OPTIMAL AGGREGATION MECHANISMS FOR AI BENCHMARKING

Andreas Haupt Anka Reuel Mykel J. Kochenderfer Sanmi Koyejo

ABSTRACT

AI benchmarks are frequently summarized by uniformly averaging item-level scores, implicitly treating every test item as equally valuable. This induces incentives to over-optimize for items that are trivial, socially irrelevant, or dominated by measurement noise. We model benchmarking as a multitask principal-agent game in which a benchmark designer chooses aggregation weights and a lab takes costly actions to improve their model. The optimal weights depend on normative welfare priorities, marginal costs of improvement, and measurement uncertainty. This analysis motivates *platinum items*: items that (i) precisely measure (ii) welfare-aligned capabilities that are (iii) comparatively cheap to improve. We propose an operational rubric and a certification workflow, implemented via expert review and LLM-based judgments, to identify platinum items and reweight benchmark items.

1 INTRODUCTION

Benchmarks and leaderboards are now a central coordination mechanism for AI development: they compress high-dimensional model behavior into a scalar score that is easy to optimize, communicate, and reward. But the choice of *aggregation* matters. Standard aggregation schemes implicitly treat all benchmark items as equally valuable. This creates a familiar pathology of metric-driven systems: when an index becomes the target, it shapes behavior in ways that can diverge from the underlying objective (Strathern, 1997). In benchmarking, equal weights can reward improvements that are socially irrelevant, poorly measured, or crowds out more important investment.

To analyze optimal benchmarking, we propose a simple game-theoretic model of benchmarking in which a designer chooses a linear aggregation rule and a lab responds by selecting costly improvement actions with measurement noise, inspired by Holmstrom & Milgrom (1991) (Section 2 and Section 3). The model yields a closed-form optimal weighting rule that depends on three primitives: (i) the normative welfare weight vector, (ii) the cost-of-improvement structure, and (iii) the evaluation noise. This motivates the notion of *platinum items*, which are items that (a) measure latent concepts that are aligned with the certifier's objective (e.g., welfare) , (b) correspond to improvements that are not prohibitively costly, and (c) can be measured with sufficient precision. In practice, the parameters needed to compute optimal weights are rarely available. We therefore translate the theory into an operational rubric and certification workflow (Section 4). We instantiate the concept set using publicly available policy-oriented taxonomies, including the OECD AI capability indicators (OECD, 2025) and the OECD framework for classifying AI systems (OECD, 2022), and connect normative weighting to multi-dimensional well-being measurement (Durand, 2015). To scale the evaluation of large item sets, we outline an LLM-as-a-judge procedure informed by recent surveys (Li et al., 2025); the prompts appear in Section A.

2 GAME-THEORETIC MODEL

We model benchmark aggregation as a welfare maximization problem in which a *designer* chooses a benchmark score (the aggregation rule) and a *lab* best responds by allocating effort across improvement activities. The model here is a reinterpretation of the multi-task model from Holmstrom & Milgrom (1991). We only claim to contribute its application to AI benchmark design. The model is deliberately stylized to highlight the contribution of different aspects to optimal aggregation in benchmarking.

Problem formulation. There are m benchmark tasks (e.g., one of the 1,000 training tasks in Chollet et al. (2026)) and n underlying improvement activities (e.g,. specialized data collection, pretraining runs, redteaming). Let $y \in \mathbb{R}^m$ denote task performance and $a \in \mathbb{R}^n$ denote effort across activities. Performance depends linearly on effort:

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

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

The designer publishes a linear benchmark score

$$s(y) = v^\top y,$$

with aggregation weights $v \in \mathbb{R}^m$. We assume the lab has constant absolute risk aversion (CARA) preferences (Pratt, 1964) with coefficient $r \geq 0$. That is, for x the score net of effort costs x, ${}^{u''(x)}/_{u'(x)} = r$ is contant.

The Lab's problem. Under CARA utility and Gaussian noise, the lab's optimization problem is equivalent to maximizing its mean score minus a risk premium given by the score variance. Given v, the lab's certainty equivalent is

$$\mathrm{CE}(a; v) = \mathbb{E}[s(y) \mid a] - c(a) - \frac{r}{2}\mathrm{Var}(s(y) \mid a) = v^\top A a - \frac{1}{2} a^\top C a - \frac{r}{2} v^\top \Sigma v.$$

The first-order condition gives

$$A^\top v - Ca = 0 \quad \Rightarrow \quad a^\star(v) = C^{-1} A^\top v.$$

Define the *output-level cost structure* as

$$M \coloneqq A C^{-1} A^\top \in \mathbb{R}^{m \times m}.$$

Then the induced expected true performance is

$$\mathbb{E}[y \mid a^\star(v)] = A a^\star(v) = Mv,$$

and the induced effort cost is

$$c(a^\star(v)) = \frac{1}{2}(a^\star(v))^\top C a^\star(v) = \frac{1}{2} v^\top M v.$$

The Certifier's problem. Society values true performance according to welfare weights $w \in \mathbb{R}^m$ and is risk-neutral. (A model where society is itself CARA risk-averse is mathematically equivalent. As the risk we consider here is one of a model failing to meet a benchmark score—an outcome that labs are more averse to than society—we present the model with lab risk aversion only.) The designer chooses v to maximize total welfare, defined as expected social value of true performance minus the lab's real effort cost and the risk premium borne by the lab. The designer chooses v anticipating the lab's best response. Total welfare under v is

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

Equivalently, the reduced problem is

$$\max_{v \in \mathbb{R}^m} w^\top M v - \frac{1}{2} v^\top M v - \frac{r}{2} v^\top \Sigma v.$$

In the following, we assume $M \succeq 0$ and $\Sigma \succ 0$. The unique optimal aggregation weights satisfy

$$(M + r\Sigma) v = Mw,$$

and hence

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

Interpretation via latent concepts. The matrix M summarizes how costly it is to improve tasks jointly: M_{ik} is large when there exist improvement activities $i \in \{1, 2, \ldots, n\}$ that raise both tasks i and k. This is analogous to multitask incentive settings in which effort reallocates across correlated tasks (Holmstrom & Milgrom, 1991). The noise covariance Σ captures both sampling variability and concept drift, and enters through the lab's risk premium.

To make the comparative statics (Samuelson, 1947; Milgrom & Shannon, 1994) more interpretable, assume for now that M and Σ share an orthonormal eigenbasis, which we also call the *concept eigenbasis*,

$$M = Q \operatorname{diag}(\lambda_1, \ldots, \lambda_m) Q^\top \qquad \Sigma = Q \operatorname{diag}(\sigma_1, \ldots, \sigma_m) Q^\top$$

with $\lambda_j \geq 0$ and $\sigma_j > 0$. Assuming simultaneous diagonalizability of M and Σ means that the same underlying dimensions determine measurement noise and cost structure.

In the rotated coordinates $\hat{v} = Q^\top v$ and $\hat{w} = Q^\top w$,

$$\hat{v}_j = \frac{\lambda_j}{\lambda_j + r\sigma_j} \hat{w}_j = \frac{\hat{w}_j}{1 + \frac{r\sigma_j}{\lambda_j}}. \tag{2}$$

Concept Alignment. We are interested in determining which items i should get a high weight in benchmarks. Let q_j be the jth column of Q, interpreted as a latent "concept" direction in task space. Define the welfare alignment of concept j by $\cos \varphi_j \coloneqq \frac{q_j^\top w}{\|w\|_2}$.

Let e_i be the ith standard basis vector (the ith task/item). Define the loading of item i on concept j by $\cos \theta_{ij} \coloneqq e_i^\top q_j$. Then the task-level incentive weight decomposes as

$$v_i = \|w\|_2 \sum_{j \in \mathcal{J}} \frac{\cos \theta_{ij} \cos \varphi_j}{1 + \frac{r\sigma_j}{\lambda_j}}.$$

The norm of the welfare vector w is up to normalization, and it is without loss to assume $\|w\|_2 = 1$. We can then bound

$$v_i \geq \max_{j \in \mathcal{J}} \frac{\cos \theta_{ij} \cos \varphi_j}{1 + \frac{r\sigma_j}{\lambda_j}} \geq \max_{j \in \mathcal{J}} \frac{\tau_w \tau_m}{1 + \frac{1}{\tau_e}}$$

if $\cos \varphi_j \geq \tau_w, \quad \cos \theta_{ij} \geq \tau_m, \frac{\lambda_j}{r\sigma_j} \geq \tau_e$.

This motivates focusing on items that load strongly on concept directions that are welfare-aligned, cheap to improve, and precisely measured.

Definition 1 (Platinum item)**.** Fix thresholds $\tau_w, \tau_m \in (0, 1)$ and a minimum signal-to-noise-cost ratio $\tau_e > 0$. An item i is *platinum* if there exists a concept $j \in \mathcal{J}$ such that $\cos \varphi_j \geq \tau_w$, $\cos \theta_{ij} \geq \tau_m$ and $\frac{\lambda_j}{r\sigma_j} \geq \tau_e$.

The definition captures a qualitative criterion: platinum items are those for which the score weight predicted by the model is high as witnessed by at least one "good" concept direction.

3 Comparative Statics

To get a better sense of the model introduced in Section 2, we consider cases in which primitives change due to innovations. This is complementary to the comparative statics we observed in our latent space.

Recall equation 2. We obtain three comparative statics:

- higher welfare relevance (larger $\hat{w}_j$),
- lower improvement cost (larger λ_j), and
- higher measurement precision (smaller σ_j)

all increase optimal aggregation weight. The first item is a normative questions that interacts, but is not primarily affected by technological innovation. The third can be improved by better design of items and measurement science. The second effect, mediated via λ_j, is not as transparent. We first consider a simple case—uniform cost or effectiveness changes—and then the more challenging case of non-uniform cost or effectiveness changes.

Uniform changes. A first change affecting M is through uniform changes. Examples are (for cost) improvements in data collection methodology and advances in hardware and infrastructure, or algorithmic improvements or synthetic data (for effectiveness). In either case, we have a scaling of the matrix C and or A, leading to a scaling of $M = AC^{-1}A^\top$. A simple comparative static is a *uniform* cost change $M \mapsto \kappa^{-1}M$, which leads to

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

Thus, uniform changes to M ($\kappa \uparrow$) are equivalent to higher effective measurement risk ($r\Sigma$), with nothing else changing. Hence uniform changes to M leads to changes toward well-measured directions.

Non-Uniform Changes. In general, innovations to A or C both change eigenvalues and rotate the concept basis. However, if an innovation is concept-separable, i.e. it only affects the weight of one concept i, the result is that weights affect only λ_i in equation 2. This is the case if a technological change affects M by a rank-one projector $\delta q_i q_i^\top$ for an eigenvector q_i of M. That is,

$$M(\delta) = M + \delta q_i q_i^\top. \tag{3}$$

Such a change only affects the corresponding eigenvalue λ_i and increases the optimal benchmark weight of tasks, weighted by the alignment of tasks with the concept. There are two natural ways to get an improvement like in equation 3.

New Activity A new task ι can achieve a change to the improvement cost like in equation 3 if it has effectiveness vector $q_i \in \mathbb{R}^m$, where q_i is the eigenvector of M associated to λ_i and costs independent of other items, $c_{\iota i} = 1_{\iota=i}$, $i = 1, 2, \ldots, n$. This means that we have a task affecting the right mixture of concepts.

Cost Changes A second version does not require the introduction of a new task. Again, let q_i be the eigenvector to λ_i in equation 3, and consider a cost decrease by $dd^\top$, where $d = (AC^{-1})^{-1}q_i$. We state and show in the appendix that this leads to an increase of the benchmark weights in the directions associated to λ_i.

4 A RUBRIC

The optimality condition in Equation (1) is useful conceptually but requires quantities (A, C and Σ) that are rarely known for real-world benchmarks. We therefore propose a *rubric* that operationalizes the key factors suggested by theory: welfare alignment, cost of improvement, and measurement precision.

Concept set. The rubric begins by selecting a set of welfare-relevant concepts $\mathcal{J} = \{1, \ldots, J\}$. One practical starting point is the OECD AI capability indicators, which define a policy-relevant taxonomy of AI capabilities (OECD, 2025) and connects to broader governance frameworks for AI systems (OECD, 2022). Normative welfare relevance can be informed by multidimensional well-being frameworks such as the OECD Better Life Initiative (Durand, 2015), which emphasizes that social value is not one-dimensional and depends on the application domain. In our pipeline, a concept $j \in \mathcal{J}$ is described by a short definition, and passed to a large language model judge.

Rubric scores. For each concept $j \in \mathcal{J}$, we elicit 1) *welfare alignment* $W_j \in \{1, 2, 3, 4, 5\}$, i.e., how directly improvements in the concept map to welfare gains; 2) *marginal improvement cost* $K_j \in \{\text{low}, \text{medium}, \text{high}\}$, i.e., how costly it is for leading labs to achieve a one-unit performance improvement on the concept over a fixed horizon; and 3) *measurement noise* $N_j \in \{\text{low}, \text{medium}, \text{high}\}$. W_j approximates $|\cos\varphi_j|$, K_j approximates the inverse of λ_j (high cost corresponds to smaller λ_j), N_j approximates σ_j.

For each benchmark item i and concept j, we elicit *measurement strength* $S_{ij} \in \{0, 1, 2, 3\}$, i.e., how strongly item i measures concept j. S_{ij} approximates $|\cos\theta_{ij}|$.

Platinum scoring and certification. Given rubric scores, define a per-concept "incentive-controlled quality factor"

$$Q_j \coloneqq \gamma(W_j)\eta(K_j)\pi(N_j),$$

where γ is increasing in welfare alignment, η is decreasing in improvement cost, and π is decreasing in noise. A simple instantiation could be

$$\gamma(W_j) = W_j \qquad \eta(K_j) = \begin{cases} 1 & K_j = \text{low} \\ \frac{1}{2} & K_j = \text{medium} \\ \frac{1}{4} & K_j = \text{high} \end{cases} \qquad \pi(N_j) = \begin{cases} 1 & N_j = \text{low} \\ \frac{1}{2} & N_j = \text{medium} \\ \frac{1}{4} & N_j = \text{high} \end{cases}$$

We say that an item i is *platinum* if there is a concept that it measures strongly and that has a high incentive-controlled quality factor,

$$P_i \coloneqq \max_{j \in \mathcal{J}} S_{ij} \cdot Q_j. \tag{4}$$

A benchmark curator can certify items as platinum when P_i exceeds a policy-chosen threshold τ_{cert}, which we do in ongoing work.

REFERENCES

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.

Martine Durand. The OECD better life initiative: How's life? and the measurement of well-being. *Review of Income and Wealth*, 61(1):4–17, March 2015. doi: 10.1111/roiw.12156. URL https://doi.org/10.1111/roiw.12156.

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.

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*, pp. 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/.

Paul Milgrom and Chris Shannon. Monotone comparative statics. *Econometrica*, 62(1):157–180, 1994. doi: 10.2307/2951479.

OECD. OECD framework for the classification of AI systems. Technical Report 323, OECD Publishing, Paris, February 2022. URL https://doi.org/10.1787/cb6d9eca-en.

OECD. Introducing the OECD AI capability indicators. Technical report, OECD Publishing, Paris, June 2025. URL https://doi.org/10.1787/be745f04-en.

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

Paul A. Samuelson. *Foundations of Economic Analysis*. Harvard University Press, Cambridge, MA, 1947.

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.

A LLM-AS-A-JUDGE PROMPTS

Large-scale item annotation is expensive. Recent work surveys how LLMs can be used as judges for scoring, ranking, and selection problems (Li et al., 2025). In our context, LLMs can be used to produce initial rubric labels, which are then calibrated and audited by human experts for a subset of items. The following are prompts we use in ongoing work to evaluate items.

Concept scoring prompt

```
Task: Score a capability concept for platinum certification.

Context:
We are designing an AI benchmark where some items should be upweighted
    . A concept should be upweighted if it is welfare-aligned, not too
     expensive for labs to improve, and can be measured precisely.

Concept:
- Name: {CONCEPT_NAME}
- Definition: {CONCEPT_DEFINITION}

Rubric:
1) Welfare alignment (1-5):
   1 = weak or unclear link to welfare improvements
   3 = moderate link; depends on deployment details
   5 = strong and direct link to welfare improvements

2) Marginal improvement cost (low/medium/high):
   low = typical frontier labs can improve materially within ~6-12
    months using
         standard data/compute/training interventions
   medium = improvement is possible but requires substantial targeted
    investment
   high = improvement appears difficult, slow, or requires
    breakthroughs

Output JSON schema:
{
  "welfare_alignment_1_to_5": int,
  "marginal_improvement_cost": "low" | "medium" | "high",
  "measurement_noise": "low" | "medium" | "high",
  "rationale": string
}

Return only JSON.
```

Item-to-concept measurement prompt

```
Task: Determine how strongly an item measures a given concept.

Item:
- Benchmark item text: {ITEM_TEXT}
- Ground-truth / reference solution (if available): {
    REFERENCE_SOLUTION}

Concept:
- Name: {CONCEPT_NAME}
- Definition: {CONCEPT_DEFINITION}

Rubric for measurement strength S_ij (0-3):
0 = does not measure the concept
1 = weakly related or incidental
2 = moderately measures the concept
```

```
3 = strongly and directly measures the concept


Output JSON schema:
{
  "measurement_strength_0_to_3": int,
  "rationale": string
}

Return only JSON.
```

We then aggregate the scores to certify an item i as *platinum* if there is a concept c such that $w_c \geq 4$, $w_c \in \{\text{low}, \text{medium}\}$, and $m_{ic} \geq 2$.