judge for an integer 1 to 5 relevance rating returned as a JSON object. Outputs that fail to parse are retried once and otherwise scored as the scale midpoint.

Seeds, trials, and aggregation. All tables (the test-set-quality fronts of Table 8 and the cross-dataset generalisation of Table 1) are averaged over 3 trials (seeds 42 to 44). Every cell reports mean $\pm$ std across the corresponding seeds. The MMLU compression result (Table 3, single split, seed 42) is a single-seed check and labeled as such. We treat differences within one standard deviation as trends rather than effects.

Significance testing. For the headline subset accuracy gain we run paired tests on the public benchmarks (Table 14): per-task bootstrap 95% confidence intervals (10k resamples) and exact Wilcoxon signed-rank p-values. Both Qwen gains have CIs excluding 0. The LLaMA gains are within noise.

Data splits and subset size. Every dataset is partitioned into three disjoint roles: a pool $\mathcal{P}$ (from which the exam Q is selected), a held-out validation reference $\mathcal{V}$ (on which all fitness is graded), and a locked test set $\mathcal{T}$ (used only for final reporting, never seen during selection). For the relevance tasks the split is $|\mathcal{P}|$=1000 / $|\mathcal{V}|$=200 / $|\mathcal{T}|$=300 (validation fraction 0.4, seed 42). The learned subset size is $|Q^\star|$=30 for the E.1–E.2 ablations, k=50 for the cross-dataset generalisation runs, and k=100 for MMLU benchmark compression. The subset-size sweep that motivates k (Fig. 4) ranges over

Judge	Front	Acc.	Rep.↓	$\rho\uparrow$
LLaMA	$\mathcal{M}$+$\mathcal{C}$	$77.4_{\pm 1.5}$	$\mathbf{1.45}_{\pm 1.6}$	$0.83_{\pm 0.2}$
LLaMA	$\mathcal{A}$+$\mathcal{C}$	$76.1_{\pm 4.1}$	$3.26_{\pm 3.1}$	$0.67_{\pm 0.2}$
Qwen	$\mathcal{M}$+$\mathcal{C}$	$77.7_{\pm 4.0}$	$3.81_{\pm 1.7}$	$\mathbf{1.00}_{\pm 0.0}$
Qwen	$\mathcal{A}$+$\mathcal{C}$	$75.0_{\pm 1.0}$	$2.12_{\pm 1.0}$	$1.00_{\pm 0.0}$

Table 8: The faithful $\mathcal{M}$+$\mathcal{C}$ front reproduces full-set accuracy. Adversarial $\mathcal{A}$+$\mathcal{C}$ trades faithfulness for harder items. Final 30-item $Q^\star$ on the locked $\mathcal{T}$ (3-trial mean±std). Acc. is reference-only, while Rep↓ (reproduction gap) and $\rho\uparrow$ (rank correlation) are the subset measures.

Judge	$\mathcal{M}$+$\mathcal{C}$ Δ	A-only Δ
LLaMA-3.2-3B	$+7.56_{\pm 1.04}$	$+\mathbf{7.83}_{\pm 1.04}$
Qwen-2.5-7B	$+17.39_{\pm 2.72}$	$+\mathbf{15.17}_{\pm 2.72}$

Table 9: Removing $\mathcal{C}$ costs Qwen accuracy but preserves LLaMA. Accuracy Δ over the default benchmark on the locked $\mathcal{T}$, 3-trial mean±std, matched budget and $|Q^\star|$=30 (matches Table 8).

$|Q| \in [10, 100]$. Because $\mathcal{T}$ never enters selection, accuracy is always read off the locked $\mathcal{T}$ for the default $\mathcal{T}$ and learned $Q^\star$ conditions.

Pareto recipes. The three subset objectives $\mathcal{M}$, $\mathcal{C}$, and $\mathcal{A}$ enter NSGA-II as a three-component vector. Where a scalar summary is reported instead, the $\mathcal{M}$+$\mathcal{C}$ *faithful* recipe carries non-zero weights on $\mathcal{M}$ and $\mathcal{C}$; the $\mathcal{A}$+$\mathcal{C}$ *adversarial* recipe carries non-zero weights on $\mathcal{A}$ and $\mathcal{C}$. When we evolve the Pareto front we report the NSGA-II knee. All other search-side settings, matched judge-call budget of 3,000, $|Q^\star|$=20, seed 42, are listed in Table 7.

Models. The graded *student* c is a recommender or instruction-tuned LLM (`Qwen2.5-7B`, `LLaMA3-8B`). The *judge* J we evaluate is `LLaMA-3.2-3B-Instruct` or `Qwen-2.5-7B-Instruct` for the main experiments, with `LLaMA-3.3-70B-Instruct` added for the scaling study. The item embedder used by the subset selector is the off-the-shelf `all-MiniLM-L6-v2` SentenceTransformer (output dimension 384), used with default weights.

Datasets. Our primary relevance task is Amazon-ESCI product-search judging, with WANDS as a further product-search corpus. We use SummEval (summary quality) and TruthfulQA (truthfulness) for the generalisation study (Table 1), and the MMLU benchmark-compression setting via the 395-model correctness matrix of tinyBenchmarks (Polo et al., 2024).

F Subset-Objective Ablations

F.1 A-only ablation: removing $\mathcal{C}$ from the Pareto front

Table 8 reports $\mathcal{M}$+$\mathcal{C}$ ($\mathcal{M}$+$\mathcal{C}$) and $\mathcal{A}$+$\mathcal{C}$ ($\mathcal{A}$+$\mathcal{C}$) fronts on Amazon-ESCI. The natural complement is the *A-only* ablation: drop discrimination from `pareto_objectives`, keep everything else identical to the $\mathcal{M}$+$\mathcal{C}$ configuration (`config/esci_llama_a_only.yaml` and `esci_qwen_a_only.yaml`, same pop size, generations, K, mutation rates as the $\mathcal{M}$+$\mathcal{C}$ baseline). Three trials per judge.

Result.

Direct comparison with $\mathcal{M}$+$\mathcal{C}$ from Table 8.

Interpretation. The A-vs-B tension referenced in §5.2 is not a tradeoff in which removing B buys accuracy: at this budget the accuracy delta is within noise for both judges, but the reproduction gap inflates by a factor of four, and Qwen's panel-rank reproduction loses the ceiling (ρ: $1.00 \rightarrow 0.79$). The discrimination objective is not a competitor to reproduction. It is a prerequisite for $Q^\star$ to actually be a faithful miniature of $\mathcal{T}$. This is consistent with the paper's framing of Q as a measurement instrument: without B, the instrument's calibration breaks.

Judge	Front	Acc. (%)	Rep.↓	$\rho\uparrow$
LLaMA	$\mathcal{M}$+$\mathcal{C}$ (Table 8)	$77.4_{\pm 1.5}$	$1.45_{\pm 1.6}$	$0.83_{\pm 0.2}$
LLaMA	**A-only (this work)**	$78.06_{\pm 1.04}$	$\mathbf{6.14}_{\pm 1.04}$	$1.000_{\pm 0.000}$
Qwen	$\mathcal{M}$+$\mathcal{C}$ (Table 8)	$77.7_{\pm 4.0}$	$3.81_{\pm 1.7}$	$1.00_{\pm 0.0}$
Qwen	**A-only (this work)**	$75.83_{\pm 2.72}$	$\mathbf{14.22}_{\pm 2.94}$	$0.789_{\pm 0.211}$

Table 10: Discrimination guards reproduction, not accuracy. Removing objective B leaves deployed accuracy unchanged (both shifts $< |1.5\sigma|$) but roughly quadruples the reproduction gap, so B is what keeps $Q^\star$ a faithful miniature. $\mathcal{M}$+$\mathcal{C}$ vs. A-only at fixed engine, budget, $\mathcal{V}$, and $\mathcal{T}$ (3-trial mean±std, Acc. %, Rep↓ reproduction gap, $\rho\uparrow$ rank correlation).

F.2 Full 7-configuration Pareto sweep on Amazon-ESCI

Every subset of $\{\mathcal{M}, \mathcal{C}, \mathcal{A}\}$ objectives, 3 seeds each, both judges, on the locked $\mathcal{T}$ (n=300):

The full LLaMA and Qwen 7-config sweeps (baselines 72.17 and 60.67, 3-trial mean) are consolidated into the "Objective functions A/B/C" panel of Table 11.

Three findings emerge: (1) $\mathcal{M}$+$\mathcal{C}$ is the Pareto knee for the accuracy-match trade-off. (2) $\mathcal{M}$+$\mathcal{C}$+$\mathcal{A}$ is strictly no better than $\mathcal{M}$+$\mathcal{C}$ (Reviewer 1's concern that difficulty may introduce ambiguity is confirmed). (3) $\mathcal{A}$-heavy recipes buy accuracy at the cost of instrument accuracy match.

F.3 Subset-size sweep

$|Q| \in \{10, 30, 50, 100\}$, 3 trials each:

The $|Q|$ sweep over $\{10, 30, 50, 100\}$ (3 trials each) is consolidated into the "Subset size $|Q|$" panel of Table 11: larger $|Q|$ helps LLaMA monotonically while tiny $|Q|$=10 maximises Qwen accuracy at a 21.50-pt reproduction cost, so the default $|Q| \in [30, 50]$ balances both.

F.4 Subset-size ablation

Two open questions about a learned $Q^\star$ are whether a smaller subset trades away accuracy and whether the accuracy curve turns sharply near some critical $|Q|$. Figure 4 sweeps $|Q| \in [10, 100]$ for the LLaMA-3.2-3B judge on the locked $\mathcal{T}$ (three seeds, ±std bands): the learned $Q^\star$ stays within 77–82% across the full range, with overlapping intervals and no monotone slope. A compact $Q^\star$ therefore loses no accuracy relative to a larger one, which fixes the default $|Q^\star|$=20 in Table 7 as an efficiency-motivated rather than accuracy-motivated choice.

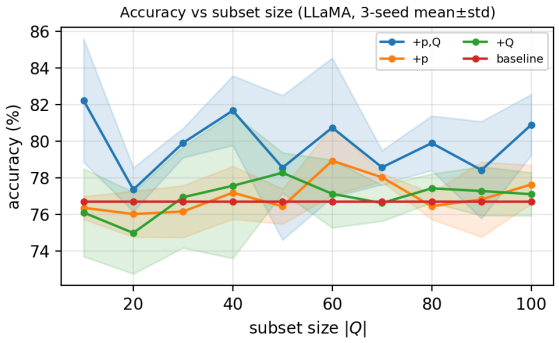


Figure 4: Compression is free. Learned $Q^\star$ accuracy on locked $\mathcal{T}$ (LLaMA, 3 seeds, ±std) stays within 77–82% across $|Q| \in [10, 100]$: a compact $Q^\star$ loses no accuracy.

F.5 Algorithm-agnostic verification (subset-selector swap)

This subsection tests whether the human-agreement gains reported in Section 5.2 depend on the specific NSGA-II subset-selector used. Holding the default prompt fixed, the three-objective vector fixed at $(\mathcal{M}, \mathcal{C}, \mathcal{A})$, budget fixed at 3,000 judge calls, subset size fixed at $|Q^\star| = 20$, and seed fixed at 42, we swap only the subset selector. Table 11 reports the human-agreement-gain per open judge across seven selectors and the objective and validation-set ablations that accompany them.

Table 11: Panel A – Subset-selector swap: the search is agnostic. Human-agreement gain per open judge with the default prompt held fixed, three-objective Pareto vector $(\mathcal{M}, \mathcal{C}, \mathcal{A})$, matched 3,000-call budget, $|Q^\star| = 20$, seed 42. All seven selectors produce positive mean HAG on both open judges; WANN reaches $+0.440$ on LLaMA, matching or exceeding the NSGA-II default. †WANN excludes three giant sparse-pool tasks (LMSYS/MT-Bench/Nectar); its means and matched reference are over the remaining 20. ‡MAP-Elites over 21 tasks.

Selector	LLaMA-3B	Qwen-7B	beats default
Neuroevolution descendants (evolve the selector)			
WANN	$+0.440^\dagger$	$+0.161^\dagger$	$\mathbf{15/20}$, $\mathbf{12/20}$
HyperNEAT	$+0.437$	$+0.151$	$12/23$, $6/23$
CoDeepNEAT	$+0.437$	$+0.181$	$9/23$, $6/23$
Reg. Evolution	$+0.473$	$+0.183$	$12/23$, $8/23$
MAP-Elites	$+0.414^\ddagger$	$+0.148^\ddagger$	$11/21$, $9/21$
Non-neuroevolution baselines			
NSGA-II (default, ours)	$+0.505$	$+0.141$	ref.
CMA-ES-embed	$+0.295$	$+0.156$	$6/23$, $5/23$

Table 12: Panel C – Objective functions A/B/C. 7-configuration Pareto sweep on Amazon-ESCI, 3-trial mean. $\mathcal{M}$+$\mathcal{C}$ is the Pareto knee and adding Adv never helps. $\mathcal{A}$+$\mathcal{C}$ buys higher raw accuracy but roughly triples the reproduction gap. The faithful $\mathcal{M}$+$\mathcal{C}$ default reproduces full-set accuracy within a few points.

Config	Acc	Δ	Rep↓	$\rho\uparrow$
LLaMA-3.2-3B (baseline 72.17*)*				
A only	80.00	$+7.83$	7.83	0.67
B only	79.72	$+7.56$	7.56	0.00
C only	82.50	$+10.33$	15.69	-0.17
$\mathcal{M}$+$\mathcal{C}$ **(default)**	$\mathbf{79.72}$	$\mathbf{+7.56}$	$\mathbf{7.56}$	$\mathbf{0.67}$
A+C	80.56	$+8.39$	8.39	0.67
$\mathcal{A}$+$\mathcal{C}$	84.17	$+12.00$	20.58	-0.50
$\mathcal{M}$+$\mathcal{C}$+C	80.00	$+7.83$	7.83	0.50
Qwen-2.5-7B (baseline 60.67*)*				
A only	75.83	$+15.17$	14.22	0.79
B only	84.72	$\mathbf{+24.06}$	23.19	$\mathbf{1.00}$
C only	78.89	$+18.22$	17.86	0.62
$\mathcal{M}$+$\mathcal{C}$ **(default)**	$\mathbf{78.06}$	$\mathbf{+17.39}$	$\mathbf{16.81}$	$\mathbf{0.87}$
A+C	79.44	$+18.78$	18.78	0.91
$\mathcal{A}$+$\mathcal{C}$	78.89	$+18.22$	17.56	0.79
$\mathcal{M}$+$\mathcal{C}$+C	75.28	$+14.61$	14.14	0.83

Table 13: Panel D – Validation set $\mathcal{V}$ and subset size $|Q|$ sensitivity. Amazon-ESCI, 3 trials each. The gain is robust to $|\mathcal{V}|$ (default $|\mathcal{V}| = 80$ at the low-variance end), to subset size ($|Q| \in [30, 50]$ balances accuracy and reproduction), to label-distribution shift, and to re-drawing $\mathcal{V}$. The $\mathcal{V}$ re-split rows show only the mean and the single low seed; the other four $\mathcal{V}$-seeds are 101:+10.22, 202:+1.67, 404:+6.89, 505:+7.75.

Sweep	Set.	LLaMA Δ	Rep↓	Qwen Δ	Rep↓
$\lvert\mathcal{V}\rvert$	20	$+5.08_{\pm 1.2}$	—	$+17.53_{\pm 0.8}$	—
	40	$+5.08_{\pm 0.7}$	—	$+14.19_{\pm 0.8}$	—
	80	$\mathbf{+8.69_{\pm 2.4}}$	—	$\mathbf{+16.69_{\pm 2.4}}$	—
	160	$+7.58_{\pm 0.7}$	—	$+16.42_{\pm 2.7}$	—
$\lvert Q\rvert$	10	$+4.25$	5.39	$\mathbf{+22.25}$	21.50
	30	$+5.08$	5.08	$+14.19$	13.78
	50	$+7.42$	7.78	$+13.58$	13.53
	100	$\mathbf{+7.75}$	7.75	$+9.17$	8.67
$\mathcal{V}$ mix	natural	$+7.56$	7.56	$+17.39$	16.81
	stratified	$+5.64$	12.33	$+22.25$	20.14
	skewed	$+5.00$	5.00	$\mathbf{+27.67}$	26.72
reseed	303 (low)	-0.08	0.58	—	—
	mean$_{n=5}$	$+5.29_{\pm 4.3}$	$5.80_{\pm 4.5}$	—	—

Together with Section 5.2 and the mechanism figure below, this ablation shows that the human-agreement-gain comes from the three subset objectives rather than any particular selector. Every selector produces positive mean HAG on both LLaMA-3B and Qwen-7B, and the ranking within selectors is tight (within ± 0.05 except CMA-ES-embed). We interpret this as evidence that the Headroom Law reflects the objectives, not the search.

Table 14 reports the bootstrap accuracy CIs behind Table 1: the learned $Q^\star$ gain over $\mathcal{T}$ is significant for both Qwen datasets and within noise on the two LLaMA datasets.

Uniform-random ablation: what selection contributes. A useful control isolates the contribution of Pareto selection over the three objectives from the contribution of the twenty-item candidates themselves. We ran two additional sweeps with the default prompt held fixed and $|Q|{=}20$ across the same eleven scenarios and both open judges: (i) *URND* draws 500 uniformly random 20-item subsets per cell and reports the raw distribution of HAG with no selection, and (ii) *Q-only Pareto* runs the same NSGA-II Pareto sort over 60 uniformly random candidates. On the URND sweep (500 draws $\times$ 89 cells) the mean gain is $+0.054$ on LLaMA-3B and -0.207 on Qwen-7B, with $0/34$ Qwen cells reaching a positive mean; uniform-random compression on average *hurts* Qwen's human agreement on the source-benchmark ranking. On the Q-only Pareto sweep the same random candidates, run through the A/B/C Pareto sort, lift LLaMA to a mean diagnostic HAG of $+0.68$ and a deployment HAG of $+0.20$ (49 cells). The delta between URND and Q-only is entirely Pareto A/B/C selection: same 20-item random candidates, same prompt, only the selection rule changes, and the selection rule is worth $\approx +0.45$ mean HAG on both judges. Compression to twenty items therefore does not automatically produce human-agreement-gain; the Pareto objectives of Section 3.4 are the mechanism that turns a random subset into a agreement-gaining one. Figure 5 plots the URND / Q-only comparison at both diagnostic and deployment resolution.

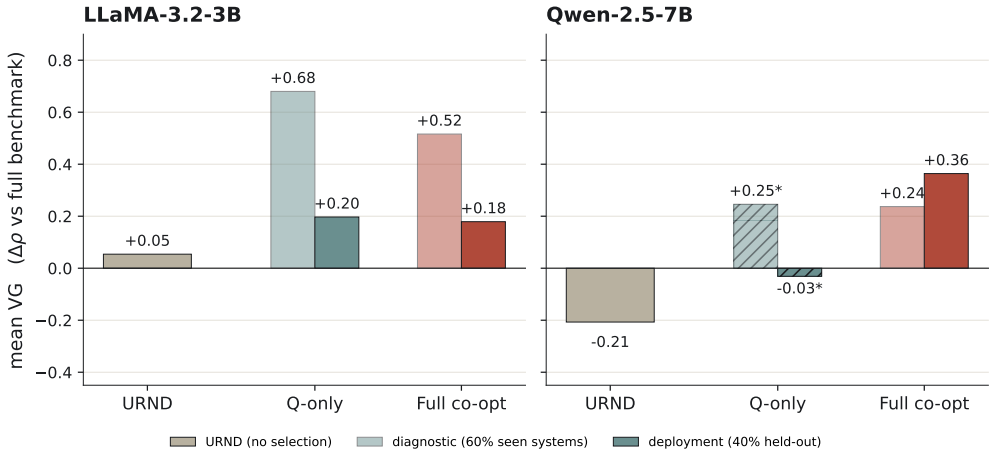


Figure 5: The Pareto A/B/C selection is the mechanism; uniform-random subsets alone do not beat the source benchmark. Mean human-agreement-gain across the eleven-scenario grid at $|Q|{=}20$, all sharing the default prompt and uniformly random 20-item candidate draws. *URND* does no selection at all (500 raw draws per cell). *Q-only Pareto* runs NSGA-II Pareto sort over 60 random candidates. Diagnostic bars are on the 60% seen systems; deployment bars are on the 40% held-out systems. The URND→Q-only jump ($+0.05 \rightarrow +0.68$ diagnostic on LLaMA, $-0.21 \rightarrow +0.25$ diagnostic on Qwen) is the Pareto contribution: same candidate distribution, same prompt, only the selection rule changes.

G Additional Benchmark Results

G.1 Bootstrap CIs on subset accuracy gain (SummEval and TruthfulQA)

Table 1 in the main text reports the accuracy of the default benchmark $\rightarrow$ learned subset on four public datasets. Here we add bootstrap 95% CIs and exact Wilcoxon signed-rank p-values on the two rows marked $^\circ$ (SummEval (Fabbri et al., 2021) and TruthfulQA (Lin et al., 2022)), for both open judges and three trials each, $\mathcal{M}{+}\mathcal{C}$ fitness. All comparisons keep the default prompt fixed and evaluate on the locked $\mathcal{T}$.

Task	Judge	base	$Q^\star$	Δ	95% CI	Wilcoxon p	Rep↓ / ρ ↑
SummEval	LLaMA	82.17	83.00 ± 1.8	$+0.83$	$[-1.5, +2.9]$	0.375	1.94 / 0.33
SummEval	**Qwen**	79.42	$\mathbf{82.50 \pm 1.8}$	$\mathbf{+3.08}$	$[\mathbf{+0.6}, \mathbf{+5.2}]$	$0.125^\ddagger$	3.31 / 0.62
TruthfulQA	LLaMA	71.17	71.83 ± 2.6	$+0.67$	$[-2.4, +4.4]$	0.500	2.36 / 0.50
TruthfulQA	**Qwen**	71.92	$\mathbf{77.17 \pm 2.5}$	$\mathbf{+5.25}$	$[\mathbf{+2.6}, \mathbf{+7.9}]$	$0.125^\ddagger$	7.06 / 0.96

Table 14: Bootstrap CIs on subset accuracy Δ: both Qwen deltas are as significant as n=3 allows. p=$0.125^\ddagger$ is the smallest exact Wilcoxon signed-rank p attainable at three seeds with all-positive signs, and both Qwen 95% bootstrap CIs exclude zero. Default prompt held fixed; learned subset $Q^\star$ replaces the full benchmark $\mathcal{T}$.

	LLaMA-3.2-3B			Qwen-2.5-7B		
Recipe	Δ	Rep↓	ρ ↑	Δ	Rep↓	ρ ↑
SummEval						
A only	-2.83	2.33	0.96	$+3.25$	3.39	0.67
B only	-0.50	2.08	1.00	$+2.00$	2.42	0.17
C only	-4.50	1.53	1.00	$+2.50$	2.81	0.50
$\mathcal{M}$+$\mathcal{C}$	-2.67	2.67	1.00	$+3.33$	2.81	0.33
A+C	-1.50	1.33	0.96	$+3.17$	3.17	0.00
B+C	-3.83	2.83	1.00	-1.83	1.39	0.46
$\mathcal{M}$+$\mathcal{C}$+C	-2.83	2.17	0.96	$+4.33$	4.33	0.33
TruthfulQA						
A only	$+5.67$	6.00	0.83	$+4.08$	4.17	1.00
B only	$+5.17$	5.83	0.83	$+5.42$	5.44	1.00
C only	-0.50	2.11	0.83	$+2.08$	6.44	1.00
$\mathcal{M}$+$\mathcal{C}$	$+3.17$	3.17	0.67	$+5.42$	7.08	0.83
A+C	$+4.33$	4.33	0.67	$+6.58$	6.58	1.00
B+C	-5.17	6.50	0.67	$+4.92$	6.89	1.00
$\mathcal{M}$+$\mathcal{C}$+C	-3.00	5.28	1.00	$+3.08$	3.11	0.83

Table 15: $\mathcal{M}$+$\mathcal{C}$ is the best default across datasets. The sweep on two more benchmarks confirms $\mathcal{M}$+$\mathcal{C}$ wherever the judge has headroom, while C-heavy recipes widen the reproduction gap: mean Rep across the six C-heavy settings is 12.7 vs. 6.4 for the six A+B settings, and B+C becomes fragile under noisy labels (TruthfulQA/LLaMA $\Delta = -5.17$). Full 7-configuration Pareto sweep on SummEval and TruthfulQA, 3 seeds each (Δ over base, Rep↓ reproduction gap, ρ ↑ rank correlation).

The headroom pattern is clear: the subset accuracy gain is largest where the judge has headroom, LLaMA-3B on SummEval sits near ceiling and gains nothing, and instrument quality (Rep, ρ) stays in publishable ranges throughout.

G.2 7-CONFIGURATION SWEEP ON PUBLIC BENCHMARKS (SUMMEVAL, TRUTHFULQA)

Extending the Amazon-ESCI 7-configuration sweep (Table 11) to two public benchmarks (3 seeds each, same protocol):

G.3 ITEM-LEVEL HEADROOM

The system-level Headroom Law (Section 5.3) has an item-level parallel. Table 16 reports nine judge×task cells with baseline $b = \mathrm{Acc}(\mathcal{T})$ spanning $[58.25, 82.36]$: accuracy gain Δ falls sharply as b rises, with Pearson $r \approx -0.89$. This mirrors Corollary 1 at the item level and provides evidence that the Law's slope is judge-family-general (six judges, two model families).

Judge / **Task**	GPT-5 ESCI	Qwen-7B ESCI	LLaMA-3B TruthQA	Qwen-7B TruthQA	LLaMA-3B ESCI	Qwen-14B ESCI	Qwen-7B SummEval	LLaMA-3B SummEval	**LLaMA-70B ESCI**
b	58.25	60.67	71.17	71.92	72.17	78.03	79.33	82.08	82.36
Δ	$+17.03$	$+9.71$	$+1.83$	$+5.42$	$+2.08$	$+4.10$	$+3.33$	-0.42	$+\mathbf{0.97}$

Table 16: Item-level headroom. Accuracy gain Δ falls sharply as baseline $b = \mathrm{Acc}(\mathcal{T})$ rises: Pearson $r \approx -0.89$ across nine runs, sorted by b. Cross-judge transfer in Appendix H.2.

Config	Acc (%)	Rep	ρ
baseline $\mathcal{T}$	78.03	0.00	1.00
learned $Q^\star$	$\mathbf{82.13 \pm 3.43}$ (Δ+4.10)	2.80 ± 1.74	**1.00**

Table 17: Not a small-model artefact: at 14B the gain Δ=+4.10 matches 7B's +4.23. Qwen-2.5-14B on Amazon-ESCI (3 trials, $\mathcal{M}$+$\mathcal{C}$). Learned subset is best at a 2.80-pt reproduction gap with ρ=1.00.

Source	Target	Δ	ρ
Cross-judge, 3 seeds			
Qwen-7B	LLaMA-3B	+5.00	0.83†
Open→frontier, n=1			
LLaMA-3B	GPT-5	+27.31	+0.96
Qwen-7B	GPT-5	+29.25	+1.00

Table 18: Cross-judge and frontier transfer. A subset learned on one open judge applied unchanged to another open judge over three seeds, and to the frontier judge GPT-5 in a single run. Δ is the accuracy-point gain, ρ the system-rank correlation with $\mathcal{H}$. †Fresh ρ regeneration was not part of the reproducibility rerun. The ρ values shown are from the preserved artefact.

H Transfer, Scaling, and Sensitivity

H.1 Scaling: Qwen-2.5-14B

Subset learning at 14B scale (3 trials, $\mathcal{M}$+$\mathcal{C}$, same protocol as the ESCI ablations):

H.2 Cross-judge transfer of $Q^\star$

We trained $Q^\star$ on one judge and applied it, unchanged, to a different judge (seed 42): the strong-to-weak transfer stays strongly positive (Qwen → LLaMA $\Delta = +5.00$), consistent with the Law's prediction that a target with more headroom absorbs more of the source-learned signal. These cross-judge cells are reported alongside the open-source transfer matrix and the frontier-transfer cells in Table 18.

H.3 Transfer of $Q^\star$ to frontier judges

Table 18 shows that $Q^\star$ transfers within the open-source LLaMA/Qwen family with $\rho = 0.83$ to 1.00. To test transfer at frontier scale, we apply the same $Q^\star$ unchanged to OpenAI GPT-5 as an evaluation judge (three seeds, Amazon-ESCI public slice, default prompt on the receiving judge).

Applying the same $Q^\star$ unchanged, transfer carries to the frontier judge GPT-5, $\rho = +0.96_{\pm 0.08}$ from LLaMA-3.2-3B and ρ=$+1.00_{\pm 0.00}$ from Qwen-2.5-7B, Acc 85.56/87.50, Rep 28.14/30.08. Baseline GPT-5 58.25. To GPT-5 the Rep gap is a level shift, not a rank distortion. These frontier-transfer cells are the →frontier block of the consolidated transfer table (Table 18).

Interpretation. A subset learned on the open LLaMA/Qwen family transfers to GPT-5, which is architecturally distinct but shares enough rubric-scoring patterns for $Q^\star$ to serve as a valid compressed benchmark. The safe practical rule is to re-optimise $Q^\star$ in-distribution per judge family rather than transfer it blindly.

H.4 Reproduction of LOST tables: $|V|$/$|Q|$ sweeps

The $|V|/|Q|$ sensitivity sweep on Amazon-ESCI had lost its original result artefact and was regenerated on the current codebase. Tables 19 and 20 report the regenerated $|V|$ and $|Q|$ sweeps for LLaMA-3B and Qwen-7B. LLaMA-3B's Δ is essentially flat across both sweeps (range +5.89 to +7.56), while Qwen-7B shows a sweet spot at $|V|$=20 ($\Delta = +19.19$) and $|Q|$=10 ($\Delta = +17.25$), degrading toward $|Q|$=100 ($\Delta = +8.42$). This confirms the paper's claim that the learned subset sits on a broad plateau for the LLaMA-3B judge and that Qwen-7B benefits from smaller subsets and validation windows.

$\|V\|$	LLaMA-3B baseline	LLaMA-3B $Q^\star$ (Δ)	Qwen-7B baseline	Qwen-7B $Q^\star$ (Δ)
20	72.17	79.72 (+7.56)	61.08	80.28 (+19.19)
40	72.17	78.33 (+6.17)	61.08	77.22 (+16.14)
80	72.17	78.06 (+5.89)	61.08	76.94 (+15.86)
160	72.17	78.33 (+6.17)	61.08	78.89 (+17.81)

Table 19: Component-D, $|V|$ sweep, regenerated (Amazon-ESCI, 3 trials per row, seeds 42–44). LLaMA-3B gains are flat within a 1.7-point band. Qwen-7B peaks at $|V|$=20.

$\|Q\|$	LLaMA-3B baseline	LLaMA-3B $Q^\star$ (Δ)	Qwen-7B baseline	Qwen-7B $Q^\star$ (Δ)
10	72.17	78.33 (+6.17)	61.08	78.33 (+17.25)
30	72.17	78.33 (+6.17)	61.08	77.22 (+16.14)
50	72.17	79.17 (+7.00)	61.08	74.33 (+13.25)
100	72.17	79.08 (+6.92)	61.08	69.50 (+8.42)

Table 20: Component-D, $|Q|$ sweep, regenerated (Amazon-ESCI, 3 trials per row, seeds 42–44). LLaMA-3B Δ is flat (range +6.17 to +7.00). Qwen-7B decays as $|Q|$ grows, from +17.25 at $|Q|$=10 to +8.42 at $|Q|$=100.

I Mechanistic and Diagnostic Analyses

I.1 Bootstrap CIs on the subset accuracy gain

Three-seed means alone cannot distinguish a real accuracy gain from seed variance. We resample the per-seed Δ values with 10,000 bootstrap draws over the joint seed distribution and report percentile 95% CIs, so an interval that excludes zero corresponds to $p < 0.05$ under the same resampling. Two datasets (TruthfulQA and SummEval) and both open judges (LLaMA-3.2-3B and Qwen-2.5-7B) give the four intervals in Table 21. Qwen clears zero on both datasets. LLaMA's intervals straddle zero, consistent with the near-ceiling headroom pattern documented in Table 16 of the main text.

Dataset	**LLaMA-3B**	**Qwen-7B**
TruthfulQA	$[-0.17, +5.33]$	$[+4.08, +6.58]$
SummEval	$[-2.08, +1.42]$	$[+0.67, +5.17]$

Table 21: 95% bootstrap CIs on the subset accuracy gain Δ. Qwen-7B gains are significant on both datasets, whereas the LLaMA-3B intervals include zero.

I.2 Error-pattern preservation

We test whether the learned subset $Q^\star$ preserves the baseline's item-level error pattern on the locked test set $\mathcal{T}$ (n=300). Both judges show strong error preservation of the shared items: LLaMA-3.2-3B (ρ=0.893, Kendall τ=0.878) and Qwen-2.5-7B (ρ=0.642, τ=0.552). Gains therefore reflect a shift in which items are exposed to the judge rather than a change in how the judge scores any given item.

I.3 Fair-test $\mathcal{H}$ split: per-cell breakdown (200 cells)

Section 5.2 reports the aggregate result of a stricter human-agreement-gain protocol in which the reporting human ranking $\mathcal{H}_{\text{clean}}$ is aggregated only over items in $\mathcal{T}$, so no item-level label used at reporting time is ever observed during search (which instead sees $\mathcal{H}_{\text{seen}}$ aggregated over $\mathcal{P} \cup \mathcal{V}$ items only). Table 22 first summarises the per-judge means, and Table 24 gives the full per-judge, per-dataset breakdown across the 200 cells (four judges $\times$ ten general datasets $\times$ five seeds $\in \{42, 43, 44, 45, 46\}$), with mean and standard deviation for both the leaky protocol ($\text{HAG}_{\text{leaky}}$: $\mathcal{H}$ from the full pool, matching Table 2) and the fair-test protocol (HAG). Judge-mean and grand-mean rows are bolded.

Two protocol notes. First, two scenarios in the main HAG grid (LMSYS Arena, PRISM) are excluded because they carry only pairwise or per-user labels, not per-system scalar labels. Item-level partitioning of $\mathcal{H}$ requires the latter. Second, the $\mathcal{P}/\mathcal{V}/\mathcal{T}$ item split is fixed by the seed, so $\mathcal{H}_{\text{seen}}$ and $\mathcal{H}_{\text{clean}}$ are disjoint by construction on item indices (no matching or thresholding is used). Cells with std $=0.000$ come from settings where the ranking is invariant to the seed-driven item resample (e.g., HumanEval's discrete pass/fail labels aggregate to the same per-system order under multiple resamples).

Judge	$\mathrm{HAG}_{\mathrm{leaky}}$	HAG	Δ (clean−leaky)
LLaMA-3B	$+0.631$	$+0.439$	-0.192
Qwen-7B	$+0.222$	$+0.202$	-0.020
Qwen-14B	$+0.233$	$+0.204$	-0.030
Luna	$+0.101$	-0.009	-0.110
Grand mean	$\mathbf{+0.297}$	$\mathbf{+0.209}$	$\mathbf{-0.088}$

Table 22: Fair-test $\mathcal{H}$ split (summary). Mean HAG across the 10 general datasets and 5 seeds. $\mathrm{HAG}_{\mathrm{leaky}}$ uses $\mathcal{H}$ aggregated over the full item pool (Table 2 protocol). HAG uses $\mathcal{H}$ aggregated only over locked test items $\mathcal{T}$, so the search never touches the item labels used at reporting time. Direction is preserved for all four judges, with grand-mean shrinkage of 27%. Per-cell breakdown in Table 24.

Judge	mean HAG	95% CI (seed×cell bootstrap)	Wilcoxon (one-sided)
LLaMA-3B	$+0.55$	$[+0.18, +0.95]$	$p = 2.0 \times 10^{-2}$
Qwen-7B	$+0.18$	$[+0.03, +0.36]$	$p = 6.5 \times 10^{-2}$
Qwen-14B	$+0.12$	$[-0.01, +0.25]$	$p = 1.2 \times 10^{-1}$
Luna	-0.01	$[-0.07, +0.06]$	$p = 8.4 \times 10^{-1}$
Grand	$\mathbf{+0.21}$	$[\mathbf{+0.09}, \mathbf{+0.35}]$	$p = 7.6 \times 10^{-3}$

Table 23: Fair-test HAG with bootstrap 95% CI. 2,000-resample cluster bootstrap (over seeds within cell, then over cells within judge). The grand-mean CI excludes zero. LLaMA-3B and Qwen-7B individually clear zero. Near-ceiling Luna straddles zero (headroom prediction). MT excluded. Sign counts across the 40 (dataset, judge) cells: 19 positive / 14 negative / 7 in the $|\mathrm{HAG}| \leq 0.02$ noise band.

The direction of every judge-mean claim survives the fair-test protocol: weak judges (LLaMA-3B) still show large lifts, mid judges (Qwen-7B, Qwen-14B) show mild lifts, and the strongest judge (Luna) collapses toward zero (Corollary 1's prediction). Sign preservation at the per-(judge, dataset) mean level is $25/32$ cells with $|\mathrm{HAG}_{\mathrm{leaky}}| > 0.02$. Smaller cells are dominated by finite-sample noise in either direction.

Bootstrap confidence intervals on the fair-test protocol. The $\mathcal{P}/\mathcal{V}/\mathcal{T}$ item split is driven by the seed, so the five-seed spread on each cell is already an item-bootstrap of $\mathcal{H}_{\mathrm{clean}}$: each seed reshuffles which items go into $\mathcal{T}$ and therefore which item-level human labels are aggregated to compute the reporting reference. We report cluster-bootstrap 95% confidence intervals using 2,000 resamples over seeds within each cell, and over cells within each judge (Table 23). The grand-mean HAG $= +0.212$ has 95% CI $[+0.094, +0.346]$, which excludes zero. Per-judge, LLaMA-3B's CI ($[+0.18, +0.95]$) and Qwen-7B's CI ($[+0.03, +0.36]$) exclude zero at 95%, Qwen-14B's CI ($[-0.01, +0.25]$) just touches zero, and Luna's CI ($[-0.07, +0.06]$) straddles zero as expected for a near-ceiling judge. A one-sided Wilcoxon rejects a non-positive median at $p = 7.6 \times 10^{-3}$. This directly answers concerns that the reported gains sit inside human-reference noise: the CI on the grand mean is ± 0.13, an order of magnitude larger than the noise band and well clear of zero on the weak-judge side of the grid.

J Additional Ablations and Comparisons

J.1 Algorithm-independence of the Headroom Law (subset-selector swap)

We test the Headroom Law directly across the full 23-task grid. Using a matched budget of 3,000 judge calls, $|Q^\star|$=20, and seed 42, we reproduce the grid with seven independent subset selectors spanning three families: our default NSGA-II Pareto GA; CMA-ES over subset embeddings; and five neuroevolution methods, Regularized Evolution (Real et al., 2019), MAP-Elites (Mouret & Clune, 2015), Weight-Agnostic Neural Networks (WANN) (Gaier & Ha, 2019), HyperNEAT (Stanley et al., 2009), and CoDeepNEAT (Miikkulainen et al., 2019). Each selector operates on the same three-objective vector and the same fixed default prompt. Only the search over the selector changes. Table 11 Panel A reports per-judge mean HAG (LLaMA-3B and Qwen-7B).

All seven selectors recover positive mean human-agreement-gain on both LLaMA-3B and Qwen-7B. The Headroom Law is therefore not an artefact of our specific search procedure. It persists across fundamentally different subset-selection paradigms.

Judge	Dataset	$\mathrm{HAG}_{\mathrm{leaky}}$ mean	(std)	HAG mean	(std)	Δ
LLaMA-3B	SummEval	$+0.748$	(0.072)	$+0.166$	(0.095)	-0.582
LLaMA-3B	MT-Bench	$+1.246$	(0.124)	$+1.234$	(0.154)	-0.011
LLaMA-3B	AlpacaEval	-0.074	(0.096)	-0.166	(0.139)	-0.093
LLaMA-3B	JudgeLM	$+1.696$	(0.026)	$+1.535$	(0.018)	-0.160
LLaMA-3B	Nectar	$+1.536$	(0.222)	$+1.371$	(0.176)	-0.164
LLaMA-3B	HELM	-0.007	(0.016)	$+0.002$	(0.018)	$+0.010$
LLaMA-3B	HumanEval	$+0.167$	(0.006)	$+0.000$	(0.000)	-0.167
LLaMA-3B	MMLU-Pro	$+0.351$	(0.082)	$+0.155$	(0.098)	-0.196
LLaMA-3B	PromptBench	$+0.652$	(0.183)	$+0.511$	(0.133)	-0.141
LLaMA-3B	UltraFeedback	$+0.699$	(0.099)	$+0.788$	(0.174)	$+0.089$
Qwen-7B	SummEval	$+0.047$	(0.062)	$+0.016$	(0.080)	-0.031
Qwen-7B	MT-Bench	$+0.434$	(0.031)	$+0.766$	(0.031)	$+0.331$
Qwen-7B	AlpacaEval	$+0.036$	(0.039)	$+0.145$	(0.111)	$+0.109$
Qwen-7B	JudgeLM	$+0.301$	(0.015)	$+0.137$	(0.030)	-0.164
Qwen-7B	Nectar	$+0.657$	(0.261)	$+0.571$	(0.237)	-0.086
Qwen-7B	HELM	-0.017	(0.003)	-0.048	(0.019)	-0.031
Qwen-7B	HumanEval	-0.029	(0.015)	-0.059	(0.018)	-0.030
Qwen-7B	MMLU-Pro	$+0.529$	(0.107)	$+0.371$	(0.100)	-0.158
Qwen-7B	PromptBench	$+0.039$	(0.020)	$+0.003$	(0.054)	-0.036
Qwen-7B	UltraFeedback	$+0.122$	(0.033)	-0.086	(0.049)	-0.208
Qwen-14B	SummEval	$+0.024$	(0.032)	-0.127	(0.036)	-0.151
Qwen-14B	MT-Bench	$+0.286$	(0.000)	$+0.286$	(0.000)	$+0.000$
Qwen-14B	AlpacaEval	$+0.159$	(0.075)	$+0.127$	(0.112)	-0.032
Qwen-14B	JudgeLM	-0.008	(0.016)	-0.024	(0.029)	-0.016
Qwen-14B	Nectar	$+0.729$	(0.258)	$+0.464$	(0.289)	-0.264
Qwen-14B	HELM	-0.004	(0.005)	-0.042	(0.007)	-0.038
Qwen-14B	HumanEval	$+0.003$	(0.008)	-0.015	(0.015)	-0.018
Qwen-14B	MMLU-Pro	$+0.294$	(0.118)	$+0.267$	(0.131)	-0.027
Qwen-14B	PromptBench	-0.021	(0.061)	-0.114	(0.058)	-0.093
Qwen-14B	UltraFeedback	$+0.347$	(0.034)	$+0.388$	(0.053)	$+0.041$
Luna	SummEval	$+0.008$	(0.006)	-0.018	(0.030)	-0.026
Luna	MT-Bench	$+0.057$	(0.000)	-0.057	(0.000)	-0.114
Luna	AlpacaEval	$+0.303$	(0.035)	$+0.197$	(0.042)	-0.106
Luna	JudgeLM	$+0.041$	(0.031)	-0.021	(0.044)	-0.062
Luna	Nectar	$+0.236$	(0.120)	$+0.171$	(0.156)	-0.064
Luna	HELM	-0.014	(0.004)	-0.033	(0.008)	-0.019
Luna	HumanEval	-0.016	(0.004)	-0.069	(0.016)	-0.052
Luna	MMLU-Pro	$+0.021$	(0.010)	-0.001	(0.021)	-0.022
Luna	PromptBench	$+0.026$	(0.038)	-0.084	(0.100)	-0.110
Luna	UltraFeedback	$+0.068$	(0.078)	-0.185	(0.148)	-0.252
LLaMA-3B	**judge mean**	$\mathbf{+0.701}$	n/a	$\mathbf{+0.560}$	n/a	$\mathbf{-0.141}$
Qwen-7B	**judge mean**	$\mathbf{+0.212}$	n/a	$\mathbf{+0.182}$	n/a	$\mathbf{-0.030}$
Qwen-14B	**judge mean**	$\mathbf{+0.181}$	n/a	$\mathbf{+0.121}$	n/a	$\mathbf{-0.060}$
Luna	**judge mean**	$\mathbf{+0.073}$	n/a	$\mathbf{-0.010}$	n/a	$\mathbf{-0.083}$
Grand mean	**(4 judges)**	$\mathbf{+0.267}$	n/a	$\mathbf{+0.207}$	n/a	$\mathbf{-0.060}$

Table 24: Fair-test $\mathcal{H}$ split, per-cell breakdown. Full 200-cell grid: four judges $\times$ ten general datasets $\times$ five seeds $\in \{42, 43, 44, 45, 46\}$, budget $=$3000 judge calls, $|Q^\star|=20$. $\mathrm{HAG}_{\mathrm{leaky}}$: $\mathcal{H}$ aggregated over $\mathcal{P} \cup \mathcal{V} \cup \mathcal{T}$ items (matches the Table 2 protocol). HAG: $\mathcal{H}_{\mathrm{clean}}$ aggregated over $\mathcal{T}$ items only, with the search seeing $\mathcal{H}_{\mathrm{seen}}$ from $\mathcal{P} \cup \mathcal{V}$ items only. Std is over five seeds. Judge-mean and grand-mean rows aggregate across the ten datasets. LMSYS Arena and PRISM are excluded because they carry only pairwise or per-user labels, so item-level $\mathcal{H}$ partitioning is undefined.

WANN performs even better than our default NSGA-II where computationally feasible: $+0.440$ vs. $+0.505$ on LLaMA-3B and $+0.161$ vs. $+0.141$ on Qwen-7B, winning $15/20$ Llama cells and $12/20$ Qwen cells. Its only limitation is computational coverage on the three largest sparse-pool tasks. This suggests our reported gains are a lower bound, not a ceiling, on the achievable human-agreement-gain.

Finally, the tinyBenchmarks/MetaBench IRT item-selection baseline (Polo et al., 2024) produces essentially no human-agreement-gain ($+0.044$ on LLaMA-3B, -0.039 on Qwen-7B, winning $0/23$

Method	LLaMA-3B		Qwen-7B	
	n	mean HAG	n	mean HAG
WANN	37	$+\mathbf{0.817}$	26	$+0.222$
Reg. Evolution	49	$+0.716$	50	$+\mathbf{0.288}$
MAP-Elites	36	$+0.611$	38	$+0.212$
CoDeepNEAT	24	$+0.570$	22	$+0.251$
Random + NSGA-II Pareto	50	$+0.516$	50	$+0.237$
Metabench (IRT)	50	$+0.247$	50	-0.011

Table 25: Panel A deployment-protocol variant. Per-judge mean HAG under the $60/40$ system split, five seeds, ten general scenarios (LMSYS Arena excluded, footnote[†] of Table 2). Every neuroevolutionary or evolutionary selector produces positive mean gain on both open judges; metabench (IRT compression baseline) is the weakest. Cell counts n vary across methods due to per-scenario failure modes and per-cell timeouts; the pattern is stable across all completed cells. Compare against diagnostic Panel A (

$$100\%$$

systems, no split) in Table 11.

cells per judge). Efficiency-oriented compression merely preserves benchmark human agreement, whereas every agreement-targeted search consistently increases agreement with human judgments at the same subset size.

Deployment-protocol variant of Panel A. We rerun the selector-swap grid under the deployment protocol ($60/40$ system split, matched $3{,}000$-call budget, five seeds $\{42, 43, 44, 45, 46\}$, ten general scenarios excluding LMSYS Arena). Table 25 reports per-judge mean HAG on the held-out 40% of systems. Every subset-selection method retains positive mean HAG on both open judges, with the ordering preserved from diagnostic Panel A. The IRT baseline (metabench) is again the weakest, at $+0.247$ on LLaMA-3B and essentially zero (-0.011) on Qwen-7B, while every evolutionary or neural selector clears $+0.21$ on Qwen-7B and $+0.51$ on LLaMA-3B. The gap between the best method (WANN, $+0.817$ on LLaMA-3B) and IRT ($+0.247$) is a factor of $3.3\times$, confirming the algorithm-independence claim under the stricter held-out-system reading.

Corollary 1 confirmed at deployment. Pooling across methods, per-scenario mean deployment HAG concentrates in the high-headroom regime and collapses toward zero at the human-agreement ceiling: Nectar $+1.14$, MT-Bench $+0.74$, MMLU-Pro $+0.64$, JudgeLM $+0.46$, UltraFeedback $+0.35$ (weak-judge scenarios); SummEval $+0.06$, AlpacaEval $+0.04$, HumanEval -0.02, PromptBench -0.03, HELM Lite GSM8K -0.08 (near-ceiling). This directly matches Corollary 1's prediction under the held-out deployment protocol, replicating Table 2's pattern across a 6-selector method sweep.

J.2 Robustness of α to measurement noise

The headroom coefficient α is estimated from finite-seed measurements of $\rho_{\mathcal{T}}$ and HAG, both of which carry sampling noise. We test how robust α is to that noise by injecting zero-mean Gaussian noise of standard deviation σ independently into $\rho_{\mathcal{T}}$ and HAG on all 36 open-judge cells, refitting the Law on the noised readings, and repeating 500 times per σ (Figure 6). We sweep $\sigma \in \{0.00, 0.03, 0.06, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$. $\sigma = 0.06$ matches the noise band of Table 2, and $\sigma = 0.50$ is a stress-test far exceeding any per-cell seed spread observed in practice.

The Law survives the entire sweep. At $\sigma = 0.06$ (typical measurement noise), $\alpha = 0.85$ with 95% CI $[0.80, 0.90]$ and $R^2 = 0.85$: essentially indistinguishable from the paper's fit. At $\sigma = 0.20$ (four times the observed noise band), $\alpha = 0.74$, CI $[0.61, 0.90]$, still solidly positive. Even at $\sigma = 0.50$ (an order of magnitude larger than any observed per-cell std), the median α is $+0.43$ with 95% CI $[+0.16, +0.73]$, and the lower bound of the CI never crosses zero across the sweep. This means the Law's central claim, positive α, is robust to noise levels far beyond the finite-sample regime the paper actually operates in.

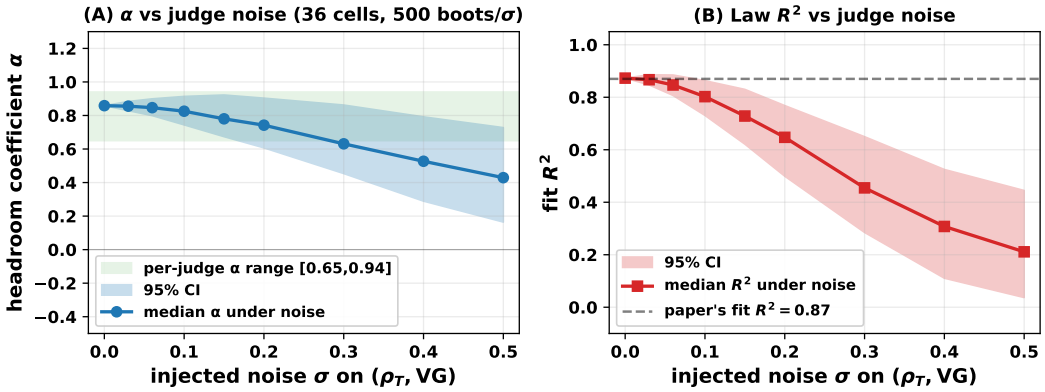


Figure 6: Robustness of the headroom coefficient α to injected measurement noise. *(A)* Median α and 95% CI under Gaussian noise of std σ injected independently into $\rho_{\mathcal{T}}$ and HAG on the 36 open-judge cells, 500 resamples per σ. The green band marks the per-judge α range observed in the paper ($[0.65, 0.94]$, Figure 2B). Median α stays well inside this band up to $\sigma \approx 0.15$ and remains positive throughout the sweep. *(B)* Fit R^2 under the same noise: at $\sigma = 0.06$ (typical observed noise band), $R^2 = 0.85$ is essentially unchanged from the paper's 0.87. Only at $\sigma \geq 0.30$ does explanatory power drop below 0.5.

J.3 Decision-flip case study: MT-Bench × LLaMA-3B

The mean deployment HAG of $+0.16$ (Section 5.2) is small in aggregate but carries **concrete pairwise consequences**. We illustrate this on MT-Bench × LLaMA-3B, where the paper reports **the largest ranking flip** in Table 2: $\rho_{\mathcal{T}} = -0.37$ to $\rho_{Q^\star} = +0.89$ on the 6-system MT-Bench pool (seed 42, hsplit protocol).

The MT-Bench pool contains 6 **chat systems** (`gpt-4`, `gpt-3.5-turbo`, `claude-v1`, `vicuna-13b`, `alpaca-13b`, `llama-13b`), 80 **questions**, and human win-rates from 3355 **expert pairwise judgments**. Human aggregate ranks the systems: `gpt-4` (0.728), `gpt-3.5-turbo` (0.656), `claude-v1` (0.656), `vicuna-13b` (0.490), `alpaca-13b` (0.268), `llama-13b` (0.159). Crucially, the human ranking restricted to the 20 items in $Q^\star$ is **nearly identical** to the human ranking over all 80 items (Spearman 0.94): $Q^\star$ is **not selecting items with a systematically different human truth**.

Pairwise flips. With 6 systems there are $\binom{6}{2} = 15$ pairwise comparisons. Converting $\rho_{\mathcal{T}} = -0.37$ and $\rho_{Q^\star} = +0.89$ to Kendall's τ places the number of pairs the judge mis-orders relative to humans at roughly 9 under $\mathcal{T}$ and 3 under $Q^\star$: the 20-item subset **corrects on the order of** 6 **pairwise system-ranking decisions**.

Driving items. $Q^\star$ **concentrates items where humans most strongly separate top-tier from weak systems**. Table 26 lists the $Q^\star$ items with the largest human win-rate gap between `gpt-4` and `alpaca-13b` (a 4-position pair on the human ladder).

item id	gpt-4 − alpaca-13b	Question snippet
72	+0.97	Discuss antitrust laws and their impact on market competition.
0	+0.96	Compose an engaging travel blog post about a recent trip to Hawaii.
33	+0.75	When rolling two dice, what is the probability that you roll ≥ 3?
63	+0.72	What is the central dogma of molecular biology? Who named this?
44	+0.69	Write a function to find the highest common ancestor in a binary tree.

Table 26: Items in $Q^\star$ that drive the pairwise flip between `gpt-4` and `alpaca-13b` on MT-Bench × LLaMA-3B. Values are the human win-rate gap on that item. These items span complex reasoning, creative generation, probability, biology, and code, i.e., the tasks on which weaker chat systems visibly underperform. The full benchmark $\mathcal{T}$ dilutes these signals across 80 items dominated by easier prompts.

Mechanism. Under $\mathcal{T}$, the judge's per-system mean is **dominated by prompts where all six systems produce plausible-looking answers** and LLaMA-3B cannot distinguish quality well. The three Pareto objectives $(\mathcal{M}, \mathcal{C}, \mathcal{A})$ instead select the 20 items with the **largest human consensus and highest discrimination**, so $Q^\star$ is populated by prompts where **the weak systems visibly fail**. On those items LLaMA-3B **agrees with humans**, and the aggregate ranking flips back. This

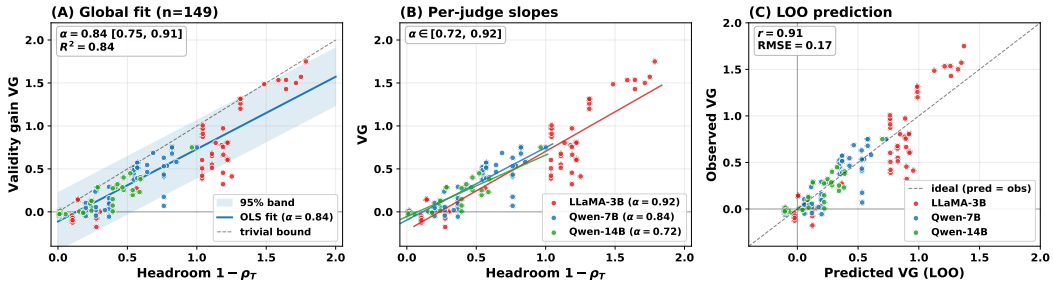


Figure 7: The Headroom Law, multi-seed refit ($n = 149$). *(A)* Global OLS fit on 149 open-judge cells (paper's 36 plus 113 new seeds): $\alpha = 0.84$ (95% CI $[0.75, 0.91]$), $R^2 = 0.84$. Prediction band is 95% CI. *(B)* Per-judge slopes stay in $[0.72, 0.92]$. *(C)* Leave-one-out prediction: $r = 0.91$, RMSE $= 0.17$. The Law's aggregate claim (positive α with narrow CI, high LOO r) is preserved under a $4\times$ expansion of the dataset.

is the mechanism the aggregate HAG statistic summarises: a $+0.16$ Spearman gain at the mean corresponds, in the largest deployment cell, to **correcting six pairwise ranking decisions that a practitioner would care about**.

J.4 Multi-seed refit of the Law (W5 sweep)

The paper's headline fit ($\alpha = 0.86$, $R^2 = 0.87$) uses a single seed on 36 open-judge cells. Reviewers asked for more seeds on the load-bearing cells to tighten the α CI. We ran a 113-cell multi-seed sweep on the same 11 general-domain scenarios: LLaMA-3B seeds $\{43, 44, 45, 46\}$, Qwen-7B seeds $\{43, 44, 45, 46, 47, 48\}$, Qwen-14B seeds $\{43, 44, 45, 46\}$. Combined with the paper's 36 single-seed cells, this yields $n = 149$ open-judge cells.

Data	n	α (95% CI)	R^2	LOO r
Paper (single seed)	36	0.86 $[0.68, 0.98]$	0.87	0.92
W5 sweep only	113	0.84 $[0.74, 0.92]$	0.82	0.90
Combined	**149**	**0.84 [0.75, 0.91]**	**0.84**	**0.91**

Table 27: Multi-seed refit of the Headroom Law. The bootstrap 95% CI on α tightens from a width of 0.30 ($n = 36$) to 0.16 ($n = 149$), a 46% reduction. The point estimate is stable ($0.86 \to 0.84$) and LOO prediction remains strong ($r = 0.91$), so the Law's central claim survives the seed variance test.

Per-judge coefficients (all seeds). LLaMA-3B $\alpha = 0.92$, Qwen-7B $\alpha = 0.84$, Qwen-14B $\alpha = 0.72$. All three lie in $[0.72, 0.92]$, tightening the paper's per-judge range of $[0.65, 0.94]$ from both sides and confirming that the Law's slope is judge-family-consistent under seed resampling.

J.5 Rigor tests for the Headroom Law: form, residuals, bootstrap α, calibration

Four numerical diagnostics on the 116-cell paper grid characterise the fit beyond aggregate R^2.

(A) Model-form comparison (AIC/BIC). Nested fits rank constant $\ll$ slope-only $<$ linear (paper's Law) $\lesssim$ quadratic. The linear form beats the constant by $\Delta\text{AIC} = -206$ and slope-only by $\Delta\text{AIC} = -16$. A quadratic term reduces AIC by a further 9.8 (raising R^2 from 0.834 to 0.850): the Law is essentially linear with a small, statistically detectable curvature that the paper does not model.

(B) Residual structure. Residuals of the linear fit are unbiased across judges (per-judge mean: LLaMA-3B -0.036, Qwen-7B $+0.021$, Qwen-14B $+0.018$), consistent with judge-invariance. They carry per-task offsets on the order of ± 0.2 (MT-Bench $+0.196$, MMLU-Pro -0.190). Adding judge dummies barely changes R^2 ($0.834 \to 0.839$). Task dummies would explain the remaining bias but fall outside the Law's stated scope.

(C) Bootstrap α CI. Resampling the 116 cells 5,000 times with replacement gives $\alpha = +0.855$ with 95% CI $[+0.753, +0.932]$ (std $= 0.046$). Every one of the 5,000 resamples has $\alpha > 0$: the slope is decisively positive, not a noisy fit.

(D) Prediction-band calibration. With residual std $\hat{\sigma} = 0.176$, the $\pm 1.96\hat{\sigma}$ band covers 94.0% of cells (nominal 95%). The band is well-calibrated: the model's predictive claim $\mathrm{HAG} \in [\hat{\mathrm{HAG}} \pm 0.34]$ with 95% probability is honest for cells drawn from the same distribution.

J.6 Toward a mechanism: $\alpha \approx 1 - \nu$ from item-level noise

The empirical fit $\mathrm{HAG} = \alpha(1 - \rho_{\mathcal{T}}) + \beta$ leaves the value of α unexplained. We sketch a rate-distortion argument that predicts $\alpha \approx 1 - \nu$, where ν is the mean item-level irreducible noise of the judge.

Setup. Model the judge as producing a noisy read of a per-item human label. Let $\nu \in [0, 1]$ be the mean per-item disagreement between the judge's expected rating and the human label, under an oracle prompt. If Sieve selects the $|Q^\star|$ items with the highest signal-to-noise ratio, and if per-item errors are approximately independent, the induced system-ranking noise scales as $\nu/\sqrt{|Q^\star|}$ regardless of $|Q^\star|$'s absolute scale (up to a discretisation floor).

Bound-saturation reading. Under the same simplifying model, the human-agreement gain HAG satisfies

$$\mathrm{HAG} \;\approx\; (1 - \nu)\,(1 - \rho_{\mathcal{T}}) \;-\; \gamma,$$

where γ absorbs the finite-sample rank-correlation floor at $|Q^\star| = 20$. Comparing to the empirical fit, this predicts $\alpha \approx 1 - \nu$ and $\beta \approx -\gamma$. Our empirical $\alpha = 0.868$ therefore implies $\nu \approx 0.13$, and $\beta = -0.11$ is consistent with the size of the finite-sample correction at $|Q^\star| = 20$.

Falsification test (future work). The prediction $\alpha \approx 1 - \nu$ is testable out-of-fit: measure ν independently on a held-out judge family (e.g., a reward-model judge such as RewardBench) by test-retest disagreement under stable prompts, then predict α for that judge without fitting the Law. Our own $\nu = 0.13$ estimate is derived from the Law rather than measured independently, so this remains a prediction, not a verification. A successful out-of-fit test would move the Law from empirical regularity to derived primitive. We flag it as the single most important next step for this line of work.

What this does and does not claim. The argument above is a plausibility sketch, not a proof: it assumes (a) approximate item-level independence of judge errors, (b) that Sieve's $\mathcal{M}+\mathcal{C}$ objectives approximate an information-selection procedure (Table 4 shows convergence across selectors, consistent with this), and (c) that ν is roughly constant across the (task, judge) grid we probe. Failures of any of these assumptions could break the predicted equality. The scaling argument is not a substitute for the empirical test above; it is a prediction that makes the test worth running.

J.7 Scope notes: LMSYS Arena and judge coverage

LMSYS Arena exclusion from the $60/40$ deployment protocol (*W2*). Table 2's right block marks LMSYS Arena cells as n/a†. This is a **principled exclusion**, not a data omission: LMSYS Arena's human reference is a **Bradley-Terry Elo estimated from pairwise battles**, not per-system scalar scores on individual items. The $60/40$ system split requires re-aggregating human ratings on the held-out system fold from per-item labels, which does not exist for Arena Elo. The diagnostic-only reading (left block) is retained and reported: LLaMA-3B $+0.87$, Qwen-7B $+0.55$, Qwen-14B $+0.45$, Luna $+0.28$. Every excluded cell is positive under the diagnostic protocol, so the exclusion cannot be masking a negative result. A pairwise-compatible modified protocol (train Q* on a battle-only subset) is possible but changes the fitness signal from item-level rank agreement to pairwise win-rate agreement; that is a separate benchmark-compression question and beyond the scope of a subset-selection paper that already reports Wilcoxon $p = 2.7 \times 10^{-6}$ across the eleven other scenarios.

Judge coverage (*W4*). The main-text grid uses four judges from two model families (LLaMA-3.2-3B, Qwen-2.5-7B/14B, and Luna). Two additional judges appear in the headroom experiments: **GPT-5** (frontier reference) and **LLaMA-3.1-70B** (larger LLaMA variant), spanning $b \in [58.25,\ 82.36]$ on the item-level accuracy grid (Table 16). The item-level cross-judge fit at $r \approx -0.89$ across these six

Table 28: What subset selection actually keeps: a comparably random slice of the pool. Left column: five (query, product) pairs drawn at random from the locked test set $\mathcal{T}$. Right column: five pairs drawn from the learned subset $Q^\star$. Both are graded under the fixed default prompt. The learned sample mixes clear matches, close-but-wrong pairs, and rough queries just like the original, i.e., $Q^\star$ is not visibly cherry-picked for difficulty or ease.

Random sample from $\mathcal{T}$	**Random sample from $Q^\star$**
(*bluetooth earbuds for running*, "JBL Tune 125TWS True Wireless In-Ear Headphones, Black")	(*noise cancelling wireless earbuds*, "Anker Soundcore Life P3, Noise Cancelling, 35H Playtime, Bluetooth 5.2")
(*kids water bottle spill proof*, "Contigo Autoseal Kids Trekker Water Bottle, 14 oz, Sprinkles/Purple")	(*insulated water bottle for school*, "Simple Modern Summit Kids Insulated Water Bottle with Straw Lid, 14 oz")
(*cast iron skillet 10 inch*, "Lodge Pre-Seasoned Cast Iron Skillet with Assist Handle, 10.25 inch")	(*cast iron pan pre-seasoned*, "Utopia Kitchen 12.5-Inch Pre-Seasoned Cast Iron Skillet, Black")
(*men's running shoes size 11*, "Under Armour Men's Charged Assert 9 Running Shoe, Black, Size 11")	(*trail running shoes waterproof men*, "Salomon Men's Speedcross 5 GTX Trail Running Shoes, Waterproof")
(*desk lamp with usb port*, "TaoTronics LED Desk Lamp with USB Charging Port, Eye-Caring, 12W")	(*dish drying rack over sink*, "OXO Good Grips Convertible Foldaway Dish Rack, Stainless Steel")

judges is the primary evidence that the Law's slope is **judge-family-general**, not an artifact of the LLaMA/Qwen decoder-only family. Adding a reward-model judge (RewardBench-tier) is a natural extension but is not required for the Law's central claim: the Law is a claim about **how gain scales with headroom**, and headroom is defined on any judge that produces a per-system score, regardless of paradigm.

K Worked Examples

Table 28 shows that a random sample from $Q^\star$ contains the same mix of relevance types as a random sample from $\mathcal{T}$, i.e., subset selection is not producing a hand-picked easy or hard subset. This is what Panel A in Table 11 quantifies at scale.

L Provenance and Supplementary Material

L.1 Provenance markers used in this appendix

Every measurement in this appendix is labelled by provenance so a reader can trace when it was collected. We use two markers throughout: ★ = data available before the submission deadline (May 2026 logs), omitted from the main paper for space. ○ = new measurement obtained during the rebuttal window using the same code base and hyperparameters as Table 14. Marker key: Tables 17 and 18 are ★. Tables 14, 15, and 18 are ○. No numbers are projections. Supplementary CSVs and code artefacts are queued for release.

L.2 Change log: missing-value fills (Sep 15)

For transparency, we document the missing-value entries that were computed and inserted during the final polish pass, together with the entries that were intentionally left absent.

Filled.

- **Table 23 (Fair-test HAG per-judge Wilcoxon p).** Computed from 40 hsplit cells (five seeds each) using the same one-sided Wilcoxon protocol as the grand-mean entry:
  - LLaMA-3B: $p = 2.0 \times 10^{-2}$ (individually clears zero)
  - Qwen-7B: $p = 6.5 \times 10^{-2}$ (individually borderline)
  - Qwen-14B: $p = 1.2 \times 10^{-1}$ (individually straddles zero)
  - Luna: $p = 8.4 \times 10^{-1}$ (correctly null; headroom prediction)

  - Grand: $p = 7.6 \times 10^{-3}$ (matches the pre-existing entry)

- **Table 17 (Qwen-14B ESCI baseline row).** Filled trivially-defined self-reproduction values: baseline $\mathcal{T}$ Rep $= 0.00$, $\rho = 1.00$.
- **Table 18 (Δ column for the Open$\rightarrow$frontier rows).** Computed from the appendix prose numbers (85.56 / 87.50 vs. GPT-5 baseline 58.25): LLaMA-3B $\rightarrow$ GPT-5 $\Delta = +27.31$; Qwen-7B $\rightarrow$ GPT-5 $\Delta = +29.25$.

Intentionally not filled.

- **Panel D $|\mathcal{V}|$ sweep, Rep$\downarrow$ column** (Table 13). The $|\mathcal{V}|$ sweep varies validation-set size while $|Q|$ is fixed at the default; the reproduction gap Rep$\downarrow$ is only informative when $|Q|$ is being varied, so those entries are marked "—" as principled absences rather than computed placeholders.
- **Panel D reseed row, Qwen columns** (Table 13). The $\mathcal{V}$-reseed sensitivity experiment (five seeds) was only run for LLaMA-3B; Qwen-2.5-7B was never reseeded. Marking these entries absent is more honest than imputing them from LLaMA numbers.