HUMAN BASELINES IN MODEL EVALUATIONS NEED RIGOR AND TRANSPARENCY

Kevin L. Wei* Harvard University
Patricia Paskov* Independent
Sunishchal Dev* Algoverse
Michael J. Byun* Independent

Anka Reuel Stanford University
Xavier Roberts-Gaal Harvard University
Rachel Calcott Harvard University

Evie Coxon Max Planck School of Cognition
Chinmay Deshpande Center for Democracy and Technology

ABSTRACT

This position paper argues that human baselines in foundation model evaluations must be more rigorous and more transparent to enable meaningful comparisons of human vs. AI performance. Human performance baselines are vital for the machine learning community, downstream users, and policymakers to interpret AI evaluations. Models are often claimed to achieve "super-human" performance, but existing baselining methods are neither sufficiently rigorous nor sufficiently well-documented to robustly measure and assess performance differences. Based on a meta-review of the measurement theory and AI evaluation literatures, we derive a framework for assessing human baselining methods. We then use our framework to systematically review 113 human baselines in foundation model evaluations, identifying shortcomings in existing baselining methods. We publish our framework as a reporting checklist for researchers conducting human baseline studies. We hope our work can advance more rigorous AI evaluation practices that can better serve both the research community and policymakers.[1]

1 INTRODUCTION

Artificial intelligence (AI) systems, and foundation models in particular, have increasingly achieved superior performance on benchmarks in natural language understanding, general reasoning, coding, and other domains (Maslej et al., 2024). These results are frequently compared to *human baselines*—reference sets of metrics intended to represent human performance on specific tasks—which has led to claims about models' "super-human" performance (Bikkasani, 2024).

Human baselines are crucial for evaluating AI systems and for understanding AI's societal impacts. For the machine learning (ML) research community, human baselines help improve benchmarks, provide context for interpreting system performance, and demonstrate concurrent validity (Hardy et al., 2024; Bowman & Dahl, 2021). For downstream users, comparisons to human performance may inform decisions about AI adoption (cf. Luo et al. 2019). And for policymakers, human baselines facilitate risk assessments (OSTP, 2022; NIST, 2023; Goemans et al., 2024; US AISI & UK AISI, 2024) and predictions of AI's economic impacts (Hatzius et al., 2023; Shrier et al., 2023). Valid and reliable human baselines thus contribute greatly to the operational value of AI evaluations.

However, despite widespread recognition in the ML community about the importance of human baselines (Reuel et al., 2024; Ibrahim et al., 2024; Tedeschi et al., 2023; Nangia & Bowman, 2019; Bender, 2015), existing human baselines are neither sufficiently rigorous nor sufficiently transparent to enable reliable claims about (the magnitude of) differences between human and AI performance. For instance, human baselines in many evaluations have small or biased samples (Liao et al., 2021;

*Equal contribution. Correspondence to: hi@kevinlwei.com
[1]Data is available at: https://github.com/kevinlwei/human-baselines

McIntosh et al., 2024), apply different instruments than those used in AI evaluation (Tedeschi et al., 2023), or fail to control for confounding variables (Cowley et al., 2022). In addition, published evaluations commonly omit study details necessary for assessing baseline validity, such as how participants were recruited or how questions were administered (Section 4.5). Measurement theory, a methodological field in the social sciences concerned with quantifying complex concepts, addresses analogous issues in human studies (Bandalos, 2018) and can inform best practices in human baselines.

Our position is that human baselines in evaluations of foundation models must be more rigorous and more transparent. Building from measurement theory, we propose a framework for assessing human baselines. Using our framework to systematically review 113 published human baselines, we find substantial shortcomings in existing human baselining methods. We hope that our framework can support researchers in developing human baselines that are more interpretable and valuable to the ML community, downstream users, and policymakers.

In defending our position, we acknowledge that there are challenges with building rigorous human baselines, such as the expense of high-quality baseline data, the evolving landscape of AI evaluations, and differences in cognition and interaction modes between humans and AI systems. *Given these complexities, our framework is not intended to be a one-size-fits-all prescription for all AI evaluations—rather, our work proposes a reference set of operational considerations to inform researchers in designing and implementing human baselines*.

We proceed as follows: Section 2 presents related work and background. Section 3 describes our methodology (full details in Appendix B). Section 4 presents our framework and the results of our systematic review, which examines the entire lifecycle of human baselining including baseline(r) design, recruitment, implementation, analysis, and documentation. Section 5 discusses results and limitations. Section 6 concludes.

2 BACKGROUND

Measurement theory is the discipline devoted to quantifying complex or unobservable *concepts* through the use of observable indicators, or *measurements* (Goertz, 2020). Since concepts are often multidimensional or impossible to measure directly, researchers usually aggregate multiple measurements and rely on proxies for quantities of interest. *Intelligence*, for instance, has sometimes been measured by aggregating multiple different cognitive tests (Deary, 2012). In the social sciences, measurement theory has also been applied to concepts such as fairness (Patty & Penn, 2019), emotion (Reisenzein & Junge, 2024), culture (Mohr & Ghaziani, 2014), personality (Drasgow et al., 2009), and language (Sassoon, 2010). Measurement theory helps build indicators for these concepts that satisfy criteria of *validity* (yielding results that support intended interpretations of measurements) and *reliability* (yielding consistent results across multiple measurements) (Bandalos, 2018).

There has been growing recognition in the AI research community that AI evaluation can draw valuable lessons from measurement theory and the social sciences (Chang et al., 2024b; Wallach et al., 2024; Eckman et al., 2025; Chouldechova et al., 2024; Blodgett et al., 2024; Xiao et al., 2023; Zhou et al., 2022; Zhao et al., 2025; Wang et al., 2023; Liao & Xiao, 2023). Like measurement theory, AI evaluation has been concerned with estimating concepts such as intelligence, fairness, emotion, and culture—though in AI models rather than in humans (Chang et al., 2024b). Research in ML has focused in particular on applying measurement theory to performance metrics (Subramonian et al., 2023; Flach, 2019) and fairness metrics (Jacobs et al., 2020; Grote, 2024; Blodgett et al., 2021). Additionally, measurement theory provides frameworks for making comparisons between (human) populations—analogous to the problem of comparing human and AI performance, which is often addressed using human baselines in AI evaluations.

We draw on measurement theory to examine human baselining in evaluations of *foundation models* (Bommasani et al., 2022), which pose unique evaluation challenges (Liao & Xiao, 2023). Applying measurement theory to the foundation model context is particularly appropriate as foundation models are exhibiting increasingly general, multidimensional capabilities (Zhong et al., 2024) and beginning to interact with the same interfaces as human users (Anthropic, 2024b; Chan et al., 2025). Specifically, we adopt the approach of Zhao et al. (2025) in drawing on measurement theory to validate the data generation process in human baselines—that is, we are particularly concerned with the validity and reliability of baselining *methods* rather than the trustworthiness of any particular baseline data.

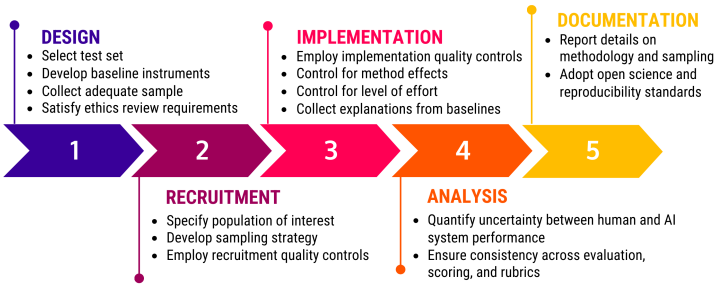
Analysis of the full pipeline of human baselining methods in the foundation model context is limited. Prior work has critiqued human annotation processes (Tedeschi et al., 2023) and offered high-level principles for human baselining (Cowley et al., 2022). Building on this literature, we provide a comprehensive discussion of operational-level methodological considerations for human baselining. We also fill a gap in the literature by systematically reviewing human baselines in foundation model evaluations, allowing us to identify shortcomings of and opportunities for improvement in existing methods for human baselining.

3 METHODOLOGY

We used a two-stage approach to develop our position, adapted from Zhao et al. (2025) and Reuel et al. (2024). First, we conducted a meta-review (review of reviews) of the measurement theory literature to construct a checklist of practices for baselining (Appendix A). Using purposive sampling and backwards snowballing, we identified 29 articles from the social sciences (psychology, economics, political science, education) and AI evaluation. The checklist was initially compiled after reviewing these articles, then refined through internal discussion and expert validation.

Second, we conducted a systematic review (see Page et al. 2021) of the AI evaluations literature to identify gaps in existing human baselining methods. From academic publications and gray literature, we identified 113 human baselines in foundation model evaluations. Inclusion criteria consisted of whether the article contained 1) an original human baseline, 2) an evaluation of a foundation model, and 3) both a human baseline-related keyword ("human baseline*", "expert baseline*", "human performance baseline*") and an AI evaluation-related keyword ("AI evaluation*", "ML benchmark*", etc.). Articles were then manually coded per the checklist from our meta-review (Appendix A), with the codebook iteratively refined during the coding process. Full methodological details are in Appendix B.

Figure 1: Our synthesized framework for human baselining. The full checklist can be found in Appendix A.



4 A FRAMEWORK FOR RIGOROUS AND TRANSPARENT HUMAN BASELINES

We organize our discussion by delineating five stages of the baselining process, as adapted from Reuel et al. (2024): design, recruitment, implementation, analysis, and documentation. For simplicity, we also synthesize our checklist (Appendix A) into an operational framework of key factors to consider at each stage of human baselining. Neither our checklist nor our framework is intended to provide hard recommendations but rather to inform researchers in developing baselines and reporting results. Our framework is presented below and summarized in Figure 1.

4.1 BASELINE DESIGN

Baseline design is the initial stage of human baseline development, at which researchers define baselines' purpose, scope, concepts, evaluation items, and metrics. (Reuel et al., 2024). We examine four considerations for baseline design.

Selecting the test set for a human baseline. Robust comparisons of human vs. AI performance require contrasting performance on the same test set. Because the cost of human baselines can make

baselining on a large number of items infeasible, researchers often construct baselines using subsets of the test sets used for AI evaluation. Baseline validity thus depends on the sampling strategy used to create the human baseline test set. Of the baselines in our review, 51% (58 of 113) used different test sets for AI evaluation and human baselines when presenting results.

Baseline test sets (where specified) were most commonly created using simple random (19% (21)), stratified (13% (15)), or purposeful sampling strategies (3% (3)). Simple random sampling from the broader evaluation dataset may be sufficient to ensure representativeness of the baseline test set, assuming the test set is sufficiently large (see Liao et al. 2021). Stratified sampling may be preferred where the baseline test set is relatively small, or where the test set must preserve important properties of the evaluation dataset such as data source (e.g., Xiang et al. (2023)), question difficulty (Tedeschi et al., 2023), or other relevant dimensions (Cowley et al. 2022; e.g., Liu et al. 2024b, Bai et al. 2024).

When reporting human baseline results, researchers should clearly indicate where human baseline test sets differ from AI evaluation test sets. To directly compare AI results with human baselines, researchers can also report AI performance on only the human baseline test set.

Using an iterative processes to develop baseline instruments. Iterative processes repeatedly test and refine the instruments (e.g., forms, surveys, prompts) used to measure constructs by applying multiple rounds of validation, feedback, and refinement of baseline instruments before data collection. 34% (38) of baselines reviewed used iterative methods during baseline design.

The feedback loops created by iteration can support construct validity (Rosellini & Brown, 2021) while ensuring clarity and consistent interpretation of instruments (Cheng et al., 2024; Cowley et al., 2022). In the social sciences, iterative processes are the gold standard in the social sciences for collecting annotations (Cheng et al., 2024), running surveys (Groves et al., 2011), and building clinical questionnaires (Rosellini & Brown, 2021). The ML community has also recognized the importance of iteration, such as when optimizing AI prompts (Hewing & Leinhos, 2024; Gao et al., 2025). Researchers often validate items in evaluation *datasets* (e.g., Nangia et al. 2021; Rein et al. 2024) but less frequently validate baseline *instruments*. An evaluation that optimizes AI prompts and validates evaluation items, but that does not validate baselining instruments, may unfairly disadvantage humans and thereby discount baseline validity.

Iteration does add complexity to the baselining process. However, it is not necessarily costly: large pilot studies and focus groups may be out of reach to budget-constrained researchers, but small-scale pre-tests or expert validation can still help improve baselining artifacts (Groves et al., 2011; Zickar, 2020).

Collecting an adequately sized sample of baseliners.[2] Baselines that are underpowered because of small sample sizes are unreliable because they cannot robustly capture the underlying distribution of human performance (Cao et al., 2024). Power analyses can help determine an appropriate sample size for human baselines, given significance levels and pre-specified minimum detectable effect sizes in the outcome metric of interest (Cohen, 2013). The importance of statistical power in ML benchmarks has been noted in prior work (Bowman & Dahl, 2021; Grosse-Holz & Jorgensen, 2024), but only 2% (2) of the human baselines we reviewed reported conducting power analyses.

If sample sizes are fixed (e.g., due to cost constraints), researchers can nevertheless calculate and report the required sample size to reliably detect practically important effects, which supports interpretation of evaluation results. Understanding the ability of human baselines to detect performance differences may be especially important to users and policymakers, who may demand added rigor and certainty in evaluation results to inform decision-making (Paskov et al., 2024). In general, considerations around statistical power reflect broader shortcomings in using statistical methods in AI evaluation, which we discuss further in Section 4.4.

Satisfying ethics review requirements for human subjects research. Ethics review protects human research participants (Page & Nyeboer, 2017), and it is legally required in many jurisdictions, including in the U.S. (U.S. Department of Homeland Security et al., 2017). Significantly, only 13% (15) of the articles we examined reported compliance with or formal exemption from ethics review requirements, near the same order of magnitude as the 2% found by Kaushik et al. (2024).

[2]"Sample" in this context refers to the subset of humans in the baseline who are drawn from an underlying population.

Reporting compliance with ethics requirements is best practice in many fields, e.g., medicine (ICMJE, 2025). Failure to report compliance in an article does not indicate failure in compliance, and some evaluations may be exempt from review (Kaushik et al., 2024). However, transparency around research ethics becomes more important as public interest in AI increases, and protection of research participants can also be critical for evaluations implicating, for example, deception, misinformation, and psychological impacts.

4.2 BASELINER RECRUITMENT

Baseliner recruitment is the stage at which human baseliners—the humans who respond to evaluation items—are found and are engaged to participate in a baseline. We examine three considerations for baseliner recruitment.

Specifying a human population of interest. Specifying a population of interest—the group of humans for whom a baseline is intended to be representative—is a cornerstone of valid human subjects research. Prior work has noted that AI evaluations rarely specify populations of interest (Subramonian et al., 2023), which is in line our review: only 31% (35 of 113) baselines explicitly or implicitly defined a population of interest along at least one axis (i.e., a population beyond "humans").

Populations of interest can be specified through axes such as geographic location, demographic characteristics (e.g., age, gender, socioeconomic status), language, cultural background, education, or domain expertise. A human baseline may seek to measure the performance of, for instance, a population of medical or legal professionals (e.g., Blinov et al. 2022; Hijazi et al. 2024). In published AI evaluations that defined a population of interest, the most commonly reported characteristics were education (20% (23 of 113)), expertise (19% (22)), language (18% (20)), and age (18% (20)). How to scope the population of interest for any given baseline will depend on the evaluation's research questions, context, and intended use.

Developing a sampling/recruitment strategy for selecting baseliners. Sampling strategy is the process by which humans are selected to participate in the baseline, such as convenience sampling or random sampling. Sampling strategy directly informs how representative the selected baseliners are of the population of interest, and representativeness is essential for external validity because it determines whether baseliners' results can be generalized to a larger population of humans (Findley et al., 2021; Stantcheva, 2023; Berinsky, 2017). Of the baselines in our review, 32% (36) used a convenience sample, 31% (35) recruited from crowdsourcing platforms, and the remainder did not report their sampling strategies.

Both convenience sampling and crowdsourced samples present sampling challenges that are rarely addressed in the AI evaluation literature. Convenience samples, such as samples of undergraduates or of an article's authors, are highly susceptible to significant biases that reduce their applicability to broader populations (cf. Mihalcea et al. 2024; Diaz & Smith 2024). Recruitment of baseliners from crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) or Prolific, also poses challenges to representativeness. Prior research has shown that MTurk workers tend to be younger, more educated, more politically liberal, and more engaged online compared to the general population (Sheehan, 2018; Shaw & Hargittai, 2021; Stantcheva, 2023). Crowdsourced baselines may thus fail to represent performance of humans who do not reflect those demographics. Even if crowdsourced samples are sufficiently large from the perspective of statistical power, results may still be biased if the sample is insufficiently representative of the underlying population of interest (Bradley et al., 2021).

Researchers designing human baselines should consider methodological adjustments to ensure baseliner representativeness. For instance, researchers could consider stratified sampling to improve representativeness along specific dimensions (Groves et al., 2011). Post hoc adjustments such as weighting may partially mitigate selection bias in non-representative samples (Solon et al., 2015). Researchers may also explore non-probability sampling approaches that aim to enhance representativeness (Couper, 2017).

At a minimum, researchers should clearly report the sampling strategy used, acknowledge the limitations of non-representative samples, discuss which populations results may apply to, and discuss implications for the validity of baseline results.

Employing quality controls for baseliner recruitment. Quality control (QC) mechanisms at the recruitment stage improve data quality by selecting for baseliners who can generate high-caliber data. Using inclusion/exclusion criteria during recruitment is considered best practice in survey research (Stantcheva, 2023) and ensures that baseliners meet appropriate evaluation criteria. QC mechanisms can include pre-testing baseliners for task-specific knowledge or general ability (see, e.g., Nangia et al. 2021) or filtering crowdsourced workers via screening questions or platform quality scores (Lu et al., 2022a).

Depending on the research question, baseliners' domain expertise may be particularly important because expert baseliners often provide higher-quality data than non-experts (Cheng et al., 2024; Liao et al., 2021). Specialized (expert) human baselines are specifically needed to enable AI evaluations that compare AI performance with the ceiling of possible human performance and that explore the possibility of "super-human" performance (e.g., Glazer et al. 2024).

Considerations for QC in recruitment include the cost and feasibility of recruiting (expert) baseliners, whether evaluations items require different QC criteria or expertise (cf. Weidinger et al. 2024), and how to establish criteria for assessing baseliners (e.g., assessing domain expertise in highly-specialized evaluations). Researchers may also wish to exclude baseliners who have been previously exposed to evaluation items to prevent data contamination, analogous to AI train/test contamination.

4.3 Baseline Implementation

Baseline implementation is the stage at which human baseline data is collected—e.g., through surveys or crowdwork platforms. We examine four considerations for baseline implementation.

Employing quality controls in baseline implementation. Quality control mechanisms at the implementation stage improve data quality by filtering out unreliable baseline responses. As with the recruitment stage, QC during implementation is considered best practice in survey research; mechanisms include checks for attention, consistency, response pattern, outliers, and time to completion (Stantcheva, 2023; Lebrun et al., 2024). Some research has also demonstrated that attention checks may improve the representativeness of crowdwork samples (Qureshi et al., 2022). Of the baselines in our review, 23% (26 of 113) reported performing QC at the implementation stage, most often using attention checks and honeypot questions.

One issue of increasing importance is the inappropriate usage of AI tools by crowdworkers, which was directly raised as a concern in one article we reviewed (Sprague et al., 2023). Empirical work has suggested that a substantial percentage of MTurk workers have used AI to complete tasks (Veselovsky et al., 2023b; Traylor, 2025), which can decrease data quality (Lebrun et al., 2024) and baseline validity. Unintentional usage of AI tools may also occur as AI adoption increases such as via AI-generated Google Search summaries. QC mechanisms to prevent AI usage may be beneficial for crowdsourced baselines, e.g., explicitly asking participants not to use LLMs (Veselovsky et al., 2023a), employing technical restrictions such as preventing copy/pasting (Veselovsky et al., 2023a), using non-standard interface elements (Gureckis, 2021), or using comprehension and manipulation checks (Frank et al., 2025). That said, AI use is not always undesirable, such as in domains where AI usage is expected, in which evaluations can compare AI capabilities with baselines of AI-augmented human capabilities (e.g., Wijk et al. 2024).

Controlling for method effects. Method effects are variations in item response attributable to data collection methods rather than to differences in underlying response distributions, and they can reduce evaluations' internal validity (Davidov et al., 2014). Our review found significant discrepancies in data collection methods between humans and AI models: 35% (40) of baselines displayed UI differences between human baselining and AI evaluation, 19% (22) displayed differences in instructions or prompts, and 7% (9) displayed differences in tool access. Note that our focus was on method effects between human and AI responses, but method effects can also occur between human baseliners (e.g., if baselines are collected from multiple platforms).

Method effects are well-documented in the social sciences, particularly in psychology and in survey methodology. Empirical research has found effects due to the mode of survey administration (Vannieuwenhuyze et al., 2010; Shin et al., 2012), question order (Engel et al., 2014), fatigue from survey length (Stantcheva, 2023), example responses provided (Eckman et al., 2025; Lu et al., 2022b), interface design (Sanchez, 1992), and question wording (Wu & Quinn, 2017; Dafoe et al., 2018).

Measurement theory offers some guidance for addressing method effects in humans. For instance, randomization of non-critical methodological details can reduce some effects (e.g., reducing order effects by randomizing question order). Fatigue can also be addressed by shortening survey length, encouraging breaks or enforcing time limits, and implementing attention checks. However, not all such guidance is applicable to AI systems.

Some method effects in AI evaluation are currently unavoidable due to differences between human and AI cognition. For instance, many AI evaluations restart the context window for each run, but it may be unrealistic to demand that baseliners are only administered one item per sitting; only 27% (30) of reviewed baselines reported instrument length, of which most reported instruments (24) were longer than one item. Similarly, although both AI systems and humans are known to be sensitive to item wording, they are sensitive in different ways (Tjuatja et al., 2024), suggesting that simply using the same data collection artifacts for humans and for AI systems may not prevent method effects.

Overall, significant additional research is needed to understand how method effects differ between humans and AI systems (and between AI systems), as well as how AI evaluations can adjust for these differences so as not to unfairly advantage humans or AI models in the evaluation process (Cowley et al., 2022; Tedeschi et al., 2023). In the absence of clear evidence, researchers can contribute to this area of research by clearly documenting evaluation methodologies.

Controlling for level of effort. Both humans' and AI systems' level of effort in responding to items can affect evaluation results, and baseliner effort can in turn be affected by training, compensation, and other factors (Tedeschi et al., 2023). Training could include tutorials, response guides, or example items; compensation structures can also affect baseline data quality (Grosse-Holz & Jorgensen, 2024) and can vary by, e.g., payment by hour vs. per task or by performance bonuses. Our review found that 22% (25) of baselines provided training to baseliners, and 39% (44) paid baseliners, with 6% (7) providing performance bonuses.

Accounting for level of effort also raises design questions about the relevant experimental unit of interest, the choice of which affects evaluations' external validity (Jackson & Cox, 2013). Most AI evaluations take humans or AI systems as the experimental unit, but some comparisons may necessitate more granularity. For instance, Wijk et al. (2024) compares performance after two human labor-hours and after two AI labor-hours. Properly scoping experimental units could make evaluations more valuable for understanding AI's broader societal effects, e.g., by enabling comparisons of labor efficiency.

Collecting explanations from baseliners. In some instances, researchers may find value in explanations of why baseliners chose particular responses. Collecting explanations is generally a best practice in survey research, as it can help surface new insights about data (Lu et al., 2022a). For AI evaluations, explanations can reveal qualitative differences between human and AI performance, explain performance gaps, and surface validity issues. Explanations may also be used for quality control, validation, and understanding the thought processes behind item responses (Lu et al., 2022a; Tedeschi et al., 2023), which may lead to improvements in questions or instrumentation. Only 10% (11) of the baselines we reviewed collected explanations from baseliners, though this finding is unsurprising since collecting explanations may increase the cost of human baselines, and it is unclear whether explanations are necessary in many baselines.

4.4 Baseline Analysis

Baseline analysis is the stage after data collection at which human baseline data is inspected and compared to AI results. We examine two considerations at the analysis stage.

Quantifying uncertainty in human vs. AI performance differences. Reporting measurements of uncertainty or applying statistical tests is necessary to rigorously assess whether measurements of performance truly reflect underlying performance distributions, as well as to interpreting evaluation results (Agarwal et al., 2022; Steinbach et al., 2022). The ML community has historically recognized these norms (e.g., Dietterich 1998; Bouckaert & Frank 2004), but many recent evaluations of large AI models have not met standards of statistical rigor (Biderman & Scheirer, 2020; Agarwal et al., 2022; Welty et al., 2019; Paskov et al., 2024). Similarly, our review finds that only 26% (29 of 113) of evaluations provided interval or distribution estimates, and only 9% (10) performed statistical tests of any type.

Lack of statistical testing is sometimes understandable given small sample sizes and other evaluation limitations (cf. Bouthillier et al. 2021). Reporting interval estimates, however, has become increasingly accessible with increased guidance (e.g., Miller 2024) and support in major evaluation frameworks (e.g., UK AISI 2025). The ML research community could also draw on established statistical methods for analyzing small samples (Neuhäuser & Ruxton, 2024; Hoyle, 1999; Schoot & Miočević, 2020). Finally, in line with recent commentary in statistics, researchers should consider reporting results of statistical tests (p-values) as one component of evidence used to judge the evaluation results, rather than as screens for statistical significance (McShane et al., 2019; Gelman & Stern, 2006).

Using consistent evaluation metrics, scoring methods, and rubrics across human and AI evaluation. Often, comparisons between AI and human baseline results are fair only when the metrics for comparison are equivalent across samples. For instance, human baseline metrics are sometimes calculated inconsistently across items, complicating baseline interpretation (Tedeschi et al., 2023). In our review, 89% (100) of baselines used the same evaluation metrics across human and AI results, but only 59% (67) and 63% (71) used the same scoring rubric and scoring method. Most commonly, researchers used majority vote for human but not for AI samples. Although these comparisons are not always inappropriate, researchers should consider adding clarifying language when reporting results, e.g., "AI evaluation metrics fell below majority-vote human performance" or "model results on each item exceeded the maximum performance across ten human baseliners."

4.5 Baseline Documentation

Baseline documentation is the provision of evaluation tasks, datasets, metrics, and experimental materials and resources (Reuel et al., 2024). We examine two considerations for baseline documentation.

Reporting key details about baselining methodology and baseliners. Documentation includes reporting information about baseliners, baselining procedures, and baseline paradata. These details can significantly affect how results are contextualized and interpreted—especially with respect to their validity—and reporting can build collective confidence in published results (Liao et al., 2021).

We believe that absent compelling reasons for confidentiality, researchers should strongly consider documenting most items in our checklist related to baseline(r) design, recruitment, implementation, and analysis (Appendix A). Researchers should also consider reporting baseliner demographics, paradata, and other study information. Baseline demographics can enable assessments of baseliner sample representativeness and reliability; paradata such as items' time to completion can offer insights into latent variables like cognitive effort (Cai et al., 2016) and into data quality, which is often correlated with response times (Traylor, 2025). Of the baselines in our review, all failed to report at least some items on our checklist, only 20% (23 of 113) provided detailed baseliner demographics, and only 24% (27) included paradata such as response times.

Adopting best practices for open science and reproducibility/replicability. Releasing human baseline data, experimental materials (e.g., forms, custom UIs), and analysis code in accessible repositories (e.g., GitHub, OSF) can facilitate research validation and reproduction/replication (Semmelrock et al., 2024; Stodden & Miguez, 2014). These open science practices also facilitate reuse of human baseline data in subsequent evaluations by other researchers, which can in turn foster more efficient use of resources within the ML community. Concerns around open science and replicability are not new in ML (Kapoor et al., 2024; Pineau et al., 2021), and our review found that most baselines (80% (90)) did not publicly release human baseline responses, experimental materials (57% (64)), and code for analyzing human baselines (52% (59)).

5 Discussion

In this section, we discuss three additional considerations for human baselines and address the limitations of our study. We discuss and respond to alternative views in Appendix D.

First, human baselines are not appropriate for all AI evaluations. Most prominently, human baselines are not meaningful for evaluations of AI tasks without human equivalents (Barnett & Thiergart, 2024). Examples include AI control evaluations, which measure an AI system's ability to monitor a more advanced AI system (Greenblatt et al., 2024), and autonomous self-replication

evaluations, which measure an AI agent's ability to create copies of itself (Pan et al., 2024). In contrast, human baselines can be valuable for evaluations that measure AI performance in domains with human equivalents, including but not limited to many question-answer benchmarks and task-based agent evaluations (e.g. Wijk et al. 2024).

Second, human baselines may also be constructed from secondary sources. Our position paper focuses on primary data collection methods in human baselining, but human performance metrics can also be derived from real-world data or pre-existing datasets. For instance, the Massive Multitask Language Understanding dataset uses the 95th percentile of human standardized test scores as a point of comparison with AI results (Hendrycks et al., 2021a); other studies (e.g., Hua et al. 2024) use human subjects data from previous work (Lewis et al., 2017). Re-use of human baselines highlights the need for transparency and documentation of baselining methods: authors should assume their datasets may be re-used by other researchers, who require significant methodological detail to design effective evaluations and draw meaningful conclusions from results. Secondary data is also subject to many other limitations (see Appendix D).

Third, human baselines can vary over time and as technology advances. Human capabilities are known to change over time (Trahan et al., 2014), and the half-life of AI-augmented human baselines (e.g., Wijk et al. 2024) may be particularly short due to the rate of progress in AI. These trends suggest that human baselines should be understood to represent measurements at specific points in time, and researchers should tread carefully when making comparisons to older human baselines. In this vein, the ML community can consider implementing regularly updated "living" baselines, analogous to how public opinion polls are regularly repeated to track variation over time. Open science practices would enhance replicability and resource efficiency for such living baselines.

Finally, we acknowledge several limitations to our work. Our methodology has generally followed best practices for systematic literature reviews, but our meta-review sample was collected purposively and could be biased as a result (see discussion in Appendix B.1). Our scope is limited to methodological considerations specific to human baselines, so we do not discuss many important aspects of AI evaluation methodology such as construct validity (Strauss & Smith, 2009). We also restricted our scope to foundation model evaluations; although we believe much of our framework is applicable to the broader research community, human baselines for evaluating other AI models may raise new methodological questions. Finally, future research can examine applications of measurement theory to human evaluation and human-AI interaction studies (e.g. human uplift), which are not explored in this position paper.

6 Conclusion

In this position paper, we argued that human baselines in foundation model evaluations should be more rigorous and transparent. Systematically reviewing 113 published evaluations, we find that many baselines lack methodological rigor across the gamut of the baselining process, from design (e.g., test set selection) to documentation (e.g., lack of study details). We provide an instructional framework for human baselining based on measurement theory. Our approach fosters validity and reliability in human baselines, enabling meaningful comparisons of human vs. AI performance and promoting research transparency. We hope that our work can guide researchers in improving baselining methods and evaluating AI systems.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically with regards to improving the quality of methods used to create and analyze human baselines in AI evaluation. We hope that by discussing methodological considerations in human baselining—and by highlighting shortcomings in existing baselining methods—our work will lead to rigorous AI evaluations that can be useful to not just the research community but also to users of AI systems and policymakers. We discuss broader implications of human baselines in Section 1, and we do not anticipate any particular negative impacts associated with our work.

REFERENCES

Almas Abdibayev, Allen Riddell, and Daniel Rockmore. BPoMP: The Benchmark of Poetic Minimal Pairs – Limericks, Rhyme, and Narrative Coherence. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 1–9, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.1/.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice, January 2022. URL http://arxiv.org/abs/2108.13264. arXiv:2108.13264.

Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13921–13937, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.828. URL https://aclanthology.org/2024.findings-acl.828/.

Joshua Albrecht, Ellie Kitanidis, and Abraham J. Fetterman. Despite "super-human" performance, current LLMs are unsuited for decisions about ethics and safety, December 2022. URL http://arxiv.org/abs/2212.06295. arXiv:2212.06295 [cs].

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. RAFT: A Real-World Few-Shot Text Classification Benchmark. In *Advances in Neural Information Processing Systems*, volume 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/ca46c1b9512a7a8315fa3c5a946e8265-Abstract-round2.html.

Heather Ames, Claire Glenton, and Simon Lewin. Purposive sampling in a qualitative evidence synthesis: a worked example from a synthesis on parental perceptions of vaccination communication. *BMC Medical Research Methodology*, 19(1):26, January 2019. ISSN 1471-2288. doi: 10.1186/s12874-019-0665-4. URL https://doi.org/10.1186/s12874-019-0665-4.

Annual Reviews. Annual Reviews, 2025a. URL https://www.annualreviews.org/. Publisher: Annual Reviews.

Annual Reviews. Journal Impact Factors, 2025b. URL https://perma.cc/S932-227W.

Anthropic. Claude 3.5 Sonnet Model Card Addendum, June 2024a. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Anthropic. Developing a computer use model, October 2024b. URL https://www.anthropic.com/news/developing-computer-use.

Daiki Asami and Saku Sugawara. PROPRES: Investigating the Projectivity of Presupposition with Various Triggers and Environments. In Jing Jiang, David Reitter, and Shumin Deng (eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 122–137, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-1.9. URL https://aclanthology.org/2023.conll-1.9/.

Mercy Asiedu, Nenad Tomasev, Chintan Ghate, Tiya Tiyasirichokchai, Awa Dieng, Oluwatosin Akande, Geoffrey Siwo, Steve Adudans, Sylvanus Aitkins, Odianosen Ehiakhamen, Eric Ndombi, and Katherine Heller. Contextual Evaluation of Large Language Models for Classifying Tropical and Infectious Diseases, January 2025. URL http://arxiv.org/abs/2409.09201. arXiv:2409.09201 [cs].

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual Minimal-Change Understanding, January 2025. URL http://arxiv.org/abs/2407.16772. arXiv:2407.16772 [cs].

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. The Power of Many: Multi-Agent Multimodal Models for Cultural Image Captioning, November 2024. URL http://arxiv.org/abs/2411.11758. arXiv:2411.11758 [cs].

Deborah L. Bandalos. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, January 2018. ISBN 978-1-4625-3213-1. Google-Books-ID: caxCDwAAQBAJ.

Peter Barnett and Lisa Thiergart. Declare and Justify: Explicit assumptions in AI evaluations are necessary for effective regulation, November 2024. URL http://arxiv.org/abs/2411.12820. arXiv:2411.12820 [cs].

David Bender. Establishing a Human Baseline for the Winograd Schema Challenge. In *Proceedings of the 26th Modern AI and Cognitive Science Conference*, volume Vol 1353, pp. 36–45. CEUR Workshop Proceedings, April 2015. URL https://ceur-ws.org/Vol-1353/paper_30.pdf.

Adam J. Berinsky. Measuring Public Opinion with Surveys. *Annual Review of Political Science*, 20 (Volume 20, 2017):309–329, May 2017. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-101513-113724. URL https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-101513-113724. Publisher: Annual Reviews.

Stella Biderman and Walter J. Scheirer. Pitfalls in Machine Learning Research: Reexamining the Development Cycle. In *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, pp. 106–117. PMLR, February 2020. URL https://proceedings.mlr.press/v137/biderman20a.html. ISSN: 2640-3498.

Dileesh Chandra Bikkasani. Navigating artificial general intelligence (AGI): societal implications, ethical considerations, and governance strategies. *AI and Ethics*, pp. 1–16, December 2024. ISSN 2730-5961. doi: 10.1007/s43681-024-00642-z. URL https://link.springer.com/article/10.1007/s43681-024-00642-z. Company: Springer Distributor: Springer Institution: Springer Label: Springer Publisher: Springer International Publishing.

Pavel Blinov, Arina Reshetnikova, Aleksandr Nesterov, Galina Zubkova, and Vladimir Kokh. RuMedBench: A Russian Medical Language Understanding Benchmark. In Martin Michalowski, Syed Sibte Raza Abidi, and Samina Abidi (eds.), *Artificial Intelligence in Medicine*, pp. 383–392, Cham, 2022. Springer International Publishing. ISBN 978-3-031-09342-5. doi: 10.1007/978-3-031-09342-5_38.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.81. URL https://aclanthology.org/2021.acl-long.81/.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. Human-Centered Evaluation of Language Technologies. In Jessy Li and Fei Liu (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 39–43, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-tutorials.6. URL https://aclanthology.org/2024.emnlp-tutorials.6/.

Martin Boeker, Werner Vach, and Edith Motschall. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Medical Research Methodology*, 13(1):1–12, December 2013. ISSN 1471-2288. doi: 10.1186/1471-2288-13-131. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-131. Number: 1 Publisher: BioMed Central.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano

Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Remco R. Bouckaert and Eibe Frank. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 3–12, Berlin, Heidelberg, 2004. Springer. ISBN 978-3-540-24775-3. doi: 10.1007/978-3-540-24775-3_3.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769, March 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.

Samuel R. Bowman and George Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385.

Valerie C. Bradley, Shiro Kuriwaki, Michael Isakov, Dino Sejdinovic, Xiao-Li Meng, and Seth Flaxman. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890):695–700, December 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04198-4. URL https://www.nature.com/articles/s41586-021-04198-4. Publisher: Nature Publishing Group.

Fan Bu, Yuhao Zhang, Xidong Wang, Benyou Wang, Qun Liu, and Haizhou Li. Roadmap towards Superhuman Speech Understanding using Large Language Models, October 2024. URL http://arxiv.org/abs/2410.13268. arXiv:2410.13268 [cs].

Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. Item Response Theory. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):297–321, June 2016. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-041715-033702. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-041715-033702. Publisher: Annual Reviews.

Ying Cao, Ronald C Chen, and Aaron J Katz. Why is a small sample size not enough? *The Oncologist*, 29(9):761–763, September 2024. ISSN 1083-7159. doi: 10.1093/oncolo/oyae162. URL https://doi.org/10.1093/oncolo/oyae162.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea. FIBER: Fill-in-the-Blanks as a Challenging Video Understanding

Evaluation Framework, March 2022. URL http://arxiv.org/abs/2104.04182. arXiv:2104.04182 [cs].

Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for AI Agents, January 2025. URL http://arxiv.org/abs/2501.10114. arXiv:2501.10114 [cs] version: 1.

Hua-Hua Chang, Chun Wang, and Susu Zhang. Statistical Applications in Educational Measurement. *Annual Review of Statistics and Its Application*, 8(Volume 8, 2021):439–461, March 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-104044. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-104044. Publisher: Annual Reviews.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL https://aclanthology.org/2023.emnlp-main.453/.

Matthew Chang, Gunjan Chhablani, Alexander Clegg, Mikael Dallaire Cote, Ruta Desai, Michal Hlavac, Vladimir Karashchuk, Jacob Krantz, Roozbeh Mottaghi, Priyam Parashar, Siddharth Patki, Ishita Prasad, Xavier Puig, Akshara Rai, Ram Ramrakhya, Daniel Tran, Joanne Truong, John M. Turner, Eric Undersander, and Tsung-Yen Yang. PARTNR: A Benchmark for Planning and Reasoning in Embodied Multi-agent Tasks, October 2024a. URL http://arxiv.org/abs/2411.00081. arXiv:2411.00081 [cs].

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45, March 2024b. ISSN 2157-6904. doi: 10.1145/3641289. URL https://dl.acm.org/doi/10.1145/3641289.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking Theory of Mind in Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.847. URL https://aclanthology.org/2024.acl-long.847/.

Xiang Cheng, Raveesh Mayya, and João Sedoc. From Human Annotation to LLMs: SILICON Annotation Workflow for Management Research, December 2024. URL http://arxiv.org/abs/2412.14461. arXiv:2412.14461 [cs].

Cyril Chhun, Fabian M. Suchanek, and Chloé Clavel. Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation. *Transactions of the Association for Computational Linguistics*, 12:1122–1142, September 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00689. URL https://doi.org/10.1162/tacl_a_00689.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: a Robust, Diverse and Challenging Benchmark on Measuring the (Lack of) Cultural Knowledge of LLMs, October 2024. URL http://arxiv.org/abs/2410.02677. arXiv:2410.02677 [cs].

Javier Chiyah-Garcia, Alessandro Suglia, and Arash Eshghi. Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11523–11542, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.643. URL https://aclanthology.org/2024.emnlp-main.643/.

Alexandra Chouldechova, Chad Atalla, Solon Barocas, A. Feder Cooper, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Matthew Vogel, Hannah Washington, and Hanna Wallach. A Shared Standard for Valid Measurement of Generative AI Systems' Capabilities, Risks, and Impacts, December 2024. URL http://arxiv.org/abs/2412.01934. arXiv:2412.01934 [cs].

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York, 2 edition, May 2013. ISBN 978-0-203-77158-7. doi: 10.4324/9780203771587.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents, July 2024. URL http://arxiv.org/abs/2406.06613. arXiv:2406.06613 [cs].

Mick P. Couper. New Developments in Survey Data Collection. *Annual Review of Sociology*, 43 (Volume 43, 2017):121–145, July 2017. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev-soc-060116-053613. URL https://www.annualreviews.org/content/journals/10.1146/annurev-soc-060116-053613. Publisher: Annual Reviews.

Hannah P. Cowley, Mandy Natter, Karla Gray-Roncal, Rebecca E. Rhodes, Erik C. Johnson, Nathan Drenkow, Timothy M. Shead, Frances S. Chance, Brock Wester, and William Gray-Roncal. A framework for rigorous evaluation of human performance in human and machine learning comparison studies. *Scientific Reports*, 12(1):5444, March 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-08078-3. URL https://www.nature.com/articles/s41598-022-08078-3. Publisher: Nature Publishing Group.

Allan Dafoe, Baobao Zhang, and Devin Caughey. Information Equivalence in Survey Experiments. *Political Analysis*, 26(4):399–416, October 2018. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2018.9. URL https://www.cambridge.org/core/journals/political-analysis/article/information-equivalence-in-survey-experiments/8D134C6387CD7D845249B0712775AB79.

Rishit Dagli, Guillaume Berger, Joanna Materzynska, Ingo Bax, and Roland Memisevic. AirLetters: An Open Video Dataset of Characters Drawn in the Air, October 2024. URL http://arxiv.org/abs/2410.02921. arXiv:2410.02921 [cs].

Eldad Davidov, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, 40(Volume 40, 2014):55–75, July 2014. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev-soc-071913-043137. URL https://www.annualreviews.org/content/journals/10.1146/annurev-soc-071913-043137. Publisher: Annual Reviews.

Ian J. Deary. Intelligence. *Annual Review of Psychology*, 63(Volume 63, 2012):453–482, January 2012. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-120710-100353. URL https://www-annualreviews-org.ezp-prod1.hul.harvard.edu/content/journals/10.1146/annurev-psych-120710-100353. Publisher: Annual Reviews.

Mark Diaz and Angela D. R. Smith. What Makes An Expert? Reviewing How ML Researchers Define "Expert". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):358–370, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31642. URL https://ojs.aaai.org/index.php/AIES/article/view/31642. Number: 1.

Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, October 1998. ISSN 0899-7667. doi: 10.1162/089976698300017197. URL https://ieeexplore.ieee.org/document/6790639. Conference Name: Neural Computation.

P. Alex Dow, Jennifer Wortman Vaughan, Solon Barocas, Chad Atalla, Alexandra Chouldechova, and Hanna Wallach. Dimensions of Generative AI Evaluation Design, November 2024. URL http://arxiv.org/abs/2411.12709. arXiv:2411.12709 [cs].

Fritz Drasgow, Oleksandr S. Chernyshenko, and Stephen Stark. Test theory and personality measurement. In *Oxford handbook of personality assessment*, Oxford library of psychology, pp. 59–80. Oxford University Press, New York, NY, US, 2009. ISBN 978-0-19-536687-7. doi: 10.1093/oxfordhb/9780195366877.013.0004.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. PIP: Physical Interaction Prediction via Mental Simulation with Span Selection. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 405–421, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19833-5. doi: 10.1007/978-3-031-19833-5_24.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. Position: insights from survey methodology can improve training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 12268–12283, Vienna, Austria, January 2025. JMLR.org.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models, August 2023. URL http://arxiv.org/abs/2303.10130. arXiv:2303.10130.

Uwe Engel, Ben Jann, Peter Lynn, Annette Scherpenzeel, and Patrick Sturgis (eds.). *Improving Survey Methods: Lessons from Recent Research*. Routledge, New York, September 2014. ISBN 978-1-315-75628-8. doi: 10.4324/9781315756288.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. MERA: A Comprehensive LLM Evaluation in Russian. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9920–9948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.534. URL https://aclanthology.org/2024.acl-long.534/.

Michael G. Findley, Kyosuke Kikuta, and Michael Denly. External Validity. *Annual Review of Political Science*, 24(Volume 24, 2021):365–393, May 2021. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-041719-102556. URL https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-041719-102556. Publisher: Annual Reviews.

Peter Flach. Performance Evaluation in Machine Learning: The Good, the Bad, the Ugly, and the Way Forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9808–9814, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33019808. URL https://ojs.aaai.org/index.php/AAAI/article/view/5055. Number: 01.

Michael C. Frank, Mika Braginsky, Julia Cachia, Nicholas Coles, and Tom E. Hardwicke. *Experimentology: An Open Science Approach to Experimental Psychology Methods*. The MIT Press, Cambridge, Massachusetts, January 2025. ISBN 978-0-262-55256-1. URL https://experimentology.io/.

Shea Fyffe, Philseok Lee, and Seth Kaplan. "Transforming" Personality Scale Development: Illustrating the Potential of State-of-the-Art Natural Language Processing. *Organizational Research Methods*, 27(2):265–300, April 2024. ISSN 1094-4281. doi: 10.1177/10944281231155771. URL https://doi.org/10.1177/10944281231155771. Publisher: SAGE Publications Inc.

Shuzheng Gao, Chaozheng Wang, Cuiyun Gao, Xiaoqian Jiao, Chun Yong Chong, Shan Gao, and Michael Lyu. The Prompt Alchemist: Automated LLM-Tailored Prompt Optimization for Test Case Generation, January 2025. URL http://arxiv.org/abs/2501.01329. arXiv:2501.01329 [cs].

Andrew Gelman and Hal Stern. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, 60(4):328–331, November 2006. ISSN 0003-1305. doi: 10.1198/000313006X152649. URL https://doi.org/10.1198/000313006X152649. Publisher: ASA Website _eprint: https://doi.org/10.1198/000313006X152649.

Gemini Team Google, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, J. D. Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, D. J. Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran

Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, R. J. Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo-yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean-baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie

Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, C. J. Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta,

Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, December 2024. URL http://arxiv.org/abs/2403.05530. arXiv:2403.05530 [cs].

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, December 2024. URL http://arxiv.org/abs/2411.04872. arXiv:2411.04872 [cs].

Arthur Goemans, Marie Davidsen Buhl, Jonas Schuett, Tomek Korbak, Jessica Wang, Benjamin Hilton, and Geoffrey Irving. Safety case template for frontier AI: A cyber inability argument, November 2024. URL http://arxiv.org/abs/2411.08088. arXiv:2411.08088 [cs].

Gary Goertz. *Social Science Concepts and Measurement: New and Completely Revised Edition*. Princeton University Press, September 2020. ISBN 978-0-691-20548-9. Google-Books-ID: EPjkDwAAQBAJ.

Josh A. Goldstein and Girish Sastry. The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:503–518, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31653. URL https://ojs.aaai.org/index.php/AIES/article/view/31653.

Yunye Gong, Robik Shrestha, Jared Claypoole, Michael Cogswell, Arijit Ray, Christopher Kanan, and Ajay Divakaran. BloomVQA: Assessing Hierarchical Multi-modal Comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14905–14918, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.885. URL https://aclanthology.org/2024.findings-acl.885/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire

Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat,

Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI Control: Improving Safety Despite Intentional Subversion, July 2024. URL http://arxiv.org/abs/2312.06942. arXiv:2312.06942 [cs].

Friederike Grosse-Holz and Ole Jorgensen. Early Insights from Developing Question-Answer Evaluations for Frontier AI | AISI Work, September 2024. URL https://www.aisi.gov.uk/work/early-insights-from-developing-question-answer-evaluations-for-frontier-ai.

Thomas Grote. Fairness as adequacy: a sociotechnical view on model evaluation in machine learning. *AI and Ethics*, 4(2):427–440, May 2024. ISSN 2730-5961. doi: 10.1007/s43681-023-00280-x. URL https://doi.org/10.1007/s43681-023-00280-x.

Robert M. Groves, Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology*. John Wiley & Sons, September 2011. ISBN 978-1-118-21134-2. Google-Books-ID: ctow8zWdyFgC.

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, and Yanghua Xiao. DetectBench: Can Large Language Model Detect and Piece Together Implicit Evidence? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 199–222, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.11. URL https://aclanthology.org/2024.findings-emnlp.11/.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation. In *Advances in Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=t1mAXb4Cop#discussion.

Himanshu Gupta, Shreyas Verma, Ujjwala Anantheswaran, Kevin Scaria, Mihir Parmar, Swaroop Mishra, and Chitta Baral. Polymath: A Challenging Multi-modal Mathematical Reasoning Benchmark, October 2024. URL http://arxiv.org/abs/2410.14702. arXiv:2410.14702 [cs].

Todd M. Gureckis. Mechanical Turk - The Poisoned Well?, June 2021. URL https://todd.gureckislab.org/2021/06/09/the-poisoned-well.

Michael Gusenbauer. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214, January 2019. ISSN 1588-2861. doi: 10.1007/s11192-018-2958-5. URL https://doi.org/10.1007/s11192-018-2958-5.

Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, and Zechang Sun. AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model, November 2024. URL http://arxiv.org/abs/2411.09012. arXiv:2411.09012 [astro-ph].

Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues, December 2023. URL https://osf.io/ey8db_v1.

Neal Robert Haddaway, Alexandra Mary Collins, Deborah Coughlin, and Stuart Kirk. The Role of Google Scholar in Evidence Reviews and Its Applicability to Grey Literature Searching. *PLOS ONE*, 10(9):e0138237, September 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0138237. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0138237. Publisher: Public Library of Science.

Gali Halevi, Henk Moed, and Judit Bar-Ilan. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *Journal of Informetrics*, 11(3):823–834, August 2017. ISSN 1751-1577. doi: 10.1016/j.joi.2017.06.005. URL https://www.sciencedirect.com/science/article/pii/S1751157717300676.

Serhii Hamotskyi, Anna-Izabella Levbarg, and Christian Hänig. Eval-UA-tion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models. In Mariana Romanyshyn, Nataliia Romanyshyn, Andrii Hlybovets, and Oleksii Ignatenko (eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pp. 109–119, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.unlp-1.13/.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M. Asmar, Sanmi Koyejo, Michael S. Bernstein, and Mykel J. Kochenderfer. More than Marketing? On the Information Value of AI Benchmarks for Practitioners, December 2024. URL http://arxiv.org/abs/2412.05520. arXiv:2412.05520 [cs].

Jan Hatzius, Joseph Briggs, Devesh Kodnani, and Giovanni Pierdomenico. The Potentially Large Effects of Artificial Intelligence on Economic Growth. Technical report, Goldman Sachs Economics Research, March 2023. URL https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html.

Fred Heiding, Simon Lermen, Andrew Kao, Bruce Schneier, and Arun Vishwanath. Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects, November 2024. URL http://arxiv.org/abs/2412.00586. arXiv:2412.00586 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021a. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Advances in Neural Information Processing Systems*, volume 1, December 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Emily A. Hennessy, Blair T. Johnson, and Ciara Keenan. Best Practice Guidelines and Essential Methodological Steps to Conduct Rigorous and Systematic Meta-Reviews. *Applied Psychology: Health and Well-Being*, 11(3):353–381, 2019. ISSN 1758-0854. doi: 10.1111/aphw.12169. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/aphw.12169. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/aphw.12169.

Michael Hewing and Vincent Leinhos. The Prompt Canvas: A Literature-Based Practitioner Guide for Creating Effective Prompts in Large Language Models, December 2024. URL http://arxiv.org/abs/2412.05127. arXiv:2412.05127 [cs].

Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini (eds.), *Proceedings of The Second Arabic Natural Language Processing Conference*, pp. 225–249, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.20. URL https://aclanthology.org/2024.arabicnlp-1.20/.

Carl Hildebrandt, Trey Woodlief, and Sebastian Elbaum. ODD-diLLMma: Driving Automation System ODD Compliance Checking using LLMs. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13809–13816, October 2024. doi: 10.1109/IROS58592.2024.10801369. URL https://ieeexplore.ieee.org/document/10801369. ISSN: 2153-0866.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. Entering Real Social World! Benchmarking the Social Intelligence of Large Language Models from a First-person Perspective, December 2024. URL http://arxiv.org/abs/2410.06195. arXiv:2410.06195 [cs].

Rick H. Hoyle. *Statistical Strategies for Small Sample Research*. SAGE, March 1999. ISBN 978-0-7619-0886-9. Google-Books-ID: 7O1gJE1X4hEC.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. Game-theoretic LLM: Agent Workflow for Negotiation Games, November 2024. URL http://arxiv.org/abs/2411.05990. arXiv:2411.05990 [cs].

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. Apathetic or Empathetic? Evaluating LLMs' Emotional Alignments with Humans. In *Advances in Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=pwRVGRWtGg.

Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks, July 2024. URL http://arxiv.org/abs/2405.10632. arXiv:2405.10632.

ICMJE. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals, January 2025. URL https://www.icmje.org/#aboutur.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015. ISBN 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. URL https://www.cambridge.org/core/books/causal-inference-for-statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB.

Igor Ivanov. BioLP-bench: Measuring understanding of biological lab protocols by large language models, October 2024. URL https://www.biorxiv.org/content/10.1101/2024.08.21.608694v4. Pages: 2024.08.21.608694 Section: New Results.

Michelle Jackson and D. R. Cox. The Principles of Experimental Design and Their Application in Sociology. *Annual Review of Sociology*, 39(Volume 39, 2013):27–49, July 2013. ISSN 0360-0572, 1545-2115. doi: 10.1146/annurev-soc-071811-145443. URL https://www.annualreviews.org/content/journals/10.1146/annurev-soc-071811-145443. Publisher: Annual Reviews.

Abigail Z. Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 706, New York, NY, USA, January 2020. Association for Computing Machinery. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3375671. URL https://doi.org/10.1145/3351095.3375671.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6750–6774, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.418. URL https://aclanthology.org/2023.emnlp-main.418/.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract Visual Reasoning with Tangram Shapes. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 582–601, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.38. URL https://aclanthology.org/2022.emnlp-main.38/.

Carlos E. Jimenez, Olga Russakovsky, and Karthik Narasimhan. CARETS: A Consistency And Robustness Evaluative Test Suite for VQA. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6392–6405, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.443. URL https://aclanthology.org/2022.acl-long.443/.

Yimin Jing, Renren Jin, Jiahao Hu, Huishi Qiu, Xiaohua Wang, Peng Wang, and Deyi Xiong. FollowEval: A Multi-Dimensional Benchmark for Assessing the Instruction-Following Capability of Large Language Models, November 2023. URL http://arxiv.org/abs/2311.09829. arXiv:2311.09829 [cs].

Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10 (18):eadk3452, May 2024. doi: 10.1126/sciadv.adk3452. URL https://www.science.org/doi/10.1126/sciadv.adk3452. Publisher: American Association for the Advancement of Science.

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. Resolving the Human-subjects Status of Machine Learning's Crowdworkers: What ethical framework should govern the interaction of ML researchers and crowdworkers? *Queue*, 21(6):Pages 60:101–Pages 60:127, January 2024. ISSN 1542-7730. doi: 10.1145/3639452. URL https://dl.acm.org/doi/10.1145/3639452.

Joshua D. Kertzer and Jonathan Renshon. Experiments and Surveys on Political Elites. *Annual Review of Political Science*, 25(Volume 25, 2022):529–550, May 2022. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051120-013649. URL https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-051120-013649. Publisher: Annual Reviews.

Prashant Kodali, Anmol Goel, Likhith Asapu, Vamshi Krishna Bonagiri, Anirudh Govil, Monojit Choudhury, Manish Shrivastava, and Ponnurangam Kumaraguru. From Human Judgements to Predictive Models: Unravelling Acceptability in Code-Mixed Sentences, May 2024. URL http://arxiv.org/abs/2405.05572. arXiv:2405.05572 [cs].

Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pp. 12493–12509, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.675. URL https://aclanthology.org/2024.acl-long.675/.

Romain Lacombe, Kerrie Wu, and Eddie Dilworth. ClimateX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?, November 2023. URL http://arxiv.org/abs/2311.17107. arXiv:2311.17107 [cs].

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=UnWhcpIyUC&noteId=OrjYu5uVxt.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring Capabilities of Language Models for Biology Research, July 2024. URL http://arxiv.org/abs/2407.10362. arXiv:2407.10362 [cs].

Benjamin Lebrun, Sharon Temtsin, Andrew Vonasch, and Christoph Bartneck. Detecting the corruption of online questionnaires by artificial intelligence. *Frontiers in Robotics and AI*, 10, February 2024. ISSN 2296-9144. doi: 10.3389/frobt.2023.1277635. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2023.1277635/full. Publisher: Frontiers.

E. D. Leeuw. To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 2005. URL https://www.semanticscholar.org/paper/To-mix-or-not-to-mix-data-collection-modes-in-Leeuw/4e09d55e25393eb22b47c4b0d82d4dfa84bc24d0.

Solim LeGris, Wai Keen Vong, Brenden M. Lake, and Todd M. Gureckis. H-ARC: A Robust Estimate of Human Performance on the Abstraction and Reasoning Corpus Benchmark, September 2024. URL http://arxiv.org/abs/2409.01374. arXiv:2409.01374 [cs].

Xiaoxuan Lei, Lucas Gomez, Hao Yuan Bai, and Pouya Bashivan. IWISDM: Assessing instruction following in multimodal models at scale. In *Proceedings of the 3rd Conference on Lifelong Learning Agents*. arXiv, July 2024a. doi: 10.48550/arXiv.2406.14343. URL https://lifelong-ml.cc/Conferences/2024/acceptedpapersandvideos/conf-2024-28. arXiv:2406.14343 [cs].

Zhikai Lei, Tianyi Liang, Hanglei Hu, Jin Zhang, Yunhua Zhou, Yunfan Shao, Linyang Li, Chenchui Li, Changbo Wang, Hang Yan, and Qipeng Guo. GAOKAO-Eval: Does high scores truly reflect strong capabilities in LLMs?, December 2024b. URL http://arxiv.org/abs/2412.10056. arXiv:2412.10056.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, Copenhagen, Denmark, 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL http://aclweb.org/anthology/D17-1259.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples, October 2024a. URL http://arxiv.org/abs/2410.14669. arXiv:2410.14669 [cs].

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. In Search of the Long-Tail: Systematic Generation of Long-Tail Inferential Knowledge via Logical Rule Guided Search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2348–2370, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.140. URL https://aclanthology.org/2024.emnlp-main.140/.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Advances in Neural Information Processing Systems*, volume 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a97da629b098b75c294dffdc3e463904-Abstract-round1.html.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 331–348, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72897-6. doi: 10.1007/978-3-031-72897-6_19.

Q. Vera Liao and Ziang Xiao. Rethinking Model Evaluation as Narrowing the Socio-Technical Gap, June 2023. URL http://arxiv.org/abs/2306.03100. arXiv:2306.03100 [cs].

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, August 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

John A. List, Sally Sadoff, and Mathis Wagner. So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14 (4):439–457, November 2011. ISSN 1573-6938. doi: 10.1007/s10683-011-9275-7. URL https://doi.org/10.1007/s10683-011-9275-7.

Jiawei Liu, Thanh Nguyen, Mingyue Shang, Hantian Ding, Xiaopeng Li, Yu Yu, Varun Kumar, and Zijian Wang. Learning Code Preference via Synthetic Evolution, October 2024a. URL http://arxiv.org/abs/2410.03837. arXiv:2410.03837 [cs].

Peilin Liu, Hongyu Lin, Meng Liao, Hao Xiang, Xianpei Han, and Le Sun. WebDP: Understanding Discourse Structures in Semi-Structured Web Documents. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10235–10258, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.650. URL https://aclanthology.org/2023.findings-acl.650/.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do Video LLMs Really Understand Videos? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

Lu Lu, Nathan Neale, Nathaniel D. Line, and Mark Bonn. Improving Data Quality Using Amazon Mechanical Turk Through Platform Setup. *Cornell Hospitality Quarterly*, 63(2):231–246, May 2022a. ISSN 1938-9655. doi: 10.1177/19389655211025475. URL https://doi.org/10.1177/19389655211025475. Publisher: SAGE Publications Inc.

Michael Lu, Hyundong Justin Cho, Weiyan Shi, Jonathan May, and Alexander Spangher. NewsInterview: a Dataset and a Playground to Evaluate LLMs' Ground Gap via Informational Interviews, November 2024. URL http://arxiv.org/abs/2411.13779. arXiv:2411.13779 [cs].

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *Proceedings of the 12th International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=KUNzEQMWU7.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*, September 2019. doi: 10.1287/mksc.2019.1192. URL https://pubsonline.informs.org/doi/abs/10.1287/mksc.2019.1192. Publisher: INFORMS.

Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *Advances in Neural Information Processing Systems*, November 2023. URL https://openreview.net/forum?id=JVlWseddak.

Nector Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Chapter 2: Technical Performance. In *The AI Index 2024 Annual Report*, pp. 73–158. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024. URL https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.

Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence, October 2024. URL http://arxiv.org/abs/2402.09880. arXiv:2402.09880.

Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon Statistical Significance. *The American Statistician*, 73(sup1):235–245, March 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1527253. URL https://doi.org/10.1080/00031305.2018.1527253. Publisher: ASA Website _eprint: https://doi.org/10.1080/00031305.2018.1527253.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking Distributional Alignment of Large Language Models, November 2024. URL http://arxiv.org/abs/2411.05403. arXiv:2411.05403 [cs].

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants. In *Proceedings of the 12th International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=fibxvahvs3.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. Why AI Is WEIRD and Should Not Be This Way: Towards AI For Everyone, With Everyone, By Everyone, October 2024. URL http://arxiv.org/abs/2410.16315. arXiv:2410.16315 [cs].

Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, November 2024. URL http://arxiv.org/abs/2411.00640. arXiv:2411.00640.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The Effect of Natural Distribution Shift on Question Answering Models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6905–6916. PMLR, November 2020. URL https://proceedings.mlr.press/v119/miller20a.html. ISSN: 2640-3498.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists?, November 2024. URL http://arxiv.org/abs/2404.01475. arXiv:2404.01475 [cs].

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4918–4929, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.442. URL https://aclanthology.org/2020.findings-emnlp.442/.

John W. Mohr and Amin Ghaziani. Problems and prospects of measurement in the study of culture. *Theory and Society*, 43(3):225–246, July 2014. ISSN 1573-7853. doi: 10.1007/s11186-014-9227-2. URL https://doi.org/10.1007/s11186-014-9227-2.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, Alham Fikri Aji, and William Chandra Tjhi. Kalahi: A handcrafted, grassroots cultural LLM evaluation suite for Filipino, December 2024. URL http://arxiv.org/abs/2409.15380. arXiv:2409.15380 [cs].

Arsenii Kirillovich Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain. *Transactions on Machine Learning Research*, May 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=8ykyGbtt2q.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Vivek Gupta, and Dan Roth. MAPWise: Evaluating Vision-Language Models for Advanced Map Queries, August 2024. URL http://arxiv.org/abs/2409.00255. arXiv:2409.00255 [cs].

Nikita Nangia and Samuel R. Bowman. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4566–4575, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1449. URL https://aclanthology.org/P19-1449/.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1221–1235, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.98. URL https://aclanthology.org/2021.acl-long.98/.

Markus Neuhäuser and Graeme D. Ruxton. *The Statistical Analysis of Small Data Sets*. Oxford University Press, Oxford, New York, December 2024. ISBN 978-0-19-887297-9.

NIST. AI Risk Management Framework: AI RMF (1.0), January 2023. URL https://perma.cc/B4VD-AL6S.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it? In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (eds.), *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–162, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.blackboxnlp-1.10. URL https://aclanthology.org/2021.blackboxnlp-1.10/.

Brian A. Nosek, Tom E. Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S. Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B. Nuijten, Julia M. Rohrer, Felipe Romero, Anne M. Scheel, Laura D. Scherer, Felix D. Schönbrodt, and Simine Vazire. Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73(Volume 73, 2022):719–748, January 2022. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-020821-114157. URL https://www.annualreviews.org/content/journals/10.1146/annurev-psych-020821-114157. Publisher: Annual Reviews.

Rasha Obeidat, Yara Al-Harahsheh, Mahmoud Al-Ayyoub, and Maram Gharaibeh. ArEntail: manually-curated Arabic natural language inference dataset from news headlines. *Language Resources and Evaluation*, April 2024. ISSN 1574-0218. doi: 10.1007/s10579-024-09731-1. URL https://doi.org/10.1007/s10579-024-09731-1.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. OpenAI o1 System Card, December 2024. URL http://arxiv.org/abs/2412.16720. arXiv:2412.16720 [cs].

OSTP. Blueprint for an AI Bill of Rights, October 2022. URL https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Right

s.pdf.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372: n71, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting.

Stacey A. Page and Jeffrey Nyeboer. Improving the process of research ethics review. *Research Integrity and Peer Review*, 2(1):14, August 2017. ISSN 2058-8615. doi: 10.1186/s41073-017-0038-7. URL https://doi.org/10.1186/s41073-017-0038-7.

Lawrence A. Palinkas, Sarah M. Horwitz, Carla A. Green, Jennifer P. Wisdom, Naihua Duan, and Kimberly Hoagwood. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5):533–544, September 2015. ISSN 1573-3289. doi: 10.1007/s10488-013-0528-y. URL https://doi.org/10.1007/s10488-013-0528-y.

Xudong Pan, Jiarun Dai, Yihe Fan, and Min Yang. Frontier AI systems have surpassed the self-replicating red line, December 2024. URL http://arxiv.org/abs/2412.12140. arXiv:2412.12140 [cs].

Patricia Paskov, Lukas Berglund, Everett Smith, and Lisa Soder. GPAI Evaluations Standards Taskforce: Towards Effective AI Governance, November 2024. URL http://arxiv.org/abs/2411.13808. arXiv:2411.13808 [cs].

John W. Patty and Elizabeth Maggie Penn. Measuring Fairness, Inequality, and Big Data: Social Choice Since Arrow. *Annual Review of Political Science*, 22(Volume 22, 2019):435–460, May 2019. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-022018-024704. URL https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-022018-024704. Publisher: Annual Reviews.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating Frontier Models for Dangerous Capabilities, April 2024. URL http://arxiv.org/abs/2403.13793. arXiv:2403.13793 [cs].

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. ISSN 1533-7928. URL http://jmlr.org/papers/v22/20-303.html.

Nabeel Qureshi, Maria Edelen, Lara Hilton, Anthony Rodriguez, Ron D. Hays, and Patricia M. Herman. Comparing Data Collected on Amazon's Mechanical Turk to National Surveys. *American Journal of Health Behavior*, 46(5):497–502, October 2022. doi: 10.5993/AJHB.46.5.1.

May Lynn Reese and Anastasia Smirnova. Comparing ChatGPT and Humans on World Knowledge and Common-sense Reasoning Tasks: A case study of the Japanese Winograd Schema Challenge. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, pp. 1–9, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0331-7. doi: 10.1145/3613905.3650975. URL https://doi.org/10.1145/3613905.3650975.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A

Benchmark. In *Proceedings of the 1st Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=Ti67584b98#discussion.

Rainer Reisenzein and Martin Junge. Measuring the intensity of emotions. *Frontiers in Psychology*, 15, September 2024. ISSN 1664-1078. doi: 10.3389/fpsyg.2024.1437843. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1437843/full. Publisher: Frontiers.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *Proceedings of the 38th Conference on Neural Information Processing Systems*, November 2024. URL https://openreview.net/forum?id=hcOq2buakM#discussion.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation, December 2024. URL http://arxiv.org/abs/2405.08807. arXiv:2405.08807 [cs].

Anthony J. Rosellini and Timothy A. Brown. Developing and Validating Clinical Questionnaires. *Annual Review of Clinical Psychology*, 17(Volume 17, 2021):55–81, May 2021. ISSN 1548-5943, 1548-5951. doi: 10.1146/annurev-clinpsy-081219-115343. URL https://www.annualreviews.org/content/journals/10.1146/annurev-clinpsy-081219-115343. Publisher: Annual Reviews.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20827–20905, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4241fec6e94221526b0a9b24828bb774-Abstract-Conference.html.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14182–14214, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.844. URL https://aclanthology.org/2024.findings-acl.844/.

Maria Elena Sanchez. Effects of Questionnaire Design on the Quality of Survey Data. *Public Opinion Quarterly*, 56(2):206–217, January 1992. ISSN 0033-362X. doi: 10.1086/269311. URL https://doi.org/10.1086/269311.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for Subpopulation Shift. In *Proceedings of the International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=mQPBmvyAuk.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. Are Machines Better at Complex Reasoning? Unveiling Human-Machine Inference Gaps in Entailment Verification, May 2024. URL http://arxiv.org/abs/2402.03686. arXiv:2402.03686 [cs].

Pooria Sarrami-Foroushani, Joanne Travaglia, Deborah Debono, Robyn Clay-Williams, and Jeffrey Braithwaite. Scoping Meta-Review: Introducing a New Methodology. *Clinical and Translational Science*, 8(1):77–81, 2015. ISSN 1752-8062. doi: 10.1111/cts.12188. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12188. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12188.

Galit Weidman Sassoon. Measurement theory in linguistics. *Synthese*, 174(1):151–180, May 2010. ISSN 1573-0964. doi: 10.1007/s11229-009-9687-5. URL https://doi.org/10.1007/s11229-009-9687-5.

Rens van de Schoot and Milica Miočević (eds.). *Small Sample Size Solutions: A Guide for Applied Researchers and Practitioners*. Routledge, London, February 2020. ISBN 978-0-429-27387-2. doi: 10.4324/9780429273872.

Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers, July 2024. URL http://arxiv.org/abs/2406.14325. arXiv:2406.14325 [cs].

Muhammad Shah Jahan, Habib Ullah Khan, Shahzad Akbar, Muhammad Umar Farooq, Sarah Gul, and Anam Amjad. Bidirectional Language Modeling: A Systematic Literature Review. *Scientific Programming*, 2021(1):6641832, 2021. ISSN 1875-919X. doi: 10.1155/2021/6641832. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6641832. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/6641832.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4717–4726, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.381. URL https://aclanthology.org/2020.emnlp-main.381/.

Aaron Shaw and Eszter Hargittai. Do the Online Activities of Amazon Mechanical Turk Workers Mirror Those of the General Population? A Comparison of Two Survey Samples. *International Journal of Communication*, 15(0):16, October 2021. ISSN 1932-8036. URL https://ijoc.org/index.php/ijoc/article/view/16942. Number: 0.

Kim Bartel Sheehan. Crowdsourcing research: Data collection with Amazon's Mechanical Turk. *Communication Monographs*, 85(1):140–156, January 2018. ISSN 0363-7751. doi: 10.1080/03637751.2017.1342043. URL https://doi.org/10.1080/03637751.2017.1342043. Publisher: NCA Website _eprint: https://doi.org/10.1080/03637751.2017.1342043.

Eunjung Shin, Timothy P. Johnson, and Kumar Rao. Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey. *Social Science Computer Review*, 30(2):212–228, May 2012. ISSN 0894-4393. doi: 10.1177/0894439311404508. URL https://doi.org/10.1177/0894439311404508. Publisher: SAGE Publications Inc.

David Shrier, Julian Emanuel, and Marc Harris. Is Your Job AI Resilient? *Harvard Business Review*, October 2023. URL https://hbr.org/2023/10/is-your-job-ai-resilient.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, September 2024. URL http://arxiv.org/abs/2409.04109. arXiv:2409.04109.

Gary Solon, Steven J. Haider, and Jeffrey M. Wooldridge. What Are We Weighting For? *The Journal of Human Resources*, 50(2):301–316, 2015. ISSN 0022-166X. URL https://www.jstor.org/stable/24735988. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

Taiga Someya and Yohei Oseki. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1581–1594, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.117. URL https://aclanthology.org/2023.findings-eacl.117/.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. ARN: Analogical Reasoning on Narratives, September 2024. URL http://arxiv.org/abs/2310.00996. arXiv:2310.00996 [cs].

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning. In *Proceedings of the 12th International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=jenyYQzue1.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W. Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras,

Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, January 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Stefanie Stantcheva. How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. *Annual Review of Economics*, 15(Volume 15, 2023):205–234, September 2023. ISSN 1941-1383, 1941-1391. doi: 10.1146/annurev-economics-091622-010157. URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-091622-010157. Publisher: Annual Reviews.

Peter Steinbach, Felicita Gernhardt, Mahnoor Tanveer, Steve Schmerler, and Sebastian Starke. Machine Learning State-of-the-Art with Uncertainties, April 2022. URL http://arxiv.org/abs/2204.05173. arXiv:2204.05173 [cs].

Victoria Stodden and Sheila Miguez. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. *Journal of Open Research Software*, 2(1):e21–e21, July 2014. ISSN 2049-9647. doi: 10.5334/jors.ay. URL https://account.openresearchsoftware.metajnl.com/index.php/up-j-jors/article/view/jors.ay. Number: 1.

Milton E. Strauss and Gregory T. Smith. Construct Validity: Advances in Theory and Methodology. *Annual Review of Clinical Psychology*, 5(Volume 5, 2009):1–25, April 2009. ISSN 1548-5943, 1548-5951. doi: 10.1146/annurev.clinpsy.032408.153639. URL https://www.annualreviews.org/content/journals/10.1146/annurev.clinpsy.032408.153639. Publisher: Annual Reviews.

Arjun Subramonian, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3234–3279, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.202. URL https://aclanthology.org/2023.findings-acl.202/.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. PhonologyBench: Evaluating Phonological Skills of Large Language Models, April 2024. URL http://arxiv.org/abs/2404.02456. arXiv:2404.02456 [cs].

Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar, and Rashedur M. Rahman. Deep learning based question answering system in Bengali. *Journal of Information and Telecommunication*, 5(2): 145–178, April 2021. ISSN 2475-1839. doi: 10.1080/24751839.2020.1833136. URL https://doi.org/10.1080/24751839.2020.1833136. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24751839.2020.1833136.

Ekaterina Taktasheva, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Valentina Kurenshchikova, Alena Spiridonova, Ekaterina Artemova, Tatiana Shavrina, and Vladislav Mikhailov. TAPE: Assessing Few-shot Russian Language Understanding. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2472–2497, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.183. URL https://aclanthology.org/2022.findings-emnlp.183/.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. TMMLU+: An Improved Traditional Chinese Evaluation Suite for Foundation Models. In *Proceedings of the 1st Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=95TayIeqJ4#discussion.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A Benchmark for Learning to Translate a New Language from One Grammar Book. In *Proceedings of the 12th International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=tbVWug9f2h.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. What's the Meaning of Superhuman Performance in Today's NLU? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12471–12491, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.697. URL https://aclanthology.org/2023.acl-long.697/.

Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, and Douwe Kiela. I am a Strange Dataset: Metalinguistic Tests for Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8888–8907, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.482. URL https://aclanthology.org/2024.acl-long.482/.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, September 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00685. URL https://doi.org/10.1162/tacl_a_00685.

Lisa H. Trahan, Karla K. Stuebing, Jack M. Fletcher, and Merrill Hiscock. The Flynn effect: A meta-analysis. *Psychological Bulletin*, 140(5):1332–1360, 2014. ISSN 1939-1455. doi: 10.1037/a0037173. Place: US Publisher: American Psychological Association.

Frederic Traylor. The threat of AI chatbot responses to crowdsourced open-ended survey questions. *Energy Research & Social Science*, 119:103857, January 2025. ISSN 2214-6296. doi: 10.1016/j.erss.2024.103857. URL https://www.sciencedirect.com/science/article/pii/S2214629624004481.

UK AISI. Scorers, 2025. URL https://inspect.ai-safety-institute.org.uk/scorers.html#scoring-metrics.

US AISI and UK AISI. US AISI and UK AISI Joint Pre-Deployment Test: OpenAI o1. Technical report, U.S. AI Safety Institute, National Institute of Standards and Technology; UK AI Safety Institute, Department of Science Innovation and Technology, December 2024. URL https://www.nist.gov/system/files/documents/2024/12/18/US_UK_AI%20Safety%20Institute_%20December_Publication-OpenAIo1.pdf.

U.S. Department of Homeland Security, Department of Agriculture, Department of Energy, National Aeronautics and Space Administration, Department of Commerce, Social Security Administration, Agency for International Development, Department of Housing and Urban Development, Department of Labor, Department of Defense, Department of Education, Department of Veterans Affairs, Environmental Protection Agency, Department of Health and Human Services, National Science

Foundation, and Department of Transportation. 45 CFR Part 46 – Protection of Human Subjects, January 2017. URL https://www.ecfr.gov/current/title-45/part-46.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the Planning Abilities of Large Language Models - A Critical Investigation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75993–76005, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/efb2072a358cefb75886a315a6fcf880-Abstract-Conference.html.

Jorre Vannieuwenhuyze, Geert Loosveldt, and Geert Molenberghs. A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, 74(5):1027–1045, January 2010. ISSN 0033-362X. doi: 10.1093/poq/nfq059. URL https://doi.org/10.1093/poq/nfq059.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1702–1717, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.95. URL https://aclanthology.org/2024.naacl-long.95/.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and prevention of large language model use in crowd work, October 2023a. URL http://arxiv.org/abs/2310.15683. arXiv:2310.15683 [cs].

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks, June 2023b. URL http://arxiv.org/abs/2306.07899. arXiv:2306.07899 [cs].

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. CONTEXTUAL: evaluating context-sensitive text-rich visual reasoning in large multimodal models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 49733–49787, Vienna, Austria, July 2024. JMLR.org.

Hanna Wallach, Meera Desai, Nicholas Pangakis, A. Feder Cooper, Angelina Wang, Solon Barocas, Alexandra Chouldechova, Chad Atalla, Su Lin Blodgett, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Evaluating Generative AI Systems is a Social Science Measurement Challenge, November 2024. URL http://arxiv.org/abs/2411.10939. arXiv:2411.10939.

Xiting Wang, Liming Jiang, Jose Hernandez-Orallo, David Stillwell, Luning Sun, Fang Luo, and Xing Xie. Evaluating General-Purpose AI with Psychometrics, December 2023. URL http://arxiv.org/abs/2310.16379. arXiv:2310.16379 version: 2.

Albert Webson, Alyssa Loo, Qinan Yu, and Ellie Pavlick. Are Language Models Worse than Humans at Following Prompts? It's Complicated. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7662–7686, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.514. URL https://aclanthology.org/2023.findings-emnlp.514/.

Laura Weidinger, John Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, Stevie Bergman, Mikel Rodriguez, Verena Rieser, and William Isaac. STAR: SocioTechnical Approach to Red Teaming Language Models, October 2024. URL http://arxiv.org/abs/2406.11757. arXiv:2406.11757 [cs].

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. Hybrid Human-LLM Corpus Construction and LLM Evaluation for Rare Linguistic Phenomena, March 2024. URL http://arxiv.org/abs/2403.06965. arXiv:2403.06965 [cs].

Chris Welty, Praveen Paritosh, and Lora Aroyo. Metrology for AI: From Benchmarks to Instruments, November 2019. URL http://arxiv.org/abs/1911.01875. arXiv:1911.01875 [cs].

Hjalmar Wijk, Tao Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Josh Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts, November 2024. URL http://arxiv.org/abs/2411.15114. arXiv:2411.15114 [cs].

Bradford D. Winters, Ayse P. Gurses, Harold Lehmann, J. Bryan Sexton, Carlyle Jai Rampersad, and Peter J. Pronovost. Clinical review: Checklists - translating evidence into practice. *Critical Care*, 13(6):210, December 2009. ISSN 1364-8535. doi: 10.1186/cc7792. URL https://doi.org/10.1186/cc7792.

Meng-Han Wu and Alexander Quinn. Confusing the Crowd: Task Instruction Quality on Amazon Mechanical Turk. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 5:206–215, September 2017. ISSN 2769-1349. doi: 10.1609/hcomp.v5i1.13317. URL https://ojs.aaai.org/index.php/HCOMP/article/view/13317.

Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning, December 2024. URL http://arxiv.org/abs/2412.02172. arXiv:2412.02172 [cs].

Yue Wu, Xuan Tang, Tom Mitchell, and Yuanzhi Li. SmartPlay : A Benchmark for LLMs as Intelligent Agents. In *Proceedings of the 12th International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=S2oTVrlcp3.

Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. CARE-MI: Chinese Benchmark for Misinformation Evaluation in Maternity and Infant Care. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42358–42381, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/84062fe53d23e0791c6dbb456783e4a9-Abstract-Datasets_and_Benchmarks.html.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10967–10982, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.676. URL https://aclanthology.org/2023.emnlp-main.676/.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. Data Contamination Can Cross Language Barriers. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17864–17875, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.990. URL https://aclanthology.org/2024.emnlp-main.990/.

Affan Yasin, Rubia Fatima, Lijie Wen, Wasif Afzal, Muhammad Azhar, and Richard Torkar. On Using Grey Literature and Google Scholar in Systematic Literature Reviews in Software Engineering. *IEEE Access*, 8:36226–36243, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2971712. URL https://ieeexplore.ieee.org/document/8984351/.

Hang Yin, Kuang-Hung Liu, Mengying Sun, Yuxin Chen, Buyun Zhang, Jiang Liu, Vivek Sehgal, Rudresh Rajnikant Panchal, Eugen Hotaj, Xi Liu, Daifeng Guo, Jamey Zhang, Zhou Wang, Shali Jiang, Huayu Li, Zhengxing Chen, Wen-Yen Chen, Jiyan Yang, and Wei Wen. AutoML for Large Capacity Modeling of Meta's Ranking Systems. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 374–382, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0172-6. doi: 10.1145/3589335.3648336. URL https://dl.acm.org/doi/10.1145/3589335.3648336.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun,

Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://ieeexplore.ieee.org/document/10656299. ISSN: 2575-7075.

Andrew Zamecnik, Abhinava Barthakur, Hanyi Wang, and Shane Dawson. Mapping Employable Skills in Higher Education Curriculum Using LLMs. In Rafael Ferreira Mello, Nikol Rummel, Ioana Jivet, Gerti Pishtari, and José A. Ruipérez Valiente (eds.), *Technology Enhanced Learning for Inclusive and Equitable Quality Education*, pp. 18–32, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72312-4. doi: 10.1007/978-3-031-72312-4_2.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A Benchmark for Compositional Visual Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29776–29788, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c08ee8fe3d19521f3bfa4102898329fd-Abstract-Datasets_and_Benchmarks.html.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–23, New York, NY, USA, May 2024a. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642647. URL https://doi.org/10.1145/3613904.3642647.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15654–15669, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.925. URL https://aclanthology.org/2024.findings-acl.925/.

Susu Zhang, Jingchen Liu, and Zhiliang Ying. Statistical Applications to Cognitive Diagnostic Testing. *Annual Review of Statistics and Its Application*, 10(Volume 10, 2023):651–675, March 2023. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-033021-111803. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-033021-111803. Publisher: Annual Reviews.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1495–1516, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.82. URL https://aclanthology.org/2024.acl-long.82/.

Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: measure dataset diversity, don't just claim it. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 60644–60673, Vienna, Austria, January 2025. JMLR.org.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149/.

Haokun Zhou and Yipeng Hong. DiffuSyn Bench: Evaluating Vision-Language Models on Real-World Complexities with Diffusion-Generated Synthetic Benchmarks, June 2024. URL http://arxiv.org/abs/2406.04470. arXiv:2406.04470 [cs].

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 314–324, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.24. URL https://aclanthology.org/2022.naacl-main.24/.

Yujun Zhou, Jingdong Yang, Kehan Guo, Pin-Yu Chen, Tian Gao, Werner Geyer, Nuno Moniz, Nitesh V. Chawla, and Xiangliang Zhang. LabSafety Bench: Benchmarking LLMs on Safety Issues in Scientific Labs, October 2024. URL http://arxiv.org/abs/2410.14182. arXiv:2410.14182 [cs].

Hao Zhu, Raghav Kapoor, So Yeon Min, Winson Han, Jiatai Li, Kaiwen Geng, Graham Neubig, Yonatan Bisk, Aniruddha Kembhavi, and Luca Weihs. EXCALIBUR: Encouraging and Evaluating Embodied Exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14931–14942, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhu_EXCALIBUR_Encouraging_and_Evaluating_Embodied_Exploration_CVPR_2023_paper.html.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, David Lo, Binyuan Hui, Niklas Muennighoff, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions, October 2024. URL http://arxiv.org/abs/2406.15877. arXiv:2406.15877 [cs].

Michael J. Zickar. Measurement Development and Evaluation. *Annual Review of Organizational Psychology and Organizational Behavior*, 7(Volume 7, 2020):213–232, January 2020. ISSN 2327-0608, 2327-0616. doi: 10.1146/annurev-orgpsych-012119-044957. URL https://www.annualreviews.org/content/journals/10.1146/annurev-orgpsych-012119-044957. Publisher: Annual Reviews.

A APPENDIX: FULL CHECKLIST

Our checklist is presented in full below, updated with slight modifications and reorganization from the version used in our coding process. Our hope is that this checklist can guide and inform researchers in building human baselines and in reporting baseline results.

Note that the following changes were made during our coding process:

- All items were open text fields unless explicitly indicated otherwise below.
- For questions on a scale of "Yes", "Partial", "No", "Unknown/Unreported", or "N/A":
  - "Yes" and "No" options were selected only if the relevant checklist item was explicitly noted in an article's main text, supplementary material/appendices, or GitHub codebase.
  - 'Partial" was selected where articles did not fully satisfy the item criterion, e.g., satisfying the criterion for some but not all of the baseline items. "Partial" labels were "rounded" up to "Yes" labels unless otherwise specified below.
  - "Unknown/Unreported": see below.
  - "N/A" was selected where the item did not apply to the baseline at hand.
- For all questions, including items with open text fields: coders indicated "Unknown/Unreported" where items were not reported or where coders were not able to determine the response based on an article's main text, supplementary material/appendices, or GitHub codebase.
  - For select items, "Unknown/Unreported" labels were resolved to default values, which are indicated below in underline and with a "(Default)" label. Default responses are selected based on our understanding of common practices in AI evaluation, and we attempt to be liberal in terms of assuming rigor in the baseline where there are is no consensus in the literature on common practices.
  - For items without default responses, "Unknown/Unreported" labels were not adjusted.

A.0 PAPER INFORMATION

0.1 **Paper Title**

0.2 **Paper Link**

0.3 **Publication Year**

0.4 **Publication Venue**

0.5 **Type of Eval**
Select all that apply

- Knowledge
- Capabilities
- Propensity
- Agent

0.6 **Mode of Eval**
Select all that apply

- Text
- Visual (photo/video)
- Audio
- Other

0.7 **Language of Eval**
Select all that apply from list

0.8 **Evaluation Dataset Size**: What is the total number of items in the evaluation dataset?

0.9 **AI Test Set Size**: What is the number of items that the AI evaluation is run on? (Default same as Q0.8)

0.10 **AI Samples per Item**: What is the number of AI responses ("samples" or "runs") that is collected for each item? (Default 1)

A.1 BASELINE DESIGN

1.1 **Number of Baseliners**: How many baseliners were there total?

1.2 **Baseline Test Set Size**: What is the number of items that the human baseline is run on? (i.e., how many of the questions do the baseliners collectively answer?) (Default same as Q0.9)

1.2.1 **Baseline Test Set Sampling Strategy**: If the baseline is only run on a sample of the total dataset: what is the sampling strategy behind how the items were selected? E.g., simple random sampling, stratified sampling, etc.

1.3 **Baseline Samples per Item**: What was the number of human baseliner responses that is collected for each item? (Default Q1.1 $*$ Q1.4 / Q1.2, or 1 if Q1.1 or Q1.4 unreported)

1.4 **Items per Baseliner**: What is the number of items that each baseliner responded to?

1.5 **Explicit Human/AI Adjustment**: Does the eval/baseline instructions and items account for both humans and AI models completing the evals items (questions/tasks)? E.g., do the authors of the eval explicitly state that the eval is designed so as not to advantage either humans or AI models?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

1.6 **Iterative Design**: Was the experimental setup of the baseline iteratively designed with participatory methods? E.g., was there a pilot study, expert validation of the items, etc.?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.7 **Amount of Effort**: Does the baseline control for the amount of effort by human baseliners and AIs? E.g., in terms of cost, time, etc.
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.8 **Power Analysis**: Did the authors conduct power analysis in order to determine baseline size?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

1.8.1 **Minimum Detectable Effect Size**: if yes, what is the minimum detectable effect size and power?

1.9 **Ethics Review**: Was the study approved or exempted by an IRB, or did it undergo other ethics review?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

1.10 **Pre-Registration**: Was the baseline/eval design pre-registered? I.e., a plan detailing the experimental setup that is publicly registered online before running the experiment (e.g., on OSF, COS, etc.)
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

A.2 BASELINER RECRUITMENT

2.1 **Population of Interest Identification**: Does the reporting identify human populations for which these results may be valid, i.e., a human population of interest?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.1.1 **Population of Interest Identification Criteria**: Which of the following factors were used to scope the target human population of interest?
Select all that apply

- Expertise
- Education
- Language
- Gender/sex
- Race
- Socioeconomic status
- Age
- Disabilities/impairments
- Political orientation
- Digital literacy (Prior experience with computers)
- AI literacy (Prior experience with AI tools)
- Baseline experience: Prior experience with AI evals/doing human baselines
- Other (specify)

2.2 **Baseliner Sampling Strategy**: How were the human baseliners recruited?
Select one of the below

- Crowdsource
- Convenience sample
- Simple random sample
- Stratified random sample
- Other (specify)
- Unknown/unreported

2.3 **Quality Control in Recruitment**: Were human baseliners pre-qualified or excluded during the recruitment process for any reason?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

2.3.1 **Quality Control Criteria for Baseliners**: If yes: please describe the inclusion/exclusion criteria for human baseliners (e.g., pre-tests, expert judgements/filtering, quality scores or ratings on crowdwork platforms, number of tasks completed on crowdwork platforms). Data quality checks that occurred after baseliners were recruited should be reported in the implementation section (e.g., attention checks in a survey).

2.3.2 **Recruitment Exclusion Rate**: If yes: how many baseliners were excluded from the final baseline based on these criteria?

2.4 **Author Baseliners**: Did the authors or members of the research team also serve as human baseliners?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.5 **Baseliner Train/Test Contamination**: Did the recruitment process exclude baseliners who had been exposed to the eval questions previously?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.6 **Baseliner Training**: Did the human baseliners receive training for the baseline? Training should be distinct from the reported data, e.g., a tutorial completed before answering baseline questions
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

2.6.1 **Baseliner Training Type**: If yes: describe the type of training received (e.g., tutorial, shown examples, etc.)

2.6.2 **Baseliner Training Compensation**: If yes: were the baseliners compensated for the training?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

2.6.2.1 **Baseliner Training Compensation Amount**: If yes: list the compensation per baseliner (preferably $ / hour, otherwise total $ amount if stated)

2.7 **Baseliner Testing Compensation**: Were the human baseliners compensated for completing the baseline?
Select one of: "Yes" (Default if Q2.2 is "Crowdsource"), "Partial", "No", "Unknown/Unreported", or "N/A"

2.7.1 **Baseliner Testing Compensation Amount**: If yes: how much was compensation? (preferably $ / hour, otherwise total $ amount if stated)

2.7.2 **Baseliner Testing Performance Bonus**: If yes: was a performance bonus offered to baseliners?
Select one of: "Yes" (Default if Q2.2 is "Crowdsource"), "Partial", "No", "Unknown/Unreported", or "N/A"

2.7.2.1 **Baseliner Testing Performance Bonus Amount**: If yes: how much was the performance bonus, and how was it determined?

2.7.3 **Baseliner Testing Compensation Structure**: If yes: were compensation rates and structures constant across baseliners? E.g., respond no if baseliners were paid differently according to expertise.
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

2.7.3.1 **Baseliner Testing Compensation Structure Details**: If not compensated equally: how were compensation amounts determined?

A.3 BASELINE IMPLEMENTATION

3.1 **Instrument Length**: How many items did the human baseliners complete in a single sitting/session? I.e., what is the length of the baseliner "context window" in units of items?

3.1.1 **Item Randomization**: If not 1: was the order of the questions randomized?

3.2 **Quality Control in Implementation**: Were quality checks implemented or data cleaned/excluded during the data collection process (i.e., after baseliners were recruited)? E.g., were there any exclusion criteria for baseliner responses due to data quality such as attention check questions, honeypot questions, filtering out responders who completed the eval too quickly, screen recording, etc.
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

3.2.1 **Quality Control in Implementation Criteria**: If yes: what factors were used to determine data quality or to exclude low-quality data?

3.2.2 **Implementation Exclusion Rate**: If yes: how many samples were excluded from the final baseline based on these criteria?

3.3 **UI Equivalence**: Did the human baseliners and AIs have access to the same UI for each item?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.3.1 **GUI vs. API**: Check this box if the humans had access to a graphical UI and the AIs only had API inputs
Checkbox item

3.3.2 **UI Equivalence Adjustment**: If no: does the eval attempt to adjust for the differences?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.4 **Instruction Equivalence**: Did the human baseliners and AIs have access to the same instructions/prompt/question for each item?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.4.1 **Instruction Equivalence Adjustment**: If no: does the eval attempt to adjust for the differences?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.5 **Tool Access Equivalence**: Did the human baseliners and AIs have access to the same (technical) tools for each item? Respond yes if neither group had access to extermal tools; respond yes if the human had internet access and the AI did not (but was trained on the internet)
Select one of: "No" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

3.5.1 **Tool Access Equivalence Enforcement**: If human baseliners' tool access was limited: was there an oversight mechanism for ensuring that the human baseliners only used the tools permitted? E.g., enforcement of AI tool use ban
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

3.6 **Explanations**: Did the eval/baseline collect explanations from the human baseliners, after the evaluation was conducted? I.e., explanations for why the human participants responded the way they did
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

A.4 BASELINE ANALYSIS

4.1 **Statistical Significance**: Did the eval test for statistically significant differences between AI and human performance?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

4.1.1 **Statistical Significance Test**: If yes: what statistical test was used?

4.2 **Uncertainty Estimate**: Did the paper present a measure of uncertainty for the AI and human baseline results? E.g., confidence intervals, variance, pooled/clustered standard errors, etc.?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

4.2.1 **Estimate Type**: Is the reported baseline a point estimate, an interval estimate, or a distribution?
Select all that apply
- Point estimate (Default)

- Interval estimate
- Distribution estimate

4.3 **Evaluation Metric Equivalence**: Was the same evaluation metric measured/compared for both humans and AIs? Respond "no" if, e.g., the human baseline is majority vote but the AI baseline is not
Select one of: "No" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.4 **Evaluation Scoring Criteria Equivalence**: Was the same scoring rubric used for both AI and human results?
Select one of: "No" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.5 **Evaluation Scoring Method Equivalence**: Was the same scoring method used for both AI and human results? E.g., human grading, LLM as a judge
Select one of: "No" (Default), "Partial", "No", "Unknown/Unreported", or "N/A"

4.6 **Quality Control Robustness**: If quality controls were implemented: are analyses robust to different choices of exclusion criteria? E.g., do the authors state that the results don't change when including/excluding incomplete data?
Select one of: "Yes", "Partial", "No", "Unknown/Unreported", or "N/A"

A.5 Baseline Documentation

5.1 **Additional Reporting**: Were the following reported?

5.1.1 **Reporting Sample Demographics**: Demographics for human baseliners, e.g., race, gender, etc. Respond yes only if within-sample demographics are reported; e.g., respond no if the paper only reports that 100% of the sample is based in the U.S.
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.2 **Reporting Baseline Instructions**: Instructions/guidelines given to human baseliners
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.3 **Reporting Time to Completion**: Time to completion for the eval items
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.4 **AI Tool Versions**: AI tools and versions (if baseliners had AI access)
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.1.5 **Completion Rate**: How many human baseliners were recruited but did not complete the tasks?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2 **Baseline Data Availability**: Is the (anonymized) human baseline data publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2.1 **Individual Baseline Data Availability**: If yes: is data available at the individual baseliner level? I.e., can you tell from the dataset which baseliners were responsible for which questions?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.2.2 **Baseline Data Non-Availability Justification**: If no: is there a reasonable justification for non-disclosure of the baseline dataset? E.g., privacy concerns, safety/security concerns, company policy, etc.

5.3 **Experimental Materials Availability**: Are experimental materials used to implement the eval/baseline publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

5.4 **Analysis Code Availability**: Is the code used to analyze the eval/baseline publicly available?
Select one of: "Yes", "Partial", "No" (Default), "Unknown/Unreported", or "N/A"

B Appendix: Methodology

We adopted a two-stage methodology as described in Section 3, adapted from the methodology of Zhao et al. (2025) and Reuel et al. (2024).

Section B.1 describes stage one, in which we conducted a meta-review of the measurement theory and AI evaluation literatures to qualitatively synthesize the checklist in Appendix A.

Section B.2 describes stage two, in which we systematically reviewed human baselines in foundation model evaluations.

B.1 Meta-Review

We begin with a scoping meta-review (a review of reviews) to qualitatively identify and synthesize literature relevant to human baselining. Meta-reviews are useful when there is little direct literature on the research question of interest (here, human baselines) but there is relevant literature from related fields (here, measurement theory) (Sarrami-Foroushani et al., 2015). As there is a wealth of literature in measurement theory, a meta-review that synthesizes the relevant evidence is appropriate to collect evidence in one place and to prevent researchers from being overwhelmed by the quantity of evidence (Hennessy et al., 2019).

Our literature search process adopted a purposive sampling approach. Although a systematic search process is normally ideal (Hennessy et al., 2019), purposive sampling is also acceptable for qualitative literature synthesis (e.g., Ames et al. 2019) and is justified here due to the broad scope of the relevant literature (Palinkas et al., 2015). Our sampling approach used theory-based inclusion criteria (Palinkas et al., 2015): we queried Google Scholar and Annual Reviews (2025a) in December 2024 for the keywords in Table 1, then filtered according to the criteria in Table 1. We also conducted backwards snowballing for the ML articles to identify further relevant literature. Finally, we added items to the sample based on our expertise, as many of the authors have experience in social science methodology and AI evaluation.

One limitation of this search strategy is that it introduces some sampling bias due to searching directly on the Annual Review website. We consider this limitation acceptable because by impact factor, Annual Reviews is a top-ranked publisher of literature reviews in the relevant social science disciplines (e.g., political science, psychology, sociology, statistics, economics) (Annual Reviews, 2025b). We thus expect our meta-review sample to be high-quality and relatively high-coverage.

Our search process yielded a total of 29 articles to be included in our meta-review (listed in Table 2). To synthesize our checklist, KW scanned these 29 articles and compiled a list of relevant methodological practices/considerations in a Google Sheet, categorizing each into the categories of baseline(r) design, recruitment, implementation, analysis, and documentation. The authors then collectively discussed the checklist and validated the checklist using expert feedback from six external experts before refining and finalizing the checklist. Finally, the checklist was also iteratively refined during the coding process.

Table 1: Inclusion criteria for meta-review articles.

Type	**Inclusion Criteria**
Document type	Literature review Position paper Synthesis article Book or book chapter (including reference texts)
Subject area	Measurement theory (including applications in statistics, economics, political science, psychology, education, sociology, or medicine) AI evaluation
Keywords (non-exhaustive)	"measurement theory" "measurement model*" "validity" "reliability" "replicability" "survey design" "survey method*" "questionnaire design" "experimental design" "causal inference"

Table 2: A complete list of the 29 articles included in our meta-review.

Subject area	**Articles**
Measurement theory ($n = 17$)	Bandalos (2018); Berinsky (2017); Cai et al. (2016); Chang et al. (2021); Couper (2017); Findley et al. (2021); Groves et al. (2011); Imbens & Rubin (2015); Jackson & Cox (2013); Kertzer & Renshon (2022); List et al. (2011); Nosek et al. (2022); Rosellini & Brown (2021); Stantcheva (2023); Strauss & Smith (2009); Zhang et al. (2023); Zickar (2020)
Machine Learning ($n = 12$)	Agarwal et al. (2022); Bowman & Dahl (2021); Cowley et al. (2022); Dow et al. (2024); Eckman et al. (2025); Ibrahim et al. (2024); Liao et al. (2021); Reuel et al. (2024); Subramonian et al. (2023); Wang et al. (2023); Xiao et al. (2023); Zhou et al. (2022)

B.2 SYSTEMATIC LITERATURE REVIEW

We conducted a systematic literature review of human baselines in AI evaluations (Page et al., 2021) to identify gaps in baselining methodology. Our review method is the similar to that of Zhao et al. (2025).

We conducted a systematic search for relevant literature. To begin, we queried Google Scholar in December 2024 for articles containing the keywords in Table 3. Our search terms were intentionally broad, as authors use a variety of different language to describe human baselines. Articles were included in the initial sample if they contained in the full text both a human baseline keyword and an AI evaluation keyword.

Table 3: Search terms for systematic literature review of human baselines

Type	**Keywords**
Human Baseline Keywords	"human baseline*" "expert baseline*" "human performance baseline*"
AI Evaluation Keywords	"LLM evaluation*" "AI evaluation*" "NLP evaluation*" "ML evaluation*" "model evaluation*" "LLM benchmark*" "AI benchmark*" "NLP benchmark*" "ML benchmark*" "evaluating LLM*" "evaluation of LLM*" "benchmark LLM" "benchmarking LLMs" "evaluation of AI models"

Google Scholar was chosen as the database of choice due to its comprehensive coverage (Gusenbauer, 2019) and its indexing of the gray literature. We included articles in the gray literature (e.g., preprints) because researchers often post preprints on arXiv prior to formal publication and because a substantial portion of ML literature is published on arXiv, including publications from many industry organizations (Shah Jahan et al., 2021). For instance, arXiv was the source of an overwhelming majority of articles in one recent systematic literature review on bidirectional language models (Shah Jahan et al., 2021).

There is debate in the methodological literature about the use of Google Scholar as a primary database in a systematic literature review. Concerns have been raised about limitations to advanced search capabilities and to the Google Scholar interface (Halevi et al., 2017), lack of precision (Boeker et al., 2013), and lack of coverage (Haddaway et al., 2015). We addressed these limitations as follows:

- To address limitations to advanced search capabilities, we did not use advanced search capabilities beyond the boolean AND and OR operators in search strings, as well as a simple date filter.
- To address interface limitations, we created workarounds by using multiple queries (to avoid the 256 character limit in search strings) and using a bookmarklet to capture reference information. In any case, we generally find that the search capabilities and interface of Google Scholar are an improvement over the search function in arXiv, making queries to Google Scholar preferable to direct queries in arXiv.
- To address limitations in precision, we adopted more stringent inclusion/exclusion criteria to filter our sample (discussed below). Furthermore, our search is necessarily imprecise due to a lack of standardization of terms in describing human baselines in the literature (e.g., we found that some literature described baselines as "human evaluation", which is normally used to describe human annotations of evaluation data).

- To address limitations in coverage, we supplemented our Google Scholar search with other sources. SD queried Elicit for articles containing human baselines, and MB identified evaluation datasets with human baselines used in industry evaluations by scanning the model/system cards of cards of OpenAI o1 (OpenAI et al., 2024), Anthropic's Claude 3.5 Sonnet (Anthropic, 2024a), Meta's Llama 3 (Grattafiori et al., 2024), and Google DeepMind's Gemini 1.5 (Gemini Team Google et al., 2024).[3] Furthermore, the most recent research has found that Google Scholar has significantly expanded its coverage (Gusenbauer, 2019), and another study found that Google Scholar indexed 96% of articles in systematic literature reviews in computer science that were conducted using other databases (Yasin et al., 2020).

Our search process yielded a sample of $n = 397$ articles (378 from Google Scholar, 13 from Elicit, and 6 from industry model/system cards), which were stored in a Google Sheet. KW then scanned the title, abstract, and main text of each article to filter the sample; the inclusion/exclusion criteria used in filtering along with rationales for each criterion are discussed in Tables 4 and 5. As Google Scholar does not always index the most authorative version of articles, KW also cross-referenced DBLP for all articles on preprint servers (including arXiv) to identify the latest version or published version of each preprint. During the coding process, all coders were also made aware of the exclusion criteria in case any invalid articles were inadvertently included for coding. The final number of articles included for analysis was $n = 109$, and these are identified in Table 6.

Table 4: Inclusion criteria for systematic review of human baselines.

Inclusion Criteria	**Rationale**
Article contains an evaluation of a foundation model	• We limited our scope to foundation models in part to make the review practically manageable • No comprehensive guidance exists for human baselines that is specific to the context of foundation models and that accounts for the most recent foundation model literature • Foundation models raise different and somewhat unique considerations for human baselines, and we aimed to narrow in on these specific considerations • Examples of qualifying articles: articles that fine-tuned or used pre-trained large (language or multi-modal) models
Article contains a human baseline (defined in Section 1)	• See exclusion criteria for examples of non-qualifying articles
Article is published in a peer-reviewed venue or is available in the gray literature (e.g., on a preprint server such as arXiv)	• See text for a discussion of arXiv

We then coded each of the included articles per our checklist, with results stored in a Google Sheet. Following the coding strategy in Zhao et al. (2025), a subset of authors each coded the same four articles, discussed results to ensure coding consistency, and refined the checklist items. The remaining articles were then split up for coding between authors. Questions that arose during the final coding process were adjudicated via discussion, and KW cleaned and standardized the final dataset.

Note that many articles conducted human baselines for multiple different datasets. During the coding process, each dataset for which a baseline was conducted was coded separately. During the process of cleaning and analyzing coded data, only baselines that contained key differences in baseline

[3]The date filter in Table 5 was not applied for these articles so that we could capture evaluations that are widely used in practice. Only one article that would have otherwise been excluded was ultimately included in our sample of baselines (Dua et al., 2019).

Table 5: Exclusion criteria for systematic review of human baselines.

Exclusion Criteria	Rationale
AI model being evaluated is not a foundation model	• See inclusion criteria for discussion of rationale • Examples of non-qualifying articles: articles that trained non-general purpose models for specific purposes
Article did not contain a human baseline	• Enforcement of analogous inclusion criteria
Human baseline in the article is not original (e.g., uses real-world data or a human baseline from a pre-existing dataset)	• Articles using real-world data are excluded as it is difficult to make direct comparisons between human and AI performance in such cases, given that the human data was not generated in a controlled laboratory setting • Articles using pre-existing human baseline data are excluded as researchers may fail to adhere to the experimental design of the previous baseline, making comparisons difficult
Article duplicates an item already included in the review	• Prevention of duplicate items • Examples of non-qualifying items: preprint or workshop version of subsequently published work
Article was published before 2020	• Most foundation model evaluation literature was published after 2020 (inclusive)
Article collected data from human annotators but not as a baseline	• Use of human data in non-baseline contexts gives rise to different methodological considerations • Examples of non-qualifying articles: articles using human evaluation (i.e., using human annotators to score or analyze evaluation data), articles collecting human data as ground truth (e.g., using annotations to determine the desired responses to evaluation items)
Article evaluates LLM-as-a-judge, i.e., compares LLM vs. human evaluation of AI models	• LLM-as-a-judge may give rise to highly idiosyncratic methodological considerations
Article is incomplete work or work-in-progress	• Quality control • Examples of non-qualifying articles: articles submitted to venues but not released as preprints (e.g., paper available on OpenReview but not on arXiv; we assume that authors of these articles do not intend to make their papers public), articles submitted to non-archival workshops intended to refine work
Article is a thesis or class work	• Quality control

instrument, construct, sample, or process were retained as distinct baselines; these were formalized as different coding for Q1.5–1.8, Q2.1–2.3, Q2.6–2.7, or Q3.2–3.5. We find four articles which contained more than one distinct baselines by this criteria (Lu et al., 2024; Meister et al., 2024; Castro et al., 2022; Verma et al., 2024), bring the total number of human baselines up to 113.

Table 6: A complete list of the 109 articles included in our systematic review of human baselines. Note that we analyze 113 individual baselines from these articles, as a single article may contain multiple baselines (see explanation in text).

	Articles
Included ($n = 109$)	Abdibayev et al. (2021); Akhtar et al. (2024); Albrecht et al. (2022); Alex et al. (2021); Asami & Sugawara (2023); Asiedu et al. (2025); Awal et al. (2025); Bai et al. (2024); Blinov et al. (2022); Bu et al. (2024); Castro et al. (2022); Chang et al. (2024a; 2023); Chen et al. (2024); Chhun et al. (2024); Chiu et al. (2024); Chiyah-Garcia et al. (2024); Costarelli et al. (2024); Dagli et al. (2024); Dua et al. (2019); Duan et al. (2022); Fenogenova et al. (2024); Fyffe et al. (2024); Gong et al. (2024); Gu et al. (2024); Guo et al. (2024); Gupta et al. (2024); Haan et al. (2024); Hackenburg et al. (2023); Hamotskyi et al. (2024); Heiding et al. (2024); Hendrycks et al. (2021b); Hijazi et al. (2024); Hildebrandt et al. (2024); Hou et al. (2024); Huang et al. (2024); Ivanov (2024); Jain et al. (2023); Ji et al. (2022); Jimenez et al. (2022); Jing et al. (2023); Kodali et al. (2024); Kruk et al. (2024); Lacombe et al. (2023); Laine et al. (2024); Laurent et al. (2024); LeGris et al. (2024); Lei et al. (2024a); Li et al. (2024a;b; 2021; 2025); Lin et al. (2022); Liu et al. (2024a; 2023; 2024b); Lu et al. (2024; 2023); Mangalam et al. (2023); Meister et al. (2024); Mialon et al. (2023); Miller et al. (2020); Mirza et al. (2024); Mizrahi et al. (2020); Montalan et al. (2024); Moskvichev et al. (2023); Mukhopadhyay et al. (2024); Norlund et al. (2021); Obeidat et al. (2024); Phuong et al. (2024); Reese & Smirnova (2024); Rein et al. (2024); Roberts et al. (2024); Ruis et al. (2023); Sakai et al. (2024); Santurkar et al. (2020); Sanyal et al. (2024); Shavrina et al. (2020); Si et al. (2024); Someya & Oseki (2023); Sourati et al. (2024); Sprague et al. (2023); Srivastava et al. (2023); Suvarna et al. (2024); Tahsin Mayeesha et al. (2021); Taktasheva et al. (2022); Tam et al. (2024); Tanzer et al. (2023); Thrush et al. (2024); Valmeekam et al. (2023); Verma et al. (2024); Wadhawan et al. (2024); Webson et al. (2023); Weissweiler et al. (2024); Wijk et al. (2024); Wu et al. (2024; 2023); Xiang et al. (2023); Yin et al. (2024); Yue et al. (2024); Zamecnik et al. (2024); Zerroug et al. (2022); Zhang et al. (2024a;b;c); Zhou & Hong (2024); Zhou et al. (2024); Zhu et al. (2023); Zhuo et al. (2024)

C Appendix: Data Availability

An up-to-date version of our checklist as well as individual annotations from our systematic review of human baselines are available at `https://github.com/kevinlwei/human-baselines`

D Appendix: Alternative Views

We discuss four alternative views to our position below.

Alternative View 1: Human baselines will soon become unnecessary or insufficient for many evaluations as AI models surpass expert human performance Goldstein & Sastry (2024). Human baselines—in addition to other AI baselines—may be useful even if AI models surpass expert human performance. For instance, they can determine the *magnitude* of human vs. AI performance differences, which is important for modeling economic impacts and making for business or policy decisions Eloundou et al. (2023). They can also help researchers understand how cognition and behavioral tendencies differ between humans and AI systems. Additionally, many authors report random baselines, which can assist interpretation of evaluation results; at the very least, a human baseline could serve as a floor for expected performance from foundation models.

Alternative View 2: Existing human baselines or real-world data may be enough to measure progress, even if only approximately. Some existing baselines may meaningfully measure performance, but many are insufficiently rigorous to draw conclusions about the pace of AI progress Tedeschi et al. (2023); Cowley et al. (2022). Moreover, stakeholders may demand additional rigor for evaluations used in, e.g., risk assessments or safety cases Goemans et al. (2024). Secondary data like standardized tests can be useful points of comparison by providing score distributions from large samples, but it may not always exist for desired use cases. Secondary data is also less well-validated for evaluating models: data contamination concerns are common Yao et al. (2024) and models can perform strangely on assessments designed for humans (e.g., Lei et al. 2024b).

Alternative View 3: Implementing all the practices in this checklist is too expensive to be realistic. We agree that collecting high-quality baseline data can be prohibitively costly. Our aim is not to provide hard requirements for authors to follow but rather an instructional framework to help researchers understand the impact of methodological design choices, allowing researchers to judge whether the rigor provided by particular design choices is justified by the marginal cost and by the evaluation's intended use case. Even where researchers decline to opt for more rigorous methods, reporting study details can nevertheless improve transparency and enable external assessments of published baselines. In general, we believe that the value of more rigorous and transparent human baselines is sufficiently high that funders and the ML community should establish more stringent norms for scientific rigor in AI evaluations.

We also believe that researchers can make many low-cost improvements to human baselining methods—for instance, carefully selecting baseline test sets, reporting uncertainty or statistical tests, and avoiding baselines performed by authors previously exposed to baseline items. Cost considerations are also not unique to ML and have been widely acknowledged in, e.g., survey methodology Leeuw (2005); methods in other fields such as phased clinical trials were developed in part to account for budgetary considerations.

Alternative View 4: This framework and checklist may not be appropriate in all cases due to differing needs in human baselines. Our framework and checklist are not meant as one-size-fits-all recommendations; different evaluations and contexts require different baselining methods. Our intention is to provide an informational guide for designing and assessing baselines (see Alternative View 3 and Section 4). Furthermore, we believe that some standardization—common in many other fields Winters et al. (2009)—is nevertheless useful for transparency, replicability, and interpretability of results (see Kapoor et al. 2024).