Welfare, Improvability, and Variance: A Principal–Agent Approach to Optimal Benchmark Item Aggregation

Andreas Haupt*
Department of Economics & Computer Science
Stanford University
h4upt@stanford.edu

Justin Hartenstein*
Institute for Computational and Mathematical Engineering
Stanford University
justinha@stanford.edu

Anka Reuel
Department of Computer Science
Stanford University
anka@cs.stanford.edu

Mykel J. Kochenderfer†
Department of Aeronautics & Astronautics
Stanford University
mykel@stanford.edu

Sanmi Koyejo†
Department of Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

AI benchmarks have well-documented limitations, with prior work examining contamination, saturation, and construct underspecification. Aggregation has received far less attention: benchmarks are typically summarized by uniformly averaging item-level scores, implicitly treating every test item as equally valuable. We model benchmarking as a multitask principal–agent game and show that the welfare loss from a benchmark is determined jointly by three item-level primitives: alignment with normative welfare priorities, marginal improvability, and performance variance. We translate the theory into an audit framework that ranks items along each of these three axes, and apply it to OLMES items using WORKBank for welfare, the EvoLM 4B suite for improvability, and the PolyPythias 410M panel for variance. The framework surfaces items that are Pareto-inferior within OLMES subject to a pro-worker welfare operationalization. All code is available at .

1 Introduction

AI benchmarks have well-documented limitations, with prior work examining contamination [65, 4], saturation [6, 28, 41], and construct underspecification [46, 51]. Aggregation has received far less attention: benchmarks are typically summarized by uniformly averaging item-level scores [35, 20], implicitly treating every test item as equally valuable. The design choices embedded in that compression—which items to include, and how to weight them—therefore determine which capabilities labs invest in next [42, 36, 29].

A growing literature asks what makes a good benchmark *item*, with two largely disjoint streams. The first applies item response theory and information-theoretic informativeness criteria, often paired

*Equal contribution.
†Equal senior authorship.

Preprint.

with construct-validity audits and item-error analysis to screen for questions that genuinely measure a coherent latent ability [35, 29, 61, 51]. A second stream documents that benchmarks are not neutral measuring instruments but incentive systems. Selective disclosure, contamination, sandbagging, and prompt-format gaming by labs systematically distort published rankings [56, 64, 49, 38]. The first literature improves benchmarks as measurement instruments; the second shows that benchmarks also function as incentive systems. What remains missing is an item-level account of benchmark quality that conditions on the strategic response induced by the benchmark itself.

This paper assesses benchmark items taking into account Lab incentives, grounded in a game-theoretic model. First, it solves a game-theoretic model of *optimal benchmark aggregation*. A benchmark designer commits to an item aggregation rule, a lab best-responds with costly and uncertain actions to improve their model—for example, additional pretraining, supervised finetuning, or reinforcement-learning-based finetuning. The model we propose is an adaptation of the multitask principal–agent framework of Holmstrom and Milgrom [23] to benchmarking. Using the framework, we show that the welfare-maximizing aggregation rule takes a closed form and is a combination of three item-level primitives: welfare alignment, improvability, and performance variance.

This theory suggests an auditing schema for benchmark items. First, an item must load on capabilities that are aligned with welfare-relevant concepts, so that effort spent improving it actually produces social value [57]. Second, the dimension must be improvable, since items along intractable axes generate effort-cost without movement. Third, it must have reliable performance in the sense that gains in true capability translate into measurable score gains, since otherwise risk-averse labs may shy away from risks of "disappointing" models [24].

We operationalize the optimal benchmark aggregation model. Welfare alignment is estimated from worker preferences over AI-affected tasks in WORKBank [55], mediated through O*NET Generalized Work Activities [39]. Improvability is estimated from per-item benchmark outcomes EvoLM 4B model suite [45] and cost estimates for investments from public sources. Variance is estimated from across-seed variance on the PolyPythias 410M pretraining panel [63].

We apply our analysis to the Open Language Model Evaluation Standard (OLMES) [17] item set, which contains, among others, the AI2 Reasoning Challenge-Challenge (ARC-Challenge) [12], HellaSwag [67], Massive Multitask Language Understanding (MMLU) [20], and Physical Interaction Question Answering (PIQA) [5]. Our analysis yields four findings. First, OLMES items very infrequently load on GWAs that workers would like to see improved. Second, math benchmarks (MATHQA, AGIEVAL-AQUA, MMLU-MATH) are welfare-dominated, none of them are on the worker welfare Pareto frontier subject to our operationalization. Third, improvability is nonuniform, and highest via supervised learning for half of items, reinforcement learning for a third, and pretraining for only 14%—this suggests that uniform aggregation would steer labs toward SFT-tractable items at the expense of pretraining-tractable ones, and that incentive-aware benchmarks would mitigate this distortion. Finally, adversarially constructed completion benchmarks (HellaSwag, PIQA, ARC-Challenge) are of meaningfully higher variance than other items in this panel [67, 5, 12, 50].

Related Work

Two literatures bear directly on the question of how benchmark items should be weighted. The first asks whether items *measure* what we claim they measure; the second asks how labs *respond* to being measured. Each is mature on its own terms, but neither speaks to the design problem we pose: choosing item weights when scores themselves induce costly, strategic capability investment by the producer.

The first strand treats the benchmark as a psychometric instrument and interrogates its construct and criterion validity. Item response theory and information-theoretic informativeness criteria have been used to ask whether items load on a coherent latent ability and to prune those that do not, with companion work on per-item scaling behavior [62, 53] and on the dimensionality of model capability more broadly [48]. Construct-validity audits, label-error analyses, and critiques of benchmark scope document the measurement-error side of the problem—noisy or mislabeled items, ambiguous targets, and ability constructs that do not survive scrutiny [40, 16, 61, 26, 46, 6, 28, 68]. This literature is essential, but its object of analysis is the item in isolation: an item is "good" if it is a precise, valid signal of an underlying ability.

The second strand takes the producer's behavior as the object of analysis. It documents the canonical Goodhart/Campbell pathology—once a measure becomes a target, it ceases to be a good measure—in the form of selective disclosure, contamination, sandbagging, and prompt-format optimization that distort published rankings [56, 19, 64, 13, 49, 69, 38, 14, 37, 59]. The lessons generalize known multitask-distortion findings from manager incentives [3], teaching-to-the-test [25], healthcare pay-for-performance [9], and policing statistics [15]: a single scalar index, applied to a multidimensional capability, predictably reallocates effort toward what is measured and away from what is valued. This work motivates calls for benchmark governance but stops short of prescribing how the index itself should be constructed.

Neither literature answers the design question. Measurement-quality work tells us which items are precise signals of ability; gaming work tells us that any aggregation will be optimized against. What is missing is an incentive-design treatment: given that the benchmark induces a costly strategic response, which items should receive weight, and how much? The criterion of a "good" item is not psychometric purity alone, nor freedom from gaming alone, but its contribution to the aggregator's welfare *conditional on* the producer's best response.

We close this gap by adapting the multitask principal–agent framework of Holmstrom and Milgrom [23]—in which a designer weighting noisy measures of effort faces an effort-substitution problem when measures vary in distortion and precision—to benchmark design. Our restriction to static, linear aggregation places us within a tradition studied in economics [22, 10]. The model identifies three primitives that jointly determine welfare loss under any linear aggregator: welfare-misalignment between item and downstream value, cost-asymmetry across items, and measurement noise. We then audit each primitive empirically. Cost is identified from per-item scaling fits on the EvoLM 4B model suite [45]. Noise—the irreducible run-to-run variation that disincentivizes producer investment—is estimated from the PolyPythias 410M multi-seed panel [63]. Welfare is anchored in worker-elicited preferences from WORKBank [55], mapped to items via human-validated loadings of generalized work activities. The result is a framework in which item-level audits and benchmark governance share a common currency: the welfare a weighted item delivers, net of the strategic response it elicits.

Outline. The remainder of the paper proceeds as follows. Section 2 develops the principal–agent model of benchmark aggregation and derives the closed-form optimal weights $v^\star = (M + r\Sigma)^{-1} M w$. Section 3 maps the model's primitives—welfare, improvability, and noise—to empirically tractable estimands. Section 4 describes the operationalization on OLMES items using WORKBank, the EvoLM 4B suite, and the PolyPythias 410M panel. Section 5 presents four findings on welfare alignment, math-vs.-general dominance, axis-level cost asymmetries, and adversarial-completion noise. Section 6 concludes with limitations and directions for future work. Notation is collected in Appendix A. Extensions of our theory are in Appendices B and C, details on cost estimation are in Appendix D and on welfare elicitation in Appendix G. Robustness checks are in Appendix F. Finally, Appendix E contains additional results on variance.

2 Game-Theoretic Analysis of Optimal Benchmark Aggregation

We present a game-theoretic model of benchmark aggregation. A Benchmark Designer chooses a benchmark *score* that linearly aggregates benchmark items $y_1, y_2, \ldots, y_n$, for example items from ARC-AGI [11]. A *Lab* best-responds by allocating effort across improvement activities $a_1, a_2, \ldots, a_m$, for example pretraining tokens and compute, supervised finetuning prompt-completion pairs and compute, or reinforcement-based finetuning. Denote $y \in \mathbb{R}^m$ the vector of items and $a \in \mathbb{R}^n$ the effort vector. The following model is simplified to be linear-quadratic for expositional purposes, an assumption we relax in Appendix B. We assume that the score is

$$y = Aa + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Sigma)$$

with a quadratic cost $c(a) = \frac{1}{2} a^\top C a$, where $C \succ 0 \in \mathbb{R}^{n \times n}$ and $\Sigma \succ 0 \in \mathbb{R}^{m \times m}$ are positive definite.

The Designer publishes a linear[3] benchmark score $s(y) \coloneqq v^\top y$, with aggregation weights $v \in \mathbb{R}^m$. The Lab has constant absolute risk aversion (CARA) preferences [44] with coefficient $r > 0$, $u(x) = -e^{-rx}$, which gives rise to utility from effort that has a mean-variance form, as we will see below, and which is a common choice for scale-invariant decisionmaking under risk. Risk aversion expresses concerns of the lab that effort may not pay off in the eventual result, as for example in unpredictable training runs [24].

Lab Effort Choice. As a feature of CARA and Gaussian noise, the Lab's optimization problem is equivalent to maximizing its mean score minus a risk premium given by the score variance [44]:

$$\mathbb{E}[s(y) \mid a] - c(a) - \frac{r}{2}\operatorname{Var}(s(y) \mid a) = v^\top Aa - \frac{1}{2}a^\top Ca - \frac{r}{2}v^\top \Sigma v.$$

The first-order condition gives $A^\top v - Ca = 0$ and hence a Lab effort of $a^\star(v) = C^{-1}A^\top v$. If we define the *item-level improvability* as $M \coloneqq AC^{-1}A^\top \in \mathbb{R}^{m\times m}$, then the induced performance of a benchmark, taking into account incentives, is $\mathbb{E}[y \mid a^\star(v)] = Aa^\star(v) = Mv$, and the induced effort cost is $\frac{1}{2}(a^\star(v))^\top Ca^\star(v) = \frac{1}{2}v^\top Mv$.

Aggregation Problem. Society values benchmark performance y according to welfare weights $w \in \mathbb{R}^m$ and is risk-neutral.[4] We take the designer to choose a benchmark weighting v to maximize total surplus: social value of induced model improvements minus real effort costs and the compensation-equivalent risk borne by the Lab. Equivalently, the designer internalizes the fact that noisy benchmarks require higher expected returns to induce the same investment from a risk-averse lab. The designer objective is:

$$W(v) \coloneqq \mathbb{E}\left[w^\top y \mid a^\star(v)\right] - c(a^\star(v)) - \frac{r}{2}v^\top \Sigma v = w^\top Mv - \frac{1}{2}v^\top Mv - \frac{r}{2}v^\top \Sigma v.$$

Hence, the reduced problem is $\max_{v\in\mathbb{R}^m} w^\top Mv - \frac{1}{2}v^\top Mv - \frac{r}{2}v^\top \Sigma v$, which is optimized by an optimizer satisfying $(M + r\Sigma)v = Mw$, yielding an optimal aggregator

$$\boxed{v^\star = (M + r\Sigma)^{-1}Mw.} \tag{1}$$

The matrix M summarizes how costly it is to improve tasks jointly: M_{jk} is large when there exist improvement activities $i \in \{1, 2, \ldots, n\}$ that raise both tasks j and k.

2.1 Interpreting the Optimal Weights.

The scalar version of (1) shows the features of the optimal benchmark aggregation. If $m = n = 1$ and the only entries of Σ and M are s and x respecitvely, it is

$$v^\star = \frac{xw}{x + rs}. \tag{2}$$

This means that the welfare w is scaled by a shrinkage factor governed by an improvability-to-noise ratio M/rs: when the activity is highly productive and effort cheap (large M), and when measurement is precise and the Lab is risk-tolerant (small rs), the Designer transmits welfare directly, $v^\star \approx w$. In the matrix model, the shrinkage factor $M/(M + rs)$ becomes the matrix operator $(M + r\Sigma)^{-1}M$.

We can formalize (2) in the matrix model through a change of variables if M and Σ share an orthonormal eigenbasis, $M = Q\operatorname{diag}(\lambda_1, \ldots, \lambda_m)Q^\top$, $\Sigma = Q\operatorname{diag}(\sigma_1, \ldots, \sigma_m)Q^\top$, the optimal equation decomposes into independent equations for what we call *concepts* [61]. In the concept coordinates $\hat{v}^\star = Q^\top v$ and $\hat{w} = Q^\top w$, the optimal weights are:

$$\hat{v}_j^\star = \frac{\hat{w}_j}{1 + \frac{r\sigma_j}{\lambda_j}}. \tag{3}$$

Equation (3) means that (i) higher welfare relevance (larger $\hat{w}_j$), (ii) lower improvement cost (larger λ_j), and (iii) higher measurement precision (smaller σ_j) all increase optimal aggregation weight.

[3] Restricting attention to linear aggregation and a one-shot game partly can be generalized. Holmström and Milgrom [22] show that when an agent chooses effort continuously over time and the Designer observes a vector of cumulative signals of the Lab's model, a linear contract in the terminal signal is optimal among all history-dependent contracts within a CARA utility Gaussian model. Linear contracts are also robust if the primitives C,A, and Σ are not known [10].

[4] We focus on the impacts of training investments that fail to result in improvements disincentivizing Lab investment. A model where society is itself CARA risk-averse leads to an identical rule.

3 Operationalization of Primitives

The primitives A, C, Σ, and w in Section 2 are stylized objects, but relate to existing literatures in item-level scaling laws, stochasticity of run results, and worker preferences on AI development.

Welfare. The welfare vector w is a normative object. In this paper, we adopt a *pro-worker* stance [1], which responds to the economic relevance of AI for work and widespread concerns regarding the impacts of AI on the labor market [8]. We do not claim that worker preferences exhaust social welfare. They omit consumers, firms, students, scientific users, safety externalities, and long-run innovation effects. Our goal is narrower: to show how a benchmark audit changes under one transparent, labor-centered welfare criterion. There are at least two ways of how AI capabilities can translate into work. AI can augment—help humans do a task—or automate them. There are many calls for human augmentation in tasks [18, 7]. At the same time there are tasks that workers prefer would be automated, as they are tedious, repetitive, or otherwise unrewarding to perform [55]. We will capture this difference by scoring items by how much they load on activities that humans prefer to be automated or augmented.

Improvability. Improvability, as captured by A and C in the model, is about how effort, for example pretraining tokens, supervised finetuning prompt-completion pairs, of reinforcement learning rollouts, translate into performance. It is intimately connected to (per-item) scaling laws [21, 62]. One object we will report is the marginal net improvement $\partial y_i / \partial a_k - c_k$, where c_k is the marginal cost of an improvement activity such as a price per prompt-completion pair.

Noise. The covariance Σ captures any result from variation from the choice of effort to a final outcome. Noise arises both in evaluation—sampled prompts, stochastic decoding, judge variation—and training variation, in particular the consequence of the random seed, which captures random data order, initialization, and stochastic gradients.

The empirical audit does not recover the full structural objects A, C and Σ and the normative object w. Instead, it estimates item-level proxies for the three forces identified by the model: welfare alignment, improvability, and measurement noise. These proxies are sufficient for identifying items that are weak along individual primitives.

4 Empirical Approach

We instantiate welfare, cost, and noise for tasks in the OLMES suite [17], which contains the Massive Multitask Language Understanding benchmark (MMLU; Hendrycks et al. 20), the AI2 Reasoning Challenge—Challenge set (ARC-Challenge; Clark et al. 12), MathQA, HellaSwag [67], CommonsenseQA [58], the Physical Interaction Question Answering benchmark (PIQA; Bisk et al. 5), and the AQuA subset of AGIEval. We instantiate concepts as automation and augmentation of generalized work activities [55], use a human-validated language-model-as-a-judge for concept loading, and take worker preferences over augmentation and automation from Shao et al. [55] as two separate concept welfare vectors. For cost we study the sample complexity of improving item-level score by a fixed amount on the EvoLM 4B model suite [45], whose models vary in pretraining tokens, supervised finetuning samples, and reinforcement-learning rollouts. For noise, we evaluate item-performance variation across PolyPythias 410M multi-seed checkpoints [63]. The following subsections detail our methodology.

4.1 Worker Welfare

To proxy w, we we use worker preferences regarding the automation and augmentation of work tasks collected in [55] from 1500 workers, yielding automation desire $A(t) \in [1, 5]$ and a level of human involvement $H(t) \in [1, 5]$ ranking from full automation (1) to essential human involvement (5). We aggregate tasks based on a hierarchical taxonomy of work maintained by the U.S. Bureau of Labor Statistics (O*Net, [39]). We use the highest level of aggregation, generalized work activities (GWAs), and consider the 27 cognitive GWAs due to their relevance to the benchmark items we consider. Examples from this category of GWAs are "Interpreting the Meaning of Information for Others" and "Scheduling Work and Activities". We drop three of the GWAs containing less than 5 tasks rated in [55]. For each GWA g, we derive an *automation share* $\mathrm{auto}(g)$ as the fraction of tasks

within GWA g on which many workers express strong automation desire. Concretely, we compute it as $\sum_{t \in g} A_w(t) \geq 3.5$, where the sum is over tasks t that belong to work activity g. For each benchmark item, we label the five GWAs the item most strongly loads on. We use a human-validated LLM-as-judge protocol [33] to scale this elicitation. Prompts and human evaluation are provided in Appendix G.

We compute a welfare weight based on the loaded GWAs

$$w_i^{\text{auto}} = \sum_{k=1}^{5} \lambda_k \text{auto}(g_{ik}).$$

Here, g_{ik} is the GWA that is loaded kth most for item i, and $\lambda_k = \frac{1}{k}$ gives a larger weight to more highly ranked GWAs. We compute the augmentation share w_i^{aug} analogously based on $H(t)$, and normalize both scores to $[1, 5]$.

4.2 Improvability

To estimate improvability, we use the EvoLM model suite [45], a family of 110 model checkpoints of transformer architectures with 0.5, 1, and 4 billion parameters that systematically vary pretraining tokens, continual pretraining composition, supervised fine-tuning, and reinforcement learning relative to a fixed anchor configuration. We use the 4 billion parameter architecture and one anchor checkpoint that uses $160\,\text{B}\,\text{token}$ tokens for pretraining, $8\,\text{B}\,\text{token}$ FineWeb-Edu and $42\,\text{B}\,\text{token}$ on FineMath, $100\,000$ supervised finetuning prompt-completion pairs trained for one epoch, and no reinforcement learning. Continual pretraining is constant, and models vary across their pretraining tokens (3 checkpoints), supervised finetuning prompt-completion pairs (9 checkpoints), and reinforcement learning rollouts (9 checkpoints). The checkpoints increase only in one parameter each; the full Cartesian product is not provided in EvoLM. We therefore will refer to *axes* $k \in \{\text{PT}, \text{SFT}, \text{RL}\}$ as the parts of EvoLM that vary pretraining, supervised finetuning, and reinforcement learning, respectively.

We evaluate OLMES on all of the checkpoints. As all of OLMES elicits single-token responses, we can evaluate the probability of correctness p_{im} of a checkpoint $m \in \text{EvoLM}$ on an item $i \in \text{OLMES}$ without sampling. To correct for difficulty of items, we fit per-item ordinary least squares regressions

$$p_{im} = \beta_{ik} + \beta_k D_{mk} + \varepsilon_{im}.$$

across the available checkpoints $m \in k$ along each axis $k \in \{\text{PT}, \text{SFT}, \text{RL}\}$, where D_m is the respective effort of m within axis k, for example, $D_{mk} = 100\,000\,\text{sample}$ within axis $k = \text{SFT}$. The different investments become commensurable by translating them into monetary units. We derive the costs of pretraining tokens, supervised-finetuning completions, and reinforcement learning rollouts in Appendix D at $\kappa_{\text{PT}}^{4\text{B}} \approx 8.2 \times 10^{-8}\,\text{USD}/\text{token}$, $\kappa_{\text{SFT}} \approx 2 \times 10^{-4}\,\text{USD}/\text{sample}$, $\kappa_{\text{RL}} \approx 5 \times 10^{-4}\,\text{USD}/\text{prompt}$. We compute the marginal improvability in monetary units as $c_k = \kappa_k / \beta_k$. The minimum cost $\arg\min_k c_k$ is the cheapest improving intervention and $\min_k c_k$ its cost. We exclude $i \in \text{OLMES}$ where $p_i \notin [0.05, 0.95]$ or $\hat{\beta}_{i,k} < 0$, leading to a total exclusion of 2917 items ($16.4\,\%$).

4.3 Variance

We estimate noise using the PolyPythias panel [63], which provides multiple independent realizations of the same pretraining run across seeds at sizes up to $410\,\text{M}$ parameters. To our knowledge, this is the only public model panel with multi-seed pretraining for multiple checkpoints. For each PolyPythias-410M seed $s = 1, 2, \ldots, S$ and checkpoint $m \in \text{PolyPythia}$ on a log-spaced trajectory, we compute the per-item probability p_{ims} as in Section 4.2. The per-item noise estimate is the pooled within-checkpoint across-seed empirical variance,

$$\hat{\sigma}_i^2 = \frac{1}{N - |\text{PolyPythia}|} \sum_{m \in \text{PolyPythia}} \sum_{s=1}^{S} \big(p_{ims} - \bar{p}_{im}\big)^2,$$

with $\bar{p}_{im}$ the across-seed cluster mean and $N = S|\text{PolyPythia}|$. Pooling across checkpoints assumes that across-seed variance is approximately constant across pretraining stages. We use $S = 8$ seeds and $K = 10$ checkpoints and exclude items with $p_{im} \notin [0.10, 0.90]$ and $\hat{\sigma}_i^2 < 1 \times 10^{-6}$ for numerical reasons. This leads to a total exclusion of 11 568 of 31 774 items ($36.4\,\%$).

5 Results

The methods described in Section 4 produce, for each OLMES item, two welfare (automation and augmentation), one cost, and one noise score. We provide four findings from them.

Finding 1: Few OLMES Items are Aligned with Worker Welfare

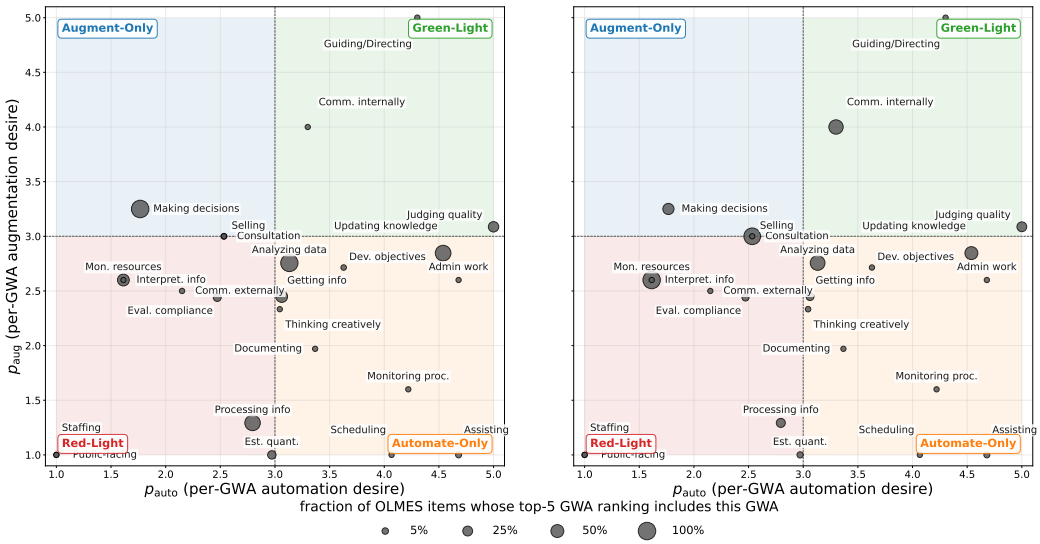


Figure 1: GWA loadings in the WORKBank welfare landscape, under automation-framed (left) and augmentation-framed (right) LLM rankings. Each point is one of 24 GWAs at its per-GWA welfare score; size encodes the fraction of OLMES items whose top-5 ranking includes that GWA. The GWAs in the green-light quadrant (top right), where workers desire both automation and augmentation, receive far fewer loadings than other quadrants.

Adapting [55], we call *green-light* those work activities where workers have both high automation and augmentation desire. This does not imply that workers simultaneously prefer full automation and human-led augmentation; rather, it indicates that workers view AI involvement favorably under either deployment mode. We map the 24 GWAs we use onto the WORKBank welfare landscape in Figure 2, Then, for each GWA we calculate the share of OLMES items that include the GWA in the item's top-5 ranking. Most items load heavily on the information-processing GWAs (*Getting Information*, *Analyzing Data*, *Processing Information*, and *Making Decisions / Solving Problems*), but only 6% of automation loading mass falls in the green-light quadrant, where workers want model improvement under either deployment mode. Under both framings, 27–29% of loading mass falls in the red-light quadrant—high-value activities where workers want little AI involvement or none at all.

Finding 2: Under Pro-Worker Welfare, General-Knowledge Dominate Math Benchmarks

Every item stemming from one of the math benchmarks (MathQA, AGIEval, MMLU-Math) is Pareto-dominated by at least one item from the general-knowledge benchmarks within our panel, and the welfare Pareto frontier contains zero math items (Figure 2, panel A). The empirical CDF dominance over the observed item set generalizes to first-order stochastic dominance on both welfare axes (panel B; cluster-bootstrap $p < 10^{-3}$ (augmentation), $p < 0.044$ (automation)): math is worse not just on average but at essentially every quantile threshold. This is consistent with math benchmarks loading on GWAs workers value less on both framings (processing information, solving problems), item-by-item, while general-knowledge benchmarks load on broader capabilities (using information, decision-making, communication) that rank high on both.

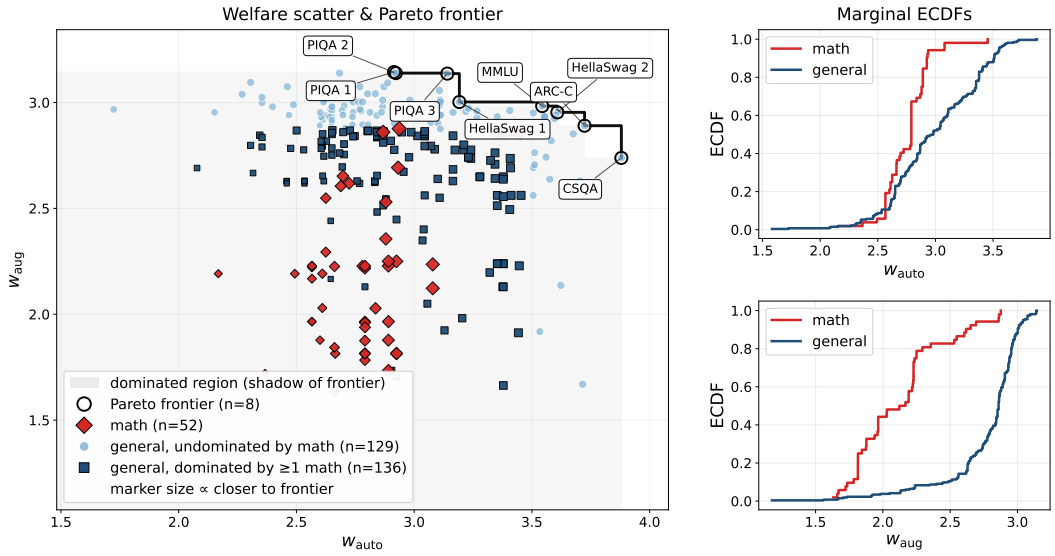


Figure 2: Under our pro-worker welfare operationalization, general-knowledge benchmark items dominate math benchmark items. Per-item welfare scores in $(w^{\text{auto}}, w^{\text{aug}})$ space (higher = higher AI desire; both axes in $[1, 5]$). *Left:* all $n = 317$ items; the global Pareto frontier ($n = 8$) contains general-knowledge benchmarks only, and $100\,\%$ of math items ($n = 52$) lie in its dominated shadow. *Right:* marginal empirical cumulative distribution functions show general items first-order stochastically dominating math on both axes.

Finding 3: SFT-Improvable Items Are Cheapest Under Our Dollar-Cost Imputation

Items vary widely in improvability (Table 1), and the cheapest axis differs across items: Supervised finetuning is the cheapest intervention for half of all items ($50.3\,\%$), reinforcement learning for a third ($35.4\,\%$), and pretraining for only $14.3\,\%$. The median cost for SFT-cheapest items is 8500 USD (interquartile range 4.5 k USD to 20 k USD), for RL-cheapest items 104 k USD (interquartile range 63 k USD to 196 k USD), for PT-cheapest items 250 k USD (interquartile range 135 k USD to 579 k USD). PT-cheapest items therefore cost a median 29 times more than SFT-cheapest items and 2.4 times more than RL-cheapest items (both pairwise Mann–Whitney $p < 10^{-200}$).

This is consistent with uniform aggregation steering resources toward SFT-tractable items at the expense of PT-tractable ones. Under uniform weights, the Lab's optimal effort allocation (per Section 2) concentrates on the items with the cheapest improving axis—predominantly SFT-tractable items in our panel.

Finding 4: Adversarial Completion Benchmarks Are Noisier Targets than Short-Form Multiple-Choice Questions in PolyPythia-410M

The per-item noise estimate $\hat{\sigma}_i$ spans almost an order of magnitude across the eligible item pool (Table 2), and correlates with benchmark membership: Within OLMES, HellaSwag, PIQA, and ARC-Challenge are overrepresented in the noisy tail, while MathQA, MMLU, and CommonsenseQA dominate the clean head. The pattern survives controlling for item difficulty via Bernoulli regression on $\hat{p}_i(1 - \hat{p}_i)$: HellaSwag retains about three times the residual noise of MMLU (Table 4 in Appendix E). One plausible mechanism is the log-probability gap between the gold answer and its closest distractor, which appears systematically larger in short-form multiple-choice questions than in adversarially-constructed benchmarks (HellaSwag, PIQA), where distractors were generated to be plausible to weak models, so small seed-induced shifts in token probabilities flip the model's preference (see Appendix E).

Table 1: Sample items from three quantiles according to improvability. Rank CI is at $2.5\,\%$ to $97.5\,\%$ based on 1000 residual bootstrap replicates of the per-item ordinary least squares. Axis denotes the cheapest intervention. Cost is the EvoLM cost required for a $10\,\%$ improvement.

Rank	Rank CI	Benchmark	Topic	Axis	Cost
10% quantile					
1489	[194, 13373]	HellaSwag	performing tricks with trained dogs	SFT	$385
1490	[1270, 1790]	MMLU/public relations	public relations communication theories	SFT	$385
1491	[1159, 1973]	MMLU/high school physics	calculating buoyancy and floating objects	SFT	$385
50% quantile					
7452	[5657, 14646]	MathQA	finding a number from a ratio and LCM	RL	$4.2k
7453	[6860, 9080]	MMLU/philosophy	Plato's theory of knowledge and forms	RL	$4.2k
7454	[5042, 14498]	ARC-C	independent variables on a graph	RL	$4.2k
90% quantile					
13415	[11484, 14719]	CSQA	finding shelter for human survival	PT	$36k
13416	[10117, 13899]	MMLU/prof. accounting	modifying report for financial stmts.	RL	$37k
13417	[11820, 14579]	MMLU/human sexuality	Freudian concepts of human sexuality	PT	$37k

Table 2: Sample items from three quantiles according to variance. Rank CI is at $2.5\,\%$ to $97.5\,\%$ based on 1000 across-seed bootstrap samples.

Rank	Rank CI	Benchmark	Topic	$\hat{\sigma}$	$\bar{p}$
10% quantile					
2019	[629, 5999]	MMLU/macroeconomics	increase in money demand factors	0.057	0.14
2020	[703, 5478]	MMLU/philosophy	Aquinas' views on pleasure and operation	0.057	0.13
2021	[404, 6542]	MMLU/nutrition	facial flushing, ALDH enzyme deficiency	0.057	0.16
50% quantile					
10100	[6456, 11376]	MMLU/college medicine	child psychology evaluation signs	0.078	0.19
10101	[1029, 12391]	MMLU/high school math	probability of divisibility by 2 and 3	0.078	0.12
10102	[5387, 11716]	CSQA	habitat and environment of crabs	0.078	0.22
90% quantile					
18183	[17279, 18875]	AGIEval-AQuA	average speed calculation for a plane	0.215	0.81
18184	[17714, 18609]	MathQA	total matches in knockout tournament	0.215	0.40
18185	[13869, 19161]	HellaSwag	treatment options for chicken pox scars	0.215	0.19

6 Conclusion

Summary. This paper develops a principal–agent theory of AI benchmark design and operationalizes it to understand the benefits of different items in the OLMES benchmark suite. We derive a closed-form expression for optimal benchmark aggregation in our theoretical model, $v^\star = (M + r\Sigma)^{-1} M w$ in which welfare alignment, improvability, and noise jointly determine how heavily each item should count. We translate the three primitives into separate empirical estimands and apply them to OLMES items using WORKBank, the EvoLM 4B suite, and the PolyPythias 410M panel. The audit surfaces item families that are weak along individual primitives and therefore would tend to receive lower weight in incentive-aware aggregation, all else equal.

Limitations. Several caveats bound the present work. First, the mapping from benchmark items to occupational exposure relies on an LLM-as-judge classification against O*NET GWAs; this inherits known pathologies of LLM raters. Second, our welfare functional w aggregates worker-side preferences only, excluding non-worker stakeholders. Third, the item-level tests underlying our rankings assume approximate independence across items conditional on model and seed; in practice, items within a benchmark share construction pipelines, annotators, and topical clusters, so reported p-values should be read as upper bounds on evidential strength. Fourth, the cost panel (EvoLM 4B) and noise panel (PolyPythias 410M) sit several orders of magnitude below frontier scale and come from different model families; the direction of bias from this scale gap is uncertain—scaling

could attenuate noise sensitivities (as capabilities saturate easy items) or amplify them (as frontier evaluations concentrate on harder items)—and we cannot adjudicate between these without a frontier-scale multi-seed replication, which is not publicly available. Finally, the model itself rests on linearity, single-Lab monopoly, and a common production technology across labs; Appendix B shows that the closed-form aggregation rule survives nonlinear μ and c as a fixed-point equation evaluated at equilibrium effort, and Appendix C shows that two-lab contest competition replaces r with a contest-implied constant but leaves the rule's algebraic shape intact.

Future work. Three directions extend the framework naturally. The most immediate is to commit to values of r and a welfare framing and report the single canonical $v^\star$, alongside its sensitivity to those choices. A second is theoretical: the model treats only a single benchmark designer, but in practice labs select which benchmarks to submit to and which scores to disclose [56], and the resulting two-sided selection problem—benchmarks compete for submissions while labs select for flattering weights—is unresolved. A third is infrastructural: the per-item welfare, cost, and noise scores compose naturally with EvalEval-style [47] eval cards as a metadata layer that downstream curators, regulators, and benchmark Designers can build on.

References

[1] Daron Acemoglu, David Autor, and Simon Johnson. Can we have pro-worker AI. *Choosing a path*, 2023.

[2] Amazon Web Services. Amazon bedrock pricing. `https://aws.amazon.com/bedrock/pricing/`, 2026. Accessed: 2026-05-06.

[3] George P. Baker. Distortion and risk in optimal incentive contracts. *Journal of Human Resources*, 37(4):728–751, 2002.

[4] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024. URL `https://aclanthology.org/2024.eacl-long.5/`.

[5] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

[6] 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)*, pages 4843–4855. Association for Computational Linguistics, 2021. URL `https://aclanthology.org/2021.naacl-main.385/`.

[7] Erik Brynjolfsson. The Turing trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287, 2022.

[8] Erik Brynjolfsson, Bharat Chandar, and Ruyu Chen. Canaries in the coal mine?: Six facts about the recent employment effects of artificial intelligence. Technical report, Stanford Institute for Economic Policy Research (SIEPR), 2025.

[9] Stephen Campbell, David Reeves, Evangelos Kontopantelis, Elizabeth Middleton, Bonnie Sibbald, and Martin Roland. Quality of primary care in England with the introduction of pay for performance. *New England Journal of Medicine*, 357(2):181–190, 2007.

[10] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–563, 2015.

[11] Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A new challenge for frontier AI reasoning systems, 2026. URL `https://arxiv.org/abs/2505.11831`.

[12] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

[13] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. In *arXiv preprint arXiv:2107.07002*, 2021.

[14] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation, 2025. URL https://arxiv.org/abs/2502.06559.

[15] John A. Eterno and Eli B. Silverman. *The Crime Numbers Game: Management by Manipulation*. CRC Press, 2012.

[16] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

[17] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5005–5033, 2025.

[18] Andreas Haupt and Erik Brynjolfsson. Ai should not be an imitation game: Centaur evaluations. In *Proceedings of the Forty-second International Conference on Machine Learning (ICML 2025)*, 2025.

[19] Chris Hays, Rachel Li, Bailey Flanigan, and Manish Raghavan. Strategic candidacy in generative ai arenas. *arXiv preprint arXiv:2603.26891*, 2026.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[21] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.15556.

[22] Bengt Holmström and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, 55(2):303–328, 1987.

[23] Bengt Holmstrom and Paul Milgrom. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(Special Issue):24–52, 1991. doi: 10.1093/jleo/7.special_issue.24. URL https://doi.org/10.1093/jleo/7.special_issue.24.

[24] Krystal Hu and Anna Tong. OpenAI and others seek new path to smarter AI as current methods hit limitations. Reuters, November 2024. URL https://www.reuters.com/technology/artificial-intelligence/openai-rivals-seek-new-path-smarter-ai-current-methods-hit-limitations-2024-11-11/.

[25] Brian A. Jacob and Steven D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3):843–877, 2003.

[26] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 375–385, 2021. doi: 10.1145/3442188.3445901.

[27] Youhe Jiang, Fangcheng Fu, Xiaozhe Yao, Taiyi Wang, Bin Cui, Ana Klimovic, and Eiko Yoneki. Thunderserve: High-performance and cost-efficient llm serving in cloud environments, 2025. URL https://arxiv.org/abs/2502.09334.

[28] 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 (NAACL-HLT)*, pages 4110–4124. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.naacl-main.324/.

[29] 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. URL https://arxiv.org/abs/2407.12844.

[30] Kai A. Konrad. *Strategy and Dynamics in Contests*. Oxford University Press, 2009.

[31] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

[32] Edward P. Lazear and Sherwin Rosen. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864, 1981.

[33] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

[34] Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://github.com/project-numina/aimo-progress-prize](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

[35] 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, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

[36] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.14992.

[37] David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

[38] 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*, 2024.

[39] National Center for O*NET Development. O*NET 30.2 database. U.S. Department of Labor, Employment and Training Administration, 2026. URL https://www.onetcenter.org/database.html.

[40] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[41] Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13, 2022.

[42] Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. URL https://arxiv.org/abs/2308.11696.

[43] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *Proceedings of Machine Learning and Systems (MLSys)*, 2023.

[44] John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1–2):122–136, 1964. doi: 10.2307/1913738.

[45] Zhenting Qi, Fan Nie, Alexandre Alahi, James Zou, Himabindu Lakkaraju, Yilun Du, Eric P. Xing, Sham M. Kakade, and Hanlin Zhang. EvoLM: In search of lost training dynamics for language model reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=B6bE2GC71a.

[46] 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*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

[47] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices, 2024. URL https://arxiv.org/abs/2411.12990.

[48] Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. In *Proceedings of NeurIPS*, 2024.

[49] 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*, pages 10776–10787, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-emnlp.722/.

[50] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. doi: 10.1145/3474381.

[51] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

[52] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[53] Rylan Schaeffer, Noam Itzhak Levi, Brando Miranda, and Sanmi Koyejo. Pretraining scaling laws for generative evaluations of language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ym33xJYINV.

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[55] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the u.s. workforce, 2025. URL https://arxiv.org/abs/2506.06576.

[56] Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=4Ae8edNqm0.

[57] Marilyn Strathern. Improving ratings: Audit in the British university system. *European Review*, 5(3):305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4. URL https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

[58] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4149–4158. Association for Computational Linguistics, 2019. URL https://aclanthology.org/N19-1421/.

[59] Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5), 2022.

[60] Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating AI for math with massive open-source instruction data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mTCbq2QssD.

[61] Sang T. Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Jude Perera, Chibuike Uwakwe, Benjamin W. Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in AI benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=SlhLRh810S.

[62] Sang T. Truong, Yuheng Tu, Rylan Schaeffer, and Sanmi Koyejo. Item response scaling laws: A measurement theory approach to generalizable neural performance prediction, 2026. URL https://openreview.net/forum?id=pIfopX18D1.

[63] Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella Biderman. Polypythias: Stability and outliers across fifty language model pre-training runs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bmrYu2Ekdz.

[64] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *arXiv preprint arXiv:2406.07358*, 2024.

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

[66] Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

[67] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4791–4800. Association for Computational Linguistics, 2019. URL `https://aclanthology.org/P19-1472/`.

[68] Hongli Zhou et al. Lost in benchmarks? Rethinking large language model benchmarking with item response theory. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[69] Kun Zhou et al. Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

A Notation Reference

Table 3 gives an overview of all notation used in the optimal benchmark aggregation problem.

Table 3: Notation used throughout the model.

Symbol	Space	Meaning
Primitives		
n	$\mathbb{N}$	Number of effort dimensions, e.g., pretraining, SFT
m	$\mathbb{N}$	Number of benchmark items
a	$\mathbb{R}^n$	Lab's effort vector across activities
y	$\mathbb{R}^m$	Vector of benchmark item outcomes
A	$\mathbb{R}^{m\times n}$	Technology matrix mapping effort to expected performance
ε	$\mathbb{R}^m$	Gaussian noise vector, $\varepsilon \sim \mathcal{N}(0,\Sigma)$
Σ	$\mathbb{R}^{m\times m}, \succ 0$	Noise covariance, benchmark sampling and seed variation
C	$\mathbb{R}^{n\times n}, \succ 0$	Quadratic effort-cost matrix
$c(a)$	$\mathbb{R}$	Lab's cost of effort, $c(a) = \frac{1}{2}a^\top Ca$
r	$\mathbb{R}_{\geq 0}$	Lab's CARA risk-aversion coefficient
w	$\mathbb{R}^m$	Society's welfare weights over items
Designer's instrument and Lab's response		
v	$\mathbb{R}^m$	Designer's linear aggregation weights
$s(y)$	$\mathbb{R}$	Published benchmark score, $s(y) = v^\top y$
$a^\star(v)$	$\mathbb{R}^n$	Lab's best-response effort, $a^\star(v) = C^{-1}A^\top v$
$W(v)$	$\mathbb{R}$	Total social welfare under weights v
$v^\star$	$\mathbb{R}^m$	Optimal aggregation weights, $v^\star = (M + r\Sigma)^{-1}Mw$
Reduced-form objects		
M	$\mathbb{R}^{m\times m}, \succeq 0$	Item-level cost structure, $M \coloneqq AC^{-1}A^\top$; M_{jk} large when shared activities raise items j and k jointly
Q	$\mathbb{R}^{m\times m}$, orthogonal	Concept eigenbasis (shared eigenbasis of M and Σ)
λ_j	$\mathbb{R}_{\geq 0}$	jth eigenvalue of M (improvability of concept j)
σ_j	$\mathbb{R}_{>0}$	jth eigenvalue of Σ (measurement noise on concept j)
$\hat{v}$	$\mathbb{R}^m$	Aggregation weights in concept coordinates, $\hat{v} = Q^\top v$
$\hat{w}$	$\mathbb{R}^m$	Welfare weights in concept coordinates, $\hat{w} = Q^\top w$

B Beyond Linear-Quadratic Utility

In Section 2 we worked with linear performance $\mu(a) = Aa$ and quadratic cost $c(a) = \frac{1}{2}a^\top Ca$. We now show that the optimal aggregation rule (1) survives essentially intact when these linear-quadratic assumptions are dropped: the only change is that the cost-structure matrix M is evaluated at the equilibrium effort, rather than being a global constant.

Setup. Fix a benchmark with m items and n activities. The lab chooses $a \in \mathbb{R}^n$ and produces $y = \mu(a) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\Sigma)$, where $\mu : \mathbb{R}^n \to \mathbb{R}^m$ is twice continuously differentiable. The Lab's effort cost is $c : \mathbb{R}^n \to \mathbb{R}$, also twice continuously differentiable and strictly convex. The designer publishes a linear score $s(y) = v^\top y$ and the Lab is CARA with risk-aversion coefficient $r \geq 0$. Society values items according to welfare weights $w \in \mathbb{R}^m$.

Lab's problem. Under CARA-Gaussian preferences the Lab's objective is

$$v^\top \mu(a) - c(a) - \frac{r}{2} v^\top \Sigma v.$$

The Lab's first-order condition is

$$\nabla \mu(a^\star)^\top v = \nabla c(a^\star). \tag{4}$$

Improvability. Define the Jacobian $J(a) \coloneqq \nabla \mu(a) \in \mathbb{R}^{m \times n}$ and the cost Hessian $H(a) \coloneqq \nabla^2 c(a) \in \mathbb{R}^{n \times n}$, both evaluated at $a^\star(v)$. Implicit differentiation of (4) gives the variation of Lab action in score,

$$\frac{\partial a^\star}{\partial v} = H_{\text{eff}}^{-1} J^\top, \qquad H_{\text{eff}} \coloneqq H - \sum_i v_i \nabla^2 \mu_i. \tag{5}$$

The induced sensitivity of expected scores to weights is

$$M(v) \coloneqq \frac{\partial\, \mathbb{E}[y]}{\partial v} = J \cdot H_{\text{eff}}^{-1} J^\top,$$

which we call the improvability matrix. In the linear-quadratic case $H_{\text{eff}} = C$ and $J = A$, so $M(v) = AC^{-1}A^\top$ recovers the matrix from Section 2—and is independent of v.

Designer's problem. Total welfare under v is

$$W(v) = w^\top \mu(a^\star(v)) - c(a^\star(v)) - \frac{r}{2} v^\top \Sigma v.$$

Differentiating in v using (5) and the Lab's FOC (4),

$$\nabla W(v) = M(v)(w - v) - r\Sigma v.$$

Setting $\nabla W(v^\star) = 0$ gives

$$\big(M(v^\star) + r\Sigma\big) v^\star = M(v^\star) w, \tag{6}$$

with solution

$$v^\star = \big(M(v^\star) + r\Sigma\big)^{-1} M(v^\star) w. \tag{7}$$

This has the same algebraic shape as Equation (1), with the global M replaced by its local counterpart $M(v^\star)$. In the linear-quadratic case $M(v^\star) = AC^{-1}A^\top$ for all v and (7) is an explicit closed form; in general it is a fixed-point equation. The concept interpretation carries through whenever $M(v^\star)$ and Σ share an orthogonal eigenbasis.

Remark. A leading non-linear case is $\mu_i(a) = \phi_i(b_i^\top a)$ for an increasing concave link ϕ_i (e.g. a sigmoid past its inflection, or a power-law-with-floor). Then $\nabla^2 \mu_i = \phi_i''(b_i^\top a) b_i b_i^\top \preceq 0$ on the concave region, so $H_{\text{eff}} \succeq H$ and $M(v^\star) \preceq J H^{-1} J^\top$. The interior-maximizer assumption holds and (7) applies. By contrast, on the *convex* half of the sigmoid—near a phase transition or emergent-capability threshold [52]—we have $\phi_i'' > 0$, so positive weight $v_i > 0$ pushes H_{eff} toward singularity and the Lab's objective can develop multiple maximizers. In that regime the first-order condition ceases to characterize lab behavior, and the framework of this section breaks down.

C Lab Competition

The baseline model treats a single Lab with payoff linear in the score. We now introduce both *competition*—two symmetric labs racing on the same benchmark—and *ordinal* payoffs, capturing the leaderboard reality that the top model captures disproportionate attention, API traffic, and follow-on contracts. The closed-form aggregation rule of Equation (1) survives, with risk aversion r replaced by a contest-implied constant.

Two symmetric Labs $k \in \{1, 2\}$ simultaneously choose effort vectors $a^k \in \mathbb{R}^n$ at quadratic cost $\frac{1}{2}(a^k)^\top C a^k$, $C \succ 0$, producing scores $y^k = Aa^k + \varepsilon^k$ with $\varepsilon^k \sim \mathcal{N}(0, \Sigma)$ i.i.d. across Labs. The Designer publishes weights v and awards prize w_H to the higher score and $w_L < w_H$ to the loser; Labs are risk-neutral. Since win probabilities are scale-invariant in v, we normalize $v^\top \Sigma v = 1$.

Lab Equilibrium. A profile (a^1, a^2) is a (pure-strategy) Nash equilibrium if each Lab's choice maximizes its own expected prize net of cost given the opponent's choice:

$$a^k \in \arg\max_{a \in \mathbb{R}^n} \mathbb{E}\left[w_H \mathbb{1}\{v^\top y^k > v^\top y^{-k}\} + w_L \mathbb{1}\{v^\top y^k \le v^\top y^{-k}\}\right] - \tfrac{1}{2} a^\top C a, \quad k = 1, 2.$$

Lab 1's score advantage $v^\top(y^1 - y^2)$ has Gaussian noise with variance $2v^\top \Sigma v = 2$, so its win probability is $\Phi(\frac{1}{\sqrt{2}} v^\top A(a^1 - a^2))$. We restrict attention to the symmetric equilibrium $a^1 = a^2 = a^\star$; at any symmetric profile $\phi(0) = 1/\sqrt{2\pi}$, so the first-order condition is

$$a^\star(v) = \beta C^{-1} A^\top v, \qquad \beta \coloneqq \frac{w_H - w_L}{2\sqrt{\pi}}. \tag{8}$$

The FOC characterizes the symmetric stationary point. Sufficiency in difference-form tournaments with normal noise requires the equilibrium argument of Φ to lie in its concave region, which holds at the symmetric profile where $v^\top A(a^1 - a^2) = 0$. For sufficient conditions for sufficiency of the first-order condition, we refer the reader to Lazear and Rosen [32] and Konrad [30, Ch. 3].

Optimal Aggregation. Only the winning Lab's model is deployed, so welfare is $W(v) = \beta w^\top M v - \beta^2 v^\top M v$ with $M \coloneqq A C^{-1} A^\top$ (one unit of expected deployment value, two units of summed effort cost). Maximizing subject to $v^\top \Sigma v = 1$ with multiplier μ, and writing $\tilde{r} \coloneqq \mu/\beta^2$,

$$v^\star = \tfrac{1}{\beta}\big(M + \tilde{r}\Sigma\big)^{-1} M w. \tag{9}$$

This has the same shape as Equation (1), with $\tilde{r}$—the shadow price of the score-scale constraint—replacing r.

D Cost Estimates

This appendix estimates the marginal cost in USD for pretraining, supervised finetuning, and reinforcement learning-based finetuning of EvoLM's $4\,\mathrm{B}$ parameter model. All values are estimates for consumer cloud and do not attempt to estimate frontier cloud training.

Cost of a FLOP. EvoLM's pretraining and reinforcement learning stages both consume GPU-hours; we anchor both to commodity A100-80GB cloud pricing of $1.5\,\mathrm{USD/h}$ [e.g., 27] at Nvidia A100 brain float 16-bit (BF16) peak throughput of 3.1×10^{14} floating-point operations per second. Assuming a typical $40\,\%$ model-FLOPs-utilization for dense training, the cost of a FLOP is

$$\frac{1.50\,\mathrm{USD/h}}{3.1 \times 10^{14}\,\mathrm{FLOP/s} \cdot 0.40 \cdot 3600\,\mathrm{s/h}} \approx 3.4 \times 10^{-18}\,\mathrm{USD/FLOP}$$

This figure prices both pretraining and reinforcement learning computation. Reinforcement learning rollouts are priced at a lower effective model FLOP utilization (see below).

Pretraining cost per token. The Chinchilla rule [21] gives training FLOPs per token of $6N$ for a model with N parameters. Combining with the cost per FLOP above,

$$\kappa_{\mathrm{pt}}^{1\mathrm{B}} = 6 \cdot 10^9 \cdot 3.4 \times 10^{-18} \approx 2.0 \times 10^{-8}\,\mathrm{USD/token}$$
$$\kappa_{\mathrm{pt}}^{4\mathrm{B}} = 6 \cdot 4 \times 10^9 \cdot 3.4 \times 10^{-18} \approx 8.2 \times 10^{-8}\,\mathrm{USD/token}$$

For reference, training EvoLM-4B on its $320\,\mathrm{B\,token}$ pretraining budget would cost about $26\,000\,\mathrm{USD}$ at this rate, consistent with academic-scale $4\,\mathrm{B}$-model training-cost estimates.

SFT cost per sample. EvoLM's supervised finetuning data is constructed from a mixture of MetaMathQA [66], OpenMathInstruct-2 [60], and NuminaMath [34], all math-targeted instruction datasets generated by frontier LLMs. The largest and most recent of these, OpenMathInstruct-2, was synthesized using Llama-3.1-405B-Instruct. We price the marginal replacement cost of one math problem-solution pair by treating its generation as an API call at frontier quality. At Amazon Bedrock pricing for Llama-3.1-405B of $2.4\,\mathrm{USD/token}$ [2] a typical math problem-solution pair of 50 input and 200 output tokens costs

$$\kappa_{\mathrm{SFT}} \approx 50 \cdot 2.40 \times 10^{-6}\,\mathrm{USD} \cdot 0.75 + 200 \cdot 2.40 \times 10^{-6}\,\mathrm{USD} \cdot 0.25 \approx 2 \times 10^{-4}\,\mathrm{USD/sample}.$$

For the $500\,000$-sample supervised finetuning pool used in EvoLM, this implies a replacement cost of about $100\,\mathrm{USD}$. We use API generation cost as the anchor because it represents what a Lab would pay at the margin to acquire one additional sample at the supervised finetuning recipe's quality level.

Reinforcement learning cost per prompt. For verifiable-reward reinforcement learning on math (the EvoLM setup [45], using Proximal Policy Optimization [54] with a binary correctness reward), each prompt-rollout-update consists of generating K rollouts of G tokens followed by a gradient update on the resulting KG tokens. With representative settings $K = 8$, $G = 512$ at the 4B scale:

$$\text{rollout (forward only)} : KG \cdot 2N \approx 3.3 \times 10^{13}\,\text{FLOP}$$
$$\text{gradient update} : KG \cdot 6N \approx 1.0 \times 10^{14}\,\text{FLOP}$$

Rollouts are autoregressive generation, which typically runs at lower model-FLOPs-utilization than dense training due to memory-bandwidth limits during decoding [43]. Pricing rollouts at the lower effective MFU and gradient updates at the dense-training MFU,

$$\kappa_{\text{RL}} \approx 1 \times 10^{-3}\,\text{USD}/\text{prompt}.^{5}$$

Verifier cost is negligible for verifiable-reward math (an answer-check function on the model's output).

E Base Rate's Influence on Variance

Table 2 showed that raw seed-to-seed noise $\hat{\sigma}_i$ is concentrated on HellaSwag, PIQA, and ARC-Challenge, and is smallest on MMLU, CommonsenseQA, and MathQA. We now ask whether this ranking reflects a genuine difference in benchmark stability, or a mechanical artifact of where the model's accuracy sits.

Setup. For each item i we observe a binary correctness score $y_{ims} \in \{0, 1\}$ across $S = 8$ seeds and $K = 10$ checkpoints. Write $\hat{p}_i$ for the pooled mean and $\hat{\sigma}_i$ for the pooled across-seed standard deviation, both as defined in Section 4.3. Items are scored by acc norm, the standard length-normalized scoring rule for multiple-choice items: the model's prediction is the candidate c that maximizes the per-character log-probability $\overline{\log p}(c) \coloneqq \log p(c \mid \text{prompt})/\text{len}(c)$, and $y_{ims} = 1$ iff this argmax is the gold answer.

Two candidate mechanisms. A binary score has variance $\hat{p}(1 - \hat{p})$ that peaks at $\hat{p} = 0.5$ and shrinks toward the tails. So benchmarks on which the model is near chance ($\hat{p} \approx 0.25$ for 4-way multiple choice) will mechanically look quieter than benchmarks on which the model is near $\hat{p} = 0.5$, regardless of any genuine stability difference. We strip this out by normalizing:

$$Z_i \coloneqq \frac{\hat{\sigma}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}.$$

Pure binomial sampling at $K = 10$ would give $Z_i \approx 1/\sqrt{K} \approx 0.32$; values above this indicate excess seed-to-seed noise beyond the sampling floor.

A second mechanism operates on items the model gets right. The gap between the gold answer's score and the best distractor's score is

$$m_i \;\coloneqq\; \text{median}\, \frac{\overline{\log p}_{\text{gold}}}{\overline{\log p}_{\text{runner-up}}}$$

measured in nats per character. Items with small m_i sit near the acc norm decision boundary, where small seed-induced shifts in continuation likelihoods can flip the argmax.

Bernoulli component. Table 4 shows that part of the raw $\hat{\sigma}$ ranking is mean-driven. PolyPythia-410M is near chance on MMLU, CSQA, and MathQA (median $\hat{p} \in [0.15, 0.24]$) but close to $\hat{p} = 0.5$ on HellaSwag and PIQA, which mechanically inflates $\hat{\sigma}$ on the latter two. Normalization collapses MMLU, CSQA, and MathQA to $Z_i \approx 0.20$, *below* the binomial floor of 0.32, so their seed-to-seed answers are essentially deterministic at this scale. HellaSwag, by contrast, sits at $Z \approx 0.62$—about twice the floor—so its excess noise survives the mean correction.

[5] We assume $1.50\,\text{USD}/\text{h}$ per GPU, A100-class peak throughput of $3.1 \times 10^{14}\,\text{FLOP}/\text{s}$ (BF16), training MFU of 0.40, and rollout MFU of 0.20 (achievable with continuous-batched inference, e.g. vLLM [31]). The settings $K = 8$ and $G = 512$ are representative for short-CoT math RL; the exact EvoLM hyperparameters are immaterial to this order-of-magnitude estimate.

Table 4: Decomposition of per-family seed-to-seed noise on the PolyPythia-410M panel ($S = 8$ seeds, $K = 10$ checkpoints). *Bernoulli component*: median $\hat{p}_i$, median $\hat{\sigma}_i$, and the Bernoulli-normalized $Z_i = \hat{\sigma}_i / \sqrt{\hat{p}_i(1-\hat{p}_i)}$. The binomial floor at $K = 10$ is $Z \approx 0.32$. *Margin component*: on items the model gets right (n_{won}), median margin m_i in nats per character, and within-family Pearson correlation r between m_i and Z_i.

		Bernoulli component			Margin component		
Family	n	$\hat{p}$	$\hat{\sigma}$	Z	n_{won}	m (nats/char)	$r(m, Z)$
HellaSwag	2029	0.512	0.261	0.615	1034	0.003	−0.08
PIQA	837	0.592	0.208	0.475	494	0.031	0.01
ARC-C	549	0.310	0.179	0.421	140	0.012	−0.18
MathQA	1889	0.235	0.092	0.219	257	0.097	−0.05
MMLU	13 732	0.161	0.071	0.198	2566	0.404	−0.57
CSQA	963	0.154	0.070	0.195	110	0.331	−0.41

Margin component. The remaining excess noise on HellaSwag, PIQA, and ARC-Challenge tracks m_i. HellaSwag's median margin is 0.003 nats per character; MMLU's is 0.404, two orders of magnitude larger. The within-family correlation is more telling than the medians. On MMLU, m_i correlates with Z_i at $r = -0.57$: tight-margin items are reliably the noisy ones, as expected if margin sets the per-item noise floor. On HellaSwag this correlation collapses to $r = -0.08$, consistent with adversarial filtering having pushed *all* items to a tight-margin ceiling, so margin no longer separates noisy from quiet items.

The residual seed-to-seed noise on HellaSwag, PIQA, and ARC-Challenge therefore traces to benchmark construction rather than to model capability or item-format artifacts: distractors are by design semantically close to gold in log-probability space, so small seed-dependent perturbations flip the acc norm argmax.

F Robustness of the Welfare Pipeline

Our welfare ranking depends on several pipeline choices: which LLM serves as the GWA-loading judge, how many GWAs are ranked per item ($k = 5$ in the headline), how we weight GWAs within the top-k ($\lambda_k = \frac{1}{k}$ in the headline), and how we aggregate GWA scores from on task-level preferences. We probe sensitivity to these choices in three ways. First, we replicate Finding 1 on a human-labeled subset of 28 OLMES items, using two independent human annotators. Second, we replicate Finding 2 on the same subset under each rater's labels. Third, we sweep over four pipeline variants—truncation depth, within-k weighting, judge identity, per-GWA scoring—and report Spearman rank correlation of the resulting welfare rankings against the headline ranking.

Replication of Finding 1 on the human-ranked subset. On the 28-item human-ranked subset, we recompute the per-item quadrant assignment under both human annotators and the LLM judge. Table 5 reports the share of items in each WORKBank quadrant under each ranker-framing combination.

The qualitative pattern—most items load on the red-light or automate-only quadrants, with green-light loadings substantially smaller—holds across all three rankers and both framings. Quadrant mass differs across rankers: the LLM places more items in the green-light quadrant than either human under augmentation (19.4 % vs. 13.4 % and 21.3 %), and A1 places more items in augment-only than A2 (21.7 % vs. 23.5 % under automation; 24.4 % vs. 23.0 % under augmentation). The differences sit within the inter-human envelope on most cells.

Replication of Finding 2 on the human-ranked subset. Under all three rankers and both framings, the finding of math items being Pareto inferior to general-knowledge items on our panel replicates on the human-ranked subset, see Table 6. The welfare Pareto frontier on this subset contains zero math items across all ranker-framing combinations, and all math items are dominated by at least one general-knowledge item. The mean welfare gap between general-knowledge and math items is positive on both axes for all rankers, ranging from 0.56 to 2.23 on w^{auto} and from 0.09 to 1.37 on w^{aug}.

Table 5: Quadrant share (%) of OLMES items under three rankers (A1, A2, three-shot-prompted Claude Opus 4.5, C), on the 28-item human-ranked subset, under both automation and augmentation framings.

Framing	Ranker	Green	Auto-only	Augment-only	Red
Automation	A1	8.5	49.1	21.7	20.8
	A2	10.2	51.0	23.5	15.3
	C	4.4	41.6	20.4	33.6
Augmentation	A1	13.4	32.3	24.4	29.9
	A2	21.3	25.4	23.0	30.3
	C	19.4	26.1	26.1	28.4

Table 6: Replication on the human-ranked subset across three rankers. Δw^{auto} and Δw^{aug} report the mean welfare gap between general-knowledge and math items. "Frontier (math/total)" counts math items on the Pareto frontier out of total frontier items on the subset. "% math dominated" is the fraction of math items Pareto-dominated by at least one general-knowledge item.

Ranker	Δw^{auto}	Δw^{aug}	Math on frontier	% math dominated
A1	2.23	0.09	0 / 2	100
A2	1.83	0.80	0 / 3	100
C	0.56	1.37	0 / 3	100

Sensitivity to pipeline choices. We sweep over four variants of the welfare aggregation pipeline: (i) varying the top-k truncation depth ($k = 3$, $k = 4$, $k = 5$ with reciprocal-rank weighting); (ii) replacing reciprocal-rank weighting with uniform within-k weighting (at $k = 5$); (iii) replacing threshold-based per-GWA scoring ($\mathbb{P}[A_w \geq 3.5]$) with mean-based scoring ($\bar{A}_w$); and (iv) substituting the LLM judge from Claude Opus 4.5 to GPT-5.2 (at $k = 5$ with reciprocal-rank weighting). Spearman rank correlation of each variant's welfare ranking against the headline ranking is reported in Table 7.

Two patterns hold. First, the ranking is stable to truncation depth: $\rho = 0.91$ at $k = 3$ and $\rho = 0.95$ at $k = 4$ on w^{auto}, with similar stability on w^{aug}. The headline finding does not depend on the specific choice of k. In addition, switching from threshold-based to mean-based per-GWA scoring retains a strong but not perfect correlation ($\rho = 0.71$ on w^{auto}, $\rho = 0.80$ on w^{aug}): the threshold preserves discriminative signal that the central-limit averaging in $\bar{A}_w$ partly compresses. Second, the ranking is sensitive to two choices we view as substantive rather than incidental: (i) uniform within-k weighting ($\rho = 0.33$ on w^{auto}) discards the rank information that reciprocal weighting treats as load-bearing under automation, while retaining it under augmentation ($\rho = 0.88$ on w^{aug}). This indicates that the rank-1 GWA carries more differentiating signal under automation than under augmentation: w^{auto} depends substantially on rank-1, while w^{aug} is distributed more evenly across the top-5. (ii) Switching to a different LLM judge ($\rho = 0.67$ on w^{auto}, $\rho = 0.50$ on w^{aug}) introduces the same annotator-level variation documented in the human-judge calibration (Appendix G). Within the family of choices that share our load-bearing assumptions (rank-inverse weighting, single judge), the welfare ranking is stable.

G Details of GWA Assignment

For concept loading, we rely on a language model judge. We validate the protocol against two independent human annotators (denoted A1 and A2), who each ranked the 24 cognitive GWAs by relevance for 30 OLMES items. We then elicit comparable rankings from Claude Opus 4.5 under a 3-example few-shot prompt and report pairwise agreement under both the automation and augmentation framings. To benchmark the LLM judge against the inter-human baseline, we compute nDCG@5 (linear and binary relevance), Precision@4, Recall@4, Jaccard@5, and Rank-Biased Overlap (RBO at $p = 0.7$) for all three ranker pairs. Macro-averages are reported in Table 8.

Two patterns are worth flagging. First, inter-human agreement on OLMES items is moderate (nDCG@5 of 0.72 under automation, 0.59 under augmentation), which is consistent with augmentation being a less sharply defined construct than automation—a pattern we also see when

Table 7: Robustness of per-item welfare scores to the ranking aggregation rule and to the LLM judge. Each row reports Spearman rank correlations between w_i^{auto} (resp. w_i^{aug}) under the variant specification and under the baseline ($k = 5$, reciprocal weights $\lambda_k = 1/k$, Claude Opus 4.5). Top three rows perturb the aggregation rule; bottom row swaps the LLM judge.

Variant	$\rho(w^{\text{auto}})$	$\rho(w^{\text{aug}})$
Aggregation rule (Claude Opus 4.5 judge)		
$k = 3$, reciprocal weights	0.91	0.87
$k = 4$, reciprocal weights	0.95	0.89
$k = 5$, uniform weights	0.33	0.88
mean-based per-GWA agg.	0.71	0.90
LLM judge ($k = 5$, reciprocal weights)		
GPT-5.2	0.67	0.50

Table 8: Pairwise agreement on GWA rankings across two human annotators (A1, A2) and an LLM judge (Claude with 3-example few-shot), on 28 OLMES items. Higher is better for all metrics.

Pair	nDCG@5 (lin.)	nDCG@5 (bin.)	Prec@4	Recall@4	Jaccard@5	RBO
Automation framing						
A1 ↔ A2	0.723	0.743	0.705	0.668	0.548	0.555
Opus 4.5 ↔ A1	0.742	0.769	0.772	0.722	0.559	0.559
Opus 4.5 ↔ A2	0.662	0.649	0.644	0.616	0.435	0.480
Augmentation framing						
A1 ↔ A2	0.593	0.673	0.667	0.585	0.493	0.431
Opus 4.5 ↔ A1	0.700	0.747	0.749	0.676	0.592	0.530
Opus 4.5 ↔ A2	0.600	0.678	0.680	0.567	0.511	0.433

comparing Claude to either human. Second, Claude's agreement with the two annotators is asymmetric: agreement with A1 exceeds agreement with A2 under both framings, with the gap largest under augmentation (0.70 vs. 0.60 nDCG@5(linear)). Taking the more conservative of the two LLM-human comparisons of Opus 4.5 vs. A2, Claude tracks human rankings approximately as well as the two humans track each other (0.66 vs. 0.72 under automation; 0.60 vs. 0.59 under augmentation), which we read as suggesting that the LLM-judge protocol recovers a signal of comparable quality to a single human annotator on this item set, while not eliminating annotator-level variation in the construct itself.

We use the following prompts for loading on GWA automation and augmentation.

Automation-focused Prompt

```
An AI system is deployed to handle this task autonomously, with no
↪  human \
involvement. Which work activities must the AI exercise to complete
↪  this \
item end-to-end?

Work in two steps inside the JSON output below:

STEP 1 - In `sub_steps`, decompose the item from first principles
↪  into the \
granular sub-tasks an AI would have to do to arrive at the correct
↪  answer. \
Think about the underlying problem, not the surface format (e.g. that
↪  it's \
multiple choice).
```

```
STEP 2 - In `ranked_gwa_ids`, rank up to 5 GWAs from the list below
↪  that \
benefit from being able to do those sub-tasks, ordered from most to
↪  least \
central.

Benchmark Item:
{task_text}

---

GWA List:
{gwa_list}

---

Example:

Benchmark Item:
An entrepreneur from State A decided to sell hot sauce to the public,
↪  \
labeling it "Best Hot Sauce." A company incorporated in State B and \
headquartered in State C sued the entrepreneur in federal court in
↪  State C. \
The complaint sought $50,000 in damages and alleged that the
↪  entrepreneur's \
use of the name "Best Hot Sauce" infringed the company's federal
↪  trademark. \
The entrepreneur filed an answer denying the allegations, and the
↪  parties \
began discovery. Six months later, the entrepreneur moved to dismiss
↪  for \
lack of subject-matter jurisdiction. Should the court grant the \
entrepreneur's motion?
  A) No, because the company's claim arises under federal law.
  B) No, because the entrepreneur waived the right to challenge
  ↪  subject-matter jurisdiction by not raising the issue initially
  ↪  by motion or in the answer.
  C) Yes, because although the claim arises under federal law, the
  ↪  amount in controversy is not satisfied.
  D) Yes, because although there is diversity, the amount in
  ↪  controversy is not satisfied.

JSON output:
{{
  "sub_steps": [
    "user pastes the fact pattern and four answer choices, asking
    ↪  which is correct",
    "AI identifies that this is a federal subject-matter jurisdiction
    ↪  question and surfaces the two grounds (federal question vs.
    ↪  diversity)",
    "AI flags the salient facts: federal trademark claim, $50,000
    ↪  amount, parties from different states",
    "AI checks the live legal rule - federal-question jurisdiction
    ↪  has no amount-in-controversy threshold; SMJ can be raised at
    ↪  any time",
    "user asks whether the timing of the motion matters; AI confirms
    ↪  it does not for SMJ and cites the relevant procedural rule",
```

```
    "AI recommends option A and explains why federal-question grounds
    ↪  jurisdiction directly"
  ],
  "ranked_gwa_ids": [
    "evaluating_compliance",
    "updating_using_knowledge",
    "interpreting_information",
    "providing_consultation",
    "communicating_internally"
  ],
  "rationale": "The interaction centers on applying a specific legal
  ↪  rule (federal-question jurisdiction) to the facts - checking
  ↪  compliance with the doctrinal requirements is the dominant AI
  ↪  move, supported by recalling the relevant rule (updating
  ↪  knowledge) and parsing what the fact pattern actually says
  ↪  (interpreting information). The user's questions and the AI's
  ↪  explanation form a consultative back-and-forth."
Example:

Benchmark Item:
Three bells ring at intervals of 36 seconds, 40 seconds and 48
↪  seconds, \
respectively. They start ringing together at a particular time. When
↪  will \
they ring together again?
  A) After 6 minutes
  B) After 12 minutes
  C) After 18 minutes
  D) After 24 minutes
  E) none

JSON output:
{{
  "sub_steps": [
    "recognize that all three bells ring together again at the least
    ↪  common multiple (LCM) of their intervals",
    "factorize each interval: 36 = 2^2 * 3^2, 40 = 2^3 * 5, 48 = 2^4
    ↪  * 3",
    "compute the LCM by taking the maximum power of each prime: 2^4 *
    ↪  3^2 * 5 = 16 * 9 * 5 = 720 seconds",
    "convert 720 seconds to minutes: 720 / 60 = 12 minutes",
    "match against the answer choices: 12 minutes corresponds to
    ↪  option B",
    "select option B"
  ],
  "ranked_gwa_ids": [
    "making_decisions_solving_problems",
    "processing_information",
    "analyzing_data",
    "scheduling_work",
    "estimating_quantifiable_characteristics"
  ],
```

```
  "rationale": "The decisive sub-task is computing the LCM of three
  ↪  intervals (processing/analyzing the prime factorizations) and
  ↪  selecting the correct multiple (problem-solving). The problem
  ↪  is fundamentally about scheduling repeating events, which makes
  ↪  that GWA salient even though the format is arithmetic.
  ↪  Estimating quantitative characteristics supports the conversion
  ↪  from seconds to minutes."
Example:

Benchmark Item:
Self-efficacy (the belief that one has control over one's situation)
↪  as it \
related to job satisfaction was studied. When a group of teachers
↪  rated their \
ability to control their situation and their satisfaction with their
↪  job, the \
two variables had a correlation of 0.30. Which statement follows from
↪  this \
correlation?
  A) If you want teachers to be happy with their job, give them more
  ↪  control over their situation.
  B) If you want teachers to take more control over their situation,
  ↪  make them happier at their jobs.
  C) These two variables show a moderate negative correlation.
  D) Higher self-efficacy and higher job satisfaction tend to occur
  ↪  together, but neither causes the other.

JSON output:
{{
  "sub_steps": [
    "parse the setup: self-efficacy and job satisfaction were
    ↪  measured in a group of teachers, with r = 0.30 between them",
    "recall the meaning of the Pearson correlation coefficient: 0.30
    ↪  is a positive linear association of moderate-to-weak strength,
    ↪  not a causal claim",
    "evaluate option A (efficacy implies satisfaction causal claim):
    ↪  unsupported by correlation alone, eliminate",
    "evaluate option B (satisfaction implies efficacy causal claim):
    ↪  also unsupported, eliminate",
    "evaluate option C (negative correlation): factually wrong
    ↪  because 0.30 is positive, eliminate",
    "evaluate option D (co-occurrence without causation): correctly
    ↪  restates what r = 0.30 implies",
    "select option D"
  ],
  "ranked_gwa_ids": [
    "analyzing_data",
    "making_decisions_solving_problems",
    "interpreting_information",
    "processing_information",
    "updating_using_knowledge"
  ],
  "rationale": "The decisive sub-task is correctly interpreting what
  ↪  a correlation coefficient does and does not imply. Eliminating
  ↪  each answer choice against that interpretation is a
  ↪  problem-solving step over the option set, and recalling the
  ↪  standard 'correlation does not imply causation' rule grounds
  ↪  the analysis in statistical knowledge."
```

```
}}
```

Augmentation-focused Prompt

```
An AI system is deployed alongside a human worker to jointly tackle
↪  this \
task - the human remains involved throughout, ranging from light
↪  oversight \
to close collaboration. Which work activities does the AI most need
↪  to be \
capable of in this collaborative setting?

Work in two steps inside the JSON output below:

STEP 1 - In `sub_steps`, imagine interacting with a user on this item
↪  and \
walk through the exchange concretely: what the user would ask, what
↪  the AI \
would do, how they go back and forth.

STEP 2 - In `ranked_gwa_ids`, rank up to 5 GWAs from the list below
↪  that \
correlate with the AI being good at this kind of interaction, ordered
↪  from \
most to least central.

Benchmark Item:
{task_text}

---

GWA List:
{gwa_list}

---

Example:

Benchmark Item:
An entrepreneur from State A decided to sell hot sauce to the public,
↪  \
labeling it "Best Hot Sauce." A company incorporated in State B and \
headquartered in State C sued the entrepreneur in federal court in
↪  State C. \
The complaint sought $50,000 in damages and alleged that the
↪  entrepreneur's \
use of the name "Best Hot Sauce" infringed the company's federal
↪  trademark. \
The entrepreneur filed an answer denying the allegations, and the
↪  parties \
began discovery. Six months later, the entrepreneur moved to dismiss
↪  for \
lack of subject-matter jurisdiction. Should the court grant the \
entrepreneur's motion?
  A) No, because the company's claim arises under federal law.
```

```
  B) No, because the entrepreneur waived the right to challenge
  ↪  subject-matter jurisdiction by not raising the issue initially
  ↪  by motion or in the answer.
  C) Yes, because although the claim arises under federal law, the
  ↪  amount in controversy is not satisfied.
  D) Yes, because although there is diversity, the amount in
  ↪  controversy is not satisfied.

JSON output:
{{
  "sub_steps": [
    "user pastes the fact pattern and four answer choices, asking
    ↪  which is correct",
    "AI identifies that this is a federal subject-matter jurisdiction
    ↪  question and surfaces the two grounds (federal question vs.
    ↪  diversity)",
    "AI flags the salient facts: federal trademark claim, $50,000
    ↪  amount, parties from different states",
    "AI checks the live legal rule - federal-question jurisdiction
    ↪  has no amount-in-controversy threshold; SMJ can be raised at
    ↪  any time",
    "user asks whether the timing of the motion matters; AI confirms
    ↪  it does not for SMJ and cites the relevant procedural rule",
    "AI recommends option A and explains why federal-question grounds
    ↪  jurisdiction directly"
  ],
  "ranked_gwa_ids": [
    "evaluating_compliance",
    "updating_using_knowledge",
    "interpreting_information",
    "providing_consultation",
    "communicating_internally"
  ],
  "rationale": "The interaction centers on applying a specific legal
  ↪  rule (federal-question jurisdiction) to the facts - checking
  ↪  compliance with the doctrinal requirements is the dominant AI
  ↪  move, supported by recalling the relevant rule (updating
  ↪  knowledge) and parsing what the fact pattern actually says
  ↪  (interpreting information). The user's questions and the AI's
  ↪  explanation form a consultative back-and-forth."

Example:

Benchmark Item:
Self-efficacy (the belief that one has control over one's situation)
↪  as it \
related to job satisfaction was studied. When a group of teachers
↪  rated their \
ability to control their situation and their satisfaction with their
↪  job, the \
two variables had a correlation of 0.30. Which statement follows from
↪  this \
correlation?
  A) If you want teachers to be happy with their job, give them more
  ↪  control over their situation.
  B) If you want teachers to take more control over their situation,
  ↪  make them happier at their jobs.
  C) These two variables show a moderate negative correlation.
```

```
  D) Higher self-efficacy and higher job satisfaction tend to occur
  ↪  together, but neither causes the other.

JSON output:
{{
  "sub_steps": [
    "user shares the question and asks the AI to help reason through
    ↪  which option is correct",
    "AI clarifies what r = 0.30 means: a modest positive linear
    ↪  association, not strong, and far from causation",
    "AI walks the user through why a correlation does not establish a
    ↪  causal direction in either direction",
    "AI helps the user evaluate options A and B as causal claims
    ↪  (both unsupported by correlation alone) and option C as
    ↪  factually wrong (0.30 is positive, not negative)",
    "user asks 'so what is left?' - AI confirms option D restates the
    ↪  correlation without overclaiming causation",
    "AI explains the underlying statistical principle (correlation
    ↪  does not imply causation) so the user retains the reasoning,
    ↪  not just the answer"
  ],
  "ranked_gwa_ids": [
    "interpreting_information",
    "providing_consultation",
    "communicating_internally",
    "analyzing_data",
    "updating_using_knowledge"
  ],
  "rationale": "The interaction is fundamentally consultative: the AI
  ↪  walks the user through the meaning of a correlation coefficient
  ↪  and the limits of causal inference. Interpreting the
  ↪  statistical concept and communicating it back to the user
  ↪  dominates the AI's contribution; analytic and knowledge-recall
  ↪  steps support that consultation."

Example:

Benchmark Item:
Fertilizer from an agricultural area runs off into a river. The river
↪  carries \
the nutrients from this fertilizer and deposits them into an ocean
↪  bay. After \
the nutrients enter the bay, scientists monitoring the water would
↪  most \
likely see a decrease in which of these dissolved gases?
  A) oxygen
  B) nitrogen
  C) carbon dioxide
  D) carbon monoxide

JSON output:
{{
  "sub_steps": [
    "user pastes the scenario and answer choices and asks the AI to
    ↪  explain the underlying mechanism and pick an answer",
    "AI walks the user through the eutrophication chain: nutrient
    ↪  runoff causes an algal bloom, the bloom dies, microbial
    ↪  decomposition consumes oxygen, and hypoxia results",
```

```
    "AI rules out the other gases for the user: nitrogen is the input
    ↪  not the output, carbon dioxide is produced by decomposition
    ↪  (would increase, not decrease), carbon monoxide is not a
    ↪  typical aquatic gas",
    "user asks why this matters in practice - AI explains the broader
    ↪  environmental impact (dead zones, fish kills) so the answer
    ↪  is grounded in real-world consequences",
    "AI confirms option A (oxygen) and offers to elaborate on related
    ↪  concepts such as algal blooms or water-quality monitoring if
    ↪  the user wants to learn more"
  ],
  "ranked_gwa_ids": [
    "providing_consultation",
    "communicating_internally",
    "interpreting_information",
    "updating_using_knowledge",
    "evaluating_compliance"
  ],
  "rationale": "The interaction is the AI explaining a multi-step
  ↪  environmental mechanism to the user - consultative teaching
  ↪  plus communication of the reasoning chain. Recalling the
  ↪  eutrophication process is the knowledge backbone; helping the
  ↪  user evaluate which dissolved gas is affected is interpretive."
}}
```

The cost of labelling one item with the (Claude) LLM judge is approximately $\$0.024$ USD.