METRIC MATCH: A SUBSET SELECTION APPROACH TO EVALUATING LLM JUDGE RELIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges are used to reduce the need for costly human labor in evaluating open-ended text generation. However, the reliability of these judges depends critically on their alignment with human raters — a property that itself depends on costly human annotations. In this work, we develop a method (**Metric Match**) for estimating correlation-based reliability metrics of LLM judges from limited annotations. **Metric Match** selects a subset of samples for human annotation such that the subset matches the population reliability metric with respect to acquired synthetic labels. We empirically show that **Metric Match** achieves a win-rate of at least 0.794 with respect to four natural baselines across four different correlation metrics and 15 datasets. Furthermore, **Metric Match** exhibits a 16.2% decrease in average estimation error and reduces annotation needs by 28.1%. We provide a cost model and highlight a medical case study where our method saves $1,083.33 on expert annotations compared to baselines. Finally, for the downstream task of classifying whether a judge is sufficiently reliable, **Metric Match** also outperforms all baselines in terms of accuracy. All project code is available[1], and we additionally provide an installable package for ease of use.

1 INTRODUCTION

Large language models (LLMs) are increasingly used for text generation tasks, but their rapid adoption has outpaced our ability to evaluate them at scale (Thirunavukarasu et al., 2023; Meyer et al., 2023; Yuan et al., 2021; Celikyilmaz et al., 2021). Accordingly, the LLM judge framework (Zheng et al., 2023), in which one LLM evaluates the outputs of another model, has emerged as a scalable alternative to human annotation. Scalability gains are particularly relevant in expert-oriented contexts such as healthcare (Asgari et al., 2025), where human labeling is slow and expensive. Recent work has explored this direction through human-labeled benchmarks (Dubois et al., 2024) and reference-free evaluation methods (Tan et al., 2025).

Evaluation of LLM-generated annotations with respect to human labels is necessary to responsibly deploy LLM judges in high-stakes domains such as healthcare (Li et al., 2026). LLM judge reliability (Bedi et al., 2025; Li et al., 2024; Croxford et al., 2025) is often measured using statistical measures in the inter-rater reliability literature (Shrout and Fleiss, 1979; Krippendorff, 1970) and standard correlation coefficients (Spearman, 1904; Kendall, 1948). A key challenge is that these metrics require human annotations in order to calculate the associated reliability score for the target LLM-judge system, creating a bottleneck for evaluation. Thus, LLM judge evaluation encounters the same scalability problem that the LLM-as-a-judge framework aims to solve in the first place: large-scale human annotation.

In order to mitigate this scalability issue, standard approaches aim to estimate judge reliability using a limited budget for human annotations. The direct approach is to collect annotations on a randomly chosen subset for estimation of the reliability metric, which produces an unbiased estimate (Cochran, 1977), and permits finite-sample analysis of the metric estimates. Another approach is to generate synthetic labels using another LLM judge and then correct the systemic bias incurred by these labels. Modern approaches such as Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023) combine bias correction with the construction of a confidence set.

[1] `anonymous.4open.science/r/MetricMatch-FEFB`

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

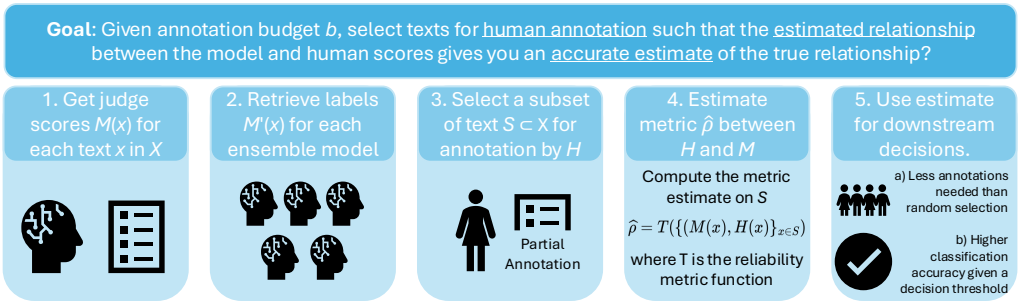


Figure 1: **Overview of the LLM judge evaluation framework and our approach.** Given text samples $\mathcal{X}$, and judge model M to evaluate, we obtain scores $\{M(x), M'(x)\}_{x\in\mathcal{X}}$ for $M' \in \mathcal{M}$. We select b texts for human annotation. Equipped with a measure of reliability, function T, we estimate a target parameter $\hat{\rho}$. We show that our improved estimation of $\hat{\rho}$ leads to downstream impacts.

In this paper, rather than aiming to obtain an unbiased estimator for judge reliability, our goal is to predict judge reliability in order to minimize estimation error. In low-annotation regimes, high variance can render an unbiased estimate practically uninformative, making minimization of total estimation error the more relevant objective. While standard approaches collect human annotations on a randomly chosen subset, we leverage the structure of the reliability metric on the synthetic labels across items to determine the subset of items to annotate.

Our main contribution is a new estimation approach (**Metric Match**) for evaluating LLM judge reliability, which combines limited human annotations with access to synthetic labels from other LLMs. **Metric Match** leverages a novel subset selection approach: we collect human annotations from a carefully constructed subset which most closely matches the full population in terms of inter-model reliability between the LLM judge scores and the synthetic labels. This approach uses the inter-model metric to guide subset selection for the estimate of the human-model metric of interest, as outlined in Figure 1. **Our results are as follows:**

1. Estimation error: We empirically show that **Metric Match** outperforms baselines such as annotating on randomly collected subsets, bias correction, and stratified selection (Figure 2). Compared to the strongest baseline (stratified selection), **Metric Match** reduces estimation error by an average of 16.2%. This results in a reduced annotation requirement of 28.1% (Figure 3). We provide a cost model to calculate the impact of improved estimation ability, showing cost savings of up to $1,083.33 for a given clinical dataset.
2. Win-rate evaluation: We then measure the win-rate of **Metric Match**, which captures the fraction of instances where **Metric Match** outperforms a given baseline. We find an average win rate of 0.794 against the strongest baseline (stratified selection) (Table 1). We perform a complementary analysis with a per-trial micro win-rate.
3. Reliability classification: Finally, we examine the downstream task of reliability classification. A practitioner may choose to use an LLM judge only if the estimated reliability coefficient is above a pre-specified deployment threshold. Shifting our task from reliability estimation to reliability classification, **Metric Match** demonstrates a classification accuracy of 0.881, which outperforms all baselines (Table 2).

Accordingly, our work constitutes a step towards scalable LLM judge evaluation, allowing for practitioners to accelerate judge development and improve early failure detection, while aligning with human preferences and maintaining reliability estimation accuracy with fewer annotations.

2 Related Works

2.1 LLM Evaluation and Human Annotation

Evaluating the outputs of large language models is a longstanding challenge, particularly as these systems are deployed in open-ended and domain-specific settings (Thirunavukarasu et al., 2023;

Meyer et al., 2023). As model success moves away from traditional multiple-choice evaluation style and into messier, real world environments, evaluation of model success becomes more nuanced than current lexical approaches can capture (Papineni et al., 2002; Lin, 2004; Novikova et al., 2017). Human evaluation has historically served as the gold standard due to its sensitivity to nuanced qualities such as coherence, and factuality while capturing the target notion of model alignment (Yuan et al., 2021; Ouyang et al., 2022). However, human annotation is costly and difficult to scale, especially in domains such as medicine and law, where annotators require expert knowledge (Bedi et al., 2025). For instance, expert oncologist labeling can be up to $500 per hour, and some tasks take hours with multiple modes of disagreement (Unell et al., 2025). Prior work has examined dimensions of annotation quality, including inter-annotator agreement, bias, and crowd-sourced label reliability, highlighting the problems that arise even if unlimited access to human annotations is available, especially internal human disagreement (Li et al., 2024; Clark et al., 2021; Bowman et al., 2022; Burns et al., 2024; Kenton et al., 2024).

2.2 LLM-AS-A-JUDGE FOR SCALABLE EVALUATION

The *LLM-as-a-judge* paradigm, popularized by Zheng et al. (2023), employs LLMs as automated evaluators to score or rank the outputs of other models. Subsequent research has identified critical failure modes, including position bias, verbosity bias, and self-preference bias (Zheng et al., 2023; Li et al., 2024). Recent benchmarks mitigate these issues by introducing length-controlled metrics via regression-based debiasing (AlpacaEval (Dubois et al., 2024)), or focusing on correctness in knowledge-heavy domains where human stylistic preferences may mislead automated judges (JudgeBench (Tan et al., 2025)). Psychometric reliability metrics, such as ICC, Krippendorff's α, Kendall's τ, and Spearman's ρ, serve as signals for judge-human alignment and reliability (Koo and Li, 2016; Artstein and Poesio, 2008; Kendall, 1948; Spearman, 1904). LLM judges are frequently compared to human outputs in order to decide if a system is acceptable to serve as a human proxy or if further iteration is needed (Chiang and Lee, 2023). This approach leaves open the question of how to use judge outputs effectively when human annotations are scarce.

2.3 EFFICIENT SAMPLING AND ANNOTATION

Our work builds on research in efficient sampling, active learning, and semi-supervised statistical inference (Cochran, 1977; MacKay, 1992; Huan et al., 2024). While active learning iteratively queries uncertain or diverse samples (Sener and Savarese, 2018), active experimental design optimizes one-step sample selection for maximum information gain (Zrnic and Candès, 2024; Wu et al., 2026). Bias correction methods, such as Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023; Feng et al., 2026), use a small set of human labels to "rectify" bias from synthetic annotations and provide valid statistical guarantees on population parameters. Bias correction is typically performed on a random subset of data for which human labels are collected, while a related line of work frames item selection instead through importance sampling (Chaganty et al., 2017; Zhao et al., 2012). A handful of works (e.g., Zrnic and Candès (2024); Wu et al. (2026)) have explored how to move beyond random selection, and construct this subset in an active manner. Our approach is similar in motivation, as we use synthetic labels to construct a subset of data points on which to collect human annotations. Finally, these frameworks rely on estimation of metrics that decompose easily as sums of samples or M-estimation, which does not extend trivially to correlation or reliability metrics, which we consider.

3 PRELIMINARIES

In this section, we outline the problem setup, describe our proposed method (**Metric Match**), provide motivating intuition, and discuss the experimental setup.

3.1 PROBLEM STATEMENT

We study the problem of LLM judge evaluation under limited annotation budgets. Assume we have a large set of text items $\mathcal{X}$, and let $H : \mathcal{X} \to \mathcal{Y} := [1, K]$ denote the ground-truth human scoring function from items to scores, where K is the greatest possible score. The human scoring function may correspond to a single human rater or an average of multiple raters. Let $M : \mathcal{X} \to \mathcal{Y}$ denote

the scoring function produced by the LLM judge. To generate synthetic labels, we have access to a family of models $M' \in \mathcal{M}$, each of which yields a synthetic scoring function $M' : \mathcal{X} \to \mathcal{Y}$.[2]

We wish to measure the reliability of judge M with respect to human annotator H on textual data. We consider a reliability metric $T : (\mathcal{Y} \times \mathcal{Y})^{|\mathcal{X}|} \to \mathbb{R}$, which operates on pairs of scores on a set of text samples and outputs a number quantifying the reliability between scorers. We want an estimate of the population-level relationship between M and H on $\mathcal{X}$, captured by $\rho = T(\{(M(x), H(x))\}_{x\in\mathcal{X}})$. With a budget of size $b \geq 1$, we can construct any $S \subseteq \mathcal{X}$ of size $|S| = b$ and obtain human annotations on the subset, $\{H(x)\}_{x\in S}$. Using the selected human annotations on S and unlimited query access to annotations from $M' \in \mathcal{M}$, we provide an estimate $\widehat{\rho}_S$ of ρ.

To exploit the availability of synthetic scoring functions $M' : \mathcal{X} \to \mathcal{Y}$, $M' \in \mathcal{M}$, we introduce *inter-model* (IM) reliability quantities as measured by T, which inform **Metric Match**. Denote the inter-model population reliability between M and M' as ρ^{IM}. Unlike the human-model population metric, the inter-model population metric is directly calculable, as we can cheaply obtain scores $M(x)$, $M'(x)$ for every $x \in \mathcal{X}$. Similarly, given any subset $S \subseteq \mathcal{X}$, let $\widehat{\rho}^{\text{IM}}_S$ signify the finite sample estimate of ρ^{IM} on S. Formally,

$$\rho^{\text{IM}} = T\big(\{M(x), M'(x)\}_{x\in\mathcal{X}}\big) \qquad \widehat{\rho}^{\text{IM}}_S = T\big(\{M(x), M'(x)\}_{x\in S}\big)$$

3.2 Method: Metric Match

We introduce our method, which we term **Metric Match**, for subset selection in order to reduce estimation error in reliability evaluation of LLM judges with respect to human raters. Given a budget b, we rely on synthetic scorers $M' \in \mathcal{M}$ to inform our subset construction. The key insight in the **Metric Match** algorithm (Algorithm 1) is to construct sets of text examples $S \subseteq \mathcal{X}$ whose inter-model empirical estimate $\widehat{\rho}^{\text{IM}}_S$ closely resembles the inter-model population metric ρ^{IM}. We characterize the intuition for and present details of our algorithm in the rest of this section.

Algorithm 1: Metric Match

Data: Metric function T; budget b; population text data $\mathcal{X}$; score functions M, M', from LLM judges M and $M' \in \mathcal{M}$, respectively; C, number of candidate subsets over which to search

Result: Subset $S^* \subseteq X$, $|S^*| = b$; Estimator $\widehat{\rho}_{S^*}$

```
Initialize δ_min ← ∞
Initialize S* ← ∅
ρ^IM ← (1/|M|) Σ_{M'∈M} T({M(x), M'(x)}_{x∈X})
for j = 1, ..., C do
    Sample S_j s.t. |S_j| = b from X without replacement
    ρ̂^IM_{S_j} ← (1/|M|) Σ_{M'∈M} T({M(x), M'(x)}_{x∈S_j})
    δ_j = |ρ̂^IM_{S_j} − ρ^IM|
    if δ_j < δ_min then
        S* ← S_j
        δ_min ← δ_j
Calculate human-model metric estimate ρ̂_{S*} = T(Y^(M)_{S*}, Y^(H)_{S*})
return S*, ρ̂_{S*}
```

In an oracle setting, where $H(x)$ is available for all $x \in \mathcal{X}$, the ideal subset $S \subseteq \mathcal{X}$ of size b is simply the subset that minimizes the estimation error $|\widehat{\rho}_S - \rho|$. While we do not operate in an oracle setting, we have access to judges $M' \in \mathcal{M}$, and can compute $M(x), M'(x)$ for every $x \in \mathcal{X}$. This motivates constructing a subset that minimizes the proxy (inter-model) estimation error $S^* := \arg\min_{S\subseteq\mathcal{X},|S|=b} |\widehat{\rho}^{\text{IM}}_S - \rho^{\text{IM}}|$. Given this subset, we compute the estimator on the human-model metric $\widehat{\rho}_{S^*} := T\big(\{M(x), H(x)\}_{x\in S^*}\big)$.

[2]Note that the average human rater may produce non-integers, whereas the LLM judge will always produce integer scores.

The success of this method implicitly assumes that the *relative* inter-model estimation errors across different subsets are good enough proxies for the relative human-model estimation errors across these subsets. We formalize this in Appendix C, which gives theoretical intuition for the observed empirical behavior. We clarify that our method does not rely on the synthetic labels acting as a proxy for human labels, and further, does not even rely on model-model agreement being a reasonable directional or magnitudinal proxy for human-model agreement. We discuss this further in Section 4.1 and provide settings under which our method performs well even under model-human misalignment in Appendix D.

There remain two important details that represent parametric choices about our algorithm:

- **Aggregation across synthetic labels:** Our selection procedure thus far has assumed there is a single synthetic scorer. However, we have access to a full ensemble of scorers $\mathcal{M}$. We compute an inter-model metric for each pair (M, M') over $M' \in \mathcal{M}$, and average the pairwise computations, resulting in:
$$\rho^{\text{IM}} = \frac{1}{|\mathcal{M}|} \sum_{M' \in \mathcal{M}} T\big(\{M(x), M'(x)\}_{x \in \mathcal{X}}\big) \qquad \widehat{\rho}^{\text{IM}}_S = \frac{1}{|\mathcal{M}|} \sum_{M' \in \mathcal{M}} T\big(\{M(x), M'(x)\}_{x \in S}\big)$$
We ablate with an alternate aggregation approach in Appendix E.1.
- **Candidate Subset Selection:** To solve $S^* := \arg\min_{S \subseteq \mathcal{X}, |S|=b} |\widehat{\rho}^{\text{IM}}_S - \rho^{\text{IM}}|$, we would need to search over $\binom{|\mathcal{X}|}{b}$ subsets, which is computationally expensive. To reduce the computational overhead, we sample C i.i.d. candidate subsets, $S_1, \ldots, S_C \sim \mathcal{X}^b$, from which S^* is selected as follows $S^* := \arg\min_{S_1,\ldots,S_C} |\widehat{\rho}^{\text{IM}}_{S_j} - \rho^{\text{IM}}|$. The number of candidate subsets to consider constitutes a hyperparameter of **Metric Match**. In our experiments, we set $C = 20$. We vary C and report performance results in Appendix E.2. We find that even at relatively small C ($C = 10$), we see gains in performance as measured by win-rate and estimation error.

3.3 Experimental Setup

Baselines. We compare to the following simple baselines: random selection, stratified selection, random selection with a bias correction component, and cluster text-based selection.[3]

- *Random selection* A classic approach to estimation of ρ is to select b samples i.i.d. from $\mathcal{X}$ to construct S, and estimate $\widehat{\rho}_S = T\big(\{M(x), H(x)\}_{x \in S}\big)$. Random selection remains the de facto method in practice (Bedi et al., 2025). However, random selection incurs high variance when b is small through finite sample error.
- *Stratified selection (Mozer et al., 2026)* Stratified selection induces a uniform distribution over the judge score space in order to construct S. In other words, given b, sample b/K text examples from each of $\{x \in \mathcal{X} \mid M(x) = k\}$ for $k = 1, \ldots, K$ to ensure full coverage of the score set $\mathcal{Y}$ in the sampled judge scores.
- *Bias correction (BC)* In bias correction, we leverage the human annotations for the purpose of debiasing the synthetic labels, such as in (Chaganty et al., 2018). We use finite sample error calculated on the inter-model reliability metric as a stand-in for finite sample error in the human-model metric. Formally, we sample $S \sim \mathcal{X}^b$, and calculate estimate $\widehat{\rho} := \underbrace{\rho^{\text{IM}}}_{\text{Inter-model metric}} + \underbrace{\widehat{\rho}_S - \widehat{\rho}^{\text{IM}}_S}_{\text{Bias correction term}}$.
- *Cluster text-based* We implement a KMedoids (Kaufman and Rousseeuw, 2009) clustering-based selection procedure on the embeddings of text examples X, drawn from standard active learning practices (Nguyen and Smeulders, 2004; Urner et al., 2013; Shuyang et al., 2018). We set the number of clusters, k, to the budget, b; the resulting $k = b$ medoids constitute the selected set S.

We use $\widehat{\rho}$ to denote a generic estimator for ρ under any method (implicitly with access to b samples). Note that in stratified selection, and cluster text-based selection, the important component of the estimator is the construction of subset S, and given S, $\widehat{\rho} = \widehat{\rho}_S$. The bias correction estimator has a different structure, due to incorporation of synthetic labels and an additive correction term. Further implementation details for baselines can be found in Appendix F.

[3] Due to the structure of the reliability metrics that we consider, we note that even random selection and bias correction may not be unbiased (Appendix A). However, we take these estimators to be baselines in the context of our goal of minimizing estimation error.

Evaluation procedure. To evaluate **Metric Match** (Section 3.2), we measure the **estimation error** and **win-rate** compared to the baselines.

- **Estimation error.** Given a method A, we directly compute estimation error as $\epsilon^A = |\widehat{\rho} - \rho|$, where $\widehat{\rho}$ indicates the estimator given by the method to be evaluated. We compute the average estimation error over $N = 40$ trials.
- **Win-rate.** Win-rate captures the fraction of instances where a method A beats a method B in terms of estimation error. We consider two variants of the win-rate.
  - **Macro-average win-rate:** This variant captures the win signal over errors averaged across the N trials, calculated as $\mathbb{1}\left[\frac{1}{N}\sum_{i=1}^{N}\epsilon_i^A < \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^B\right] \in \{0, 1\}$ (win-rate of A over B). We compute a win signal for each of the 75 contexts, and average the win signal to produce the macro-average win-rate as a summary metric.
  - **Micro-average win-rate:** This variant captures the win signal at a per-trial level. Over N trials (with estimation errors ϵ_i, $i = 1, \ldots, N$), the micro-average win-rate is equal to $\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\epsilon_i^A < \epsilon_i^B]$ (win-rate of A over B). We compute this statistic for each of the 75 dataset/evaluation axis/judge model contexts. We also report the average of the micro-average win-rate over the 75 contexts as a summary metric.

Reliability metrics. To capture notions of reliability for evaluation of an LLM judge, we consider the following reliability metrics: intraclass correlation coefficient (ICC) Shrout and Fleiss (1979), Krippendorff's α Krippendorff (1970), Spearman's ρ rank correlation Spearman (1904), and Kendall's τ rank correlation Kendall (1948). Based on previous literature Bedi et al. (2025); Chiang and Lee (2023); Liu et al. (2023), these metrics encompass a simple suite of correlative measures between LLM judge scores and human scores. We provide formulae for each metric computation and comparative exposition on these metrics in Appendix A.

Datasets and axes. We provide empirical evaluations on 4 diverse real-world datasets: MSLR (Wang et al., 2023), HANNA (Chhun et al., 2024), MedVAL (Aali et al., 2025), and SummEval (Fabbri et al., 2021). Each dataset is annotated along multiple axes (e.g., coherence, relevance, surprise) with a total of 15 sets of human annotations by which to evaluate sampling performance, with additional dataset details in Appendix B. All datasets contain samples that are annotated by more than one human; we take the average score of a sample over human raters as the true score. All datasets use Likert scaling with a range of 1–5 for SummEval and HANNA, 0–2 for MSLR, and 1–4 for MedVAL (Joshi et al., 2015). We truncate each dataset to contain 300 human-labeled samples and perform evaluations on subsets of this annotation set. We perform the evaluations over 40 trials. Therefore, we evaluate in a total of $5 \times 15 = 75$ contexts over all datasets, axes and target judge models.

Judges and synthetic labels. Our model suite consists of Claude-3.5-Sonnet Anthropic (2024), GPT-4.1 OpenAI et al. (2024), GPT-5 OpenAI (2025), DeepSeek-R1 Guo et al. (2025), and Gemini-2.5-pro Comanici et al. (2025), with additional model ablations in Appendix G. We run experiments where each model serves as the reference judge model M and the remaining models comprise the synthetic label collection $\mathcal{M}$.

4 Estimation Evaluation

In this section, we discuss empirical results for **Metric Match**. First, we compare the estimation error of **Metric Match** against baselines and compute annotation savings of using **Metric Match** instead of stratified selection (Section 4.1). Then, we compute a win-rate of **Metric Match** against stratified selection and discuss a connection between win-rate and subset estimation error rank (Section 4.2). Finally, we evaluate **Metric Match** on the task of reliability classification given a deployment threshold, highlighting the practical benefit of improved reliability estimation (Section 4.3).

4.1 Estimation Error

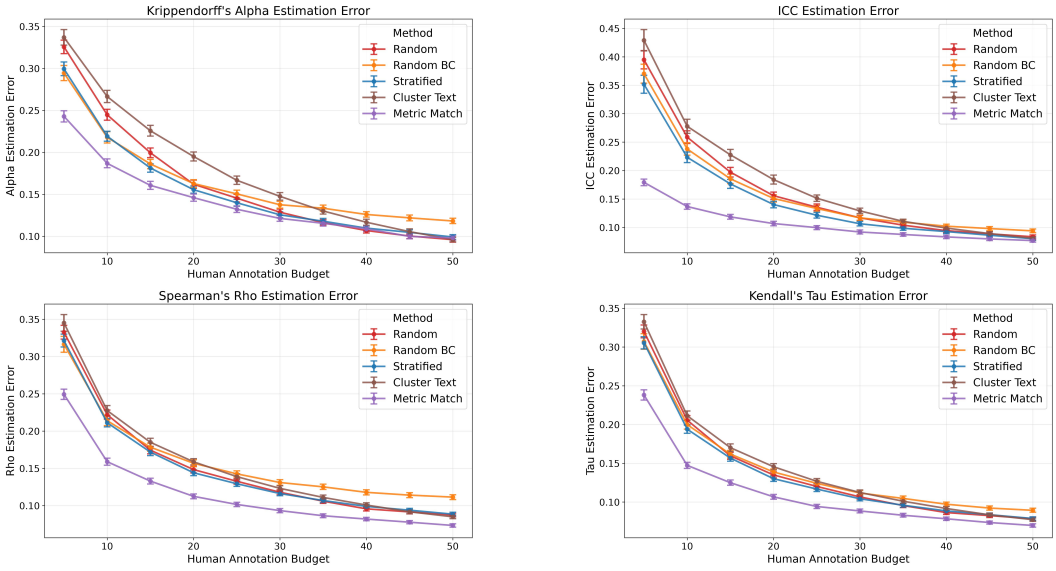


Figure 2: **Metric Match selection outperforms baselines.** We report the estimation error by annotation budget across our suite of target metrics, averaged over all datasets and compare between selection methods. The relative improved estimation error between **Metric Match** and the most competitive baseline, stratified selection, at a given budget is 16.2%.

Raw Estimation Improvement. We highlight in Figure 2, the average estimation error of **Metric Match** and the suite of comparable baselines. Across metrics, we see improvement in estimation error with **Metric Match**, most significantly at low budgets, demonstrating the utility of **Metric Match** in labor-constrained domains, where human annotation is extremely costly. Notably, we do not observe a degradation in our performance at higher budgets, which we see in the bias-correction baseline (*Random BC*). We calculate the relative improved estimation error between **Metric Match** and each of the baselines. We find that stratified selection is the strongest baseline, yielding the smallest relative improvement in estimation error for **Metric Match** (an average of 16.2% across all metrics and budgets). Disaggregated relative improvement results across metrics, budgets, and baselines are present in Appendix H.

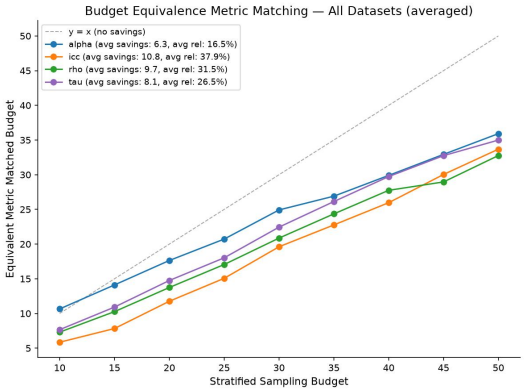


Figure 3: **Metric Match selection achieves the same estimation error as stratified selection with less annotations required.** We highlight the average relative savings over each reliability metric between **Metric Match** and stratified selection, with an average relative improvement of 28.1%.

Importantly, we emphasize that our method does not depend on the synthetic labels serving as a close proxy for human labels. In fact, it does not even rely on model-model agreement being a proxy for human-model agreement in terms of magnitude. To show this, we provide explicit examples where the model-model population metric is much higher than the human-model population metric, but our method still outperforms baselines (Appendix D). We provide a table of the model-model metrics on the population and compare the values with the human-model metrics on the population. We observe several instances of low agreement measured via the human-model metrics on the population with high agreement in the analogous dataset/axis settings measured via the model-model metrics on the population, but still highlight the high performance of our method.

Relative Annotation Savings. We highlight in Figure 3 the average annotations saved from using **Metric Match**; we again focus on comparing against stratified selection, since this method yielded the lowest annotation savings from **Metric Match** across baselines. For each budget given to stratified selection, we compute the required budget needed for **Metric Match** to match stratified selection estimation error according to Figure 2. We plot the relationship between the stratified selection budget and our required budget. As such, a line with a slope of 1 would indicate that **Metric Match** performs as well as stratified selection, while points below that line indicate **Metric Match** outperforms stratified selection. We see gains across all metrics observed, with an average relative improvement of 28.1%. Disaggregated results are reported in Appendix J.1, highlighting annotation savings across datasets and budget sizes.

Annotation Cost Savings. Finally, we perform an analysis that translates reduced annotations to cost savings, again focusing on comparing against stratified selection. Let S be the stratified selection budget, t be the estimated time per annotation (in hours), and c be the cost per unit time of annotation (in dollars per hour). Given S, we determine the required annotation budget B by identifying the point at which the estimation error of **Metric Match** selection equals the estimation error at S. We quantify cost savings as: $C(S) = (S - E_b(S)) \cdot t \cdot c$, where $E_b(S)$ denotes the **Metric Match** annotation budget that achieves equivalent estimation accuracy to stratified selection at budget S. For MedVAL, assuming 5 minutes per annotation at a cost of \$500 per annotator hour (as described in Unell et al. (2025)), this becomes for ICC: $\$1{,}083.33 = (50 - 24) \cdot \frac{1}{12} \cdot 500$.

We contribute an additional cost analysis in Appendix K, further highlighting the utility of **Metric Match** while demonstrating the associated break-even point of when human annotation would be more cost effective than multiple LLM calls.

4.2 Win-rate Analysis

We show in Table 1 the win-rate over trials between **Metric Match**'s estimation error with the target parameter and baseline estimation error. In terms of macro win-rate, we achieve an average win-rate of at least 0.794 with respect to all four baselines. This highlights how **Metric Match** wins against all four baselines on a significant fraction of instances. In terms of micro win-rate, we achieve an average win-rate of at least 0.538 against all four baselines. This serves as a demonstration that even on a single trial basis, the odds are in favor of using **Metric Match**. We include disaggregated win rate tables for individual axis performance in Appendix J.3, as well as win rates disaggregated by budget in Appendix I.

	Macro Win Rate				Micro Win Rate			
Method	Alpha	ICC	Rho	Tau	Alpha	ICC	Rho	Tau
Random	0.724	0.892	0.879	0.856	0.508	0.553	0.553	0.539
Random BC	0.757	0.873	0.933	0.889	0.545	0.557	0.593	0.576
Stratified	0.716	0.825	0.829	0.805	0.526	0.561	0.581	0.562
Cluster Text	0.699	0.868	0.843	0.805	0.563	0.600	0.585	0.577

Table 1: **Stratified has the lowest average macro win rate across metrics, while random has the lowest average micro win rate.** Values are averages over budgets 5–50 and across all 15 datasets.

Both macro win-rate and micro win-rate calculate the fraction of instances where **Metric Match** beats the baseline method of interest. However, the conceptual difference is that macro win-rate compares estimation error on each instance but averages over trials, while micro win-rate counts the individual wins on individual trials. As a result, macro win-rate captures wins regardless of the margin, whereas micro win-rate depends on how the variance coming from randomness in the method interacts with the margin. **Metric Match** achieves a macro win-rate and micro win-rate of greater than 50% with respect to all of the baselines, demonstrating it outperforms the baselines along both metrics.

Improved performance with **Metric Match** generalizes across model scales (see Appendix G for small model results) and outperforms random selection on alternative metrics including mean squared error (Appendix L). We additionally show results when matching on complementary met-

rics in Appendix M, highlighting that alternative signals of inter-model reliability may be helpful for optimal subset selection, as well.

4.3 Reliability Classification

In practice, a key motivation for estimating judge reliability is to determine whether the judge is sufficient to be used for the downstream task. This corresponds to a classification task of determining whether judge reliability is above a certain threshold t (Feng et al., 2026). Given an estimate ρ, we derive a prediction based on $1[\rho > t]$. We evaluate the performance of **Metric Match** for the purposes of this classification task in Table 2. We compute the accuracy of **Metric Match** at thresholds of $t \in \{0.1, 0.2, 0.3, ..., 0.8, 0.9\}$. We present averaged accuracies over thresholds in Table 2, with per-threshold performance reported in Appendix N. We highlight that **Metric Match** is best able to leverage its improved estimation precision, yielding higher overall accuracies for reliability classification as compared to baselines.

Method	Alpha	ICC	Rho	Tau
Metric Match	**0.881**	**0.887**	**0.876**	**0.884**
Random	0.869	0.850	0.850	0.863
Random BC	0.866	0.853	0.847	0.862
Stratified	0.872	0.864	0.850	0.862
Cluster Text	0.864	0.845	0.843	0.858

Table 2: **Metric Match has highest average classification accuracy over 9 thresholds.** When using judge reliability as a metric for classification, **Metric Match**'s improved estimation yields higher classification accuracy than baselines.

5 Discussion

We develop a new method, **Metric Match**, for estimating the reliability of LLM judges using synthetic labels to best select a representative subset for human annotation. **Metric Match** reduces estimation error by an average of 16.2% and annotation requirements by 28.1% relative to stratified selection, beating all evaluated baselines. In a case study on the MedVAL dataset, our cost model translates this reduction into savings of up to $1,083.33. We observe an average win rate against stratified selection of 0.794 on estimation error when averaged across contexts and budgets. We provide theoretical justification for this estimation approach, proving that under rank alignment assumptions between human-model and inter-model metrics on subsets, **Metric Match** provably outperforms random selection. Finally, we show that for a complementary task of reliability classification, where the judge reliability coefficient can be used as a statistic for deployment, **Metric Match** has the highest accuracy across baselines of 0.881. Together, these results demonstrate that improved estimation of the population-level target parameter enables practitioners to make more accurate decisions using fewer human annotations.

Future work may engage with a wider set of metrics beyond correlation-based targets. We include preliminary analysis for Mean Squared Error as a target parameter in Appendix L, showing that **Metric Match** outperforms random selection on non-correlation based metrics but with smaller absolute estimation error savings. We do not address how to improve the judge score or select the optimal judge from the ensemble. This work does not address how to incorporate human feedback online to improve estimation actively. Integrating active learning approaches may improve performance beyond our current batch setting, though this constitutes a different problem formulation.

More broadly, **Metric Match** offers a promising approach for practitioners to validate LLM judges more efficiently, reducing both the cost and time required for expert annotation. We hope that our work serves as a starting point for organizations to iterate faster on judge development, identify failure modes earlier, and to detect misalignment with human preferences while achieving equivalent estimation accuracy with fewer labeled examples.

AI USE STATEMENT

In this work, we used generative AI tools for generation of LLM judge annotations/scores as well as some result interpretation and data analysis. Additionally, we used generative AI tools for assistance in identifying relevant literature, plotting code generation, and improving readability and grammatical errors within our final manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Asad Aali, Vasiliki Bikia, Maya Varma, Nicole Chiou, Sophie Ostmeier, Arnav Singhvi, Magdalini Paschali, Ashwin Kumar, Andrew Johnston, Karimar Amador-Martinez, Eduardo Juan Perez Guerrero, Paola Naovi Cruz Rivera, Sergios Gatidis, Christian Bluethgen, Eduardo Pontes Reis, Eddy D. Zandee van Rilland, Poonam Laxmappa Hosamani, Kevin R Keet, Minjoung Go, Evelyn Ling, David B. Larson, Curtis Langlotz, Roxana Daneshjou, Jason Hom, Sanmi Koyejo, Emily Alsentzer, and Akshay S. Chaudhari. Medval: Toward expert-level medical text validation with language models, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

Anthropic. Claude 3.5 sonnet model card addendum. Technical report, Anthropic, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008. ISSN 0891-2017. doi: 10.1162/coli.07-034-R2. URL https://doi.org/10.1162/coli.07-034-R2.

E. Asgari, N. Montaña-Brown, M. Dubois, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8:274, 2025. doi: 10.1038/s41746-025-01670-7.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi, Asad Aali, Ashwin Nayak, Shivam Vedak, Sneha S. Jain, Birju Patel, Oluseyi Fayanju, Shreya Shah, Ethan Goh, Dong han Yao, Brian Soetikno, Eduardo Reis, Sergios Gatidis, Vasu Divi, Robson Capasso, Rachna Saralkar, Chia-Chun Chiang, Jenelle Jindal, Tho Pham, Faraz Ghoddusi, Steven Lin, Albert S. Chiou, Christy Hong, Mohana Roy, Michael F. Gensheimer, Hinesh Patel, Kevin Schulman, Dev Dash, Danton Char, Lance Downing, Francois Grolleau, Kameron Black, Bethel Mieso, Aydin Zahedivash, Wen wai Yim, Harshita Sharma, Tony Lee, Hannah Kirsch, Jennifer Lee, Nerissa Ambers, Carlene Lugtu, Aditya Sharma, Bilal Mawji, Alex Alekseyev, Vicky Zhou, Vikas Kakkar, Jarrod Helzer, Anurang Revri, Yair Bannett, Roxana Daneshjou, Jonathan Chen, Emily Alsentzer, Keith Morse, Nirmal Ravi, Nima Aghaeepour, Vanessa Kennedy, Akshay Chaudhari, Thomas Wang, Sanmi Koyejo, Matthew P. Lungren, Eric Horvitz, Percy Liang, Mike Pfeffer, and Nigam H. Shah. Medhelm: Holistic evaluation of large language models for medical tasks, 2025. URL https://arxiv.org/abs/2505.23802.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron

McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022. URL https://arxiv.org/abs/2211.03540.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. URL https://arxiv.org/abs/2006.14799.

Arun Tejasvi Chaganty, Ashwin Paranjape, Percy Liang, and Christopher D. Manning. Importance sampling for unbiased on-demand evaluation of knowledge base population. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Arun Tejasvi Chaganty, Stephen Mussman, and Percy Liang. The price of debiasing automatic metrics in natural language evaluation, 2018. URL https://arxiv.org/abs/1807.02202.

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do language models enjoy their own stories? Prompting large language models for automatic story evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142, 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00689. URL https://doi.org/10.1162/tacl_a_00689.

Cheng-Han Chiang and HungYi Lee. Can large language models be an alternative to human evaluations?, 2023. URL https://arxiv.org/abs/2305.01937.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. All that's 'human' is not gold: Evaluating human evaluation of generated text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.565. URL https://aclanthology.org/2021.acl-long.565/.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 3 edition, 1977.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks,

Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pil-

lai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Sham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy,

Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye,

Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu,

Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed,

Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzebski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, Hyun-Jeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan

Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen,

Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha,

Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang,

George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

E. Croxford, Y. Gao, N. Pellegrino, K. Wong, G. Wills, E. First, M. Schnier, K. Burton, C. Ebby, J. Gorski, M. Kalscheur, S. Khalil, M. Pisani, T. Rubeor, P. Stetson, F. Liao, C. Goswami, B. Patterson, and M. Afshar. Development and validation of the provider documentation summarization quality instrument for large language models. *Journal of the American Medical Informatics Association*, 32(6):1050–1060, June 2025. doi: 10.1093/jamia/ocaf068.

Yann Dubois et al. Alpacaeval 2.0: Automatic evaluation of instruction-following models. *arXiv preprint arXiv:2401.04088*, 2024.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation, 2021. URL https://arxiv.org/abs/2007.12626.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hEhxreaLdU.

Google. Gemini API pricing — Gemini 2.5 Pro. https://ai.google.dev/gemini-api/docs/pricing#gemini-2.5-pro, 2026. Accessed: 2026-09-01.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan,

Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo

Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Xun Huan, Jayanth Jagalur, and Youssef Marzouk. Optimal experimental design: Formulations and computations. *Acta Numerica*, 33:715–840, 2024. ISSN 1474-0508. doi: 10.1017/s0962492924000023. URL http://dx.doi.org/10.1017/S0962492924000023.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Maurice G Kendall. *Rank correlation methods*. Griffin, London, 1948.

Zachary Kenton, Noah Yamamoto Siegel, Janos Kramar, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, and Rohin Shah. On scalable oversight with weak LLMs judging strong LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=O1fp9nVraj.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. ISSN 1556-3707. doi: https://doi.org/10.1016/j.jcm.2016.02.012. URL https://www.sciencedirect.com/science/article/pii/S1556370716000158.

Klaus Krippendorff. Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970.

Chenyu Li, Zohaib Akhtar, Mingu Kwak, Yuelyu Ji, Hang Zhang, Tracey Obi, Yufan Ren, Xizhi Wu, Sonish Sivarajkumar, Harold P. Lehmann, Shyam Visweswaran, Michael J. Becich, Danielle L. Mowery, Renxuan Liu, Haoyang Sun, and Yanshan Wang. A scoping review of llm-as-a-judge in healthcare and the medjudge framework, 2026. URL https://arxiv.org/abs/2604.25933.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024.

David Liljequist, Britt Elfving, and Kirsti Skavberg Roaldsen. Intraclass correlation – a discussion and demonstration of basic features. *PLOS ONE*, 14:1–35, 07 2019. doi: 10.1371/journal.pone.0219854. URL https://doi.org/10.1371/journal.pone.0219854.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, July 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

Jayson G. Meyer, Ryan J. Urbanowicz, Pedro C. N. Martin, et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData Mining*, 16(1):20, 2023. doi: 10.1186/s13040-023-00339-9.

Microsoft Azure. DeepSeek models — pricing — AI Foundry models. https://azure.microsoft.com/en-us/pricing/details/ai-foundry-models/deepseek/, 2026. Accessed: 2026-09-01.

Reagan Mozer, Nicole E. Pashley, and Luke Miratrix. Stratified sampling for model-assisted estimation with surrogate outcomes. *arXiv preprint arXiv:2602.12992*, 2026.

H. T. Nguyen and A. W. M. Smeulders. Active learning using pre-clustering. In *International Conference on Machine Learning*, pages 623–630, 2004.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. URL https://aclanthology.org/D17-1238/.

OpenAI. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/, August 2025.

OpenAI. API pricing. https://developers.openai.com/api/docs/pricing, 2026. Accessed: 2026-09-01.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila

Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2203.02155.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. An active learning method using clustering and committee-based sample selection for sound event classification. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 116–120, 2018. doi: 10.1109/IWAENC.2018.8521336.

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

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges, 2025. URL https://arxiv.org/abs/2410.12784.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Anuraj J. Thirunavukarasu, Daniel S. W. Ting, Kishore Elangovan, et al. Large language models in medicine. *Nature Medicine*, 29(9):1930–1940, 2023. doi: 10.1038/s41591-023-02448-8.

Alyssa Unell, Noel Codella, Sam Preston, Peniel Argaw, Wen-wai Yim, Zelalem Gero, Cliff Wong, Rajesh Jena, Eric Horvitz, Amanda K. Hall, Ruican Rachel Zhong, Jiachen Li, Shrey Jain, Mu Wei, Matthew P Lungren, and Hoifung Poon. Cancerguide: Cancer guideline understanding via internal disagreement estimation. arXiv, September

2025. URL https://www.microsoft.com/en-us/research/publication/cancerguide-cancer-guideline-understanding-via-internal-disagreement-estimation/.

Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 376–397, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

Lucy Lu Wang, Yulia Otmakhova, Jay DeYoung, Thinh Hung Truong, Bailey E Kuehl, Erin Bransom, and Byron C Wallace. Automated metrics for medical multi-document summarization disagree with human evaluations. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Toronto, Canada, 2023. Association for Computational Linguistics.

Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient evaluation of llm performance with statistical guarantees, 2026. URL https://arxiv.org/abs/2601.20251.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. Importance-weighted label prediction for active learning with noisy annotations. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, 2012.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A METRIC FUNCTIONS

Here, we show equations and population information about our chosen metrics.

Intraclass Correlation Coefficient Intraclass Correlation Coefficient (ICC) was originally proposed as an extension to *interclass* correlation coefficient (Pearson's correlation coefficient (PCC)), and measures the extent to which the total variance in observed data is due to differences between groups, rather than within groups. In this perspective, the ICC is understood within the analysis of variance (ANOVA) framework. As opposed to PCC, the data are pooled in the mean calculation.

In the generic version, ICC is considered a measure that quantifies inter-rater reliability between k raters on n subjects, first introduced as an application of the metric in Shrout and Fleiss (1979). ICC measures reliability by decomposing the total variance in human evaluations into between-subjects variance and within-subjects error variance. The ICC determines the reliability of ratings by comparing the variability of different ratings of the same individuals to the total variation across all ratings and all individuals. As we only consider two raters, the human and the LLM, we consider the case $k = 2$.

Analogously, modern ICC estimators derive ICC through the random effects model framework. In the random effects model, X_{ij}, rating j on subject i, $i \in [n], j \in [k]$, is modeled as

$$X_{ij} = \mu + \alpha_i + c_j + \varepsilon_{ij}$$

such that μ is an unobserved overall mean, α_i is an unobserved random effect shared by all ratings on subject i, c_j is an unobserved random effect shared by all subject ratings by rater j, and ε_{ij} is an unobserved noise term. Each class of terms is assumed to be respectively identically distributed with expected value 0, and the terms are assumed to be uncorrelated. For certain random effects models, either α_i or c_j is neglected or considered fixed. We refer to Liljequist et al. (2019) for a comprehensive overview of ICC definitions and derivations relating classical estimators to the random effects model.

In our specific use case, we use a two-way consistency average, i.e. $\mathrm{ICC}(3, k)$ as this formulation treats *raters* as fixed effects, (i.e. c_j is fixed), meaning the same evaluation panel assesses all LLM outputs, and estimates reliability for the average rating across k evaluators rather than individual rater consistency. The numerator $(MS_R - MS_E)$ captures the true variance between different LLM responses after removing measurement error, while the denominator represents the total variance in averaged ratings, making $\mathrm{ICC}(3, k)$ particularly sensitive to systematic differences in how evaluators rate different model outputs while accounting for random measurement error within the evaluation process. With random effects model for $\mathrm{ICC}(3, k)$, the population ICC

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2/k}$$

We utilize the associated formula as the ICC metric for our experiments due to the appropriateness of the setting, random effects model, and use in previous empirical work (Bedi et al., 2025; Croxford et al., 2025; Li et al., 2024). See table below for reproduced formulas from (Liljequist et al., 2019) for intraclass correlation coefficient. We use $\mathrm{ICC}(3, k)$ in all of our analyses.

Name	Notation	Rater Model	Use Case	Formula
One-way single	ICC(1,1)	Random	Agreement of 1 random rater	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + (k-1)\mathrm{MS_E}}$
One-way average	ICC(1,k)	Random	Agreement of average random raters	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R}}$
Two-way absolute single	ICC(2,1)	Random	Absolute agreement of 1 random rater	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + (k-1)\mathrm{MS_E} + \frac{k}{n}(\mathrm{MS_C} - \mathrm{MS_E})}$
Two-way absolute average	ICC(2,k)	Random	Absolute agreement of average raters	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + \frac{1}{n}(\mathrm{MS_C} - \mathrm{MS_E})}$
Two-way consistency single	ICC(3,1)	Fixed	Consistency of 1 fixed rater	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R} + (k-1)\mathrm{MS_E}}$
Two-way consistency average	ICC(3,k)	Fixed	Consistency of average fixed raters	$\frac{\mathrm{MS_R} - \mathrm{MS_E}}{\mathrm{MS_R}}$
Pearson correlation	r	N/A	Correlation only (not agreement)	$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$

Notation:

- $\mathrm{MS_R}$: Mean square between targets (rows)
- $\mathrm{MS_C}$: Mean square between raters (columns)
- $\mathrm{MS_E}$: Residual mean square (error)
- n: Number of targets
- k: Number of raters

Formulas:

$$\mathrm{MS_R} = \frac{k}{n-1}\sum_{i=1}^{n}(S_i - \overline{X}_{\mathrm{tot}})^2$$

$$\mathrm{MS_C} = \frac{n}{k-1}\sum_{j=1}^{k}(M_j - \overline{X}_{\mathrm{tot}})^2$$

$$\mathrm{MS_E} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - M_j)^2 - k\sum_{i=1}^{n}(S_i - \overline{X}_{\mathrm{tot}})^2}{(n-1)(k-1)}$$

$$S_i = \frac{1}{k}\sum_{j=1}^{k} x_{ij}$$

$$M_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\overline{X}_{\mathrm{tot}} = \frac{1}{k \cdot n}\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$$

Krippendorff's α Krippendorff's alpha, introduced in Krippendorff (1970) is a statistical measure for reliability common in inter-rater reliability literature. The metric operates on the same reliability matrix as intraclass correlation coefficient, $X \in \mathbb{R}^{n\times k}$, such that $X_i j$ denotes the rating from rater j on subject i. Unlike ICC, Krippendorff's alpha is not generally positioned in the context of the random effects model, nor the analysis of variance (ANOVA) framework. The measure is highly versatile, designed to handle a variety of data types (nominal, ordinal, interval, and ratio) and common research obstacles including missing values.

The general form of the coefficient is based on the ratio of observed disagreement amongst raters to the disagreement expected by chance,

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement and D_e is the disagreement expected by chance. To define D_o and D_e, define a unit u as the ratings of all raters for item i:

- n be the total number of values assigned to all units.
- m_i be the number of coders who responded to subject i.
- v, k represent specific values (categories) in the data.

Then, the observed disagreement, D_o is the average of the distances between all pairs of values assigned to the same units. It is defined as:

$$D_o = \frac{1}{n}\sum_{i}\frac{1}{m_i - 1}\sum_{v}\sum_{k} n_{iv} n_{ik} \delta^2_{vk}$$

where

- n_{iv} and n_{ik} are the number of times values v and k were assigned to unit i.
- δ^2_{vk} is a distance function (metric) that depends on the level of measurement (Nominal, Ordinal, Interval, or Ratio).

The expected disagreement, D_e, is the disagreement that would occur if the values were assigned to units completely at random, constrained only by the overall frequency of each value in the dataset:

$$D_e = \frac{1}{n(n-1)} \sum_v \sum_k n_v n_k \delta^2_{vk}$$

where

- n_v and n_k are the total number of times values v and k were used across all units.

Finally, we turn to the distance functions (δ^2_{vk}). The power of α lies in the metric δ^2_{vk}, which accounts for the magnitude of disagreement:

- **Nominal:** $\delta^2_{vk} = \begin{cases} 0 & \text{if } v = k \\ 1 & \text{if } v \neq k \end{cases}$
- **Ordinal:** $\delta^2_{vk} = \left(\sum_{g=v}^{k} n_g - \frac{n_v + n_k}{2} \right)^2$
- **Interval:** $\delta^2_{vk} = (v-k)^2$
- **Ratio:** $\delta^2_{vk} = \left(\frac{v-k}{v+k} \right)^2$

Substituting D_o and D_e back into the original formula:

$$\alpha = 1 - (n-1) \frac{\sum_i \frac{1}{m_i - 1} \sum_v \sum_k n_{iv} n_{ik} \delta^2_{vk}}{\sum_v \sum_k n_v n_k \delta^2_{vk}}$$

This formula ensures that when agreement is perfect, $D_o = 0$ and $\alpha = 1$. When disagreement matches chance, $D_o = D_e$ and $\alpha = 0$.

Spearman r Spearman's correlation coefficient, which we denote as r, evaluates the monotonic relationship between two rank variables. In particular, let X and Y be random variables with paired samples (X_i, Y_i) and rank variables $R[X_i], R[Y_i]$, $i = 1, \ldots, n$. Spearman's correlation coefficient is the Pearson correlation coefficient on the *rank* variables. Unlike Pearson correlation, which measures linear relationships, Spearman correlation measures the strength and direction of the association between two ranked variables. The equation representing Spearman's rank correlation is shown below:

$$r = \frac{\text{Cov}[R[X], R[Y]]}{\sigma_{R[X]} \sigma_{R[Y]}}$$

If there are no ties in ranks, the coefficient can be calculated as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = R[X_i] - R[Y_i]$, $i = 1, \ldots, n$.

Kendall's τ Kendall's tau rank correlation measures the ordinal association between two variables based on the similarity of the orderings of the data. The measure is a rank correlation metric as in Spearman's r, and operates on ranks of pairs of random variables, (X_i, Y_i), $i = 1, \ldots, n$. Differing from Spearman's r, Kendall's τ considers number of concordant and discordant pairs (of pairs) for calculation of rank correlation. A pair of pairs (X_i, Y_i), (X_j, Y_j), such that $i < j$ is *concordant* if the sort order matches, on X_i, X_j and Y_i, Y_j, i.e. if $\text{sign}\big(R[X_i] - R[X_j]\big) = \text{sign}\big(R[Y_i] - R[Y_j]\big)$. Otherwise, the pair is discordant. Then τ is calculated as

$$\tau = \frac{n_C - n_D}{\frac{1}{2} n(n-1)} \tag{1}$$

where n_C is the number of concordant pairs, and n_D is the number of discordant pairs. The value ranges from -1 (perfect inversion) to $+1$ (perfect agreement).

B Dataset Details

The axes of evaluation for each dataset and number of raters per data point are detailed in Table 3.

Table 3: Datasets, evaluation axes, rating scales, and number of raters per datapoint.

Dataset	Axes of Evaluation	Scale	Raters per datapoint
SummEval	Coherence, Consistency, Fluency, Relevance	1–5	8
HANNA	Relevance, Coherence, Empathy, Surprise, Engagement, Complexity	1–5	3
MedVAL	Risk	1–4	1–3
MSLR	Fluency, Population, Intervention, Outcome	0–2	1–2

C Rank Estimation Error Derivation

As noted in Section 3.2, the performance of metric matching relies on the joint distribution over the ranks of the inter-model and human-model estimation errors in subsets $S \subseteq \mathcal{X}$.

Here, we clarify the precise relationship between micro-average win-rate and the expected rank of subset S^* over possible subsets with respect to the human-model estimation error. Consider the candidate subsets $S_1, \ldots, S_C \sim \mathcal{X}$: let $\epsilon_j := |\widehat{\rho}_{S_j} - \rho|$ and $\epsilon_j^{\text{IM}} := |\widehat{\rho}_{S_j}^{\text{IM}} - \rho^{\text{IM}}|$, and j^* be the index of S^* in $\{1, \ldots, C\}$. The expected micro-average win-rate of metric matching is exactly $\mathbb{E}\Big[\mathbb{1}\big[\epsilon_{j^*} < \epsilon^{\text{rand}}\big]\Big] = \Pr(\epsilon_{j^*} < \epsilon^{\text{rand}})$. Let R denote the rank operator with respect to human-model estimation error. Recall that the rank of S^* with respect to inter-model estimation error is 1 by definition. The expected micro-average win-rate of metric matching depends linearly on the expected rank of the j^*th subset in the human-model, $\mathbb{E}[R[\epsilon_{j^*}]]$.

The performance of metric matching relies on the human-model estimation error rank of S^*, which is, by definition the subset with rank 1 in the inter-model estimation error. We develop the intuition more formally below.

As in the main paper, consider candidate subsets $S_1, \ldots, S_C \sim \mathcal{X}$. Let $\epsilon_j = |\widehat{\rho}_{S_j} - \rho|$ be the human-model estimation error for subset S_j, $j = 1, \ldots, C$. Similarly, let $\epsilon_j^{\text{IM}} = |\widehat{\rho}_{S_j}^{\text{IM}} - \rho^{\text{IM}}|$. For initial understanding, consider the best-case scenario, when the rank with respect to human-model estimation error of each subset exactly matches the rank with respect to inter-model estimation error of each subset. For now, we are eliding the probabilistic component and simply assuming that this holds for all subsets $S \subseteq \mathcal{X}$ of size b. Formally, $R[\epsilon_j] = R[\epsilon_j^{\text{IM}}]$. This ensures that when we choose the subset S^* with the smallest difference $|\widehat{\rho}_{S^*}^{\text{IM}} - \rho^{\text{IM}}|$, we choose the optimal subset w.r.t. human-model metric as well, i.e.

$$\begin{aligned} S^* &= \arg \min_{S \subseteq X, |S|=b} |\widehat{\rho}_S^{\text{IM}} - \rho^{\text{IM}}| \\ &\equiv \arg \min_{S \subseteq X, |S|=b} |\widehat{\rho}_S - \rho| \end{aligned}$$

This is an extremely strong condition and likely never satisfied in practice. Furthermore, we don't even need this to hold in order for metric matching to select the optimal subset with respect to the human-model metric. We solely require that the *top* rank matches, i.e. that $R[\epsilon_j] = R[\epsilon_j^{\text{IM}}]$ for any $j = 1, \ldots, C$ such that the rank equals 1. In this case, metric matching will always choose the optimal subset (over possible candidate subsets $S_1, \ldots, S_C$, not over all subsets).

Again, in practice, this is unlikely to occur but it gives some intuition for how to understand when metric matching should work (and in particular, when metric matching should beat random selection). On a high-level, when the rank relationship between human-model estimation error and inter-model estimation error with respect to subsets of size b is as close as possible to that condition, we can expect to perform well. We formalize this more below by comparison to random.

In particular, consider the *expected* micro win-rate of our method. This depends both on the expectation of estimation errors of C random subsets, and the estimation error of a $C+1$th subset, selected by random selection. We provide a brief derivation that shows: when the expected rank $R[\epsilon_{j^*}]$ of the human-model estimation error of the *best* subset (rank 1) w.r.t. the inter-model estimation is less than $\frac{C+1}{2}$, we beat random the majority of the time. Furthermore, our expected micro win-rate is linear in the expected rank $R[\epsilon_{j^*}]$.

Lemma 1. *Given the notation introduced in this section, the expected micro win rate:*

$$\begin{aligned}\mathbb{E}\big[\mathbb{1}[\epsilon_{j^*} < \epsilon^{rand}]\big] &= \Pr[\epsilon_{j^*} < \epsilon^{rand}] \\ &= 1 - \frac{\mathbb{E}[R[\epsilon_{j^*}]]}{C+1}\end{aligned}$$

This admits a straightforward derivation. Given budget b, randomly sampled candidate subsets $S_1, \ldots, S_C$ from which $S^* = S_{j^*}$ is selected by the condition that $R[\epsilon^{\text{IM}}_{j^*}] = 1$, consider the rank of subset j^* w.r.t. the human-model estimation error over subsets $S_1, \ldots, S_C$ **and** the subset selected by random selection S_r. Since all subsets are selected i.i.d. from the same distribution over $\mathcal{X}^b$, we can concretely analyze the chance that S_r has a better rank than S_{j^*} in the human-model estimation error.

$$\begin{aligned}\Pr[\epsilon_{j^*} < \epsilon^{\text{rand}}] &= \sum_{j=1}^{C} \Pr[R[\epsilon^{\text{rand}}] > j, R[\epsilon_{j^*}] = j] \\ &= \sum_{j=1}^{C} \Pr[R[\epsilon^{\text{rand}}] > j \mid R[\epsilon_{j^*}] = j] \cdot \Pr[R[\epsilon_{j^*}] = j] \\ &= \sum_{j=1}^{C} \Pr[R[\epsilon^{\text{rand}}] > j] \cdot \Pr[R[\epsilon_{j^*}] = j] \\ &= \sum_{j=1}^{C} \Big(1 - \frac{j}{C+1}\Big) \Pr[R[\epsilon_{j^*}] = j] \\ &= 1 - \frac{\mathbb{E}[R[\epsilon_{j^*}]]}{C+1}\end{aligned}$$

The third to fourth line in the above derivation follows from the formula for the random variable ϵ^{rand} having greater than rank j out of $C+1$ variables (estimation errors ϵ_j on each of $S_1, \ldots, S_C$ and S_r).

Condition 1 (Expected rank condition)**.** $\mathbb{E}\big[R[\epsilon_{j^*}]\big] < \frac{C+1}{2}$

Theorem 2. *If the expected rank condition (Condition 1) is satisfied, then the expected micro win-rate over random selection is greater than* 0.5*, i.e. metric matching beats random the majority of the time.*

This follows directly from the lemma. We include empirical results validating the theory in Figures 4 and 5.

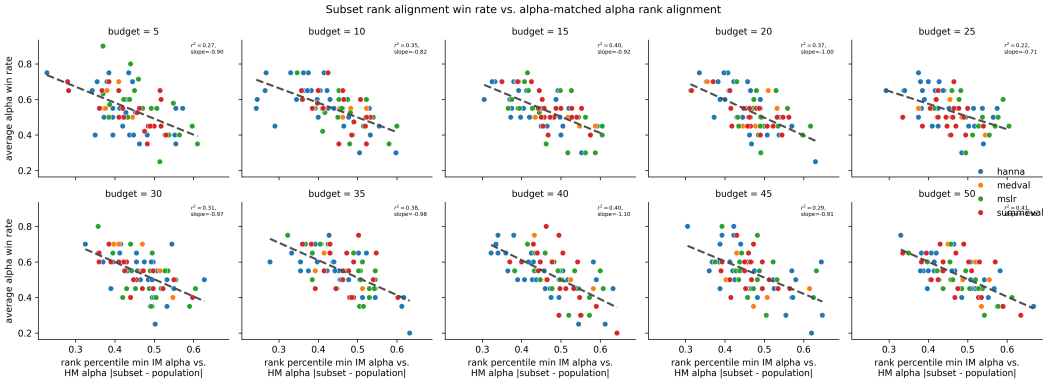


Figure 4: Empirical micro win-rate versus empirical average human-model estimation error rank percentile, $\frac{\sum_{t=1}^{N} R[\epsilon_{j^*(t)}]}{C+1}$, over $N = 40$ trials in metric-matched α estimation. As expected, the relationship is linear across budgets.

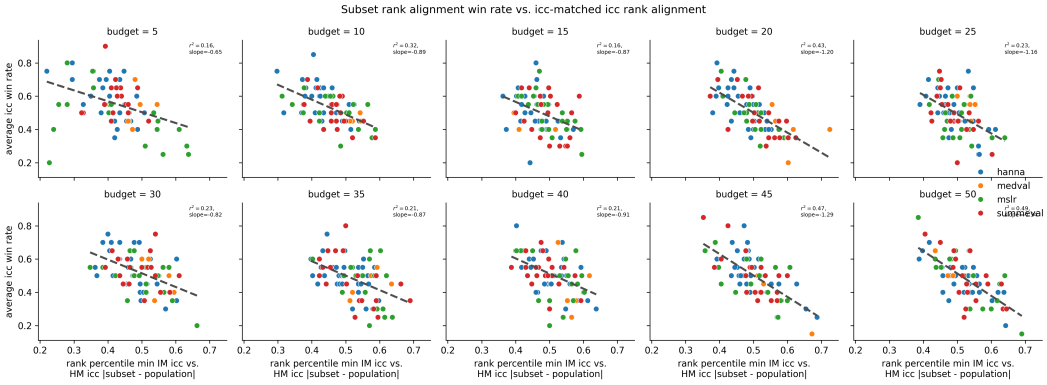


Figure 5: Observed micro win-rate versus observed average human-model estimation error rank percentile, $\frac{\sum_{t=1}^{N} R[\epsilon_{j^*(t)}]}{C+1}$, in metric-matched ICC estimation. As expected, the relationship is linear across budgets.

D Model-Human Misalignment Examples

In this section, we illustrate how **Metric Match** does not rely on model-model agreement acting as a reasonable proxy for human-model agreement. We show examples where the model-model correlation (i.e., the correlation between the LLM judge labels M and synthetic labels from a different model M') is much higher than human-model correlation (i.e., the correlation between the human labels H and the LLM judge labels M), but **Metric Match** still significantly outperforms random selection in terms of win-rate (Table 4). This suggests that **Metric Match** performs well even in the face of correlated errors across models.

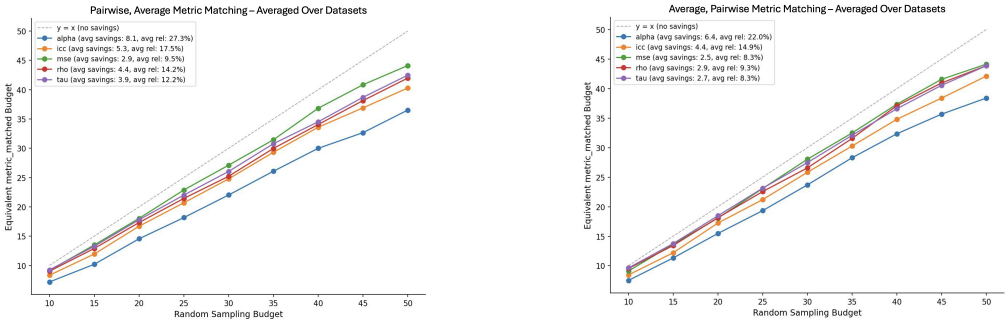
Table 4: Comparison between Model–Model and Human–Model Estimation Error and Win Rate

Dataset	Axis	Metric	Inter-model	Human-model	Macro win-rate
hanna	Complexity	α	0.603	0.427	1.000
		ICC	0.812	0.652	1.000
		ρ	0.671	0.453	1.000
		τ	0.613	0.376	1.000
	Empathy	α	0.680	0.415	1.000
		ICC	0.854	0.623	1.000
		ρ	0.715	0.406	1.000
		τ	0.661	0.338	0.800
	Engagement	α	0.743	0.293	1.000
		ICC	0.891	0.712	1.000
		ρ	0.676	0.484	1.000
		τ	0.640	0.411	1.000
	Surprise	α	0.585	0.148	0.800
		ICC	0.753	0.422	1.000
		ρ	0.636	0.303	1.000
		τ	0.559	0.245	1.000

E Metric Match Algorithmic Ablations

E.1 Average Aggregation

We show that the way we aggregate model information impacts the performance of our algorithm. Given a model suite $\mathcal{M} = \{M_1, M_2, \ldots, M_n\}$ and a target model $T \in \mathcal{M}$, we present an alternative aggregation approach in Figure 6b. We calculate the target metric by $\hat{\rho} = f\left(T, \frac{1}{|\mathcal{M}|}\sum_{M_i \in \mathcal{M}} M_i\right)$. This essentially constructs an average score function from the scorers in $\mathcal{M}$, $M^{\text{avg}} : \mathcal{X} \to \mathcal{Y}$ s.t. $M^{\text{avg}}(x) = \frac{1}{|\mathcal{M}|}\sum_{M' \in \mathcal{M}} M'(x)$. We posit that this fails to capture all necessary signal from individual judges, thus causing the alternative aggregation method that we employ in the main body to perform better.



(a) (b)

Figure 6: Aggregation approaches impact effectiveness of **Metric Match**, with the optimal approach being to calculate the reliability of our target model with each ensemble member individually and average these results (a) as opposed to averaging the ensemble labels and calculating the reliability coefficient with a single "averaged" ensemble point (b). Comparison is against random selection.

Estimation error win rates as shown in Table 5 are higher for this aggregation method, indicating that although the gains are less significant, they are equally if not more reliable.

Table 5: Average-Pairwise Metric Matching Performance: Estimation Error and Threshold Classification Macro Win Rate. Comparisons are made against random selection baseline.

	Estimation Error (Win Rate)				*Threshold Classification (Win Rate)*			
Budget	α	ICC	τ	ρ	α	ICC	τ	ρ
5	0.867	0.760	0.587	0.507	0.631	0.625	0.797	0.671
10	0.920	0.840	0.760	0.667	0.623	0.667	0.739	0.611
15	0.893	0.787	0.773	0.733	0.750	0.653	0.661	0.791
20	0.867	0.733	0.693	0.680	0.644	0.612	0.571	0.714
25	0.853	0.733	0.680	0.707	0.667	0.585	0.630	0.774
30	0.773	0.720	0.733	0.773	0.795	0.516	0.730	0.655
35	0.840	0.680	0.627	0.733	0.675	0.477	0.649	0.623
40	0.760	0.720	0.640	0.587	0.694	0.550	0.630	0.700
45	0.853	0.747	0.640	0.653	0.611	0.567	0.600	0.660
50	0.880	0.800	0.720	0.720	0.581	0.613	0.556	0.651
Avg	0.851	0.752	0.685	0.676	0.667	0.587	0.656	0.685

E.2 Number of Candidate Subsets C

Here, we vary the number of candidate subsets considered for metric matching between $C = 10$ to $C = 200$, and report the estimation errors with comparison to random baseline in the red horizontal line in Figure 8 for ICC and Figure 7 for alpha estimation. We note that these values correspond to macro-average win-rate, and thus may not represent the results we would expect to see with micro-average win-rate. We do not observe an interpretable trend across datasets or budgets between macro-average and number of candidate subsets. Additionally, due to computational constraints, we only study up to 200 candidate subsets, which may be well below the number required to observe any relationship (recall total possible subsets is $\binom{|\mathcal{X}|}{b}$).

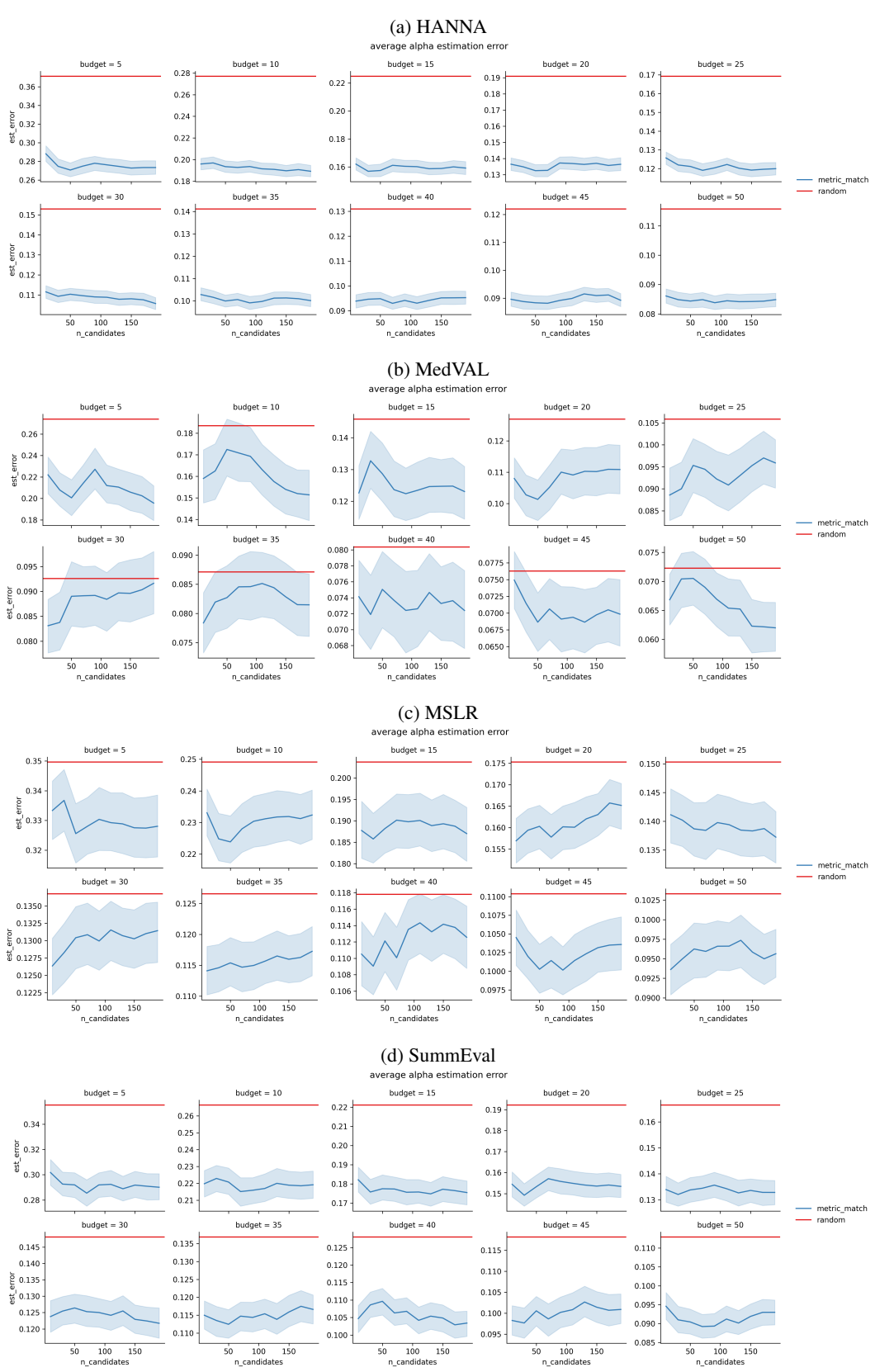


Figure 7: Ablation study over number of candidate subsets considered: average estimation error in α metric matching at each budget for each dataset for runs of metric matching with $C = 10$ up to $C = 200$ candidate subsets with $N = 40$ trials each.

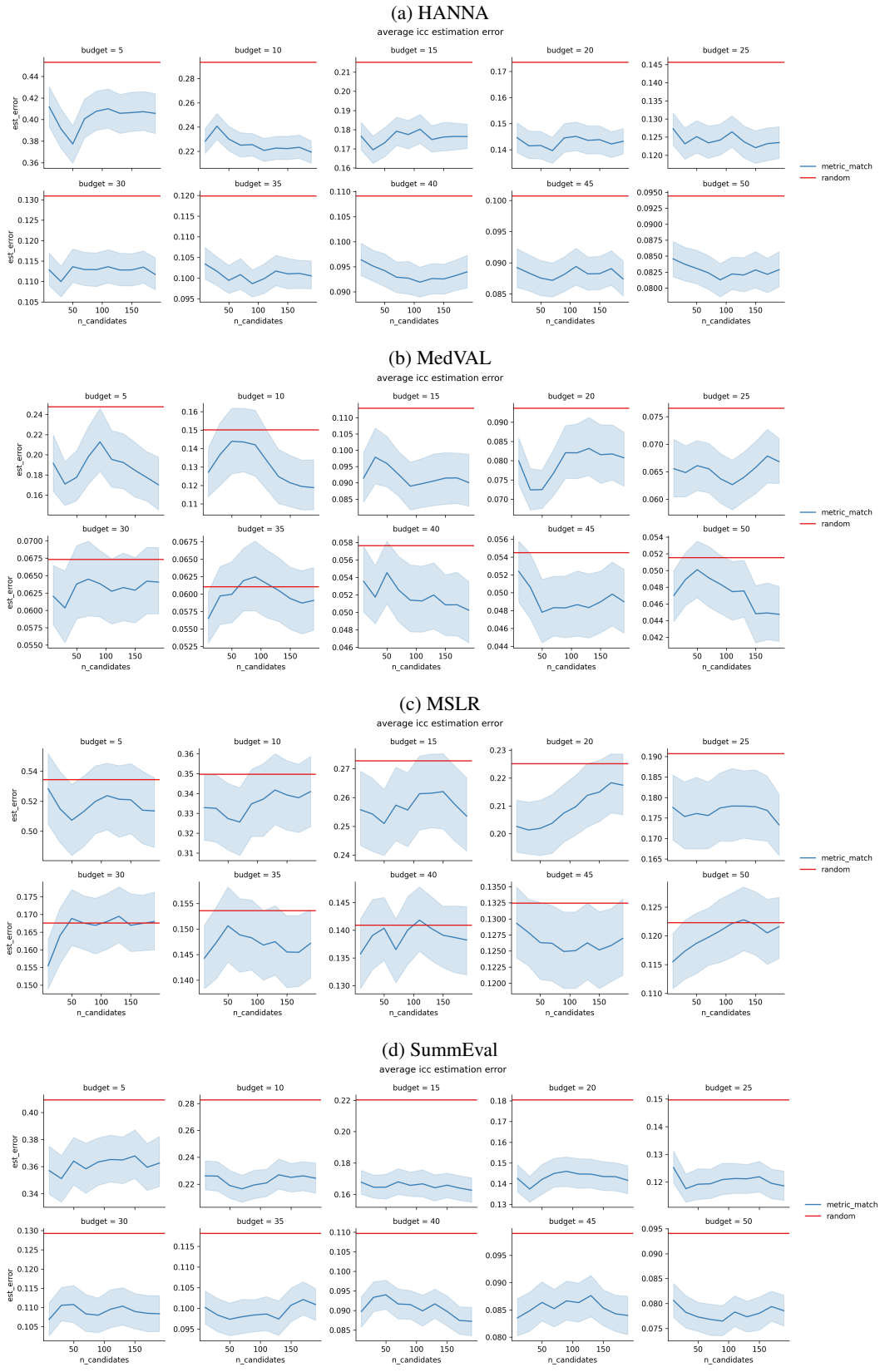


Figure 8: Ablation study over number of candidate subsets considered: average estimation error in ICC metric matching at each budget for each dataset for runs of metric matching with $C = 10$ up to $C = 200$ candidate subsets with $N = 40$ trials each.

F Additional Baseline Implementation Details

Cluster text-based In general, the text samples X for which an LLM judge produces a rating include the relevant prompt and the LLM-generated text, the latter of which is judged with respect to the evaluation axis given the prompt. For the cluster-based selection, we feed both the prompt text and the LLM-generated text through an encoder model, specifically `all-MiniLM-L6-v2` from the HuggingFace SentenceTransformers repository Reimers and Gurevych (2019). The resulting text embeddings are clustered using the KMedoids algorithm, with k set to the budget size b. The $k = b$ medoids selected constitute the subset S. Given the constructed S, we compute $\widehat{\rho}_S$ as the standard sample estimate on S.

Stratified Selection For the stratified selection baseline, we partition the evaluation pool into $m = 4$ strata $\{Q_1, \ldots, Q_m\}$ defined by the empirical quantiles of the target model's scores $f(X)$, so that Q_j contains the examples whose scores fall between the $(j-1)/m$ and j/m quantiles. We then allocate the budget as evenly as possible across strata, drawing $b_j = \lfloor b/m \rfloor + \mathbb{1}[j \leq b \bmod m]$ examples uniformly at random without replacement from Q_j. The union of the per-stratum draws constitutes the subset S, with $|S| = b$. By construction, S spans the full range of target-model scores rather than concentrating on high-density score regions. Given the constructed S, we compute $\widehat{\rho}_S$ as the standard sample estimate on S.

G Target and Ensemble Ablations

G.1 Small models

Here we present the same analysis but with a smaller, less powerful suite of models to show generalizability. Models include GPT-4o-mini (OpenAI et al., 2024), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Google-gemma-3-1b-it (Team et al., 2025), and Qwen2.5-7B-Instruct (Bai et al., 2023).

	Estimation Error					Thresholding				
Budget	Alpha	ICC	Rho	Tau	MSE	Alpha	ICC	Rho	Tau	MSE
5	0.92	0.90	0.91	0.95	0.85	0.74	0.48	0.77	0.81	0.68
10	0.80	0.92	0.81	0.84	0.72	0.92	0.59	0.89	0.91	0.79
15	0.80	0.88	0.82	0.84	0.68	0.84	0.62	0.94	0.96	0.79
20	0.78	0.85	0.83	0.81	0.62	0.88	0.61	0.83	0.96	0.66
25	0.78	0.83	0.88	0.86	0.62	1.00	0.58	0.95	1.00	0.80
30	0.78	0.85	0.86	0.84	0.53	0.92	0.85	0.92	1.00	0.73
35	0.82	0.87	0.84	0.79	0.55	1.00	0.77	0.88	1.00	0.81
40	0.73	0.92	0.84	0.79	0.50	1.00	0.76	0.94	1.00	0.76
45	0.70	0.90	0.78	0.79	0.47	1.00	0.84	0.67	1.00	0.73
50	0.73	0.85	0.74	0.74	0.45	1.00	0.75	1.00	—	0.81
Average	0.78	0.88	0.83	0.83	0.60	0.93	0.69	0.88	0.96	0.76

Table 6: Estimation Error and Thresholding win rate by Budget w/ small ensemble against random selection baseline.

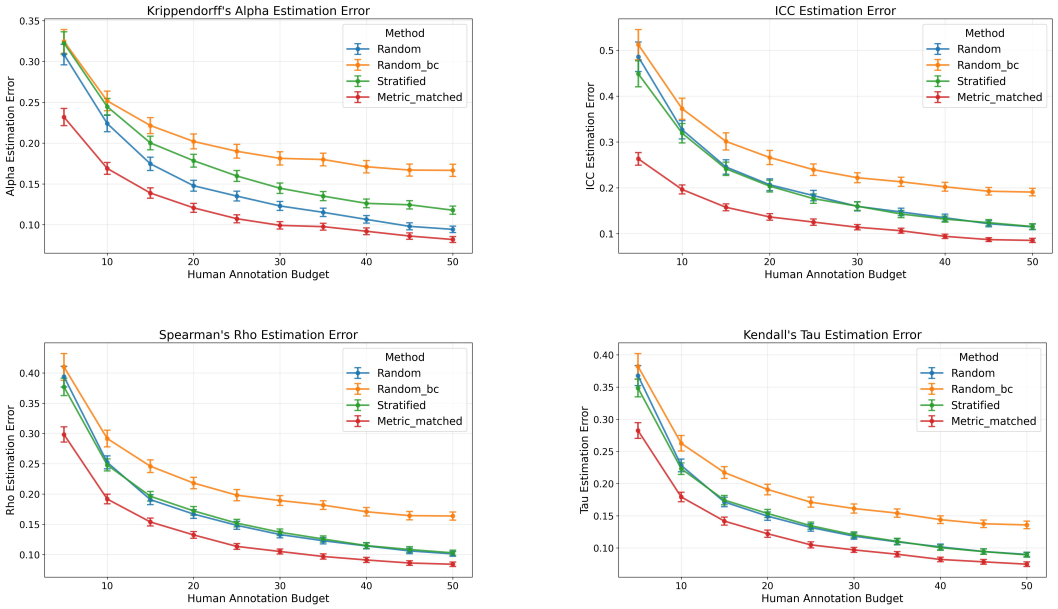


Figure 9: **Small models show similar performance gains across metrics and budgets.** We report results from a suite of small models to highlight that this approach generalizes across model scale.

G.2 Single Model Ensemble Ablations

We provide additional ablations for changing the ensemble itself that provides synthetic labels for selection of points to annotate. We see that the largest gains over random are present when an ensemble of models is used as opposed to individual models, which are more prone to reliance on the correlation of any individual model with human feedback for a given dataset. Ensembling in this setting helps improve robustness. Further, we identify that the target models must be roughly aligned capability-wise with their ensemble, and having an ensemble that is miscalibrated to the target results in worse performance.

Table 7: **Percent Relative Improvement over Random by Method and Metric**. These results highlight that single model ensembles don't generalize as well as multi-model ensembles.

Metric	Strong Ens.+ Strong Target	Claude-Only	GPT-5-Only	Weak Ens.+ Strong Target	Weak Ens.+ Weak Target
ICC	26.65%	15.71%	23.56%	−34.49%	33.09%
Alpha	9.05%	-2.88%	0.80%	−46.87%	21.56%
Rho	20.95%	9.61%	13.57%	−19.26%	21.67%
Tau	17.93%	5.74%	11.12%	−18.36%	20.11%

H Relative Estimation-Error Improvement over the Best Baseline

For each evaluation metric and annotation budget, the relative estimation error improvement of Metric Match over a baseline is calculated as

$$\text{Improvement}(\%) = 100 \times \frac{\overline{E}_{\text{baseline}} - \overline{E}_{\text{MetricMatch}}}{\overline{E}_{\text{baseline}}}, \tag{2}$$

where $\overline{E}$ is the mean estimation error pooled across datasets, models, evaluation axes, and trials. Positive values indicate lower error for Metric Match, while negative values indicate higher error. The average row reports the unweighted mean across annotation budgets.

Table 8: Relative estimation error improvement (%) of Metric Match over four baseline selection methods across annotation budgets. Positive values indicate that Metric Match achieves lower estimation error.

(a) vs. Stratified selection

Budget	ICC	Alpha	Rho	Tau
5	49.06	18.91	22.42	22.04
10	38.78	14.76	24.75	24.06
15	32.86	11.40	22.66	20.28
20	23.85	6.12	21.94	17.97
25	17.98	5.55	21.48	19.26
30	13.61	3.42	19.69	15.10
35	11.08	2.11	19.10	13.49
40	10.11	0.76	17.17	11.49
45	7.57	4.62	17.16	12.10
50	4.04	1.06	17.10	11.37
Avg.	**20.89**	**6.87**	**20.35**	**16.72**

(b) vs. Random selection

Budget	ICC	Alpha	Rho	Tau
5	54.63	25.44	25.03	25.60
10	47.17	23.65	28.42	28.35
15	39.91	19.48	23.79	21.72
20	31.47	9.88	24.19	20.90
25	26.48	9.10	23.44	21.53
30	21.33	5.77	20.89	16.99
35	15.72	0.79	18.43	13.10
40	11.67	−1.90	14.33	9.26
45	10.12	0.46	15.07	10.83
50	8.00	−2.22	15.91	11.07
Avg.	**26.65**	**9.05**	**20.95**	**17.93**

(c) vs. Random BC

Budget	ICC	Alpha	Rho	Tau
5	51.61	17.53	21.19	22.46
10	42.55	14.25	25.71	26.29
15	36.10	13.69	25.50	22.80
20	29.27	10.31	28.30	23.08
25	25.23	12.17	28.78	24.06
30	21.27	11.85	28.77	20.74
35	19.78	13.47	30.96	20.76
40	18.44	13.59	30.43	19.33
45	18.67	18.13	31.82	20.25
50	18.14	17.05	34.11	21.88
Avg.	**28.11**	**14.20**	**28.56**	**22.16**

(d) vs. Cluster Text

Budget	ICC	Alpha	Rho	Tau
5	58.24	27.93	27.74	28.38
10	50.80	29.89	30.15	30.30
15	47.95	28.84	28.14	26.57
20	42.00	25.15	29.18	26.66
25	34.19	20.72	26.89	25.64
30	28.75	17.83	24.23	21.21
35	20.84	11.28	22.15	17.92
40	15.73	6.73	18.70	14.46
45	10.56	5.37	15.23	11.88
50	5.07	−2.07	13.62	9.65
Avg.	**31.41**	**17.17**	**23.60**	**21.27**

I ESTIMATION ERROR WIN RATE PERFORMANCE DIS-AGGREGATED BY BUDGET

We show in Table 9 the estimation error macro and micro win rates between **Metric Match** and baselines.

J RESULTS BY DATASET AXIS

Here we present disaggregated results at the dataset level in order to highlight the robustness of this method across datasets, metrics, and budgets. We present annotation savings by dataset with Metric Match compared to both random and stratified baselines (Figure 10). We highlight that poor performance for certain (metric, dataset) pairs are empirically driven by low true reliability parameters, reporting true reliability parameters (Table 10) and estimation error analysis disaggregated by dataset (Section J.3). Finally, we include a post-hoc failure case analysis for the instances where our method underperforms (Section J.4).

Table 9: **Baseline metric matching performance: estimation error.** Macro and micro win rates of Metric Match against each baseline, by annotation budget.

(a) Stratified selection

	Macro				*Micro*			
Budget	α	ICC	ρ	τ	α	ICC	ρ	τ
5	0.827	0.933	0.947	0.867	0.563	0.624	0.597	0.595
10	0.760	0.960	0.947	0.933	0.546	0.609	0.602	0.588
15	0.733	0.907	0.853	0.880	0.542	0.595	0.585	0.569
20	0.720	0.880	0.867	0.813	0.515	0.560	0.582	0.565
25	0.693	0.907	0.813	0.840	0.514	0.553	0.582	0.575
30	0.733	0.813	0.800	0.787	0.515	0.561	0.579	0.563
35	0.707	0.787	0.827	0.787	0.517	0.543	0.576	0.544
40	0.680	0.747	0.720	0.707	0.506	0.529	0.565	0.534
45	0.640	0.667	0.747	0.707	0.528	0.519	0.565	0.539
50	0.667	0.653	0.773	0.733	0.513	0.520	0.575	0.547
Avg.	**0.716**	**0.825**	**0.829**	**0.805**	**0.526**	**0.561**	**0.581**	**0.562**

(b) Random selection

	Macro				*Micro*			
Budget	α	ICC	ρ	τ	α	ICC	ρ	τ
5	0.853	1.000	0.947	0.987	0.574	0.641	0.580	0.575
10	0.827	0.987	0.947	0.973	0.569	0.616	0.595	0.584
15	0.747	0.947	0.907	0.907	0.537	0.591	0.562	0.552
20	0.733	0.920	0.960	0.920	0.507	0.559	0.556	0.546
25	0.720	0.907	0.880	0.907	0.511	0.554	0.560	0.555
30	0.733	0.893	0.867	0.893	0.501	0.548	0.551	0.535
35	0.653	0.853	0.893	0.773	0.476	0.522	0.540	0.521
40	0.653	0.840	0.800	0.760	0.467	0.511	0.529	0.505
45	0.667	0.800	0.800	0.747	0.476	0.505	0.527	0.509
50	0.653	0.773	0.787	0.693	0.465	0.484	0.528	0.509
Avg.	**0.724**	**0.892**	**0.879**	**0.856**	**0.508**	**0.553**	**0.553**	**0.539**

(c) Random selection with bias correction

	Macro				*Micro*			
Budget	α	ICC	ρ	τ	α	ICC	ρ	τ
5	0.880	0.960	0.853	0.867	0.540	0.595	0.547	0.562
10	0.787	1.000	0.947	0.987	0.537	0.569	0.572	0.591
15	0.773	0.973	0.947	0.933	0.540	0.579	0.572	0.575
20	0.707	0.907	0.973	0.947	0.532	0.554	0.583	0.588
25	0.720	0.880	0.933	0.960	0.545	0.558	0.590	0.589
30	0.733	0.840	0.920	0.867	0.532	0.546	0.598	0.569
35	0.733	0.813	0.920	0.827	0.543	0.536	0.613	0.581
40	0.747	0.827	0.947	0.867	0.548	0.539	0.602	0.559
45	0.760	0.800	0.947	0.813	0.563	0.551	0.622	0.569
50	0.733	0.733	0.947	0.827	0.569	0.540	0.631	0.579
Avg.	**0.757**	**0.873**	**0.933**	**0.889**	**0.545**	**0.557**	**0.593**	**0.576**

(d) Cluster text

	Macro				*Micro*			
Budget	α	ICC	ρ	τ	α	ICC	ρ	τ
5	0.813	1.000	0.933	0.893	0.611	0.666	0.590	0.601
10	0.813	0.973	0.920	0.907	0.618	0.662	0.608	0.610
15	0.813	0.987	0.907	0.880	0.621	0.655	0.599	0.603
20	0.773	0.933	0.947	0.933	0.596	0.642	0.616	0.607
25	0.693	0.853	0.853	0.827	0.578	0.616	0.604	0.596
30	0.707	0.907	0.840	0.827	0.563	0.595	0.586	0.576
35	0.653	0.840	0.787	0.787	0.528	0.576	0.585	0.566
40	0.600	0.773	0.760	0.693	0.517	0.546	0.564	0.547
45	0.587	0.747	0.747	0.667	0.512	0.533	0.551	0.533
50	0.533	0.667	0.733	0.640	0.489	0.505	0.545	0.531
Avg.	**0.699**	**0.868**	**0.843**	**0.805**	**0.563**	**0.599**	**0.585**	**0.577**

J.1 ANNOTATION SAVINGS.

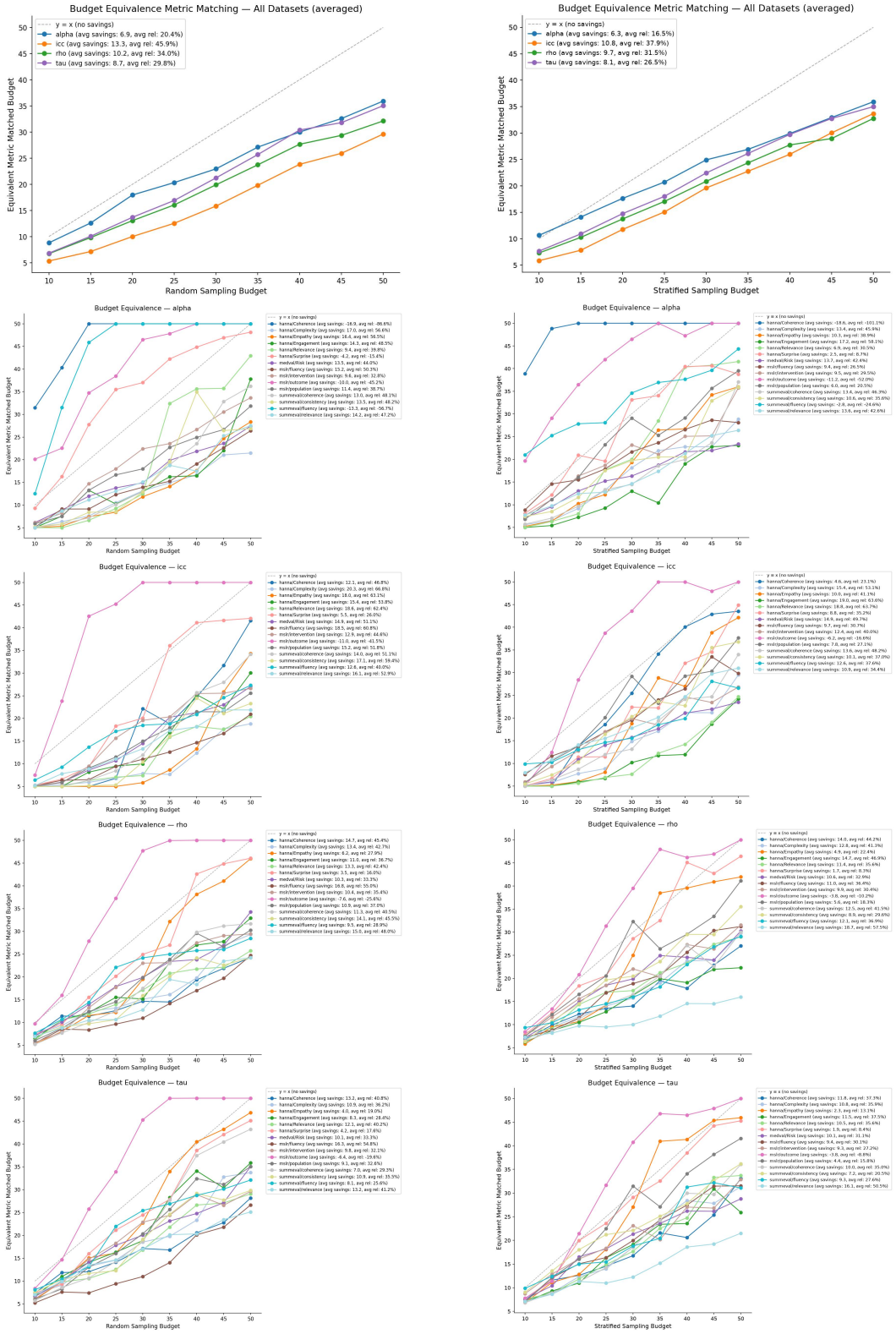


Figure 10: Annotation savings for **Metric Match** compared to Random Baseline (left) as well as stratified selection baseline (right).

J.2 TRUE RELIABILITY COEFFICIENT.

We highlight the true reliability coefficient value per (dataset,axis) pair in Table 10.

Dataset	Axis	Alpha	ICC	Rho	Tau
hanna	Coherence	-0.137	0.714	0.460	0.394
hanna	Complexity	0.427	0.652	0.453	0.376
hanna	Empathy	0.415	0.623	0.406	0.338
hanna	Engagement	0.293	0.712	0.484	0.411
hanna	Relevance	0.482	0.778	0.596	0.506
hanna	Surprise	0.148	0.422	0.303	0.245
medval	Risk	0.668	0.818	0.707	0.624
mslr	Fluency	0.545	0.716	0.597	0.585
mslr	Intervention	0.499	0.675	0.510	0.466
mslr	Outcome	-0.050	0.113	0.070	0.064
mslr	Population	0.330	0.506	0.346	0.317
summeval	Coherence	0.573	0.736	0.600	0.493
summeval	Consistency	0.617	0.798	0.639	0.592
summeval	Fluency	-0.425	0.543	0.393	0.352
summeval	Relevance	0.543	0.758	0.579	0.487

Table 10: True population values of disaggregated datasets across metrics, averaged over our model suite.

J.3 ESTIMATION ERROR RESULTS FOR BASELINES BY DATASET.

Here we highlight the estimation error by sampling scheme (Table 11), as well as macro (Table 12) and micro (Table 13) win rate percentages disaggregated by dataset and metric

Table 11: Estimation error by sampling scheme, dataset, axis, and agreement metric. Lower values indicate better estimation. Averaged over budgets and trials.

		Random				Random BC				Stratified				Cluster				Metric Match			
Dataset	Axis	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ
Hanna	Coherence	0.204	0.149	0.149	0.135	0.154	0.122	0.145	0.138	0.193	0.146	0.156	0.140	0.240	0.209	0.170	0.150	0.340	0.087	0.112	0.108
	Complexity	0.161	0.170	0.162	0.145	0.169	0.161	0.169	0.151	0.141	0.145	0.160	0.144	0.237	0.242	0.179	0.155	0.098	0.087	0.120	0.112
	Empathy	0.180	0.203	0.171	0.149	0.169	0.191	0.182	0.155	0.156	0.168	0.167	0.148	0.211	0.255	0.193	0.168	0.114	0.105	0.134	0.129
	Engagement	0.175	0.133	0.140	0.128	0.148	0.117	0.136	0.126	0.189	0.138	0.144	0.129	0.249	0.184	0.157	0.142	0.136	0.071	0.103	0.099
	Relevance	0.184	0.109	0.138	0.127	0.152	0.093	0.130	0.115	0.159	0.106	0.132	0.122	0.185	0.147	0.149	0.134	0.127	0.058	0.100	0.093
	Surprise	0.143	0.168	0.172	0.148	0.163	0.169	0.181	0.152	0.161	0.179	0.166	0.140	0.154	0.199	0.183	0.157	0.153	0.138	0.151	0.129
MedVal	Risk	0.115	0.088	0.109	0.107	0.108	0.083	0.108	0.101	0.105	0.077	0.101	0.098	0.093	0.061	0.090	0.091	0.079	0.048	0.080	0.077
MSLR	Fluency	0.191	0.177	0.168	0.165	0.177	0.148	0.154	0.145	0.176	0.133	0.149	0.145	0.234	0.212	0.197	0.201	0.131	0.102	0.116	0.115
	Intervention	0.144	0.143	0.145	0.137	0.163	0.148	0.158	0.137	0.137	0.131	0.139	0.132	0.163	0.131	0.163	0.159	0.110	0.087	0.109	0.106
	Outcome	0.164	0.300	0.165	0.153	0.192	0.319	0.194	0.175	0.168	0.350	0.184	0.169	0.144	0.345	0.165	0.153	0.209	0.356	0.191	0.174
	Population	0.179	0.241	0.185	0.174	0.207	0.251	0.225	0.190	0.156	0.186	0.164	0.153	0.147	0.183	0.154	0.143	0.134	0.142	0.141	0.133
SummEval	Coherence	0.131	0.118	0.131	0.120	0.137	0.115	0.136	0.119	0.117	0.102	0.128	0.120	0.126	0.097	0.126	0.129	0.085	0.065	0.092	0.090
	Consistency	0.188	0.161	0.123	0.116	0.183	0.154	0.132	0.125	0.155	0.102	0.118	0.110	0.174	0.129	0.121	0.117	0.123	0.070	0.094	0.095
	Fluency	0.143	0.165	0.144	0.132	0.187	0.198	0.176	0.151	0.187	0.171	0.171	0.150	0.185	0.164	0.149	0.137	0.191	0.111	0.116	0.108
	Relevance	0.128	0.108	0.124	0.111	0.158	0.119	0.153	0.127	0.127	0.079	0.137	0.123	0.131	0.095	0.122	0.115	0.087	0.061	0.085	0.081
Mean		**0.162**	**0.162**	**0.148**	**0.137**	**0.165**	**0.159**	**0.159**	**0.140**	**0.155**	**0.148**	**0.148**	**0.135**	**0.178**	**0.177**	**0.155**	**0.143**	**0.141**	**0.106**	**0.116**	**0.110**

J.4 POST-HOC ANALYSIS OF FAILURE CASES

We conduct a post-hoc analysis of the principal failure cases to understand when **Metric Match** underperforms compared to baselines and why. All estimation errors reported below are averaged over repeated runs, five target models, and annotation budgets from 5 to 50. Lower values indicate better estimation.

Model-label collapse on MSLR Outcome. The clearest general failure occurs on the MSLR Outcome axis, where Metric Match is worse than random selection for all four agreement metrics (Table 14). MSLR Outcome and MSLR fluency are 2 of the top 3 dataset axes out of all 15 evaluated with the highest model score collapse, with MSLR outcome having 62.31% of judge labels being label 0 and fluency having 79.92% with label 2. For MSLR Outcome, the regime in which **Metric Match** underperforms compared to baselines, the human labels have mode label 1 and accounts for

Table 12: Macro win rates (%): percentage of comparisons in which Metric Match attains lower estimation error than each baseline, by sampling scheme.

		Random				Random BC				Stratified				Cluster			
Dataset	Axis	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ
Hanna	Coherence	0	76	94	94	0	56	94	90	0	56	98	90	0	80	96	82
	Complexity	100	100	100	94	100	100	98	98	100	96	94	90	100	100	86	80
	Empathy	100	90	72	66	90	92	96	78	90	76	66	62	96	86	86	80
	Engagement	82	92	90	84	70	82	96	86	86	96	98	94	88	94	96	90
	Relevance	70	98	96	94	60	90	94	92	70	96	94	96	60	92	98	92
	Surprise	38	68	76	70	60	70	94	88	62	72	60	54	48	78	84	78
MedVal	Risk	98	100	100	94	96	100	96	88	94	100	94	96	76	82	66	70
MSLR	Fluency	100	100	98	100	90	94	90	90	68	80	80	72	100	100	100	100
	Intervention	90	100	94	94	98	100	98	94	90	96	90	90	88	88	88	84
	Outcome	10	20	18	30	30	26	50	52	14	32	34	40	8	38	34	34
	Population	96	100	96	94	100	100	100	100	84	86	76	76	58	70	54	52
SummEval	Coherence	100	96	94	84	100	100	100	86	96	96	96	96	100	98	100	100
	Consistency	96	100	98	98	100	100	96	96	86	90	78	76	94	100	92	84
	Fluency	8	98	94	94	42	100	100	96	44	74	88	78	32	96	88	88
	Relevance	98	100	98	94	100	100	98	100	90	92	98	98	100	100	96	94
Mean		**72.4**	**89.2**	**87.9**	**85.6**	**75.7**	**87.3**	**93.3**	**88.9**	**71.6**	**82.5**	**82.9**	**80.5**	**69.9**	**86.8**	**84.3**	**80.5**

Table 13: Micro win rates (%): percentage of comparisons in which Metric Match attains lower estimation error than each baseline, by sampling scheme.

		Random				Random BC				Stratified				Cluster			
Dataset	Axis	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ	α	ICC	ρ	τ
Hanna	Coherence	19.30	48.45	58.42	54.57	10.55	45.95	59.14	59.50	18.80	48.40	60.59	59.01	28.35	56.95	62.26	59.05
	Complexity	59.00	60.30	56.95	55.85	64.80	57.15	57.90	58.71	61.10	61.65	60.60	57.80	76.05	70.95	58.58	56.28
	Empathy	60.05	58.20	53.41	49.22	60.40	54.60	56.48	53.64	57.95	53.70	54.73	51.75	64.70	61.70	57.33	54.94
	Engagement	52.30	55.30	56.99	54.76	48.95	54.40	59.47	57.19	55.95	60.65	61.10	57.51	65.25	61.90	61.88	59.28
	Relevance	50.15	56.85	57.41	57.16	48.95	49.70	57.01	55.61	50.15	60.80	58.45	57.05	52.25	62.00	58.39	58.49
	Surprise	42.35	49.00	49.85	50.30	50.95	50.35	54.80	53.85	50.55	53.10	52.30	51.80	49.70	54.80	55.53	56.68
MedVal	Risk	56.45	60.45	55.71	56.18	55.90	56.45	56.21	56.28	58.25	60.70	56.63	55.66	54.90	55.00	52.28	54.13
MSLR	Fluency	58.68	60.89	60.27	59.85	62.23	57.66	59.68	58.66	60.30	54.70	59.01	58.32	71.03	75.67	71.83	72.25
	Intervention	55.10	58.50	56.58	56.33	63.80	59.40	61.04	57.58	56.20	58.70	57.52	56.79	65.35	66.95	66.52	66.06
	Outcome	37.90	29.89	39.87	40.32	44.10	32.90	48.83	48.28	39.40	36.35	45.05	44.49	33.70	35.07	43.75	43.14
	Population	57.55	58.75	55.96	56.25	65.45	62.30	66.31	62.33	54.00	55.85	53.04	53.38	53.20	52.93	50.75	49.67
SummEval	Coherence	60.25	59.75	57.34	54.54	62.35	58.40	59.30	56.79	60.10	59.45	61.05	57.85	64.75	65.20	62.95	63.25
	Consistency	58.13	59.93	56.70	54.67	61.57	64.51	61.53	61.31	53.90	54.05	56.87	54.47	60.10	60.71	59.20	57.25
	Fluency	36.70	54.15	54.26	52.46	48.40	64.60	63.76	61.03	47.50	65.05	63.97	61.16	44.25	58.48	56.82	56.57
	Relevance	58.70	59.15	59.34	55.88	69.70	66.80	69.35	64.20	64.75	58.80	70.40	65.75	62.25	61.50	61.16	59.61
Mean		**50.84**	**55.30**	**55.27**	**53.89**	**54.54**	**55.68**	**59.39**	**57.66**	**52.59**	**56.13**	**58.09**	**56.19**	**56.39**	**59.99**	**58.62**	**57.78**

Table 14: Average estimation error for the two MSLR axes exhibiting strong, shared model-label collapse. Despite stronger model collapse on Fluency, Metric Match succeeds because the model and human label distributions are aligned.

Axis	Method	α	ICC	ρ	τ
Outcome	Random	0.164	0.300	0.165	0.153
	Metric Match	0.209	0.356	0.191	0.174
Fluency	Random	0.191	0.177	0.168	0.165
	Metric Match	0.131	0.102	0.116	0.115

Table 15: Alpha-specific Metric Match failures. Each cell reports Random / Metric Match estimation error. Metric Match is worse for alpha but better for the other agreement metrics.

Dataset–axis	α	ICC	ρ	τ
Hanna Coherence	0.204 / 0.340	0.149 / 0.087	0.149 / 0.112	0.135 / 0.108
Hanna Surprise	0.143 / 0.153	0.168 / 0.138	0.172 / 0.151	0.148 / 0.129
SummEval Fluency	0.143 / 0.191	0.165 / 0.111	0.144 / 0.118	0.132 / 0.110

only 34.89% of examples, indicating that the human distribution is not collapsed, whereas the LLM judges incorrectly concentrate on 0

However, model-label collapse is not necessarily harmful. In MSLR Fluency the human distribution collapses in the same location and to the same degree: the human mode is also label 2, with frequency 81.87%. Metric Match consequently improves all four metrics on Fluency.

Collapse should be interpreted as a *risk indicator*, rather than a deterministic failure condition. Metric Match can perform well when model collapse reflects genuine concentration in the human labels, but can fail across all metrics when the models share the same incorrect collapse. Because alignment with the human distribution is unavailable before annotation, strong shared model collapse represents an intrinsically riskier operating regime.

Alpha-specific failures. Three additional failures are specific to Krippendorff's alpha metric: Hanna Coherence, Hanna Surprise, and SummEval Fluency. On these axes, Metric Match is worse than random selection for alpha but improves ICC, Spearman's rho, and Kendall's tau. Table 15 summarizes the estimation errors.

These failures can be understood through the structure of Krippendorff's interval alpha. Alpha is defined as

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{3}$$

where D_o is observed disagreement and D_e is disagreement expected under the pooled marginal label distribution. For interval alpha, disagreement is measured using squared label distances. Consequently, alpha depends not only on whether two raters order examples similarly, but also on whether they use similar locations and spreads on the label scale.

For two raters X and Y, observed interval disagreement can be expressed approximately as

$$D_o \approx \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{Cov}(X, Y) + (\mu_X - \mu_Y)^2. \tag{4}$$

The final term makes alpha directly sensitive to systematic differences in rater means. Its denominator D_e is additionally sensitive to marginal label concentration and dispersion. By contrast, Spearman's rho and Kendall's tau are invariant to monotonic changes in the label scale, while the $\mathrm{ICC}(3, k)$ consistency coefficient used in our experiments largely removes systematic rater-level mean differences.

On Hanna Coherence, SummEval Fluency, and Hanna Surprise, the model and human label distributions differ in location or spread: pooled model means sit 1.63 and 1.84 scale points from the human means on Hanna Coherence and SummEval Fluency, and model scores on Hanna Surprise have 1.82 times the human standard deviation. Spearman's rho and Kendall's tau depend only on

Dataset	Axis	Comparison	$(\mu_X - \mu_Y)^2$	σ_X^2, σ_Y^2
Hanna	Coherence	$X, Y =$ model	0.06	1.03
		$X, Y =$ human, model	2.64	0.60
	Complexity	$X, Y =$ model	0.28	1.35
		$X, Y =$ human, model	0.19	0.65
	Empathy	$X, Y =$ model	0.20	1.12
		$X, Y =$ human, model	0.12	0.54
	Engagement	$X, Y =$ model	0.19	0.97
		$X, Y =$ human, model	0.70	0.66
	Relevance	$X, Y =$ model	0.04	1.31
		$X, Y =$ human, model	0.49	1.00
	Surprise	$X, Y =$ model	0.12	1.83
		$X, Y =$ human, model	0.44	0.55
MedVal	Risk	$X, Y =$ model	0.01	1.28
		$X, Y =$ human, model	0.08	1.55
MSLR	Fluency	$X, Y =$ model	0.02	0.24
		$X, Y =$ human, model	0.01	0.23
	Intervention	$X, Y =$ model	0.04	0.62
		$X, Y =$ human, model	0.03	0.58
	Outcome	$X, Y =$ model	0.02	0.45
		$X, Y =$ human, model	0.27	0.65
	Population	$X, Y =$ model	0.07	0.63
		$X, Y =$ human, model	0.03	0.51
SummEval	Coherence	$X, Y =$ model	0.06	1.06
		$X, Y =$ human, model	0.03	1.15
	Consistency	$X, Y =$ model	0.05	1.49
		$X, Y =$ human, model	0.16	1.06
	Fluency	$X, Y =$ model	0.03	0.69
		$X, Y =$ human, model	3.38	0.55
	Relevance	$X, Y =$ model	0.18	0.85
		$X, Y =$ human, model	0.16	0.79

Table 16: Comparison of the mean difference, $(\mu_X - \mu_Y)^2$, between average evaluation scores for each axis and dataset, when X, Y are both models (inter-model regime); and when X, Y are target model and human (human-model regime). The column σ_X^2, σ_Y^2 denotes the average variance in model evaluation score in rows where X, Y are both models; and the average variance in human evaluation score in rows where X, Y are target model and human.

ranks, and ICC$(3, k)$ removes rater-level mean differences, so these metrics are largely insensitive to the shift, and Metric Match still improves them. Observe in Table 16, indeed the dataset/axis pairs with the greatest difference in $(\mu_X - \mu_Y)^2$ (Hanna Coherence and SummEval Fluency) *or* σ_X^2 and σ_Y^2 (Hanna Surprise) between the inter-model and human-model regimes are exactly the settings we observe poor performance on Krippendorff's alpha.

K Cost Analysis

The cost of Metric Match is driven by three factors: the number of models in the ensemble, the per-token price of those models, and the number of tokens required for annotation, which itself depends on the number of items to annotate and the length of each item. Human annotation cost, in contrast, is determined by the hourly rate of labor and the estimated time required for the task. This framing places the break-even regime for our method in settings where human annotation is fast, readily accessible, and applied to long-context inputs.

Formally, let

1. H = human hourly cost (USD/hour)
2. t = estimated hours for the task

3. M_i = API cost of model i (USD per 10^6 tokens)
4. n = number of models in the ensemble
5. m = number of prompts (the same set is used for every model)
6. p_j = number of tokens in prompt j

Metric Match is more cost-effective than human annotation whenever

$$\frac{P}{10^6}\sum_{i=1}^{n} M_i \;\le\; Ht, \qquad P \equiv \sum_{j=1}^{m} p_j.$$

Table 17: Input token pricing for selected large language models (USD per 1M tokens).

Model	Price (USD / 1M tokens)
Claude 3.5 Sonnet	3.00 Anthropic (2024)
GPT-4.1	2.00 OpenAI (2026)
GPT-5	1.25 OpenAI (2026)
DeepSeek-R1	1.485 Microsoft Azure (2026)
Gemini 2.5 Pro	1.25 Google (2026)

Table 18: Datasets used in evaluation, with 300 prompts sampled per dataset.

Dataset	n dimensions	n prompts	Mean chars
HANNA	6	300	2241.2
MedVAL	1	300	2586.8
MSLR	4	300	1303.4
SummEval	4	300	3591.8

Using these values, we approximate the relative cost of human and model annotation. Taking an average prompt length of 2374 characters, a ratio of 4 characters per token, and the average API cost across our five-model ensemble ($\bar{M} = 1.797$ USD/10^6 tokens), the per-dataset cost of running Metric Match is:

$$\frac{2374}{4} \times \frac{1.797}{10^6} \times 5 \times 300 \approx \$1.60 \text{ per dataset.}$$

Assuming 60 seconds per item for a human annotator, the corresponding human-annotation cost for 300 items is $300H \cdot \frac{1}{60}$. Setting this equal to the Metric Match cost gives a break-even hourly wage of:

$$300H \cdot \frac{1}{60} \ge 1.60 \quad \Longrightarrow \quad H \ge \$0.32/\text{hr.}$$

In other words, Metric Match is more cost-efficient than human annotation as long as the human annotator's effective wage exceeds \$0.32/hour — a threshold far below any realistic labor cost, underscoring the practical economic advantage of our approach even under generous assumptions about annotation speed.

L Mean Squared Error

We introduce results with non-correlation-based metrics, such as Mean Squared Error (MSE). We show that variance matching as opposed to **Metric Match** on the MSE metric provides larger gains, although **Metric Match** still outperforms random for estimation error and threshold classification at the macro level for small budgets.

Table 19: MSE Win Rates by Method: Macro and Micro Win Rates

	Estimation Error				*Threshold Classification*			
	Metric Matching		Variance Matching		Metric Matching		Variance Matching	
Budget	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
5	0.827	0.475	0.733	0.481	0.568	0.481	0.512	0.511
10	0.720	0.492	0.813	0.517	0.524	0.469	0.550	0.531
15	0.693	0.486	0.800	0.517	0.565	0.448	0.676	0.520
20	0.587	0.469	0.733	0.503	0.462	0.445	0.567	0.508
25	0.613	0.459	0.800	0.512	0.415	0.439	0.550	0.505
30	0.587	0.450	0.760	0.500	0.398	0.434	0.547	0.495
35	0.613	0.447	0.827	0.522	0.387	0.442	0.498	0.508
40	0.573	0.449	0.640	0.502	0.392	0.435	0.512	0.481
45	0.520	0.442	0.680	0.499	0.378	0.504	0.454	0.497
50	0.587	0.446	0.627	0.492	0.419	0.450	0.467	0.485
Avg	0.632	0.462	0.741	0.504	0.451	0.455	0.533	0.505

M VARIANCE MATCHING

We present an alternative algorithm as an extension to metric matching called *Variance Matching*. We find that Variance Matching performs worse than metric matching on average, but in cases where metric matching performs very poorly, variance matching is more robust to these outlying cases.

Algorithm 2: Weighted Variance Matching

Data: Metric function T; MSB function $\texttt{MSB}$, MSE function $\texttt{MSE}$; weights α, β; budget b; generated text data $X = \{x_i\}_{i=1}^n \in \mathcal{X}^n$; scores $\{y_i^{(M)}\}_{i=1}^n$, $\{y_i^{(M')}\}_{i=1}^n$ from LLM judges M and M', respectively; C, number of candidate subsets over which to search

Result: Subset $S^* \subseteq X$, $|S^*| = b$

Initialize $\delta_{\min} \leftarrow \infty$

Initialize $S^* \leftarrow \emptyset$

$\rho^{\text{IM}} \leftarrow \alpha \cdot \texttt{MSB}\big(\{y_i^{(M)}\}_{i=1}^n, \{y_i^{(M')}\}_{i=1}^n\big) + \beta \cdot \texttt{MSE}\big(\{y_i^{(M)}\}_{i=1}^n, \{y_i^{(M')}\}_{i=1}^n\big)$

for $j = 1, \ldots, C$ **do**

  Sample $S_j = \{x_{i_1}, \ldots, x_{i_b}\}$ uniformly from X without replacement

  $\widehat{\rho}_{S_j}^{\text{IM}} \leftarrow \alpha \cdot \texttt{MSB}\big(\{y_{i_k}^{(M)}\}_{k=1}^b, \{y_{i_k}^{(M')}\}_{k=1}^b\big) + \beta \cdot \texttt{MSE}\big(\{y_{i_k}^{(M)}\}_{k=1}^b, \{y_{i_k}^{(M')}\}_{k=1}^b\big)$

  $\delta_j = |\widehat{\rho}_{S_j}^{\text{IM}} - \rho^{\text{IM}}|$

  if $\delta_j < \delta_{\min}$ **then**

    $S^* \leftarrow S_j$

    $\delta_{\min} \leftarrow \delta_j$

return S^*

We find that the optimal hyperparameters for this mixture are $\alpha = 0.9$ and $\beta = 0.1$. We highlight results of this algorithm in Table 20.

Table 20: Variance Matching Performance: Estimation Error and Threshold Classification

	Estimation Error (Win Rate)				*Threshold Classification (Win Rate)*			
Budget	α	ICC	τ	ρ	α	ICC	τ	ρ
5	0.867	0.840	0.760	0.800	0.655	0.505	0.585	0.572
10	0.920	0.947	0.880	0.880	0.728	0.539	0.710	0.653
15	0.973	0.947	0.773	0.827	0.667	0.548	0.587	0.627
20	0.933	0.907	0.827	0.840	0.506	0.496	0.333	0.511
25	0.867	0.867	0.773	0.787	0.574	0.609	0.500	0.640
30	0.880	0.947	0.800	0.840	0.620	0.583	0.720	0.640
35	0.933	0.907	0.787	0.800	0.563	0.642	0.464	0.488
40	0.760	0.827	0.707	0.733	0.466	0.539	0.452	0.449
45	0.907	0.893	0.773	0.800	0.433	0.586	0.461	0.504
50	0.853	0.867	0.733	0.760	0.679	0.522	0.558	0.442
Avg	0.889	0.895	0.781	0.807	0.589	0.557	0.537	0.552

N CLASSIFICATION PERFORMANCE DIS-AGGREGATED

We report the dis-aggregated accuracy scores for model reliability classification in Table 21. We show results for each threshold $t \in (.1, .9)$ across evaluated reliability metrics.

Table 21: Performance by sampling method and threshold.

Method	Threshold	Alpha	ICC	Rho	Tau
Metric Match	0.1	0.875	0.968	0.946	0.944
	0.2	0.859	0.927	0.917	0.910
	0.3	0.881	0.916	0.892	0.860
	0.4	0.873	0.907	0.830	0.764
	0.5	0.794	0.900	0.756	0.745
	0.6	0.796	0.870	0.755	0.837
	0.7	0.890	0.779	0.853	0.923
	0.8	0.968	0.755	0.947	0.977
	0.9	0.992	0.957	0.987	0.994
	Average	**0.881**	**0.887**	**0.876**	**0.884**
Random	0.1	0.875	0.920	0.920	0.915
	0.2	0.859	0.904	0.887	0.876
	0.3	0.842	0.886	0.846	0.810
	0.4	0.807	0.860	0.779	0.751
	0.5	0.790	0.823	0.734	0.741
	0.6	0.811	0.779	0.747	0.821
	0.7	0.890	0.739	0.833	0.905
	0.8	0.961	0.792	0.929	0.959
	0.9	0.989	0.953	0.975	0.987
	Average	**0.869**	**0.851**	**0.850**	**0.863**
Random BC	0.1	0.864	0.926	0.889	0.889
	0.2	0.850	0.909	0.856	0.847
	0.3	0.807	0.885	0.789	0.774
	0.4	0.747	0.853	0.712	0.730
	0.5	0.762	0.801	0.725	0.766
	0.6	0.859	0.756	0.805	0.859
	0.7	0.936	0.723	0.903	0.933
	0.8	0.982	0.850	0.964	0.972
	0.9	0.992	0.974	0.985	0.987
	Average	**0.867**	**0.853**	**0.847**	**0.862**
Stratified	0.1	0.898	0.937	0.928	0.924
	0.2	0.866	0.923	0.904	0.893
	0.3	0.855	0.911	0.876	0.847
	0.4	0.823	0.894	0.821	0.748
	0.5	0.810	0.871	0.723	0.718
	0.6	0.781	0.801	0.706	0.781
	0.7	0.869	0.735	0.793	0.895
	0.8	0.960	0.757	0.919	0.960
	0.9	0.991	0.951	0.977	0.990
	Average	**0.873**	**0.864**	**0.850**	**0.862**
Cluster Text	0.1	0.882	0.915	0.918	0.913
	0.2	0.862	0.913	0.887	0.873
	0.3	0.855	0.891	0.835	0.810
	0.4	0.813	0.856	0.779	0.785
	0.5	0.817	0.823	0.773	0.769
	0.6	0.795	0.766	0.738	0.800
	0.7	0.845	0.768	0.797	0.873
	0.8	0.933	0.753	0.905	0.933
	0.9	0.976	0.922	0.958	0.969
	Average	**0.864**	**0.845**	**0.843**	**0.858**