# ONEBENCH to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities

Adhiraj Ghosh[1,2*] Sebastian Dziadzio[1,2*] Ameya Prabhu[1,2°] Vishaal Udandarao[1,2,3°]
Samuel Albanie Matthias Bethge[1,2]

[1]Tübingen AI Centre, University of Tübingen [2]Open-Ψ Collective [3]University of Cambridge

adhiraj.ghosh@bethgelab.org

Project Page Code

## Abstract

Traditional fixed test datasets fall short in evaluating the open-ended capabilities of foundation models. To address this, we propose ONEBench (**O**pe**N**-**E**nded **Bench**marking), a new paradigm that consolidates individual evaluation datasets into a unified, ever-expanding sample pool. ONEBench enables custom benchmarks for specific capabilities while reusing and aggregating samples, mitigating overfitting and dataset bias for broader capability assessment. It reframes model evaluation as selecting and aggregating sample-level tests. Transitioning from task-specific benchmarks to ONEBench introduces two challenges: *heterogeneity* (aggregating diverse metrics) and *incompleteness* (comparing models tested on different data subsets). To address these, we propose an aggregation algorithm that ensures identifiability—asymptotically recovering ground-truth scores—and rapid convergence, enabling accurate model comparisons with relatively little data. On homogenous datasets, our algorithm produces rankings that highly correlate with average scores. Moreover, it remains robust to over 95% missing measurements, reducing evaluation costs by up to 20 times. We introduce ONEBench-LLM for language models and ONEBench-LMM for vision-language models, enabling targeted model testing across diverse capabilities.

## 1 Introduction

Deep learning has arrived in the post-dataset era[1]. As foundation models rapidly expand their zero-shot capabilities, the focus of evaluation has moved beyond singular, dataset-specific performance measurements that rely on dividing fixed collections of data into train and test sets. Instead, foundation models are employed as general knowledge and reasoning engines across a wide range of domains. This creates a pressing need to characterize their open-ended capabilities using diverse metrics in zero-shot settings (Ge et al., 2024). However, static benchmarks, which test generalization on fixed test splits, cannot probe the ever-evolving capabilities of foundation models effectively. This raises an important question: *How can benchmarking adapt to measure an open-ended set of capabilities?*

We propose a solution based on dynamic, sample-level evaluation, which we call ONEBench (**O**pe**N**-**E**nded **Bench**marking). In this approach, test sets for particular capabilities are generated ad-hoc from a large pool of individual annotated data samples. These sample-level evaluations act as atomic units of measurement that can be flexibly aggregated into an exponential number of configurations. Thanks to this flexibility, the sample pool and corresponding annotation metrics can be continuously updated to incorporate new evaluations. Additionally, this approach can reduce *dataset bias*—systematic quirks in the data arising from its collection process (Liu and He, 2024). Finally, by combining samples across test sets, ONEBench captures real-world diversity (Ni et al., 2024).

The most important feature of ONEBench is its potential to democratize evaluation. Unlike traditional benchmarks, typically created by individual groups based on their own criteria for data collection and evaluation procedures (Bansal and Maini, 2024), ONEBench integrates test sets from multiple sources reflecting a wide range of perspectives, use cases, and objectives. This flexibility allows different interest groups to collaboratively define their own evaluations by selecting the most appropriate combination of tests that best suit their specific requirements. Moreover, the design of ONEBench challenges the dominant approach of chasing single benchmark scores, which fail to account for the difficulty of individual data instances (Ethayarajh et al., 2022), in favor of a plurality of rankings and a dynamic, granular, multi-faceted evaluation.

[*]Equal contribution, random order, [°] core contributors
[1]From a talk by Alexei Efros at ICML 2020



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32445–32481


Figure 1: **The ONEBench Framework.** *Left*: ONEBench comprises a set of models, a pool of data samples spanning multiple test sets, metadata describing models and data samples, and a collection of sample-level measurements. *Right*: the user formulates a query to capture the desired model capability, using a mix of structured metadata filters and semantic search. Selected models are then ranked on a subset of data samples that meet the specified criteria.

**Challenges.** Building ONEBench requires addressing two key challenges: *heterogeneity* and *incompleteness*. *Heterogeneity* arises because model evaluations span diverse metric types, such as binary (correct/incorrect), numeric (BLEU scores), and ordinal (preference rankings), making aggregation difficult. *Incompleteness* occurs when models are tested on non-overlapping subsets of data, preventing fair and direct comparisons. Traditional benchmarks sidestep these issues by using a multi-task setup, where all models are evaluated on the same samples using a single metric.

**Solution and Theoretical Guarantees.** We address these challenges using social choice theory, treating data samples as voters expressing preferences over models. By converting all measurements into ordinal rankings, we leverage established principles to robustly aggregate heterogeneous and incomplete data. Our approach assumes a random utility model based on the Plackett-Luce framework (Plackett, 1975; Luce, 1959), which provides guarantees for accurately recovering ground-truth utility scores. This approach ensures that our model rankings are both theoretically sound and practical, with rapid convergence guarantees enabling accurate rankings from limited data.

**Why rankings?** While converting cardinal measurements to ordinal ones results in information loss, two key considerations justify this choice in benchmarking: *external validity* and *benchmark longevity*. *External validity* (Liao et al., 2021) refers to how well evaluation results generalize across different settings. Recht et al. (2019) found that although accuracy shifts across test sets, model rankings remain stable. Salaudeen and Hardt (2024) extended this by comparing ImageNet models to those trained and tested on a new dataset, arguing that ordinal comparisons are more robust, e.g., injecting random labels into two otherwise identical benchmarks causes accuracy to vary while preserving relative order. Given the noise in LLM and LMM evaluation (from prompt design, label choices, in-context examples, etc.), rankings offer a more reliable signal than absolute values. *Benchmark longevity* is another growing challenge, as internet-scale datasets lead to faster saturation. Contamination and adaptive overfitting threaten evaluation integrity. While ONEBench mitigates contamination by filtering compromised samples, its use of rankings helps guard against overfitting—an idea inspired by the Ladder algorithm (Hardt and Recht, 2022).

**Empirical Validation.** We create ONEBench for two domains: ONEBench-LLM for language models and ONEBench-LMM for vision-language models. These benchmarks unify evaluations by aggregating data from diverse sources, including preference data (arenas) and heterogeneous multi-task leaderboards. Our empirical results demonstrate that the Plackett-Luce model effectively aggregates real-world benchmarks, showing a high correlation with ground-truth score-based rankings over homogeneous datasets. Notably, this strong correlation persists even when up to 95% of the data is missing, enabling a 20 times reduction in evaluation costs with minimal impact on performance. Finally, we compare Plackett-Luce rankings to widely adopted methods such as ELO (Elo, 1967) and Bradley-Terry (Bradley and Terry, 1952), demonstrating superior accuracy and robustness to missing data. We additionally compare our Plackett-Luce rank aggregation approach to popular computational social choice theory methods like Borda Count and the Dowdall System (section 3.2.4).

**Personalized Aggregation.** Imagine you are a biochemist seeking an LLM to assist with designing experiments related to antibodies. With ONEBench, you can input a query, such as "`immunology`" or "`antibodies`" to generate a dynamically constructed benchmark that ranks models based on their performance in *this specific domain*. While the optimal selection of personalized capability sets remains an open research challenge, we present a proof of concept by distinguishing between *tasks* (e.g., reading comprehension) and *concepts* (e.g., Clostridium bacteria). By combining structured filters and flexible semantic search, users can define their capability of interest along these dimensions and conduct targeted evaluations, resulting in personalized rankings.

To summarize, ONEBench is a democratized, open-source collection of diverse evaluation samples enriched with detailed metadata. Its robust aggregation method ranks models across heterogeneous metrics and incomplete evaluation data. Users can perform semantic searches and apply structured query filters to dynamically generate benchmarks tailored to their needs. They can also contribute new evaluation samples and model measurements, which are instantly aggregated to refine rankings. This framework enables lifelong aggregation of arbitrary test sets with unprecedented flexibility and precision.

## 2 ONEBench: Formulation

### 2.1 Components

The goal of ONEBench is to evaluate a set of models $\{m_k\}_{k=1}^M$ using a continuously expanding pool of test data samples $\mathcal{D}$ drawn from multiple benchmarks $\{\mathcal{B}_k\}_{k=1}^B$. Each data sample may include metadata specifying the capabilities it is testing. To handle the diversity of data from different benchmarks, we generate sample-level rankings ($\mathcal{S}$) for all samples in the test pool. Figure 1 provides a schematic overview of ONEBench, with each component described below.

**(i) Data Pool.** The data pool $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^D$ consists of data samples $x_k$ with reference answers $y_k$. An example of a data sample is the question *"What was the dominant strain of flu in 2010? Select among four choices."* with a reference answer *"H1N1/09"*. Each instance can also include metadata specifying tested capabilities, for example as a list of keywords like *temporal QA, pandemics, history, biology, virology, multiple-choice QA*.

**(ii) Models.** The set of models is defined as $\mathcal{M} = \{m_{base}\} \cap \{m_k\}_{k=1}^M$, where $m_{base}$ serves as a baseline for evaluating the capabilities of the other models. A common choice for $m_{base}$ is a random model. Since the original benchmarks evaluate different sets of models, each benchmark $\mathcal{B}_k$ considers a subset of models $\mathcal{M}_{\mathcal{B}_k} \subseteq \mathcal{M}$.

**(iii) Sample-level Rankings.** For each data sample $(x_j, y_j) \in \mathcal{D}$, we construct a sample-level ranking $s_j \in \mathcal{S}$ over model subset $\mathcal{M}_j \subseteq \mathcal{M}_{\mathcal{B}_k}$, where $k$ denotes the index of the benchmark from which the sample $(x_j, y_j)$ was collected (more in appendix C). Crucially, these rankings depend only on the evaluation metrics used by each benchmark, abstracting away the specifics of those metrics. This abstraction is central to our approach, as it enables aggregation across heterogeneous evaluation metrics.

**(iv) Capabilities.** To enable selective retrieval of relevant sample-level rankings in $\mathcal{B}$ based on user queries, each ranking can be associated with a *capability*. We distinguish between *tasks* (e.g., question answering, captioning) and *concepts* (e.g., makeup, geometry). Since capabilities are inherently open-ended, we only tag data samples with task information, while concept-based retrieval is performed dynamically at test time using semantic search.

**Lifelong Expansion.** Data pool $\mathcal{D}$ and model set $\mathcal{M}$ are stored as tables, while sample-level model evaluations are represented as a relational database linking these tables. Expanding ONEBench over time requires augmenting $\mathcal{D}$, $\mathcal{M}$, and $\mathcal{S}$ through the following operations: $\texttt{insert}_\mathcal{D}$, $\texttt{insert}_\mathcal{M}$, $\texttt{insert}_\mathcal{S}$. The first two operations simply add new samples and models, while $\texttt{insert}_\mathcal{S}$ registers a new sample-level ranking.

### 2.2 Capability Querying

To evaluate a given capability, ONEBench takes a dynamic approach. First, we retrieve samples that match the query. Then, we aggregate the sample-level rankings to produce the overall ranking.

**Retrieve ($\texttt{retrieve}_\mathcal{D}$).** Here, the system selects relevant data instances based on a user's query. The query language is flexible and allows retrieving data instances that semantically relate to a specific topic or match certain criteria. Retrieval combines k-nearest neighbors (kNN) search over dense embeddings, with the query as input, and structured queries that leverage the unified data schema.

**Aggregate ($\texttt{Aggregate}_{\mathcal{S},\mathcal{D}}$).** Measurements over the retrieved subset are combined using the random utility modelling approach (Xia, 2019), defining a joint probability distribution over all measurements (sample rankings $s_j$ and model scores $\gamma_j$), given model permutations $\sigma_j$ and binary sequence of pairwise performance relations $\pi_j$ assuming statistical independence:

$$p(s_1, \ldots, s_{n_\infty} | \gamma_1, \ldots, \gamma_M) = \prod_{j=1}^{n_\infty} p(s_j = [.]_{(\sigma_j, \pi_j)} | \gamma_1, \ldots, \gamma_M).$$

The Placket-Luce framework assumes the following probability model:

$$p\left(s_j = [\cdot]_{(\sigma_j,\pi_j)}\right) = \underbrace{\frac{\gamma_{\sigma_j(1)}}{\sum_{k=1}^{m_j} \gamma_{\sigma_j(k)}}}_{f_{\sigma_j(1)}} \times \cdots \times \underbrace{\frac{\gamma_{\sigma_j(m_j-1)}}{\gamma_{\sigma_j(m_j-1)} + \gamma_{\sigma_j(m_j)}}}_{f_{\sigma_j(m_j)}},$$

defining one parameter $\gamma_k$ for each model $m_k$ that determines its performance relative to all other models. To aggregate model performances over sample rankings, we estimate parameters

$$\hat{\gamma} = \operatorname*{argmax}_{\gamma \in \mathbb{R}^m} \log p(\mathbf{s} \mid \gamma)$$

with maximum likelihood estimation (MLE). The global ranking follows the permutation $\sigma_\infty$ where $\hat{\gamma}_{\sigma_\infty(1)} > \cdots > \hat{\gamma}_{\sigma_\infty(m)}$. The ML condition uniquely determines all performance parameters $\{\hat{\gamma}_k\}_{k=1}^M$, as the likelihood function is strictly concave. The parameters of the Plackett-Luce model are identifiable up to an arbitrary additive constant. Consistency and asymptotic normality can also be shown under certain assumptions about the comparison graph (Han and Xu, 2023). We refer to the estimated latent variables $\{\hat{\gamma}_k\}_{k=1}^M$ as *model scores*. A model with a higher score likely performs better on a randomly picked sample-level task than one with a lower score. To fix the additive constant, we set the baseline model score $\hat{\gamma}_{baseline}$ to zero.

## 3 ONEBench: Aggregation

We view aggregating sparse ordinal preferences over models through a computational social choice lens, where samples are voters, models are candidates, and the aggregation algorithm is the voting mechanism (Brandt et al., 2016). We aggregate ordinal comparisons with partial data to produce a global ranking and analyze its properties.

### 3.1 Theoretical Foundations

We begin by postulating a ground-truth statistical model generating the data, which is converted into ordinal comparisons ($\mathcal{S}$). This is in contrast to Zhang and Hardt (2024), who view aggregation as classical voting and analyse tradeoffs in aggregating voter preferences rather than uncover an underlying ranking. Specifically, we use a random-utility model (Thurstone, 1927), where model $m_i$ is associated with utility distribution $\mathcal{U}_{m_i}$. Preferences between models $m_i$ and $m_j$ are based on comparing sampled utilities, i.e., $m_i \prec m_j := u(m_i) < u(m_j)$, where $u_m \sim \mathcal{U}_m$. Since computing maximum likelihood estimates over general random-utility models is computationally hard (Xia, 2019), we focus on the Plackett–Luce model (Plackett, 1975; Luce, 1977), the only known exception that allows for tractable MLE.

**Property 1: Identifiability.** We first ask: *Are the utility distributions for all models recoverable?* The Plackett-Luce model allows identifying the utility distribution (up to an arbitrary additive constant) if all models are compared via a directed path (Xia, 2019). Using reference model $m_{\text{base}}$ removes additive ambiguity. Consistency and asymptotic normality hold under specific assumptions about the comparison graph (Han and Xu, 2023).

**Property 2: Sample-Efficient Convergence from Sparse Data.** Given that identifiability is asymptotic, we ask: *How sample-efficient is the algorithm for recovering the utility distribution?*

With partial rankings of size $k$, the MLE is surprisingly sample efficient while being minmax-optimal (Maystre and Grossglauser, 2015). Sampling $k$ model comparisons from the model set $|\mathcal{M}|$ uniformly at random induces an expander graph with high probability, giving guarantees for sample-efficient recovery, with $\Omega(|\mathcal{M}|)/k$ samples being necessary, and $\Omega(|\mathcal{M}|\log|\mathcal{M}|)/k$ samples being sufficient. Efficient algorithms like Maystre and Grossglauser (2015) achieve these bounds. Rank-breaking techniques, used in our evaluation, offer near-optimal solutions (Soufiani et al., 2014).

**Property 3: Social Properties.** The Plackett-Luce model offers computational efficiency and recoverability of the underlying ranking. However, designing democratic decision-making systems also requires fair aggregation. Ensuring fairness involves trade-offs (Zhang and Hardt, 2024), as different notions of fairness often conflict. Moreover, depending on the intended application areas, differing or even opposing preferences may be valid (Arrow, 1950). Plackett-Luce offers "procedural fairness" (List, 2022), satisfying:

**(i) Anonymity.** All voters (samples) are treated equally, ensuring the system does not over-rely on any single vote. Rankings remain unchanged if the input sample set is permuted.

**(ii) Neutrality.** The ranking is invariant to model identities, ensuring fairness among alternatives. This means permuting the models similarly permutes the resulting ranking.

**(iii) Independence from Irrelevant Alternatives.** The relative ranking of two models is unaffected by other alternatives in a given sample, as guaranteed by Luce (1959). This provides grounding for incomplete model evaluations.

In Section 7, we discuss separability and pairwise majority consistency—two properties violated by the Plackett-Luce model.

### 3.2 Translating Theory to Practice

Here, we show that: (i) the Plackett-Luce model works well on real-world data, (ii) our aggregation method is sample-efficient, and (iii) it handles high levels of incompleteness. Below, we describe our setup and address these points.

#### 3.2.1 Setup

**Benchmarks.** We conduct experiments using four popular benchmarks with established model rankings based on benchmark-specific average scores: HELM (Liang et al., 2023) and Open LLM Leaderboard (Beeching et al., 2023) for LLMs, and VHELM (CRFM, 2024) and LMMs-Eval (Zhang et al., 2024c) for LMMs. We define our data pool as the sum of all samples in the constituent datasets. To test the faithfulness of our aggregation strategy we compare the resulting rankings to the original leaderboards. These leaderboards evaluate models across varied tasks with different metrics, serving as good indicators of real-world performance.

**Ground Truth.** The current system of benchmarking involves evaluating models on individual test sets and measuring the mean score per model. This holds even for benchmarks that combine test sets. We consider these scores as the ground truth measurement and generate a ground truth model ranking from these scores. Since we aggregate multiple measurement metrics, we implement a min-max normalization of numeric measurements to bring all benchmark samples to the same 0-1 score range. Our final ground truth refers to the model rankings derived from the mean score across all benchmarks. While this procedure is suboptimal, and indeed we propose to replace it with our method, there is no other ground truth available for real data. We show results on synthetic data in appendix D.

**Methods.** We evaluate three ranking methods:

**(i) Elo Score** (Elo, 1967): A competitive game rating system adapted to rank models through pairwise comparisons, adjusting scores based on wins or losses to reflect win-rate reliability.

**(ii) LMArena Ranking**: A ranking method based on the Bradley-Terry model (Bradley and Terry, 1952), using a Maximum Likelihood Estimation (MLE) based on pairwise comparisons with an underlying ELO model for rank aggregation.

**(iii) Ours**: We leverage the Plackett-Luce model (Maystre and Grossglauser, 2015) to aggregate pairwise comparisons using partial rank breaking, speeding up rank estimation.

**Metrics.** We compare the rankings generated by each method to the ground-truth from the leaderboards using Kendall's $\tau$, a standard correlation metric for rankings. Each method is tested three times and we report the mean and variance. We also check that the top-$k$ models are reliably recovered.

#### 3.2.2 Plackett-Luce on Real-World Data

**Q1. Is it suitable?** We evaluate the Plackett-Luce model on large-scale benchmarks by comparing the rankings produced by our aggregation algorithm to the leaderboard rankings. We achieve strong alignment with the ground truth rankings(Table 1).

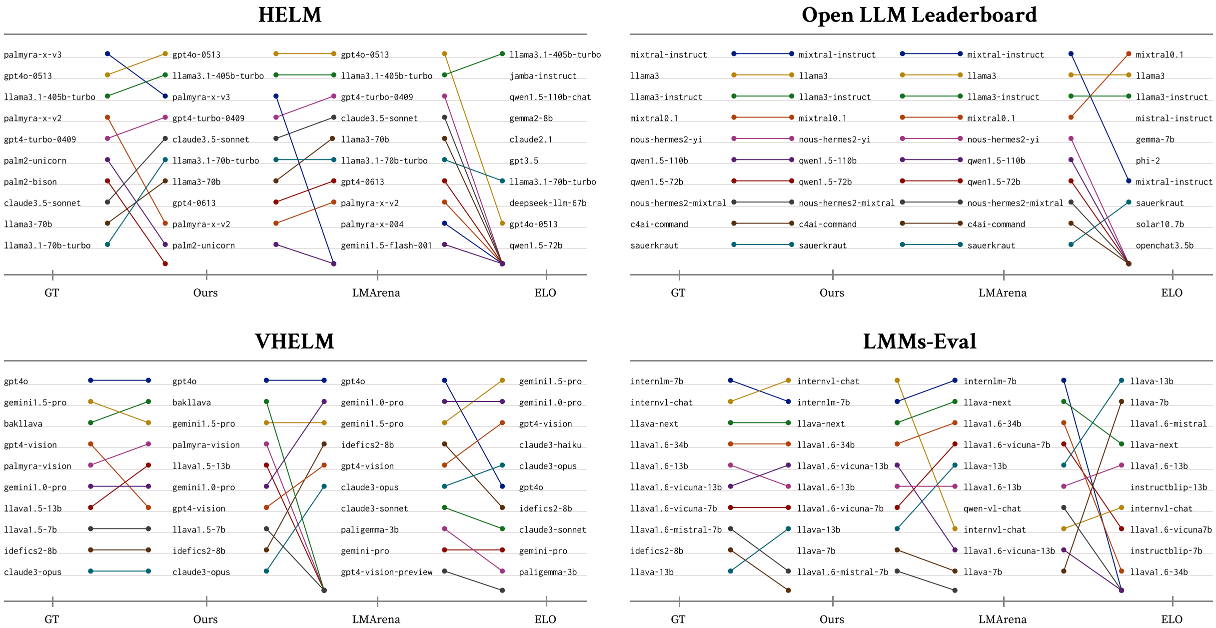


Figure 2: **Top-10 model ranking changes across different aggregation methods.** Plackett-Luce (`Ours`) shows the most similarity to the Ground Truth model rankings (`GT`). However, there is a progressive degradation in ranking accuracy for LMArena (`LMArena`) and Elo (`ELO`).

| Dataset | Elo | LMArena | Ours |
|---|---|---|---|
| HELM | 0.35 ± 0.13 | 0.85 ± 0.00 | **0.88 ± 0.00** |
| Leaderboard | 0.21 ± 0.07 | 0.97 ± 0.00 | **0.99 ± 0.00** |
| VHELM | 0.63 ± 0.02 | 0.69 ± 0.00 | **0.80 ± 0.00** |
| LMMs-Eval | 0.33 ± 0.11 | 0.42 ± 0.00 | **0.64 ± 0.00** |

Table 1: **Comparison to other aggregation algorithms.** Kendall's $\tau$ correlations to ground-truth ranking across individual benchmarks.

**Q2. Is it better than current metrics?** In addition to evaluating fit, we also compare our method to popular algorithms like Elo and LMArena. Table 1 shows that our algorithm consistently outperforms these methods, demonstrating its superior performance for large real-world datasets.

**Q3. Are the top-k models preserved?** A key concern for practitioners is whether the top models are ranked correctly. Figure 2 shows that our algorithm preserves the ground truth top-10 model rankings, with other rank aggregation methods currently used showing poor and unreliable correlation to the ground truth.

**Conclusion.** The Plackett-Luce model fits real-world data well, outperforming other methods in both overall Kendall's $\tau$ and top-10 rankings, proving its effectiveness for large-scale benchmarks. The underlying reason is that we avoid using Elo distributions, which rely on assumptions that do not apply to foundation models (Boubdir et al., 2023).

#### 3.2.3 Sample Efficiency and Robustness

**Q1. Is Our Algorithm Sample-Efficient?** We systematically reduce the number of samples and re-rank the models using various methods, calculating Kendall's $\tau$ for each. Missing data is simulated from 0% to 99%, with 10% intervals until 90%, followed by 1% increments. As shown in fig. 3, our method maintains stable performance even with up to 95% samples missing, demonstrating that it can achieve accurate rankings with up to 20x less data points than current benchmarks.

**Q2. Can our Algorithm Aggregate Highly Sparse Rankings?** We assess our method's ability to handle incomplete data by randomly removing a fraction of model measurements from each sample and re-ranking using the three aggregation methods. We simulate data removal from 0% to 99%, with increments as before. As shown in fig. 3, our method remains effective even when 95% of model comparisons are missing, proving it can recover accurate rankings with highly sparse data. This is crucial for ONEBench, which is structured to continually expand to accommodate more models and datasets, where models cannot be expected to be evaluated on the entire data pool.

**Conclusion.** Our method is sample efficient and robust to sparse input rankings, maintaining accurate rankings with 20x fewer data points.

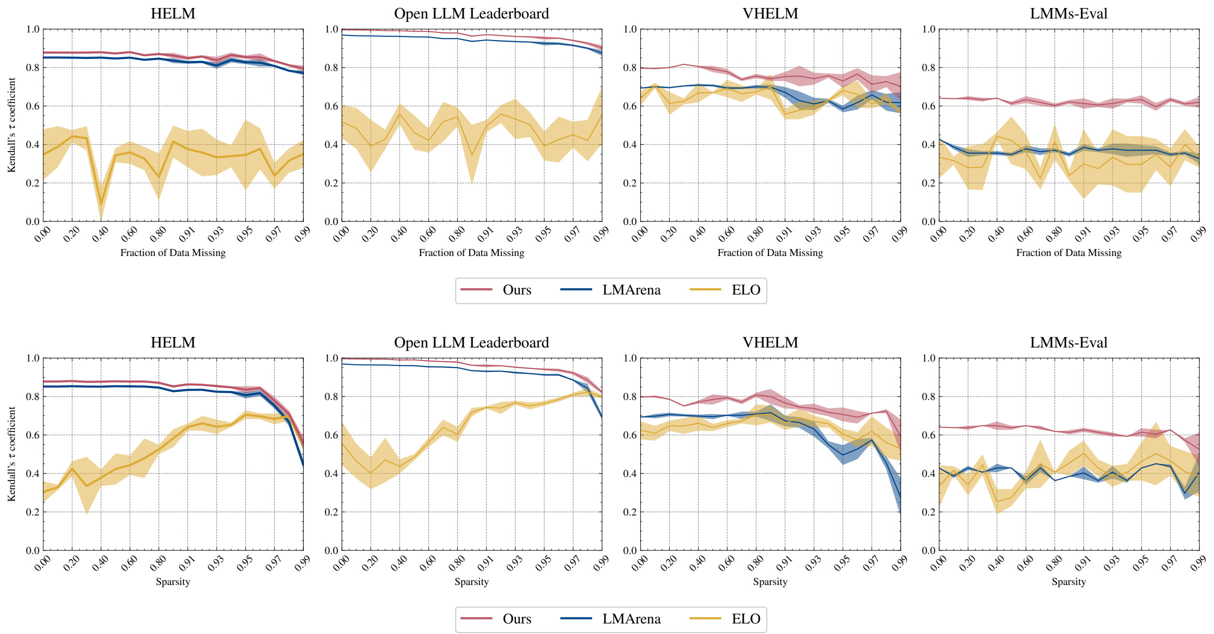


Figure 3: **Sample-efficient convergence and robustness to sparsity.** Kendall $\tau$ between ground-truth ranking and different ranking methods as random individual data samples are dropped (top) and model measurements are randomly removed (bottom). Methods typically remain robust to missing data, with Plackett-Luce consistently achieving higher correlation, even with 95% measurements missing. Querying the sample pool is deterministic and the strongly convex Plackett-Luce model reliably converges to the same ranking.

#### 3.2.4 Comparison to other Social Choice Theory Frameworks

Following the evaluation protocol used in Rofin et al. (2022), we compare our rank aggregation method with the Borda Count and the Dowdall system to provide a broader perspective on its performance against social choice theory methods. As shown in Table 2, our method generally outperforms all the social choice theory frameworks, with drastic boosts on the more heterogeneous and incomplete benchmarks (VHELM and LMMs-Eval). We attribute this to the fact that the Plackett-Luce method is a probabilistic scoring model that explicitly considers the uncertainty in rankings and captures inter-model agreement patterns, while Borda Count(by amplifying noise by assigning uniform linear weights to ranks) and the Dowdall System are both deterministic point-based methods.

| Dataset | Borda | Dowdall | Ours |
|---|---|---|---|
| HELM | 0.81 ± 0.00 | 0.83 ± 0.00 | **0.88 ± 0.00** |
| Leaderboard | 0.95 ± 0.00 | **0.99 ± 0.00** | **0.99 ± 0.00** |
| VHELM | 0.35 ± 0.00 | 0.21 ± 0.00 | **0.79 ± 0.00** |
| LMMs-Eval | 0.08 ± 0.00 | 0.18 ± 0.00 | **0.64 ± 0.00** |

Table 2: **Comparison to other social choice theory methods.** Kendall's $\tau$ correlations to ground-truth ranking across individual benchmarks.

## 4 ONEBench: Creation & Capability Probing

### 4.1 ONEBench-LLM

**Data Pool $\mathcal{D}$.** For ONEBench-LLM, we source data from the Open LLM Leaderboard, HELM, and LMArena. Open LLM Leaderboard and HELM aggregate several individual benchmarks, such as MMLU (Hendrycks et al., 2021a) and HellaSwag (Zellers et al., 2019), while LMArena uses pairwise model comparisons based on user-generated prompts. Metrics include F1-Score, Exact Match (EM), and Quasi-Exact Match (QEM), as well as pairwise preferences.

**Models $\mathcal{M}$.** For ONEBench-LLM, we use the 100 most downloaded models from Open LLM Leaderboard and all 79 models from HELM (as of v1.9.0), including both proprietary models like `GPT-4o` (OpenAI, 2024) and open-weights ones like `LLaMA-3` (Meta, 2024).

### 4.2 ONEBench-LMM

**Data Pool $\mathcal{D}$.** For ONEBench-LMM, data is sourced from VHELM, LMMs-Eval, and WildVisionArena. Similar to ONEBench-LLM, VHELM and LMMs-Eval aggregate individual datasets like MMMU (Yue et al., 2024) and VQAv2 (Goyal et al., 2017), while WildVisionArena uses pairwise tests for LMMs through image-based chats. Measurements include binary metrics like EM, QEM, and real-valued scores like ROUGE (Lin, 2004). We augment pairwise comparisons from WildVisionArena with LLM-as-a-Judge preferences generated using Prometheus-2 (Kim et al., 2024), which correlate highly with human judgments.

**Models $\mathcal{M}$.** For ONEBench-LMM, we use 14 models from LMMs-Eval and 25 models from VHELM, including proprietary models like `Gemini Pro Vision` (Team et al., 2023) and open-weights models like `LLaVA` (Liu et al., 2023a).

### 4.3 Capability Probing

Given a query, the system retrieves relevant data samples using a combination of semantic and metadata search. This *capability probing* provides a personalized comparison of foundation models.

**(i) Semantic search.** We perform k-NN lookup in the embedding space of `all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) for language tasks and `SigLIP-B16` (Zhai et al., 2023) for vision-language tasks, using cosine similarity. We retrieve the top $k$ samples for a given concept with tuned cut-off similarity scores of 0.3 (ONEBench-LLM) and 0.7 (ONEBench-LMM)

**(ii) Metadata search.** We verify that per-sample metadata satisfies the constraints defined in the query. Some benchmarks, such as MMMU, are equipped with detailed metadata, including categories like image type, question type, area, etc.

Using these search mechanisms, we retrieve relevant samples from the data pool and aggregate ordinal model rankings using the Plackett-Luce model. We test ONEBench with a curated set of 50 concepts ranging from domain-specific knowledge, such as the Coriolis effect, to broader academic disciplines like neuroscience, and objects like the iPad (Fig. 4 and Appx. F).

| Metric | LLM | LMM |
|---|---|---|
| Number of concepts | 40 | 50 |
| Cohen-$\kappa$ | 0.79 | 0.91 |
| mAP | 0.85 | 0.73 |
| CMC@1 | 0.95 | 0.94 |
| CMC@10 | 1.00 | 0.96 |

Table 3: **Capability Probing (Quantitative)**: summary of accuracy and retrieval metrics.

**Insight 1. Retrieved samples are accurate and well-aligned with target concepts.** Manual filtering out of incorrect samples by expert annotators[2] confirms high retrieval precision, with a mean Average Precision (mAP) of 0.85 (LLM) and 0.73 (LMM). These results, shown in table 3, indicate that our system reliably retrieves samples matching the intended capabilities, with room for improvement on some underrepresented concepts. Please refer to the per-concept AP in table 10 for a better indicator. Note that the retrieval mechanism is expected to only improve with better models and larger test sets covering more diverse capabilities.

**Insight 2. Top models vary significantly across queries.** A key aspect of evaluating capability-specific performance is checking whether the set of top-performing models remains consistent across different queries. If rankings were largely stable, with the same models dominating regardless of the capability assessed, fine-grained querying would offer limited practical value. However, our results in fig. 4 and fig. 6 reveal substantial variation in the top-$k$ models across domains and concepts. This variation highlights that relying solely on global leaderboards can obscure meaningful differences. These findings confirm that ONEBench enables more targeted model selection by surfacing the most appropriate candidates for a given capability query.

| Samples Retrieved | Random $\tau$ | Query $\tau$ |
|---|---|---|
| 100 | 0.78 ± 0.04 | 0.68 ± 0.10 |
| 1,000 | 0.93 ± 0.01 | 0.83 ± 0.05 |
| 10,000 | 0.98 ± 0.00 | 0.91 ± 0.03 |

Table 4: **Concept-wise probing vs. random retrieval:** Average Kendall's $\tau$ with global ranking.

**Insight 3. More specific queries lead to more distinct model rankings.** We conduct an experiment analyzing the divergence between global and capability-specific model orderings to assess how domain specificity affects ranking. Similar rankings would suggest that domain-specific evaluation adds limited value. For each of the 40 concepts in appendix F, we retrieve $n$ relevant samples and compute the average Kendall rank correlation with the global ranking. As a baseline, we perform five random retrievals of $n$ samples from the full pool and compute their average correlation. Results in table 4 show that concept-specific rankings are

[2] Inter-annotator agreement in table 3 shows strong consistency between annotators. Annotators volunteered and provided informed consent. IRB approval was not obtained.

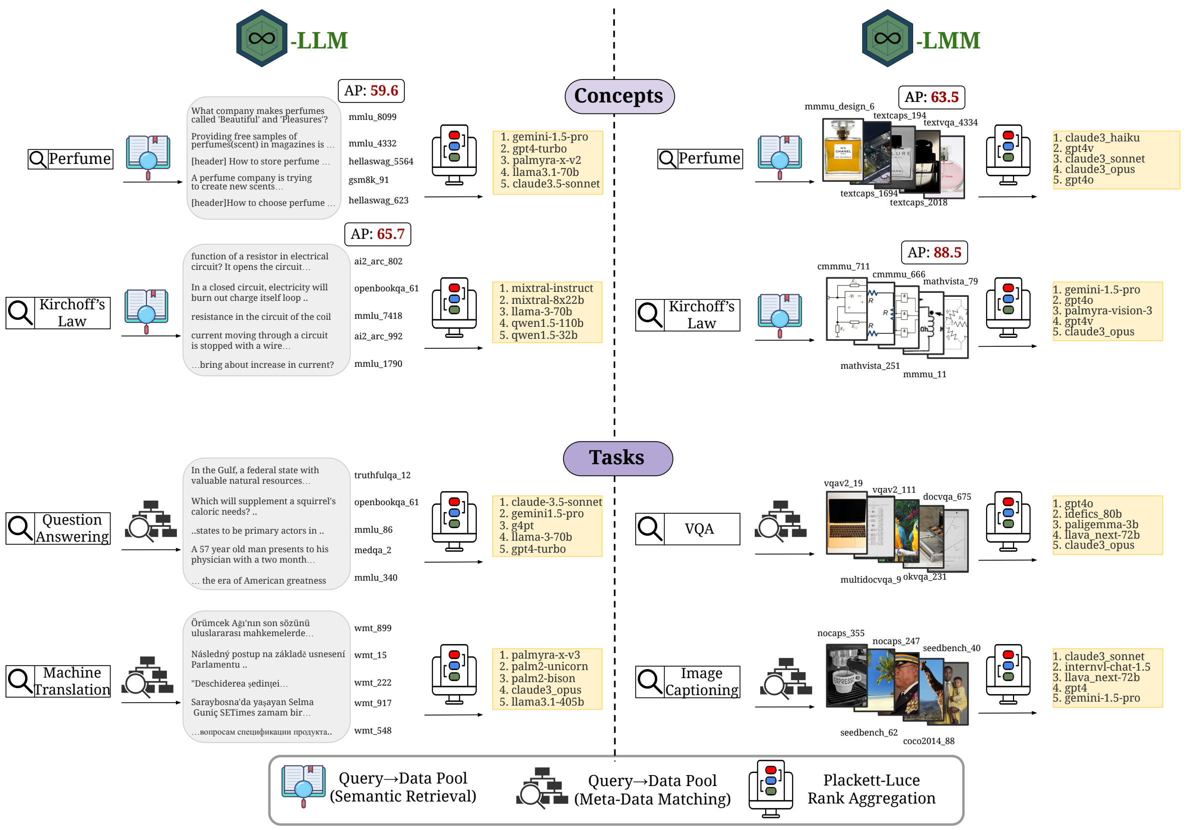


Figure 4: **Capability Probing (Qualitative):** we provide six sample retrieval results for a set of queries covering a diverse set of topics and report the top-5 models for each query.

consistently less correlated with the global ranking than random ones, even as $n$ increases. This suggests that models specialize across domains and should be evaluated accordingly.

## 5 Related Works

Recent multi-task benchmarks, such as GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a), and BigBench (Srivastava et al., 2023), test the broad capabilities of foundation models. However, these benchmarks use arithmetic mean for task aggregation (Beeching et al., 2023) which can distort rankings (Zhang and Hardt, 2024) and is sensitive to outliers (Agarwal et al., 2021) or missing scores (Himmi et al., 2023). ONEBench addresses these by enabling sample reuse, avoiding task selection bias (Dominguez-Olmedo et al., 2024). Inspired by social choice theory, ONEBench employs ordinal rankings and the Plackett-Luce model for aggregation, which is robust to irrelevant alternatives and outliers. Moreover, ONEBench reduces evaluation costs, similar to compressed subsets (Zhao et al., 2024) and lifelong benchmarks (Prabhu et al., 2024). Further, by flexibly integrating diverse sample and measurements contributions, ONEBench can be more inclusive than traditional benchmarks dominated by well-funded institutions (Pouget et al., 2024; Nguyen et al., 2024). We provide an expanded review in appendix G.

## 6 Conclusions and Open Problems

We introduce ONEBench, an open-ended benchmarking framework for foundation models. Our open, democratized benchmarking methodology allows various stakeholders to contribute evaluation samples and model measurements with detailed metadata. This affords creating customized benchmarks and testing arbitrary capabilities with semantic and structured searches. We provide an aggregation mechanism that is both theoretically grounded and empirically validated to be robust to incomplete data and heterogeneous measurements across evaluations. We demonstrate the utility of ONEBench for LLMs and LMMs, showing how dynamic probing reveals new insights into model performance on specific tasks and concepts. This combination of theoretical rigour, empirical results, and practical flexibility makes ONEBench a valuable tool for comprehensive evaluation.

## 7 Limitations

Our approach, while promising, comes with its share of challenges. We highlight three key issues.

**Effects of Aggregation.** Combining different types of evaluation data into a single ranking risks oversimplifying important performance differences. We mitigate this by introducing flexible querying. Furthermore, conversion to pairwise ranking leads to loss of information which could hurt aggregation algorithms. However, in real-world scenarios pairwise measurements perform better, despite information loss (Shah et al., 2014).

**Reproducibility.** The dynamic nature of capability querying and the expanding sample pool, though useful, makes it harder to maintain consistency and can introduce bias during data collection and aggregation. However, our framework supports consistent evaluation, as each sample has a unique identifier and the querying mechanism is deterministic. By simply specifying which samples are included, one can ensure the reproducibility of any evaluation.

**Statistical Modeling Assumptions.** Our reliance on statistical models like Plackett–Luce might make assumptions about data distribution that may not always hold, affecting the reliability of our results. This is not specific to our work, but holds for any work which makes modeling assumptions, and we demonstrate strong empirical performance. Nevertheless, Noothigattu et al. (2020) show that Plackett–Luce based models do not satisfy the following axioms:

**(i) Separability.** If model a is higher than model b in MLE estimate scores in two input sets, a must be higher than b in MLE estimate scores of their combined set. We show empirically that violating separability is not a problem in practice. We present results in Appendix E.

**(ii) Pairwise Majority Consistency.** If pairwise preference order across models are consistent: $a > b$, $b > c$ and $a > c$, then ranking should preserve the consistency: $a > b > c$.

Overall, we believe democratic, open-ended benchmarking is an impactful direction to explore, despite the apparent limitations.

## 8 Broad Impacts

Our work could have a meaningful impact on efficacy of benchmarking for foundation models. With ONEBench, we offer a benchmarking framework that can adapt to different domains, allowing for more inclusive and transparent evaluation practices, empowering researchers and downstream practitioners. By making benchmarking more accessible, we hope to encourage fairness, reproducibility, and innovation in how evaluation frameworks are designed. In the long run, this approach can help build a deeper understanding of foundation models across both language and vision–language tasks. We do not believe that there are any immediate negative societal consequences as a result of this work, but caution that all findings are preliminary and need additional evaluation before deployment.

## Acknowledgements

The authors would like to thank (in alphabetic order): Nikhil Chandak, Karel D'Oosterlinck, Shashwat Goel, Shyamgopal Karthik, Bálint Mucsányi and Tim Weiland for their helpful suggestions and feedback. Special thanks to Bo Li for (promptly and consistently) providing sample-level model evaluation logs of LMMs-Eval, Yujie Lu for providing early access to the battle data of Vision Arena and Palzer Lama for creating the logo of ONEBench and helping with figures. The authors would also like to thank Federico D'Agostino for recommending the title of the paper and the name of the benchmark. AP and MB acknowledge financial support via the Open Philanthropy Foundation funded by the Good Ventures Foundation. VU and SD thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. VU is supported by a Google PhD Fellowship in Machine Intelligence. MB acknowledges support via the CRC 1233 on Robust Vision and from the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

## References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Amith Ananthram, Elias Stengel-Eskin, Carl Vondrick, Mohit Bansal, and Kathleen McKeown. 2024. See it from my perspective: Diagnosing the western cultural bias of large vision-language models in image understanding. *arXiv preprint arXiv:2406.11665*.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, and 1 others. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Kenneth J Arrow. 1950. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346.

Hritik Bansal and Pratyush Maini. 2024. Peeking behind closed doors: Risks of llm evaluation by private data curators. `https://pratyushmaini.github.io/blog/2024/risks-private-evals`. Accessed November 27, 2024.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Alessio Benavoli, Giorgio Corani, and Francesca Mangili. 2016. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2021. Are we done with imagenet? In *Conference on Neural Information Processing Systems (NeurIPS)*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, and 1 others. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Samuel R Bowman and George E Dahl. 2021. What will it take to fix benchmarking in natural language understanding? *arXiv preprint arXiv:2104.02145*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. Introduction to computational social choice. *Handbook of Computational Social Choice*, pages 1–29.

Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachandran. 2024. Benchagents: Automated benchmark creation with agent interaction. *arXiv preprint arXiv:2410.22584*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022a. What are the best systems? new perspectives on nlp benchmarking. *Advances in Neural Information Processing Systems*, 35:26915–26932.

Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. 2022b. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10554–10562.

CRFM. 2024. The first steps to holistic evaluation of vision-language models. `https://crfm.stanford.edu/helm/vhelm/latest/`. Accessed: 2024-06-15.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

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

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. 2024. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization: Quantifying data leakage in nlp performance evaluation. *arXiv preprint arXiv:2102.01818*.

Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with v-usable information. In *International Conference on Machine Learning (ICML)*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. *arXiv preprint arXiv:2009.13888*.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. 2024. Muffin or chihuahua? challenging large vision-language models with multipanel vqa. *arXiv preprint arXiv:2401.15847*.

Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Kathleen Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and 1 others. 2023. Datacomp: In search of the next generation of multimodal datasets. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, and 1 others. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36.

Shahriar Golchin and Mihai Surdeanu. 2023. Data contamination quiz: A tool to detect and estimate contamination in large language models. *arXiv preprint arXiv:2311.06233*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. 2024. Exploring why object recognition performance degrades across income levels and geographies with factor annotations. *Advances in Neural Information Processing Systems*, 36.

Melissa Hall, Samuel J Bell, Candace Ross, Adina Williams, Michal Drozdzal, and Adriana Romero Soriano. 2024. Towards geographic inclusion in the evaluation of text-to-image models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 585–601.

Melissa Hall, Bobbie Chern, Laura Gustafson, Denisse Ventura, Harshad Kulkarni, Candace Ross, and Nicolas Usunier. 2023a. Towards reliable assessments of demographic disparities in multi-label image classifiers. *arXiv preprint arXiv:2302.08572*.

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. 2023b. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *arXiv preprint arXiv:2308.06198*.

Ruijian Han and Yiming Xu. 2023. A unified analysis of likelihood-based estimators in the plackett–luce model. *arXiv preprint arXiv:2306.02821*.

Moritz Hardt and Benjamin Recht. 2022. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press.

Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. 2024. Improving geo-diversity of generated images with contextualized vendi score guidance. *arXiv preprint arXiv:2406.04551*.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Anas Himmi, Ekhine Irurozki, Nathan Noiry, Stephan Clemencon, and Pierre Colombo. 2023. Towards more robust nlp system evaluation: Handling missing scores in benchmarks. *arXiv preprint arXiv:2305.10284*.

Yuheng Huang, Jiayang Song, Qiang Hu, Felix Juefei-Xu, and Lei Ma. 2024. Active testing of large language model via multi-stage sampling. *arXiv preprint arXiv:2408.03573*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Disi Ji, Robert L. Logan, Padhraic Smyth, and Mark Steyvers. 2021. Active bayesian assessment of black-box classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7935–7944.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, and 1 others. 2021. Dynabench: Rethinking benchmarking in nlp. *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Alex Kipnis, Konstantinos Voudouris, Luca M Schulze Buschoff, and Eric Schulz. 2024. metabench—a sparse benchmark to measure general ability in large language models. *arXiv preprint arXiv:2407.12844*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. 2022. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *Conference on Neural Information Processing Systems (NeurIPS)*.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning (ICML)*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

LAION-AI. 2024. Clip_benchmark. Accessed: 2024-06-15.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and 1 others. 2022. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. 2024c. Autobencher: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Christian List. 2022. Social Choice Theory. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2022 edition. Metaphysics Research Lab, Stanford University.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Zhuang Liu and Kaiming He. 2024. A decade's battle on dataset bias: Are we there yet? *arXiv preprint arXiv:2403.08632*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Wang, Yejin Choi, and Bill Yuchen Lin. 2024b. Wildvision arena: Benchmarking multimodal llms in the wild.

Alexandra Sasha Luccioni and David Rolnick. 2023. Bugs in the data: How imagenet misrepresents biodiversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14382–14390.

R Duncan Luce. 1959. *Individual choice behavior*, volume 4. Wiley New York.

R Duncan Luce. 1977. The choice axiom after twenty years. *Journal of mathematical psychology*, 15(3):215–233.

Netta Madvil, Yonatan Bitton, and Roy Schwartz. 2023. Read, look or listen? what's needed for solving a multimodal dataset. *arXiv preprint arXiv:2307.04532*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Lucas Maystre and Matthias Grossglauser. 2015. Fast and accurate inference of plackett–luce models. *Advances in neural information processing systems*, 28.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. Accessed: 2024-06-15.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Swaroop Mishra and Anjana Arunkumar. 2021. How robust are model rankings: A leaderboard customization approach for equitable evaluation. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 13561–13569.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *arXiv preprint arXiv:2405.16915*.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*.

Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. 2020. Axioms for learning from pairwise comparisons. *Advances in Neural Information Processing Systems*, 33:17745–17754.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-06-15.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793.

Lorenzo Pacchiardi, Lucy G Cheke, and José Hernández-Orallo. 2024. 100 instances is all you need: predicting the success of a new llm on unseen data by testing on a few instances. *arXiv preprint arXiv:2409.03563*.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of nlp systems. *arXiv preprint arXiv:2110.10746*.

Matúš Pikuliak and Marián Šimko. 2023. Average is not enough: Caveats of multilingual evaluation. *arXiv preprint arXiv:2301.01269*.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. Civics: Building a dataset for examining culturally-informed values in large language models. *arXiv preprint arXiv:2405.13974*.

Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversityin contrastive vision-language models. *arXiv preprint arXiv:2405.13777*.

Ameya Prabhu, Vishaal Udandarao, Philip Torr, Matthias Bethge, Adel Bibi, and Samuel Albanie. 2024. Lifelong benchmarks: Efficient model evaluation in an era of rapid progress. *arXiv preprint arXiv:2402.19472*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*.

Nils Reimers and Iryna Gurevych. 2019. 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.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinskiy, Andrey Kravchenko, Elena Tutubalina, Tatiana Shavrina, Daniel Karabekyan, and Ekaterina Artemova. 2022. Vote'n'rank: Revision of benchmarking with social choice theory. *Annual Meeting of the Association for Computational Linguistics (EACL)*.

Oscar Sainz, Iker García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, and 1 others. 2024. Data contamination report from the 2024 conda shared task. *arXiv preprint arXiv:2407.21530*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Olawale Salaudeen and Moritz Hardt. 2024. Imagenot: A contrast with imagenet preserves model rankings. *arXiv preprint arXiv:2404.02112*.

Gayathri Saranathan, Mahammad Parwez Alam, James Lim, Suparna Bhattacharya, Soon Yee Wong, Martin Foltin, and Cong Xu. 2024. Dele: Data efficient llm evaluation. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. 2024. Benchmarks as microscopes: A call for model metrology. *arXiv preprint arXiv:2407.16711*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. 2014. When is it better to compare than to score? *arXiv preprint arXiv:1406.6618*.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

Tatiana Shavrina and Valentin Malykh. 2021. How not to lie with a benchmark: Rearranging nlp leaderboards. In *I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*.

Valeriy Shevchenko, Nikita Belousov, Alexey Vasilev, Vladimir Zholobov, Artyom Sosedka, Natalia Semenova, Anna Volodkevich, Andrey Savchenko, and Alexey Zaytsev. 2024. From variability to stability: Advancing recsys benchmarking practices. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5701–5712.

Aditya Siddhant, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the International Conference on Machine Learning 2020*, pages 4411–4421.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hossein Azari Soufiani, David Parkes, and Lirong Xia. 2014. Computing parametric ranking models via rank-breaking. In *International Conference on Machine Learning*, pages 360–368. PMLR.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, and 1 others. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Abhishek Sureddy, Dishant Padalia, Nandhinee Periyakaruppa, Oindrila Saha, Adina Williams, Adriana Romero-Soriano, Megan Richards, Polina Kirichenko, and Melissa Hall. 2024. Decomposed evaluations of geographic disparities in text-to-image models. *arXiv preprint arXiv:2406.11988*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish V Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729.

MTCAJ Thomas and A Thomas Joy. 2006. *Elements of information theory*. Wiley-Interscience.

Louis Leon Thurstone. 1927. Three psychophysical laws. *Psychological Review*, 34(6):424.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J Hénaff. 2024a. Active data curation effectively distills large-scale multimodal models. *arXiv preprint arXiv:2411.18674*.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024b. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022. Ildae: Instance-level difficulty analysis of evaluation data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3412–3425.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Conference on Neural Information Processing Systems (NeurIPS)*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. 2024a. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, and 1 others. 2024b. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. 2024. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. 2024. Top leaderboard ranking= top coding proficiency, always? evoeval: Evolving coding benchmarks via llm. *arXiv preprint arXiv:2403.19114*.

Lirong Xia. 2019. *Learning and decision-making from rank data*. Morgan & Claypool Publishers.

H Peyton Young. 1988. Condorcet's theory of voting. *American Political science review*, 82(4):1231–1244.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. 2023. Skill-mix: A flexible and expandable family of evaluations for ai models. *arXiv preprint arXiv:2310.17567*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*.

Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Hui Xue, Wenhai Wang, Kui Ren, and Jingyi Wang. 2024. S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. *arXiv preprint arXiv:2405.14191*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024a. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Guanhua Zhang and Moritz Hardt. 2024. Inherent trade-offs between diversity and stability in multi-task benchmarks. In *Forty-first International Conference on Machine Learning*.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. 2024b. Task me anything. *arXiv preprint arXiv:2406.11775*.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2024c. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.

Lin Zhao, Tianchen Zhao, Zinan Lin, Xuefei Ning, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Flasheval: Towards fast and accurate evaluation of text-to-image diffusion generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16122–16131.

Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*.

# Part I
# Appendix

## Table of Contents

## A Datasets used in ONEBench: Further Details

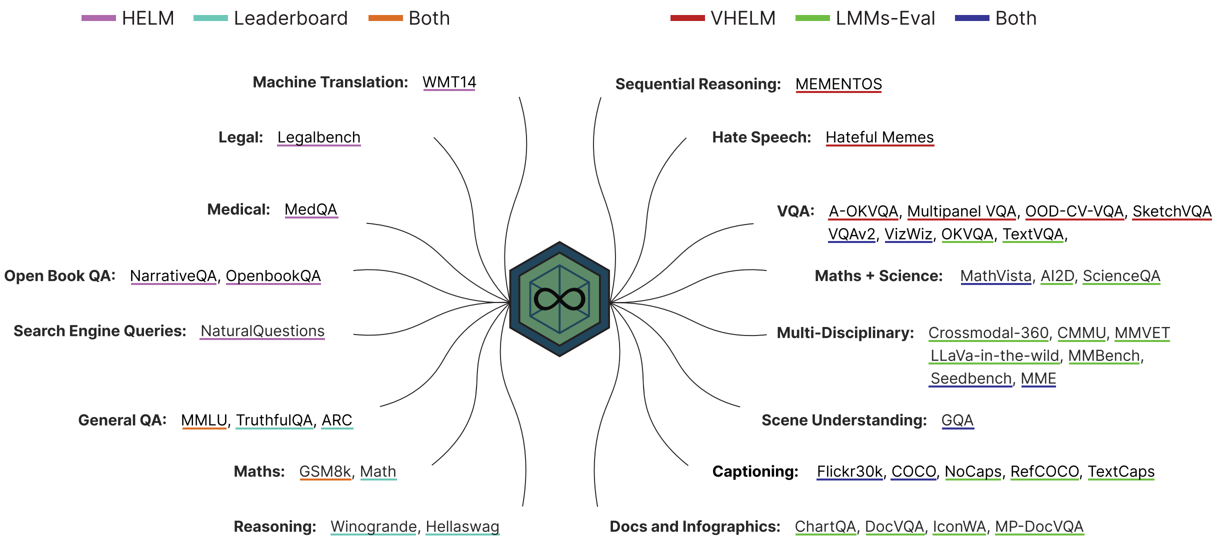


Figure 5: **Constituent datasets of ONEBench-LLM (left) and OneBench-LMM (right)**. We provide task type, metric, and license about each dataset in table 5 and table 6.

| Dataset | Source | Task | Size | Metric | License |
|---|---|---|---|---|---|
| **Cardinal** | | | | | |
| LegalBench (Guha et al., 2024) | HELM | Legal | 1K | QEM | Unknown |
| MATH (Hendrycks et al., 2021b) | HELM | Maths | 1K | QEM | MIT |
| MedQA (Jin et al., 2021) | HELM | Medical | 1K | QEM | MIT |
| NarrativeQA (Kočiskỳ et al., 2018) | HELM | Openbook QA | 1K | F1 | Apache-2.0 |
| NaturalQuestions (Kwiatkowski et al., 2019) | HELM | Search Engine Queries | 1K | F1 | CC BY-SA 3.0 |
| OpenbookQA (Mihaylov et al., 2018) | HELM | Openbook QA | 1K | EM | Apache-2.0 |
| WMT 2014 (Bojar et al., 2014) | HELM | Machine translation | 1K | BLEU | CC-BY-SA-4.0 |
| ARC (Clark et al., 2018) | Leaderboard | General QA | 1.1K | EM | CC-BY-SA-4.0 |
| HellaSwag (Zellers et al., 2019) | Leaderboard | Reasoning | 10K | EM | MIT |
| TruthfulQA (Lin et al., 2022) | Leaderboard | General QA | 817 | EM | Apache-2.0 |
| Winogrande (Sakaguchi et al., 2021) | Leaderboard | Reasoning | 1.2K | EM | Apache-2.0 |
| GSM8K (Cobbe et al., 2021) | HELM + Leaderboard | Maths | 1.3K | QEM | MIT |
| MMLU (Hendrycks et al., 2021a) | HELM + Leaderboard | General QA | 13.8K | EM | MIT |
| **Ordinal** | | | | | |
| Chatbot Arena (Chiang et al., 2024) | Chatbot Arena | Pairwise Battles | 51K | - | CC BY 4.0 |

Table 5: **Datasets in ONEBench-LLM**. A diverse collection of benchmarks testing the abilities of LLMs in areas such as law, medicine, mathematics, question answering, reasoning and instruction following, as well as the performance of LLMs in pairwise battles.

| Dataset | Source | Task | Size | Metric | License |
|---|---|---|---|---|---|
| **Cardinal** | | | | | |
| A-OKVQA (Schwenk et al., 2022) | VHELM | VQA | 7.2K | QEM | Apache-2.0 |
| Bingo (Cui et al., 2023) | VHELM | Bias+Hallucination | 886 | ROUGE | Unknown |
| Crossmodal-3600 (Thapliyal et al., 2022) | VHELM | Captioning | 1.5K | ROUGE | CC BY-SA 4.0 |
| Hateful Memes (Kiela et al., 2020) | VHELM | Hate Speech | 1K | QEM | Custom(Meta) |
| Mementos (Wang et al., 2024b) | VHELM | Sequential Reasoning | 945 | GPT | CC-BY-SA-4.0 |
| MultipanelVQA (Fan et al., 2024) | VHELM | VQA | 200 | QEM | MIT |
| OODCV-VQA (Tu et al., 2023) | VHELM | VQA | 1K | QEM | CC-BY-NC-4.0 |
| PAIRS (Fraser and Kiritchenko, 2024) | VHELM | Bias | 508 | QEM | Unknown |
| Sketchy-VQA (Tu et al., 2023) | VHELM | VQA | 1K | QEM | CC-BY-NC-4.0 |
| AI2D (Kembhavi et al., 2016) | LMMs-Eval | Maths+Science | 3.09K | QEM | Apache-2.0 |
| IconQA (Lu et al., 2021) | LMMs-Eval | Docs and Infographics | 43K | ANLS | CC BY-SA 4.0 |
| InfoVQA (Mathew et al., 2022) | LMMs-Eval | Docs and Infographics | 6.1K | ANLS | Unknown |
| LLaVA-in-the-Wild (Liu et al., 2023a) | LMMs-Eval | Multi-disciplinary | 60 | GPT4 | Apache-2.0 |
| ChartQA (Masry et al., 2022) | LMMs-Eval | Docs and Infographics | 2.5K | QEM | GPL-3.0 |
| CMMMU (Zhang et al., 2024a) | LMMs-Eval | Multi-disciplinary | 900 | QEM | CC-BY-4.0 |
| DocVQA (Mathew et al., 2021) | LMMs-Eval | Docs and Infographics | 10.5K | ANLS | Unknown |
| MMBench (Liu et al., 2023b) | LMMs-Eval | Multi-disciplinary | 24K | GPT | Apache-2.0 |
| MMVET (Yu et al., 2024) | LMMs-Eval | Multi-disciplinary | 218 | GPT | Apache-2.0 |
| MP-DocVQA (Tito et al., 2023) | LMMs-Eval | Docs and Infographics | 5.2K | QEM | MIT |
| NoCaps (Agrawal et al., 2019) | LMMs-Eval | Captioning | 4.5K | ROUGE | MIT |
| OK-VQA (Marino et al., 2019) | LMMs-Eval | VQA | 5.1K | ANLS | Unknown |
| RefCOCO (Kazemzadeh et al., 2014; Mao et al., 2016) | LMMs-Eval | Captioning | 38K | ROUGE | Apache-2.0 |
| ScienceQA (Lu et al., 2022) | LMMs-Eval | Maths+Science | 12.6K | EM | CC BY-NC-SA 4.0 |
| TextCaps (Sidorov et al., 2020) | LMMs-Eval | Captioning | 3.2K | ROUGE | CC BY 4.0 |
| TextVQA (Singh et al., 2019) | LMMs-Eval | VQA | 5K | EM | CC BY 4.0 |
| COCO (Lin et al., 2014) | VHELM+LMMs-Eval | Captioning | 45.5K | ROUGE | CC-BY-4.0 |
| Flickr30k (Young et al., 2014) | VHELM+LMMs-Eval | Captioning | 31K | ROUGE | CC-0 Public Domain |
| GQA(Hudson and Manning, 2019) | VHELM+LMMs-Eval | Scene Understanding | 12.6K | QEM | CC-BY-4.0 |
| MathVista (Lu et al., 2024a) | VHELM+LMMs-Eval | Maths+Science | 1K | QEM/GPT4 | CC-BY-SA-4.0 |
| MME (Fu et al., 2023) | VHELM+LMMs-Eval | Multi-disciplinary | 2.4K | QEM/C+P | Unknown |
| MMMU (Yue et al., 2024) | VHELM+LMMs-Eval | Multi-disciplinary | 900 | QEM | CC BY-SA 4.0 |
| POPE (Li et al., 2023b) | VHELM+LMMs-Eval | Hallucination | 9K | QEM/EM | MIT |
| SEED-Bench (Li et al., 2023a, 2024a) | VHELM+LMMs-Eval | Multi-disciplinary | 42.5K | QEM/EM | Apache |
| VizWiz (Gurari et al., 2018) | VHELM+LMMs-Eval | VQA | 4.3K | QEM/EM | CC BY 4.0 |
| VQAv2 (Goyal et al., 2017) | VHELM+LMMs-Eval | VQA | 214K | QEM/EM | CC BY 4.0 |
| **Ordinal** | | | | | |
| Vision Arena (Lu et al., 2024b) | - | Pairwise Battles | 9K | - | MIT |
| LMMs-Eval(Prometheus2) (Kim et al., 2024) | - | Pairwise Battles | 610K | - | MIT |

Table 6: **Datasets in ONEBench-LMM**: a diverse collection of benchmarks testing the abilities of LLMs in tasks such as general VQA, image captioning, hate speech detection, bias and hallucination understanding, maths and science, documents and infographics, scene understanding and sequential reasoning as well as the performance of LMMs in pairwise battles. Additional preference comparisons are sampled randomly from LMMs-Eval, which are excluded from the cardinal measurement sample pool.

## B Models used in ONEBench: Further Details

In this section, we provide a deeper insight into the models used in the creation of ONEBench. It is important to note that ONEBench-LLM and ONEBench-LMM have complementary characteristics: while ONEBench-LLM has fewer data samples $\mathcal{D}_k$, they are evaluated on more models $\mathcal{M}_k$, while ONEBench-LMM contains (significantly) more data samples but they are evaluated on less models.

### B.1 ONEBench-LLM: Open LLM Leaderboard

The Open LLM Leaderboard (Beeching et al., 2023) was created to track progress of LLMs in the open-source community by evaluating models on the same data samples and setup for more reproducible results and a trustworthy leaderboard where all open-sourced LLMs could be ranked.

However, due to the abundance of models found on the leaderboard and the lack of adequate documentation, and therefore reliability, of many of these models being evaluated, we rank the models based on the number of downloads, as a metric of adoption of these models by the community. We provide the total list of models as an artefact and list the top 100 models below:

1. `01-ai/Yi-34B-200K`
2. `AI-Sweden-Models/gpt-sw3-126m`
3. `BioMistral/BioMistral-7B`
4. `CohereForAI/c4ai-command-r-plus`
5. `CohereForAI/c4ai-command-r-v01`
6. `Deci/DeciLM-7B-instruct`
7. `EleutherAI/llemma_7b`
8. `EleutherAI/pythia-410m`
9. `Felladrin/Llama-160M-Chat-v1`
10. `Felladrin/Llama-68M-Chat-v1`
11. `FreedomIntelligence/AceGPT-7B`
12. `GritLM/GritLM-7B`
13. `Intel/neural-chat-7b-v3-1`
14. `JackFram/llama-160m`
15. `Nexusflow/NexusRaven-V2-13B`
16. `Nexusflow/Starling-LM-7B-beta`
17. `NousResearch/Hermes-2-Pro-Mistral-7B`
18. `NousResearch/Meta-Llama-3-8B-Instruct`
19. `NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO`
20. `NousResearch/Nous-Hermes-2-SOLAR-10.7B`
21. `NousResearch/Nous-Hermes-2-Yi-34B`
22. `OpenPipe/mistral-ft-optimized-1227`
23. `Qwen/Qwen1.5-0.5B`
24. `Qwen/Qwen1.5-0.5B-Chat`
25. `Qwen/Qwen1.5-1.8B`
26. `Qwen/Qwen1.5-1.8B-Chat`
27. `Qwen/Qwen1.5-110B-Chat`

28. Qwen/Qwen1.5-14B
29. Qwen/Qwen1.5-14B-Chat
30. Qwen/Qwen1.5-32B-Chat
31. Qwen/Qwen1.5-4B
32. Qwen/Qwen1.5-4B-Chat
33. Qwen/Qwen1.5-72B-Chat
34. Qwen/Qwen1.5-7B
35. Qwen/Qwen1.5-7B-Chat
36. SeaLLMs/SeaLLM-7B-v2
37. TinyLlama/TinyLlama-1.1B-Chat-v1.0
38. TinyLlama/TinyLlama-1.1B-intermediate-step-3T
39. VAGOsolutions/SauerkrautLM-Mixtral-8x7B
40. abhishekchohan/mistral-7B-forest-dpo
41. ahxt/LiteLlama-460M-1T
42. ai-forever/mGPT
43. alignment-handbook/zephyr-7b-sft-full
44. augmxnt/shisa-gamma-7b-v1
45. bigcode/starcoder2-15b
46. bigcode/starcoder2-3b
47. bigcode/starcoder2-7b
48. cloudyu/Mixtral_7Bx4_MOE_24B
49. codellama/CodeLlama-70b-Instruct-hf
50. cognitivecomputations/dolphin-2.2.1-mistral-7b
51. cognitivecomputations/dolphin-2.6-mistral-7b-dpo
52. cognitivecomputations/dolphin-2.9-llama3-8b
53. daekeun-ml/phi-2-ko-v0.1
54. deepseek-ai/deepseek-coder-1.3b-instruct
55. deepseek-ai/deepseek-coder-6.7b-base
56. deepseek-ai/deepseek-coder-6.7b-instruct
57. deepseek-ai/deepseek-coder-7b-instruct-v1.5
58. deepseek-ai/deepseek-math-7b-base
59. deepseek-ai/deepseek-math-7b-instruct
60. deepseek-ai/deepseek-math-7b-rl
61. google/codegemma-7b-it
62. google/gemma-1.1-7b-it
63. google/gemma-2b
64. google/gemma-2b-it
65. google/gemma-7b

66. google/gemma-7b-it
67. google/recurrentgemma-2b-it
68. h2oai/h2o-danube2-1.8b-chat
69. hfl/chinese-alpaca-2-13b
70. ibm/merlinite-7b
71. meta-llama/Meta-Llama-3-70B
72. meta-llama/Meta-Llama-3-70B-Instruct
73. meta-llama/Meta-Llama-3-8B
74. meta-llama/Meta-Llama-3-8B-Instruct
75. meta-math/MetaMath-Mistral-7B
76. microsoft/Orca-2-7b
77. microsoft/phi-2
78. mistral-community/Mistral-7B-v0.2
79. mistral-community/Mixtral-8x22B-v0.1
80. mistralai/Mistral-7B-Instruct-v0.2
81. mistralai/Mixtral-8x22B-Instruct-v0.1
82. mistralai/Mixtral-8x7B-Instruct-v0.1
83. mistralai/Mixtral-8x7B-v0.1
84. openai-community/gpt2
85. openai-community/gpt2-large
86. openchat/openchat-3.5-0106
87. openchat/openchat-3.5-1210
88. openchat/openchat_3.5
89. sarvamai/OpenHathi-7B-Hi-v0.1-Base
90. speakleash/Bielik-7B-Instruct-v0.1
91. speakleash/Bielik-7B-v0.1
92. stabilityai/stablelm-2-1_6b
93. stabilityai/stablelm-2-zephyr-1_6b
94. stabilityai/stablelm-zephyr-3b
95. teknium/OpenHermes-2.5-Mistral-7B
96. tokyotech-llm/Swallow-70b-instruct-hf
97. upstage/SOLAR-10.7B-Instruct-v1.0
98. upstage/SOLAR-10.7B-v1.0
99. wenbopan/Faro-Yi-9B
100. yanolja/EEVE-Korean-Instruct-10.8B-v1.0

### B.2 ONEBench-LLM: HELM

Similar to the Open LLM Leaderboard, the goal of HELM was to provide a uniform evaluation of language models over a vast set of data samples (termed as `scenarios` in Liang et al. (2023)). HELM, however, has a broader scope of models used for evaluation, employing open, limited-access, and closed models. All models currently used in ONEBench-LLM is listed below:

1. 01-ai_yi-34b
2. 01-ai_yi-6b
3. 01-ai_yi-large-preview
4. ai21_j2-grande
5. ai21_j2-jumbo
6. ai21_jamba-1.5-large
7. ai21_jamba-1.5-mini
8. ai21_jamba-instruct
9. AlephAlpha_luminous-base
10. AlephAlpha_luminous-extended
11. AlephAlpha_luminous-supreme
12. allenai_olmo-7b
13. anthropic_claude-2.0
14. anthropic_claude-2.1
15. anthropic_claude-3-5-sonnet-20240620
16. anthropic_claude-3-haiku-20240307
17. anthropic_claude-3-opus-20240229
18. anthropic_claude-3-sonnet-20240229
19. anthropic_claude-instant-1.2
20. anthropic_claude-instant-v1
21. anthropic_claude-v1.3
22. cohere_command
23. cohere_command-light
24. cohere_command-r
25. cohere_command-r-plus
26. databricks_dbrx-instruct
27. deepseek-ai_deepseek-llm-67b-chat
28. google_gemini-1.0-pro-001
29. google_gemini-1.0-pro-002
30. google_gemini-1.5-flash-001
31. google_gemini-1.5-pro-001
32. google_gemini-1.5-pro-preview-0409
33. google_gemma-2-9b-it
34. google_gemma-2-27b-it

35. google_gemma-7b
36. google_text-bison@001
37. google_text-unicorn@001
38. meta_llama-2-7b
39. meta_llama-2-13b
40. meta_llama-2-70b
41. meta_llama-3-8b
42. meta_llama-3-70b
43. meta_llama-3.1-8b-instruct-turbo
44. meta_llama-3.1-70b-instruct-turbo
45. meta_llama-3.1-405b-instruct-turbo
46. meta_llama-65b
47. microsoft_phi-2
48. microsoft_phi-3-medium-4k-instruct
49. microsoft_phi-3-small-8k-instruct
50. mistralai_mistral-7b-instruct-v0.3
51. mistralai_mistral-7b-v0.1
52. mistralai_mistral-large-2402
53. mistralai_mistral-large-2407
54. mistralai_mistral-medium-2312
55. mistralai_mistral-small-2402
56. mistralai_mixtral-8x7b-32kseqlen
57. mistralai_mixtral-8x22b
58. mistralai_open-mistral-nemo-2407
59. nvidia_nemotron-4-340b-instruct
60. openai_gpt-3.5-turbo-0613
61. openai_gpt-4-0613
62. openai_gpt-4-1106-preview
63. openai_gpt-4-turbo-2024-04-09
64. openai_gpt-4o-2024-05-13
65. openai_gpt-4o-mini-2024-07-18
66. openai_text-davinci-002
67. openai_text-davinci-003
68. qwen_qwen1.5-7b
69. qwen_qwen1.5-14b
70. qwen_qwen1.5-32b
71. qwen_qwen1.5-72b
72. qwen_qwen1.5-110b-chat

73. qwen_qwen2-72b-instruct
74. snowflake_snowflake-arctic-instruct
75. tiiuae_falcon-7b
76. tiiuae_falcon-40b
77. writer_palmyra-x-004
78. writer_palmyra-x-v2
79. writer_palmyra-x-v3

### B.3 ONEBench-LMM: LMMs-Eval

LMMs-Eval is the first comprehensive large-scale evaluation benchmark for Large Multimodal models, meant "to promote transparent and reproducible evaluations" (Zhang et al., 2024c). The models supported by LMMs-Eval are primarily open-sourced and the full list of currently used models are listed below:

1. idefics2-8b
2. internlm-xcomposer2-4khd-7b
3. instructblip-vicuna-7b
4. instructblip-vicuna-13b
5. internVL-Chat-V1-5
6. llava-13b
7. llava-1.6-13b
8. llava-1.6-34b
9. llava-1.6-mistral-7b
10. llava-1.6-vicuna-13b
11. llava-1.6-vicuna-7b
12. llava-7b
13. llava-next-72b
14. qwen_vl_chat

### B.4 ONEBench-LMM: VHELM

Finally, ONEBench-LMM comprises VHELM, an extension of HELM for Vision-Language models. The models currently used by us, spanning open, limited-access, and closed models, are as follows:

1. anthropic_claude_3_haiku_20240307
2. anthropic_claude_3_opus_20240229
3. anthropic_claude_3_sonnet_20240229
4. google_gemini_1.0_pro_vision_001
5. google_gemini_1.5_pro_preview_0409
6. google_gemini_pro_vision
7. google_paligemma_3b_mix_448
8. huggingfacem4_idefics2_8b
9. huggingfacem4_idefics_80b
10. huggingfacem4_idefics_80b_instruct
11. huggingfacem4_idefics_9b

12. huggingfacem4_idefics_9b_instruct
13. llava_1.6_mistral_7b
14. llava_1.6_vicuna_13b
15. llava_1.6_vicuna_7b
16. microsoft_llava_1.5_13b_hf
17. microsoft_llava_1.5_7b_hf
18. mistralai_bakllava_v1_hf
19. openai_gpt_4_1106_vision_preview
20. openai_gpt_4_vision_preview
21. openai_gpt_4o_2024_05_13
22. openflamingo_openflamingo_9b_vitl_mpt7b
23. qwen_qwen_vl
24. qwen_qwen_vl_chat
25. writer_palmyra_vision_003

# C Sample-level Rankings

## C.1 Further Details

In our ONEBench formulation, $s_j \in \mathcal{S}$ represents an ordinal ranking over the models $\mathcal{M}_j$ for sample $(x_j, y_j)$ represented by a permutation $\sigma_j$ such that $f_{\sigma_j(1)} \succeq \cdots \succeq f_{\sigma_j(m_j)}$ where $m_j = |\mathcal{M}_j|$ is the number of models compared in the $j$-th sample-level ranking. In addition, for each $k$ we distinguish the case $f_{\sigma(k-1)} \succ f_{\sigma(k)}$ if $f_{\sigma(k-1)}$ performs better than $f_{\sigma(k)}$ and $f_{\sigma(k-1)} \sim f_{\sigma(k)}$ in case of indistinguishable performance. Thus, each sample-level ranking $s_j \in \mathcal{S}$ can be uniquely determined by a mapping $\sigma_j : \{1, \ldots, m_j\} \to \{1, \ldots, m\}$ with $\sigma_j(k)$ providing the index of the model in $\mathcal{M}$ that is on the $k$-th place in the ordering for the $j$-th sample-level ranking and $\pi_j \in \{\succ, \sim\}^{m_j-1}$ defining the corresponding binary sequence of pairwise performance relations.

## C.2 Ordinal Rankings and Information Loss

Using ordinal measurements leads to information loss, which can impede downstream aggregation algorithms due to the data processing inequality (Thomas and Joy 2006, Section 2.8). This principle asserts that any estimation made from processed data cannot outperform estimation based on the original, unprocessed data. However, cardinal measurements frequently suffer from calibration issues, even within a single metric (Shah et al., 2014). Consequently, in practice, ordinal measurements can paradoxically outperform cardinal ones despite the inherent information loss.

## D Synthetic data experiments

To further verify the ability of our method to recover the ground truth ranking, we create a simulation of the data-generating process. Following Colombo et al. (2022a), we generate synthetic scores using Gumbel-distributed random variables. Specifically, we simulate $N = 100$ systems, each modeled by a Gumbel random variable $G_n$ centered at $\phi \cdot n$ with scale $\beta = 1$, where $\phi \in [0, 1]$ is a dispersion parameter. We then generate $n = 1000$ measurements, where the score of each system is sampled from its corresponding Gumbel distribution. The measurements are in equal proportions numerical (raw scores), binary (whether the score is over a randomly selected threshold), and ordinal (whether the score of one system is higher than another). In Table 7, we present the Kendall rank correlation to the ground truth ranking for different values of $\phi$, which controls the noise in the ranking (a lower value is more noise).

| **Dispersion $\phi$** | **ELO** | **LMArena** | **Ours** |
|---|---|---|---|
| 0.01 | 0.28 ± 0.02 | 0.88 ± 0.00 | **0.92 ± 0.00** |
| 0.02 | 0.39 ± 0.05 | 0.92 ± 0.01 | **0.96 ± 0.01** |
| 0.05 | 0.74 ± 0.01 | 0.95 ± 0.00 | **0.98 ± 0.00** |
| 0.10 | 0.85 ± 0.03 | 0.93 ± 0.01 | **0.99 ± 0.00** |

Table 7: Kendall rank correlation to the ground truth ranking at different dispersion levels of synthetic model measurements. Lower dispersion means more noise.

| **Missing data** | **ELO** | **LMArena** | **Ours** |
|---|---|---|---|
| 50% | 0.87 ± 0.02 | 0.92 ± 0.00 | **0.99 ± 0.00** |
| 90% | 0.84 ± 0.01 | 0.90 ± 0.00 | **0.97 ± 0.00** |
| 95% | 0.84 ± 0.03 | 0.88 ± 0.01 | **0.95 ± 0.01** |
| 99% | 0.86 ± 0.01 | 0.83 ± 0.01 | **0.89 ± 0.01** |

Table 8: Kendall rank correlation to the ground truth ranking with different percentages of synthetic model measurements missing ($\phi = 0.1$).

# E Separability

To study separability, we follow the definition in Noothigattu et al. (2020) and randomly split the data in each benchmark into two equal-sized subsets 5 times. Then, we use all three aggregation methods to produce sub-rankings, as well as a combined ranking. We identify model pairs that are consistently ranked in both sub-rankings and compute the percentage of cases where that ordering is maintained in the combined ranking. The results, showing the mean and standard deviation of these percentages, are presented in Table 9.

| **Benchmark** | **ELO (mean ± std)** | **LMArena (mean ± std)** | **Plackett-Luce (mean ± std)** |
|---|---|---|---|
| HELM | 91.84 ± 0.25 | 100.00 ± 0.00 | 100.00 ± 0.00 |
| Leaderboard | 83.11 ± 0.73 | 100.00 ± 0.00 | 100.00 ± 0.00 |
| VHELM | 95.81 ± 0.77 | 100.00 ± 0.00 | 100.00 ± 0.00 |
| LMMs-Eval | 91.85 ± 0.88 | 100.00 ± 0.00 | 100.00 ± 0.00 |
| Synthetic | 97.37 ± 0.40 | 100.00 ± 0.00 | 100.00 ± 0.00 |

Table 9: **Separability analysis:** Percentage of consistent pairwise rankings from sub-rankings preserved in the combined ranking.

# F Capability Testing Across Arbitrary Queries

## F.1 Queries: List and Additional Results

| **Concept** | ONEBench-LLM AP | ONEBench-LMM AP |
|---|---|---|
| Common Queries | | |
| apple ipad | 0.7435 | 0.1985 |
| architecture | 0.7683 | 0.8981 |
| beach | 0.7152 | 0.5698 |
| biochemistry | 0.9778 | 0.7303 |
| boat | 0.7728 | 0.8829 |
| botany | 0.9876 | 0.7556 |
| bus | 0.9035 | 0.9739 |
| car | 0.9140 | 0.8477 |
| cell(biology) | 0.9937 | 0.5075 |
| china tourism | 0.6392 | 1.0000 |
| cigarette advertisment | 0.7249 | 0.6590 |
| coffee maker | 0.8426 | 0.4057 |
| components of a bridge | 0.9222 | 0.5865 |
| decomposition of benzene(organic chemistry) | 0.6745 | 0.7623 |
| epidemiology | 0.9316 | 0.7991 |
| kirchoff's law(electrical engineering) | 0.6572 | 0.4824 |
| food chain | 0.5405 | 1.0000 |
| game of football | 0.8221 | 1.0000 |
| german shepherd (dog) | 0.9359 | 0.3078 |
| gothic style (architecture) | 0.7829 | 1.0000 |
| law | 0.8566 | 0.4138 |
| literary classics | 0.9869 | 1.0000 |
| macroeconomics | 1.0000 | 0.9570 |
| makeup | 1.0000 | 0.2247 |
| microwave oven | 0.7979 | 1.0000 |
| neuroscience components | 0.9844 | 0.2854 |
| pasta | 0.5678 | 0.2142 |
| perfume | 0.5996 | 0.6355 |
| photosynthesis | 0.9848 | 0.3665 |
| plants | 1.0000 | 0.6488 |
| political diplomacy | 0.9529 | 0.9561 |
| python code | 0.8850 | 0.9444 |
| renaissance painting | 0.9270 | 0.9799 |
| shareholder report | 1.0000 | 0.8317 |
| sheet music | 0.8322 | 0.9750 |
| solar cell battery | 0.8853 | 0.8082 |
| thermodynamics | 0.9567 | 0.8852 |
| united states of america | 0.8096 | 0.8642 |
| vaccines | 0.8572 | 0.3411 |
| volcanic eruption | 0.7905 | 0.9229 |
| Queries testing Visual Capabilities | | |
| bike leaning against a wall | - | 0.8271 |
| child playing baseball | - | 0.9638 |
| coriolis effect | - | 0.7063 |
| dijkstra's shortest path algorithm | - | 0.9135 |
| empty bridge overlooking the sea | - | 0.5934 |
| judo wrestling | - | 0.6092 |
| man in a suit | - | 0.5611 |
| musical concert | - | 0.9879 |
| sine wave | - | 0.4232 |
| woman holding an umbrella | - | 0.8821 |

Table 10: Aggregate Average Precision(AP) for ONEBench-LLM and ONEBench-LMM concepts.

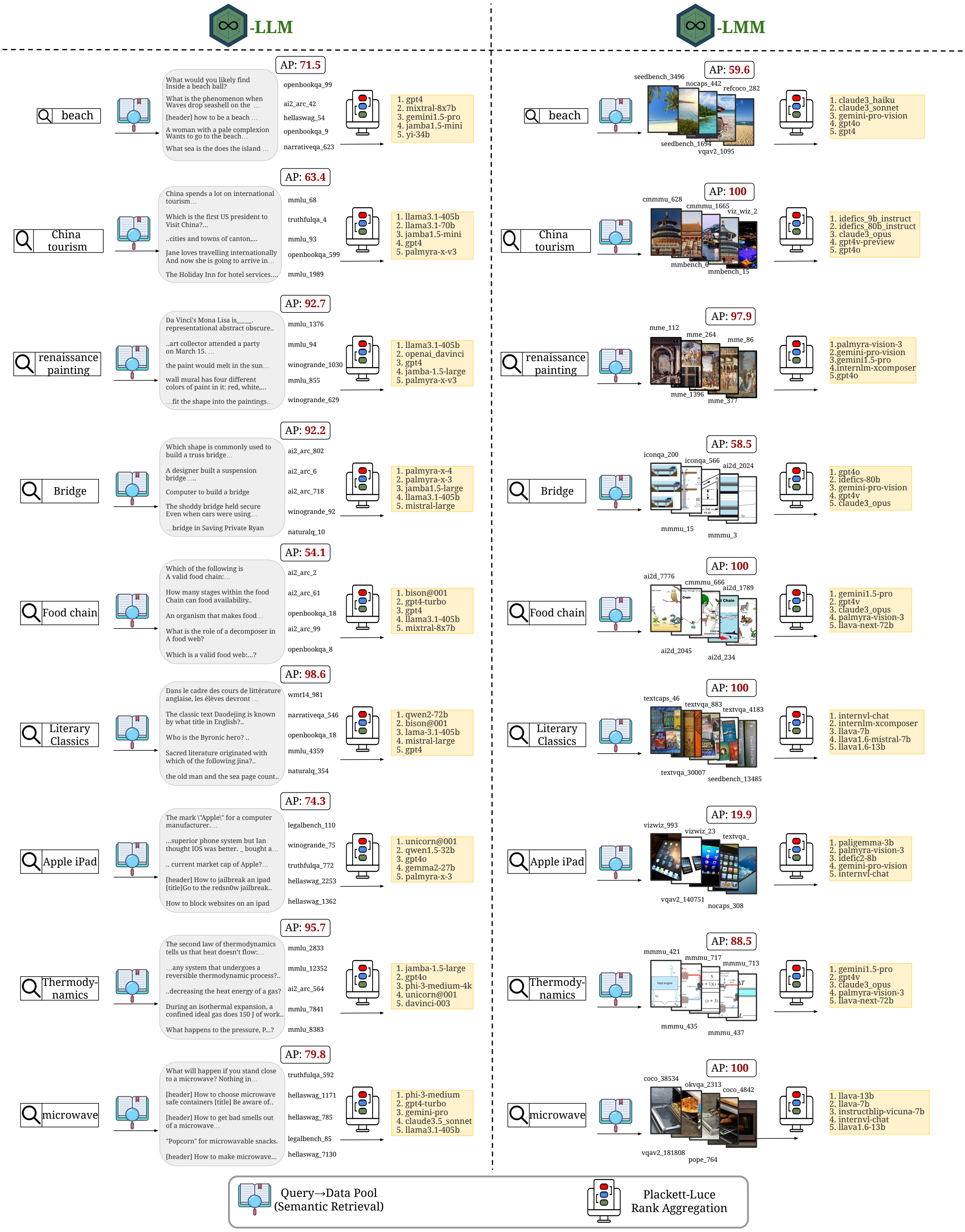


Figure 6: Additional qualitative analysis for ONEBench's capability probing for selected queries.

## F.2 Variability in Concept-specific Rankings

We extend the analysis conducted in Section 4.3 demonstrating that the performance gap is non-trivial while comparing model rankings for specific capabilities to randomly sampling datapoints and aggregating model performance. Showcasing this reiterates the practical value of ONEBench - such that practioners who are interested in specific capabilities may extract the best LLM/LMM for their needs.

We already studied how the number of samples determines model ranking variation over the entire concept pool. Now, we check whether the top-k model rankings vary between the global ranking over the whole data pool and for specific capabilities.

Regarding the top-k models, we do a specific study into ONEBench-LMM, as we test on more concepts than ONEBench-LMM(50 vs 40). The overall top-1 model using the Plackett-Luce algorithm, evidenced as the state-of-the-art rank aggregation method, combining both LMMs-Eval and VHELM is `openai_gpt_4o_2024_05_13`. Hence, we conduct our analysis on the top-5 and top-10 model rankings.

For all the concepts collected for ONEBench-LMM (Table 10), we provide aggregated results of correlations between the overall ranking obtained by the Plackett-Luce algorithm and specific benchmark rankings. We indicate the overall rank differences for each concept with respect to the overall ranking and the average Kendall-$\tau$ correlation of model rankings for all the capability-specific benchmarks. We also conduct a Wilcoxon signed-rank test as a hypothesis test to check the systematic difference between paired observations as shown in Table 11.

| **Metric** | **Top-5** | **Top-10** |
|---|---|---|
| Mean absolute rank difference | 7.900 | 7.920 |
| Mean Kendall's $\tau$ correlation | 0.220 | 0.124 |
| Wilcoxon test statistic | 53.000 | 623.000 |
| Wilcoxon $p$-value | $9.66 \times 10^{-7}$ | $1.63 \times 10^{-8}$ |
| Conclusion | Significant difference | Significant difference |

Table 11: Comparison of Overall vs. Capability-Specific Rankings

We infer from the results that there is a significant difference between overall and specific rankings. Hence, across experiments, we observe variations across the capabilities we test and conclude that we cannot have a one-model-fits-all system for different concepts and domains.

## G Extended Related Works

**Multi-task Benchmarks as Broad Capability Evaluators.** Multi-task leaderboards have been the standard for benchmarking foundation models. Examples include GLUE (Wang et al., 2019b), decaNLP (McCann et al., 2018), SuperGLUE (Wang et al., 2019a), BigBench (Srivastava et al., 2023), Dynabench (Kiela et al., 2021), Open LLM Leaderboard (Beeching et al., 2023), CLIP-Benchmark (LAION-AI, 2024), ELEVATOR (Li et al., 2022), StableEval (Udandarao et al., 2024a) and DataComp-38 (Gadre et al., 2023), as well as massive multitask benchmarks like XTREME (Siddhant et al., 2020) and ExT5 (Aribandi et al., 2021). However, concerns have arisen regarding the limitations of multi-task benchmarks (Bowman and Dahl, 2021). Issues include saturation and subsequent discarding of samples (Liao et al., 2021; Beyer et al., 2021; Ott et al., 2022; Ethayarajh and Jurafsky, 2020; Xia et al., 2024), susceptibility to dataset selection (Dehghani et al., 2021), obscuring progress by evaluation metrics (Schaeffer et al., 2023; Colombo et al., 2022b), training on test tasks (Udandarao et al., 2024b; Dominguez-Olmedo et al., 2024; Nezhurina et al., 2024; Mirzadeh et al., 2024; Srivastava et al., 2024; Wang et al., 2024a), and data contamination (Elangovan et al., 2021; Magar and Schwartz, 2022; Deng et al., 2023; Golchin and Surdeanu, 2023; Sainz et al., 2024). ONEBench tackles these challenges by enabling extensive reuse of samples for broader model comparisons, avoiding task selection bias through democratized sourcing of samples, and using ordinal rankings to avoid evaluation minutia. Sample-level evaluation with sparse inputs also allows selective removal of contaminated data for fairer comparisons. Moreover, by supporting over-ended, evolving evaluation, it makes it harder to train on all test tasks, as opposed to fixed leaderboards that are easier to game.

**On Aggregation across Benchmarks.** The dominant approach to benchmarking has traditionally been multi-task benchmarks, where the most common aggregation strategy is the arithmetic mean of scores across individual tasks. However, this approach assumes that the scoring metrics are homogeneous and scaled correctly, and treat tasks of different complexities equally (Mishra and Arunkumar, 2021; Pikuliak and Šimko, 2023). In consequence, simple normalization preprocessing influences the rankings (Colombo et al., 2022a), and makes them nearly entirely dependent on outlier tasks (Agarwal et al., 2021). Simply changing the aggregation method from arithmetic to geometric or harmonic mean can change the ranking (Shavrina and Malykh, 2021). Similarly, including irrelevant alternative models can change statistical significance or even change the ranking entirely (Benavoli et al., 2016; Zhang and Hardt, 2024). Mean-aggregation also has significant failure modes in handling missing scores in benchmarks (Himmi et al., 2023). The benchmarking paradigm is hence shifting towards adopting evaluation principles from other fields, such as non-parametric statistics and social choice theory (Brandt et al., 2016; Rofin et al., 2022). We use ordinal rankings instead of scores, similar to LMArena. However, unlike Arena, we use the pairwise variant of the Plackett-Luce model, which has been shown to have advantages both theoretically and empirically (Peyrard et al., 2021). We benefit from some of its theoretical properties like identifiability, sample-efficient convergence, provable robustness to irrelevant alternatives, non-dominance of outliers and empirical robustness across a wide range of real-world factors which affect ranking. Moreover, we do not aggregate over benchmarks in the first place—our primary proposal is to avoid monolithic benchmarks and consider aggregation on a sample level, needing to tackle incomplete and heterogeneous measurements. We note that several other social-choice theory-based models such as score-based models (Shevchenko et al., 2024) based on the Condorcet-winner criterion (Young, 1988) have been proposed, yet they were primarily applied for aggregation on multi-task benchmarks, whereas a crucial component of our proposal is to break down the benchmark boundaries and aggregate heterogeneous samples.

**Dynamic Evaluation and Active Testing.** Some previous works like (Ji et al., 2021; Kossen et al., 2021, 2022; Saranathan et al., 2024; Huang et al., 2024; Zhu et al., 2023) tackle the 'active testing' problem, where the goal is to identify small "high-quality" test data-subsets, from a large pool of uncurated evaluation data. These works typically assume that the cost of unlabeled test data acquisition is low whereas the cost of acquiring per-instance labels is high. However, as pointed out by Prabhu et al. (2024), these assumptions are unrealistic for foundation models, as both the acquisition of test data and label annotations can be tedious in general. Hence, in our work, we tackle a broader problem: given a large testing data pool, how can we curate and query to produce a consistent and targeted set of model rankings?

**Efficient Evaluation.** As evaluation suites have grown, associated inference costs have also increased. Recent research has focused on creating compressed subsets of traditional benchmarks to address this issue (Varshney et al., 2022; Zhao et al., 2024; Perlitz et al., 2024; Kipnis et al., 2024; Pacchiardi et al., 2024). Popular approaches include subsampling benchmarks to preserve correlations with an external source like LMArena (Ni et al., 2024), sample clustering to gauge sample difficulty and then sub-sampling (Vivek et al., 2024), item-response-theory based methods for informatively sampling a subset of samples for evaluation (Polo et al., 2024), or designing evolving sample-level benchmarks (Prabhu et al., 2024). While the work of Prabhu et al. (2024) is similar to us in principle, it requires binary metrics as input and does not handle incomplete input matrices, which is necessary for aggregation over multiple time steps. We precisely address these limitations by showing efficient evaluation while accommodating incomplete data and extending it to ordinal ranks.

**Democratizing Evaluation.** Standard image classification and retrieval benchmarks are collected from platforms like Flickr, which are predominantly Western-centric (Ananthram et al., 2024; Shankar et al., 2017). This has raised the important question: "Progress for whom?", with many seminal works showcasing large disparities in model performance on concepts (Hemmat et al., 2024), tasks (Hall et al., 2024, 2023b,a), and even input samples (Pouget et al., 2024; Sureddy et al., 2024; Gustafson et al., 2024) from the Global South. In response, works have developed benchmarks tailored to diverse cultures and demographics to include their voice in measuring progress (Pistilli et al., 2024; Pouget et al., 2024; Nguyen et al., 2024; Luccioni and Rolnick, 2023). Further works have tried to create personalized, task-specific benchmarks for flexibly evaluating models based on user-preferences (Butt et al., 2024; Saxon et al., 2024; Yuan et al., 2024; Li et al., 2024c)— Zhang et al. (2024b) created Task-Me-Anything that enables users to input specific queries that then get processed to provide model rankings or responses to the query. However, their system is entirely procedurally generated, thereby not reflecting the real-world use-cases that models are typically subjected to in practice. Further, they are restricted to the fixed set of instances in their task generator pool. We take a different approach by creating flexible benchmarks where individuals, and contributing entities, can add their own samples and preferences collected from both real-world benchmarks and live model arenas like LM-Arena, thereby providing users with a realistic overview of model rankings on practical scenarios. Further, during capability testing, users can select similar preferences, making ONEBench more inclusive than traditional test sets.

## H Open Problems and Future Directions

In this section, we highlight some promising directions for improvement below:

1. Testing Limits and Scaling Up ONEBench: currently, our prototype comprises less than 100K samples in ONEBench-LLM and under 1M in ONEBench-LMM. These pools can be greatly expanded and diversified by expanding to incorporating *all existing* LLM and LMM benchmarks. Our retrieval mechanisms are designed to scale efficiently as the test pool grows in size and diversity.

2. Exploring Other Aggregation Algorithms: while we use the Plackett-Luce model for aggregating diverse measurements, there exist other algorithms from computational social choice theory with different trade-offs. A comprehensive evaluation of these alternatives could offer new insight for aggregating model performance.

3. Structured Querying and Enhanced Retrieval: One can improve retrieval by better querying mechanisms using models like ColBERT (Khattab and Zaharia, 2020) and ColPALI (Faysse et al., 2024), further optimized using DSPy (Khattab et al., 2023). A particularly interesting direction is allowing compositional queries, where users combine multiple queries to test behaviour in foundation models, similar to works like ConceptMix (Wu et al., 2024) and SkillMix (Yu et al., 2023).

4. On the Limits of Capability Probing: While we currently allow broad, open-ended inputs to probe capabilities, some are easier to assess than others (Madvil et al., 2023; Li et al., 2024b). As foundation models become more generalizable, a thorough analysis identifying which capabilities can be *easily, reliably evaluated*, which are *possible to evaluate but challenging*, and which are in principle *impossible to evaluate* is needed—this will help improve benchmarking effectiveness.