Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations

Kevin L. Wei [*1 2] **Patricia Paskov** [*1 3] **Sunishchal Dev** [*1 4] **Michael J. Byun** [*1 3] **Anka Reuel** [2 5]
Xavier Roberts-Gaal [2] **Rachel Calcott** [2] **Evie Coxon** [6] **Chinmay Deshpande** [7]

Executive Summary

This paper finds that existing human baselines are neither sufficiently rigorous nor transparent to enable meaningful comparisons of human vs. AI performance. We provide recommendations and a reporting checklist to increase rigor and transparency in human baselines.

Human baselines are reference sets of metrics intended to represent human performance on specific tasks. They are used in AI evaluations to compare human vs. AI performance on evaluation items, adding important context to results and helping inform stakeholders in the broader AI ecosystem (e.g., downstream users, policymakers).

Specifically, this paper makes three contributions:

1. **Methodological recommendations**: Based on a meta-review of the measurement theory and AI evaluation literatures, we provide methodological recommendations for evaluators to build rigorous human baselines in AI evaluations. Recommendations are summarized in Figure I, with more details in Table II.
2. **Reporting checklist**: We provide a reporting checklist for evaluators to increase transparency when publishing human baselines. The full checklist is in Appendix B.
3. **Literature review**: We review 115 human baselines (studies) to identify methodological gaps in existing AI evaluations, and we find substantial shortcomings in the rigor and transparency of existing human baselines. Summary statistics of our review are in Table III, with more statistics and figures in Appendix A.

Maximal rigor may not be possible in all human baselines due to resource limitations. In these cases, we hope to help researchers make informed tradeoffs, discuss/acknowledge methodological limitations, narrow interpretation of results, and transparently report methods and results.

Data, code, and Word/LaTeX versions of our recommendations and reporting checklist are available at: `https://github.com/kevinlwei/human-baselines`.

Readers' Guide[1]

We recommend the following reading strategies for different types of readers:

- **2-minute read**: Read Figure I and Tables I–III.
- **10-minute read**: Read Figure I and Tables I–III. If needed, skim the relevant parts of Section 4 to understand the rationale behind specific recommendations.
- **Stakeholders seeking to assess the quality of human baselines but who are not building human baselines** (e.g., policymakers, AI governance researchers, AI researchers who don't work on evaluations): Read Figure I and Tables I–III. If needed, skim the relevant parts of Section 4 for clarity about the meaning of or the rationale behind specific recommendations. Read Appendix C for examples of better/worse human baselines.
- **AI evaluators building human baselines**: Start with Figure I and Tables I–II. Read Section 4 to understand the rationale behind our recommendations, and read Appendix D if considering an expert human baseline. Use the reporting checklist in Appendix B when writing up results. Optionally, read the discussion in Sections 5 and 6, and read the case studies in Appendix C for examples of better/worse human baselines.

[1]Inspired by Weidinger et al. (2021; 2023); Wei et al. (2024).

[*]Equal contribution [1]Technology & Security Policy Fellow, RAND, Santa Monica, CA, USA. Views, opinions, findings, conclusions, & recommendations contained herein are the authors' alone and not those of RAND or its research sponsors, clients, or grantors. [2]Harvard University, Cambridge, MA, USA [3]Independent [4]Algoverse [5]Stanford University, Stanford, CA, USA [6]Max Planck School of Cognition, Leipzig, Germany [7]Center for Democracy & Technology, Washington, D.C., USA. Correspondence to: Kevin L. Wei <kevinwei@acm.org>.

A version of this paper has been accepted for publication at the 2025 International Conference on Machine Learning with the title "Position: Human Baselines in Model Evaluations Need Rigor and Transparency (With Recommendations & Reporting Checklist)." *Proceedings of the 42^{nd} International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025.

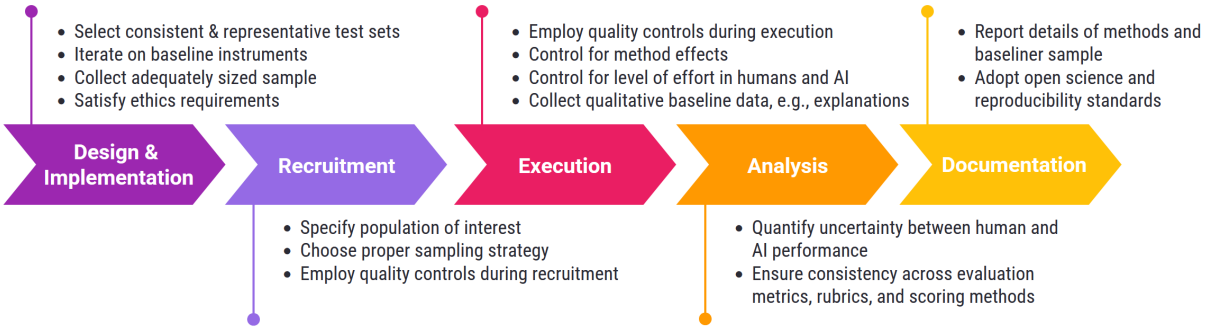

Figure I: A summary of our recommendations for robust and transparent human baselines. Definitions of each stage of the baseline lifecycle are provided in Table I, and more details about our recommendations are provided in Table II. Full recommendations are in Section 4 and full checklist is in Appendix B.

Human Baseline Stage	**Definition**
Baseline Design & Implementation	Baseline design is the initial stage of human baseline development, at which researchers define baselines' purpose, scope, concepts, evaluation items, and metrics; baseline implementation is the selection and construction of tools and datasets for evaluation.
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.
Baseline Execution	Baseline execution is the stage at which the human baseline is conducted and result data is collected—e.g., through surveys or crowdwork platforms.
Baseline Analysis	Baseline analysis is the stage after data collection at which human baseline data is inspected and compared to AI results.
Baseline Documentation	Baseline documentation is the provision of evaluation tasks, datasets, metrics, and experimental materials and resources to relevant audiences.

Table I: Definition of different stages in the human baseline lifecycle.

Recommendation	Details
Baseline Design & Implementation	
Use consistent & representative test sets for human baselines and AI results	• Use the same test set for human baselines and AI results. • If using a subset of the full test set for human baselines, calculate AI results on that subset and make comparisons only on that subset. • If using a subset of the full test set for human baselines, select the subset randomly or stratify on relevant criteria (e.g., based on question difficulty, topic, dataset source, etc.).
Iteratively develop baseline instruments	• Validate, collect feedback, and refine human baseline instruments such as survey questions, instructions, training materials (analogous to refining AI prompts or other materials). • Examples of iterative development processes include (roughly in ascending order of cost/effort): expert validation, pre-tests of baseline instruments, focus groups, and pilot studies.
Collect an adequately sized sample of baseliners	• For generalist (non-expert) baselines, conduct a statistical power analysis to ensure that your baseliner sample size is sufficiently large to represent the human population of interest for your baseline. • A rule of thumb is that a sample of 1,000 respondents is needed to represent the U.S. adult population • If baseliner samples must be smaller due to resource limitations, consider: 1) narrowing population of interest, 2) calculating and reporting the required sample size to reliably detect effects (even if researchers are unable to collect a sample of that size), and 3) narrowing interpretations of baseline results. • For expert baselines, see discussion in Appendix D.
Satisfy ethics requirements for human subjects research	• Ensure ethics requirements are followed, e.g., collecting informed consent, ethics (IRB) review, and other human subjects protections. • Report which ethics requirements were necessary/satisfied for your baseline.
Baseliner Recruitment	
Specify a human population of interest	• Specify which subset of humans the baseline is intended to represent. • Define your population using dimensions such as geographic location, demographic characteristics (e.g., age, gender, socioeconomic status), language, cultural background, education, or domain expertise. • Narrow your population of interest so that you aren't attempting to estimate metrics for all humans.
Use an appropriate sampling strategy for selecting baseliners	• For generalist (non-expert) baselines, random sampling is ideal, but crowdwork samples are often the norm due to cost considerations. When using crowdwork samples, consider methodological adjustments to make your sample more representative of the underlying human population of interest—e.g., stratified sampling, survey weights, the "representative sample" option in Prolific, etc. • For expert baselines, convenience samples are acceptable. Clearly define criteria for baseliner eligibility, and consider snowball sampling. See Appendix D for additional discussion and case studies in Appendix C.

Table II: Methodological recommendations for rigorous human baselines (cont'd on next page)

Recommendation	Details
Baseliner Recruitment (Cont'd)	
Employ quality controls for baseliner recruitment	• Use inclusion/exclusion criteria for baseliners to ensure data quality. • Consider using pre-qualification tests or screening questions, quality scores (on crowdwork platforms), and excluding baseliners previously exposed to evaluation items. • Consider excluding the authors or other members of the research team as baseliners if they have been previously exposed to evaluation items (or would otherwise bias results).
Baseline Execution	
Employ quality controls during baseline execution	• Exclude unreliable baseliner responses to improve data quality—e.g., filtering baseliner responses by outliers, time to completion, or other paradata. • Consider including attention checks, comprehension/manipulation checks, consistency checks, or honeypot questions during the baseline itself. • If baseliners aren't supposed to use AI tools during baselining, consider including instructions asking participants not to use AI tools, employing technical restrictions such as preventing copy/pasting, using non-standard interface elements, or using comprehension/manipulation checks.
Control for method effects and use identical tasks	• Method effects are variations in item response attributable to data collection methods rather than to differences in underlying response distributions (e.g., due to instructions, option order, or mode of data collection). • Use the same tasks for both human and AI responses (i.e., identical instructions, examples, context, etc.). • Randomize question order, response option order, and other non-critical methodological details. • Some method effects may be inevitable due to differences between human and AI cognition; consider clearly documenting methodological details and discussing these limitations. See additional discussion in Section 4.3.
Control for level of effort	• Make fair comparisons by comparing human and AI results at similar levels of effort—e.g., when given the same amount of time to solve a task, or at similar levels of financial cost. • AI effort can be affected by inference cost, task time limits, sampling or elicitation strategy, and other factors. • Human baseliner effort can be affected by training, compensation, task time limits, and other factors.
Collect qualitative data from baseliners	• Collect qualitative data from baseliners to help surface new insights, e.g., failure modes. • Qualitative data can include baseliner explanations (about why they chose particular responses), task trajectories, etc. • Qualitative data may not be necessary for all baselines, especially since collecting such data may increase the cost of baselines.

Methodological recommendations for rigorous human baselines (Table II, cont'd)

Recommendation	Details
Baseline Analysis	
Quantify uncertainty in human vs. AI performance differences	• Report measurements of uncertainty rather than just point estimates—e.g., statistical tests, interval estimates, or distributions of performance.
Use consistent evaluation metrics, scoring methods, and rubrics across human and AI evaluation	• Use the same evaluation metrics, scoring methods, and scoring rubrics for both human and AI results. • When using aggregate metrics such as pass@k, majority vote, etc., consider using the same aggregate metrics for both human and AI performance (or explaining why it may be appropriate to use different metrics).
Baseline Documentation	
Report key details about baselining methodology and baseliners	• Report information about baseliners, baselining procedures, and baseline paradata. These details are important for interpreting and assessing baseline results. • Consider using the reporting checklist we provide in Appendix B.
Adopt best practices for open science and reproducibility/replicability	• Where possible, release (anonymized) human baseline data, experimental materials such as forms or custom UIs, and analysis code. • Releasing such data helps validate research and may promote re-use of your human baseline in future work.

Methodological recommendations for rigorous human baselines (Table II, cont'd)

Question	All Baselines ($n = 115$)			Model Card Baselines ($n = 7$)		
	Yes	No	Unknown	Yes	No	Unknown
Baseline Design & Implementation						
Test Set Equivalence: Were human and AI test sets identical? (Default: No)	59.13% 68	40.87% 47		57.14% 4	42.86% 3	
1.6 **Iterative Design**: Was the experimental setup of the baseline iteratively designed with participatory methods?	34.78% 40	12.17% 14	63.04% 61	42.86% 3	28.57% 2	28.57% 2
1.7 **Amount of Effort**: Does the baseline control for the amount of effort by human baseliners and AIs	13.91% 16	35.65% 41	50.43% 58	28.57% 2	57.14% 4	14.29 1
1.8 **Power Analysis**: Did the authors conduct power analysis in order to determine baseline size? (Default: No)	1.74% 2	98.26% 113		0.00% 0	100.00% 7	
1.9 **Ethics Review**: Was the study approved or exempted by an IRB, or did it undergo other ethics review?	13.91% 16	2.61% 3	83.48% 96	0.00% 0	14.29% 1	85.71% 6
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? (Default: No)	42.61% 49	57.39% 66		57.14% 4	42.86% 3	
2.3 **Quality Control in Recruitment**: Were human baseliners pre-qualified or excluded during the recruitment process for any reason? (Default: Yes)	28.70% 33	71.30% 82		28.57% 2	71.43% 5	
Baseline Execution						
3.2 **Quality Control in Execution**: Were quality checks implemented or data cleaned/excluded during the data collection process (i.e., after baseliners were recruited)? (Default: No)	23.48% 27	76.52% 88		28.57% 2	71.43% 5	
3.4 **Instruction Equivalence**: Did the human baseliners and AIs have access to the same instructions/prompt/question for each item? (Default: No)	24.35% 28	75.65% 87		14.29% 1	85.71% 6	
3.5 **Tool Access Equivalence**: Did the human baseliners and AIs have access to the same (technical) tools for each item? (Default: Yes)	89.57% 103	10.43% 12		71.43% 5	28.57% 2	
3.6 **Explanations**: Did the eval/baseline collect explanations from the human baseliners, after the evaluation was conducted? (Default: No)	11.30% 13	97.39% 112		42.86% 3	57.14% 4	
Baseline Analysis						
4.1 **Statistical Significance**: Did the eval test for statistically significant differences between AI and human performance? (Default: No)	8.70% 10	91.30% 105		0.00% 0	100.00% 7	
4.2 **Uncertainty Estimate**: Did the paper present a measure of uncertainty for the AI and human baseline results? (Default: No)	33.04% 38	66.96% 77		14.29% 1	85.71% 7	
4.3 **Evaluation Metric Equivalence**: Was the same evaluation metric measured/compared for both humans and AIs? (Default: Yes)	93.91% 108	6.09% 7		100.00% 7	0.00% 0	
Baseline Documentation						
5.1.1 **Reporting Sample Demographics**: Were demographics for human baseliners, e.g., race, gender, etc. reported? (Default: No)	22.61% 26	77.39% 89		28.57% 2	71.43% 5	
5.2 **Baseline Data Availability**: Is the (anonymized) human baseline data publicly available? (Default: No)	21.74 % 25	78.26% 90		28.57% 2	71.43% 5	

Table III: Summary statistics from our literature review of 115 existing human baselines, with the same results on a subset of 7 evaluations commonly used in industry model cards (MMMU, GPQA, MATH, DROP, ARC, CONCEPTARC, and EGOSCHEMA). "Unknown" values mean that the item was not reported in text; rows without unknowns are imputed to default values (specified per-question above). More figures and statistics are in Appendix A.

Contents

List of Figures

List of Tables

1. Introduction

Artificial intelligence (AI) systems, 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 used currently to assess human performance on a wide array of AI evaluation tasks (including reasoning, coding, visual perception, etc.) 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; 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 recommendations for producing more rigorous human baselines. We also synthesize our recommendations into a reporting checklist, which we use to systematically review 115 published human baselines, finding substantial shortcomings in existing human baselining methods. We hope that our recommendations and reporting checklist can support researchers in developing and documenting human baselines that are more interpretable and valuable to the ML community, downstream users, and policymakers.

In defending our position, we believe evaluators should be expected to conduct baselines with significant rigor to enable performance comparisons. However, we acknowledge that there are often barriers to rigor, including the expense of high-quality baseline data, the evolving evaluation landscape, and differences in cognition and interaction modes between humans and AI systems. Where maximal rigor is infeasible, evaluators should discuss limitations and narrow their interpretations of baseline comparisons. Our work highlights some of these limitations where applicable and aims to support evaluators in making conscious decisions about tradeoffs between experimental rigor and practical considerations such as cost and efficiency.

We proceed to discuss background in Section 2. Section 3 describes our methodology (details in Appendix E). Section 4 presents our recommendations (full reporting checklist in Appendix B) and results of our systematic review, which examines the entire lifecycle of human baselines: baseline(r) design, recruitment, execution, analysis, and documentation. Section 5 contains discussion and limitations, and Section 6 surveys alternative views. Section 7 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). Concepts are often multidimensional or impossible to measure directly, so 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 like 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 many measurements) (Bandalos, 2018; Salaudeen et al., 2025).

There has been growing recognition in the AI research community that AI evaluation can learn 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; Saxon et al., 2024). Like measurement theory, AI evalua-

tion 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). Recent 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*.

Analysis of the full pipeline of human baselining methods in the foundation model context is limited. Cowley et al. (2022) adapts best practices from psychology to human baselines in computer vision studies but does not examine the entire baseline lifecycle or provide operational-level recommendations. Tedeschi et al. (2023) critiques baseline practices as part of a larger commentary on LLMs but does not examine related disciplines. And research in human-computer interaction methods has similarly been concerned with concept measurement (Lazar et al., 2017), though rarely with human baselines in particular. Building on this literature, we draw on measurement theory across the social sciences to both conceptual and operational methodological recommendations for human baselining in the context of foundation models. 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 best practices for baselining (Appendix B). Using purposive sampling and backwards snowballing, we identified 29 articles from the social sciences (psychology, economics, political science, education) and AI evaluation. We synthesized these recommendations into a more detailed reporting checklist, which was initially compiled after reviewing these articles and later refined through internal discussion and expert validation.

Second, using our reporting checklist, 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 115 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.). "Baselines" from observational data were excluded. Articles were then manually coded and validated per the checklist from our meta-review (Appendix B), with the codebook iteratively refined during the coding process. Full methodological details are in Appendix E.

4. A Framework for Rigorous and Transparent Human Baselines

In this section, we provide high-level recommendations for conducting human baselines. We organize our discussion by delineating five stages of the baselining process, as adapted from Reuel et al. (2024) and Paskov et al. (2025): design, recruitment, execution, analysis, and documentation. We summarize these stages and recommendations in Figure 1, and we examine both positive and negative examples of human baseline studies in Appendix C.

We also discuss results of our systematic review. Appendix B has the full reporting checklist used in our review, and Appendix A contains select per-question summary statistics. Appendix F has additional resources and practical guidance.

4.1. Baseline Design & Implementation

Baseline design is the initial stage of human baseline development, at which researchers define baselines' purpose, scope, concepts, evaluation items, and metrics; baseline implementation is the selection and construction of tools and datasets for evaluation (Reuel et al., 2024; Paskov et al., 2025). We examine four considerations for this stage.

Use consistent & representative test sets for human baselines and AI results. Robust comparisons of human vs. AI performance require comparing performance on the same test set. Where the human baseline's test set is a subset of the AI test set, performance comparisons should only be made on the subset, and the subset should also be representative of the underlying set. Of the baselines in our review, 41% used different test sets for AI vs. human baselines.

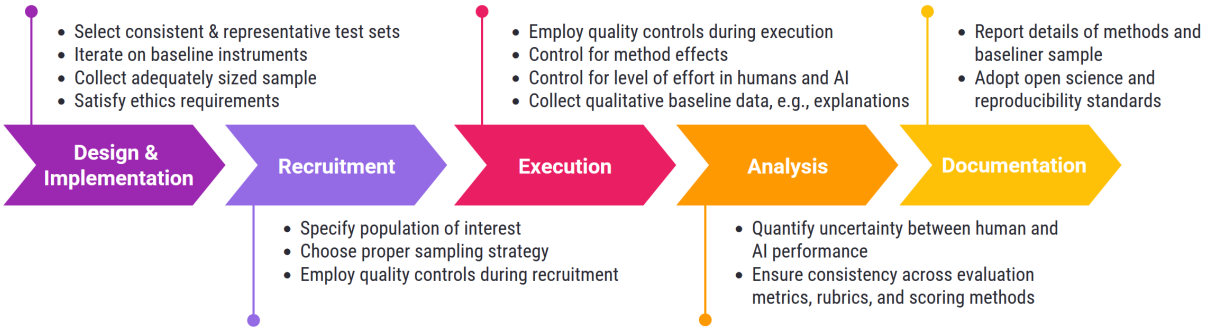

Figure 1: A summary of our recommendations for robust and transparent human baselines. Full recommendations in Section 4 and full checklist in Appendix B.

Because the cost of human baselines can make baselining on a large dataset 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. Simple random sampling from the broader evaluation dataset may be sufficient to ensure representativeness of the baseline test set when 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; see also Siska et al. 2024). Baseline test sets (where specified and distinct from AI test sets) were most commonly created using simple random (43%), stratified (38%), or purposeful sampling strategies (6%).

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

Iteratively develop baseline instruments. Iterative processes repeatedly test and refine the measurement instruments (e.g., forms, surveys) by applying multiple rounds of validation, feedback, and refinement of before final data collection. Although researchers often iterate on prompts for AI systems, only 35% of baselines reviewed reported iteratively developing human baseline instruments.

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 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 also 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 could still improve measurement instruments (Groves et al., 2011; Zickar, 2020).

Collect 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 across a population (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 (McNulty, 2021; Cohen, 2013). The importance of statistical power in ML benchmarks has been noted in prior work (Card et al., 2020; Bowman & Dahl, 2021; Grosse-Holz & Jorgensen, 2024; Beyer et al., 2025), but only 2% of the human baselines we reviewed reported conducting power analyses. A rule of thumb is that a sample size of 1,000 is needed to represent the population of U.S. adults with a reasonable margin of error (Gelman, 2004). By this standard, baselines are vastly underpowered since the median sample size in our review was 8 (mean 90, though with high variance).

If sample sizes are fixed (e.g., due to cost constraints), researchers can nevertheless calculate and report the required

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

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.

Satisfy ethics requirements for human subjects research. Ethics requirements—such as ethics review and collecting informed consent—protect human research participants (Page & Nyeboer, 2017); ethics review is legally required in many jurisdictions, including the U.S. (U.S. Department of Homeland Security et al., 2017). Significantly, only 14% 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) (see also McKee 2024).

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 to comply, 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, e.g., 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.

Specify a human population of interest. Specifying a population of interest—the group of humans for whom a baseline is intended to be representative—is necessary to interpret *which* group of humans a baseline represents. Defining the population is important since population sizes can affect sampling reliability and statistical power, and researchers can choose more targeted populations to save on sample size (though cost of targeting may also increase). Prior work has noted that AI evaluations rarely specify populations of interest (Subramonian et al., 2023), which is in line our review: only 43% baselines explicitly or implicitly defined a population of interest along at least one axis (i.e., a population beyond "humans," which is too large for most baselines to represent).

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). Of baselines in our review, among AI evaluations that defined a population of interest, the most commonly reported characteristics were education (21%), language (19%), age (19%), and expertise (18%). How to scope the population of interest for any given baseline will depend on the evaluation's research questions, context, and intended use.

Use an appropriate sampling strategy for selecting baseliners. Sampling strategy–the methodology by which baseliners are selected–directly informs how representative the baseliner sample is of the population of interest. Representativeness is essential for external validity because it determines whether baseliners' results can be generalized to that underlying population (Findley et al., 2021; Stantcheva, 2023; Berinsky, 2017; Lohr, 2022; Valliant et al., 2018). In our review, 31% of human baselines used a convenience sample, 32% recruited from crowdsourcing platforms, none used a random sample, and 37% did not report sampling strategies.

Random samples are ideal when baselines are meant to mirror broad human populations (e.g., generalist baselines), since other sampling strategies like convenience sampling are susceptible to significant biases that reduce generalizability (cf. Mihalcea et al. 2024; Diaz & Smith 2024; Brown 2023). Most generalist baselines are conducted through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) or Prolific, which are not random samples and can pose challenges to representativeness. Crowdsourced samples could be demographically biased—MTurk workers tend to more educated, politically liberal, online, and younger than the general population (Sheehan, 2018; Shaw & Hargittai, 2021; Stantcheva, 2023)—or biased due to expertise if crowdworkers have been exposed to extensive AI evaluation/training tasks.[3] Crowdsourced baselines may thus fail to represent performance of humans not of those demographics; even very large samples may be biased if insufficiently representative of the population of interest (Bradley et al., 2021). In expert baselines, however, convenience sampling may be justified because expert populations can be very small (see discussion in Appendix D).

When random sampling is infeasible, as is often the case due to cost, researchers designing human baselines can consider methodological adjustments to improve representativeness. For instance, stratified sampling can improve representativeness along specific dimensions (Groves et al., 2011), and researchers building generalist baselines on Prolific can consider the "representative sample" option (Prolific,

[3]Prolific samples are less well-studied than MTurk samples, so less is known about their representativeness.

2025).[4] Post hoc adjustments such as weighting (which may require collecting baseliners' demographic information) or other sampling adjustments may also partially mitigate selection bias in non-representative samples (Solon et al., 2015; Couper, 2017; Valliant et al., 2018).

At a minimum, researchers should report their sampling strategies, acknowledge sampling limitations, discuss which populations results may generalize to, and discuss implications for the validity of baseline results.

Employ 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 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). Of the baselines in our review, 29% reported using quality control measures for recruitment, with most of these using pre-testing such as qualification tests or thresholds.

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). Expert 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 Execution

Baseline execution is the stage at which the human baseline is conducted and result data is collected (Paskov et al., 2025)—e.g., through surveys or crowdwork platforms. We examine four considerations for baseline execution.

Employ quality controls during baseline execution. QC at the execution stage improves data quality by filtering out unreliable baseline responses. As in the recruitment stage, QC during execution 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, only 23% reported performing QC at the execution 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 more than a third of MTurk and Prolific workers have used AI to complete tasks (Veselovsky et al., 2023b; Zhang et al., 2025; 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 to prevent AI usage may be beneficial for crowdsourced baselines: mechanisms may include 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). Depending on the research question, AI use may be appropriate for baseliners, such as for evaluations in domains where AI usage is expected. In these cases, researchers may carefully define protocols for AI use and evaluations can compare AI capabilities with baselines of AI-augmented human capabilities (e.g., Wijk et al. 2024).

Control for method effects and use identical tasks. Method effects are variations in item response attributable to data collection methods rather than to differences in underlying response distributions (e.g., due to instructions, option order, or mode of data collection). Method effects can reduce the internal validity of evaluations (Davidov et al., 2014). Evaluators should control for method effects wherever possible, and AI and human results should use the same tasks (i.e., identical instructions, context, etc.). Our review revealed significant discrepancies in data collection methods between humans and AI models. Of the baselines in our review, 88% displayed UI differences between human baselining and AI evaluation, 76% displayed differences in instructions or prompts, and 10% displayed differences in tool access.[5]

Method effects are well-documented in the social sciences, particularly in psychology and in survey methodology. Em-

[4] This option helps with representativeness *as long as the axes on which Prolific samples are also those that define population of interest*. As of this writing, MTurk has no comparable option.

[5] 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).

pirical research has found effects in humans 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). AI systems are also subject to method effects such as prompt sensitivity and other biases (Anagnostidis & Bulian, 2024; Ye et al., 2024).

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.

Some method effects in AI evaluation, however, are currently inevitable due to differences between human and AI cognition (cf. McCoy et al. 2024); evaluators should discuss these limitations where they could significantly affect results. 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 25% of reviewed baselines reported instrument length, of which most reported instruments 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 even using the same data collection artifacts for humans and for AI systems may not prevent all method effects.

Without clear evidence, we suggest for now that evaluators default to using identical setups for AI and human evaluation, including provision of identical instructions, examples, context, role information, and other supplementary materials or details that could affect performance (e.g., documentation, images).[6] Researchers should also clearly document evaluation methodologies and differences in measurement instruments between AI and human results. 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 measurement instruments 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).

Control for level of effort. Both humans' and AI systems' level of effort in responding to items can affect evaluation results. For AI systems, "effort" could be proxied by inference cost, task time limits, sampling or elicitation strategy, and other factors; analogously, baseliner effort can be affected by training, compensation, task time limits, and other factors (Tedeschi et al., 2023; Kapoor et al., 2024b). 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. Of the baselines in our review, 23% provided training to baseliners, and 41% reported paying baseliners, with 8% providing performance bonuses.

Accounting for level of effort also raises design questions about the choice of the experimental unit of interest, 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 vs. 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.

Collect qualitative data from baseliners. Qualitative data from baseliners—including but not limited to explanations of why baseliners chose particular responses—may help interpret differences in human and AI performance, explain performance gaps, and surface validity issues. Collecting explanations is generally a best practice in survey research, as it can help surface new insights (Lu et al., 2022a); 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% 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 not all baselines need explanations.

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.

Quantify uncertainty in human vs. AI performance differences. Reporting measurements of uncertainty, such as result distributions or 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; Ying et al., 2025). 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; Marie et al., 2021). Similarly, our review finds that only 37% of evaluations provided

[6]These details may be particularly important to monitor when using validation data as a baseline (e.g., GPQA (Rein et al., 2024)).

interval or distribution estimates, and only 8% performed statistical tests of any type.

Lack of statistical testing is sometimes understandable given small sample sizes and other limitations (cf. Bouthillier et al. 2021).[7] Reporting interval estimates, however, has become increasingly accessible with increased guidance (e.g., Miller 2024; Bowyer et al. 2025) and support in major evaluation frameworks (e.g., UK AISI 2025). 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).

Use 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); 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 to relevant audiences (Reuel et al., 2024). We examine two considerations for baseline documentation.

Report key details about baselining methodology and baseliners. Documentation includes reporting information about baseliners, baselining procedures, and baseline paradata. Documenting methodology in particular is crucial to enable reproducibility/replicability and external assessments of baseline results. These details can significantly affect how results are contextualized, interpreted, and operationalized—especially with respect to their validity—and reporting can build collective confidence in published results (Liao et al., 2021; Biderman et al., 2024).

Absent compelling reasons for confidentiality, researchers should document most of the items in our checklist that are related to baseline(r) design, recruitment, execution, and analysis (Appendix B; see also McKee 2024). 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; West, 2011) 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 23% provided detailed baseliner demographics, and only 21% included paradata such as response times.

Adopt 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); annotator-level data is also important to gain a fuller picture of baseliner performance (Prabhakaran et al., 2021). In addition, these open science practices facilitate reuse of human baseline data in subsequent evaluations by other researchers, which in turn fosters more efficient use of resources within the ML community. Concerns around open science and replicability are not new in ML (Kapoor et al., 2024a; Pineau et al., 2021), and our review found that most baselines (78%) did not publicly release human baseline responses, experimental materials (56%), and code for analyzing human baselines (59%).

5. Discussion

In this section, we discuss three additional considerations for human baselines and address the limitations of our study.

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; Laine et al., 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 observational/real-world data or pre-existing datasets. For instance, the Massive Multitask Language Understanding dataset uses the 95$^{\text{th}}$ percentile of human standardized test

[7] Addressing the small sample size challenge is an ongoing area of research (Xiao et al., 2025; Luettgau et al., 2025). See generally, Neuhäuser & Ruxton 2024; Hoyle 1999; Schoot & Miočević 2020.

scores as a point of comparison with AI results (Hendrycks et al., 2020); 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 Section 6).

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 may be particularly short due to the rate of progress in AI. These trends suggest that human baselines should be interpreted as 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 followed best practices for systematic literature reviews, but our meta-review sample was collected purposively and could be biased as a result (see Appendix E.1). Our scope is limited to methodological considerations specific to human baselines, so we do not discuss in depth 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 different methodological questions. Finally, future research can examine applications of measurement theory to human evaluation and human-AI interaction studies, which are not explored in this position paper (e.g. human uplift, LLM-as-a-judge).

6. Alternative Views

We discuss four alternative views to our position below.

Alternative View 1: Implementing all these recommendations is too expensive to be realistic. We believe that researchers have a responsibility to ensure experimental rigor and reasonable interpretation of results; however, we also acknowledge that collecting high-quality baseline data can be prohibitively costly. Our hope is not that *all* baselines will be maximally rigorous but rather that 1) *all* baselines should be transparent even if not maximally rigorous; and 2) *some* baselines should be both transparent and maximally rigorous. Our framework is intended 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. Where researchers decline to opt for more rigorous methods, reporting study details is nevertheless important for transparency and can enable external assessments of published baselines. By narrowing interpretations of less rigorous baseline results and discussing limitations, researchers can also prevent readers from misunderstanding or over-hyping results.

Moreover, we believe that many low-cost improvements can be made to existing human baselining methods, even though these improvements may not suffice for maximal rigor: e.g., using consistent test sets, satisfying ethics requirements, specifying a population of interest, using recruitment/execution QC such as excluding authors/baseliners previously exposed to evaluation items, reporting uncertainty or statistical tests, and documenting results. We also note that cost considerations are not unique to ML and have been widely acknowledged in, e.g., survey methodology (Leeuw, 2005); ML can learn from methods in other fields such as phased clinical trials that were developed in part to account for cost. ML researchers can also collaborate with social scientists to reduce the administrative burden of gaining new expertise in measurement theory.

Finally, some baselines will need to be both transparent and maximally rigorous for use in risk management and AI governance. 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. The community should also encourage and accept individual baseline results such as Nangia & Bowman (2019) or LeGris et al. (2024) as substantial and stand-alone technical contributions.

Alternative View 2: Human baselines will soon become unnecessary or insufficient for many evaluations as AI systems surpass expert human performance (Goldstein & Sastry, 2024). Human baselines—in addition to other AI baselines—may be useful even if systems 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 for making 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. At the very least, a human baseline could serve as a floor for expected performance from foundation models, similar to random baselines currently. Moreover, we note that existing methods using AI to simulate human participants are subject to substantial limitations (Wang et al., 2025; Liu et al., 2025; Anthis et al., 2025), so current AI systems are unlikely to

be able to simulate human baseliners validly and reliably in many contexts.

Alternative View 3: 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 4: This framework and checklist may not be appropriate in all cases due to differing needs in human baselines. We agree that different evaluations and contexts require different methods. Our intention is to provide a starting point for designing and assessing baselines, not necessarily a one-size-fits-all solution. Furthermore, we believe that *some* standardization—common in other fields (Winters et al., 2009)—is useful for transparency, replicability, and interpretability of results (see Kapoor et al. 2024a).

7. Conclusion

In this position paper, we argue that human baselines in foundation model evaluations should be more rigorous and transparent. Systematically reviewing 115 published evaluations, we find that many baselines lack methodological rigor across the gamut of the baselining process, from design (e.g., different human vs. AI test sets) to documentation (e.g., lack of study details). We provide recommendations for human baselining based on measurement theory to foster validity and reliability, enable meaningful comparisons of human vs. AI performance, and promote research transparency. We hope that our work can guide researchers in improving baselining methods and evaluating AI systems.

Acknowledgements

We are extremely grateful for comments and feedback on drafts of this paper from (in randomized order): Ella Guest, three anonymous reviewers from the ICLR Workshop on Building Trust in LLMs and LLM Applications, Laura Weidinger, Michael Chen, Megan Kinniment, Fred Heiding, four anonymous ICML reviewers, Marius Hobbhahn, Christopher Summerfield, Ryan Ritterson, Lujain Ibrahim, Jacy Reese Anthis, Jacob Haimes, and Sarah Gebauer. The highly detailed comments from each of these readers have substantially improved the quality—and, we hope, the impact—of this work, and we are thankful for the time that they took to engage with our research. In addition, Anka Reuel acknowledges support from the Stanford Interdisciplinary Graduate Fellowship.

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

Abdibayev, A., Riddell, A., and Rockmore, D. BPoMP: The Benchmark of Poetic Minimal Pairs – Limericks, Rhyme, and Narrative Coherence. In Mitkov, R. and Angelova, G. (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/.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A., and Bellemare, M. G. Deep Reinforcement Learning at the Edge of the Statistical Precipice, January 2022. URL http://arxiv.org/abs/2108.13264. arXiv:2108.13264.

Akhtar, M., Subedi, N., Gupta, V., Tahmasebi, S., Cocarascu, O., and Simperl, E. ChartCheck: Explainable Fact-Checking over Real-World Chart Images. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Albrecht, J., Kitanidis, E., and Fetterman, A. J. 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].

Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C., Hine, E., Ashurst, C., Sedille, P., Carlier, A., Noetel, M., and Stuhlmüller, A. 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.

Ames, H., Glenton, C., and Lewin, S. 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.

Anagnostidis, S. and Bulian, J. How Susceptible are LLMs to Influence in Prompts? In *Proceedings of the First Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=y7JnjDcIQa#discussion.

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.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., and Bernstein, M. LLM Social Simulations Are a Promising Research Method, April 2025. URL http://arxiv.org/abs/2504.02234. arXiv:2504.02234 [cs].

Anthropic. Claude 3.5 Sonnet Model Card Addendum, June 2024a. URL https://perma.cc/5S2B-SLVW.

Anthropic. Developing a computer use model, October 2024b. URL https://perma.cc/ES79-73B8.

Asami, D. and Sugawara, S. PROPRES: Investigating the Projectivity of Presupposition with Various Triggers and Environments. In Jiang, J., Reitter, D., and Deng, S. (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/.

Asiedu, M., Tomasev, N., Ghate, C., Tiyasirichokchai, T., Dieng, A., Akande, O., Siwo, G., Adudans, S., Aitkins, S., Ehiakhamen, O., Ndombi, E., and Heller, K. 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].

Awal, R., Ahmadi, S., Zhang, L., and Agrawal, A. VisMin: Visual Minimal-Change Understanding, January 2025. URL http://arxiv.org/abs/2407.16772. arXiv:2407.16772 [cs].

Bai, L., Borah, A., Ignat, O., and Mihalcea, R. 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].

Bai, L., Borah, A., Ignat, O., and Mihalcea, R. The Power of Many: Multi-Agent Multimodal Models for Cultural Image Captioning. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2970–2993, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.152/.

Bandalos, D. L. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications, January 2018. ISBN 978-1-4625-3213-1.

Barnett, P. and Thiergart, L. 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].

Bender, D. 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.

Berinsky, A. J. 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.

Beyer, T., Xhonneux, S., Geisler, S., Gidel, G., Schwinn, L., and Günnemann, S. LLM-Safety Evaluations Lack Robustness, March 2025. URL http://arxiv.org/abs/2503.02574. arXiv:2503.02574 [cs].

Biderman, S. and Scheirer, W. J. 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.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J. Z., Foster, C., Hsu, J., Jaiswal, M., Lee, W. Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K. A., Winata, G. I., Yvon, F., and Zou, A. Lessons from the Trenches on Reproducible Evaluation of Language Models, May 2024. URL http://arxiv.org/abs/2405.14782. arXiv:2405.14782 [cs].

Bikkasani, D. C. 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.

Blinov, P., Reshetnikova, A., Nesterov, A., Zubkova, G., and Kokh, V. RuMedBench: A Russian Medical Language Understanding Benchmark. In Michalowski, M., Abidi, S. S. R., and Abidi, S. (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.

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Zong, C., Xia, F., Li, W., and Navigli, R. (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/.

Blodgett, S. L., Cheung, J. C. K., Liao, V., and Xiao, Z. Human-Centered Evaluation of Language Technologies. In Li, J. and Liu, F. (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/.

Boeker, M., Vach, W., and Motschall, E. 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.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. v., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C.,

Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Bouckaert, R. R. and Frank, E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In Dai, H., Srikant, R., and Zhang, C. (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.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., and Vincent, P. 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.

Bowman, S. R. and Dahl, G. What Will it Take to Fix Benchmarking in Natural Language Understanding? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (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.

Bowyer, S., Aitchison, L., and Ivanova, D. R. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints, March 2025. URL http://arxiv.org/abs/2503.01747. arXiv:2503.01747 [cs].

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. 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.

Brodeur, P. G., Buckley, T. A., Kanjee, Z., Goh, E., Ling, E. B., Jain, P., Cabral, S., Abdulnour, R.-E., Haimovich, A. D., Freed, J. A., Olson, A., Morgan, D. J., Hom, J., Gallo, R., McCoy, L. G., Mombini, H., Lucas, C., Fotoohi, M., Gwiazdon, M., Restifo, D., Restrepo, D., Horvitz, E., Chen, J., Manrai, A. K., and Rodman, A. Superhuman performance of a large language model on the reasoning tasks of a physician, June 2025. URL http://arxiv.org/abs/2412.10849. arXiv:2412.10849 [cs].

Brown, A. When surveying small populations, some approaches are more inclusive than others, May 2023. URL https://perma.cc/3VL8-425Q.

Bu, F., Zhang, Y., Wang, X., Wang, B., Liu, Q., and Li, H. Roadmap towards Superhuman Speech Understanding using Large Language Models, October 2024. URL http://arxiv.org/abs/2410.13268. arXiv:2410.13268 [cs].

Burden, J., Tešić, M., Pacchiardi, L., and Hernández-Orallo, J. Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture, February 2025. URL http://arxiv.org/abs/2502.15620. arXiv:2502.15620 [cs].

Cai, L., Choi, K., Hansen, M., and Harrell, L. 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.

Cao, Y., Chen, R. C., and Katz, A. J. 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.

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. With Little Power Comes Great Responsibility. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.745. URL https://aclanthology.org/2020.emnlp-main.745/.

Castro, S., Wang, R., Huang, P., Stewart, I., Ignat, O., Liu, N., Stroud, J. C., and Mihalcea, R. 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].

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., and Anderljung, M. Infrastructure for AI Agents. *Transactions on Machine Learning Research*, February 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ckh17xN2R2.

Chang, H.-H., Wang, C., and Zhang, S. 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.

Chang, K., Cramer, M., Soni, S., and Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In Bouamor, H., Pino, J., and Bali, K. (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/.

Chang, M., Chhablani, G., Clegg, A., Cote, M. D., Desai, R., Hlavac, M., Karashchuk, V., Krantz, J., Mottaghi, R., Parashar, P., Patki, S., Prasad, I., Puig, X., Rai, A., Ramrakhya, R., Tran, D., Truong, J., Turner, J. M., Undersander, E., and Yang, T.-Y. 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].

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. 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.

Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., Cao, Y., Hu, M., Lai, Y., Xiong, Z., and Huang, M. ToMBench: Benchmarking Theory of Mind in Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

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

Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., and Choi, Y. 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].

Chiyah-Garcia, J., Suglia, A., and Eshghi, A. Repairs in a Block World: A New Benchmark for Handling User Corrections with Multi-Modal Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Chouldechova, A., Atalla, C., Barocas, S., Cooper, A. F., Corvi, E., Dow, P. A., Garcia-Gathright, J., Pangakis, N., Reed, S., Sheng, E., Vann, D., Vogel, M., Washington, H., and Wallach, H. 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].

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

Costarelli, A., Allen, M., Hauksson, R., Sodunke, G., Hariharan, S., Cheng, C., Li, W., Clymer, J., and Yadav, A. GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents, July 2024. URL http://arxiv.org/abs/2406.06613. arXiv:2406.06613 [cs].

Couper, M. P. 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.

Cowley, H. P., Natter, M., Gray-Roncal, K., Rhodes, R. E., Johnson, E. C., Drenkow, N., Shead, T. M., Chance, F. S., Wester, B., and Gray-Roncal, W. 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/s415

98-022-08078-3. Publisher: Nature Publishing Group.

Dafoe, A., Zhang, B., and Caughey, D. 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.

Dagli, R., Berger, G., Materzynska, J., Bax, I., and Memisevic, R. AirLetters: An Open Video Dataset of Characters Drawn in the Air, October 2024. URL http://arxiv.org/abs/2410.02921. arXiv:2410.02921 [cs].

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., and Billiet, J. 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.

de Haan, T., Ting, Y.-S., Ghosal, T., Nguyen, T. D., Accomazzi, A., Wells, A., Ramachandra, N., Pan, R., and Sun, Z. 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] version: 1.

de Haan, T., Ting, Y.-S., Ghosal, T., Nguyen, T. D., Accomazzi, A., Wells, A., Ramachandra, N., Pan, R., and Sun, Z. Achieving GPT-4o level performance in astronomy with a specialized 8B-parameter large language model. *Scientific Reports*, 15(1):13751, April 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-97131-y. URL https://www.nature.com/articles/s41598-025-97131-y. Publisher: Nature Publishing Group.

Deary, I. J. 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.

Diaz, M. and Smith, A. D. R. 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.

Dietterich, T. G. 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.

Dow, P. A., Vaughan, J. W., Barocas, S., Atalla, C., Chouldechova, A., and Wallach, H. Dimensions of Generative AI Evaluation Design, November 2024. URL http://arxiv.org/abs/2411.12709. arXiv:2411.12709 [cs].

Drasgow, F., Chernyshenko, O. S., and Stark, S. 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.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In Burstein, J., Doran, C., and Solorio, T. (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/.

Duan, J., Yu, S., Poria, S., Wen, B., and Tan, C. PIP: Physical Interaction Prediction via Mental Simulation with Span Selection. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (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.

Eckman, S., Plank, B., and Kreuter, F. 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.

Elicit. Elicit: The AI Research Assistant. URL https://elicit.com/welcome.

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. 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.

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

Fenogenova, A., Chervyakov, A., Martynov, N., Kozlova, A., Tikhonova, M., Akhmetgareeva, A., Emelyanov, A., Shevelev, D., Lebedev, P., Sinev, L., Isaeva, U., Kolomeytseva, K., Moskovskiy, D., Goncharova, E., Savushkin, N., Mikhailova, P., Minaeva, A., Dimitrov, D., Panchenko, A., and Markov, S. MERA: A Comprehensive LLM Evaluation in Russian. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Findley, M. G., Kikuta, K., and Denly, M. 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.

Flach, P. 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.

Frank, M. C., Braginsky, M., Cachia, J., Coles, N., and Hardwicke, T. E. *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/.

Fyffe, S., Lee, P., and Kaplan, S. "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.

Gao, S., Wang, C., Gao, C., Jiao, X., Chong, C. Y., Gao, S., and Lyu, M. 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].

Gelman, A. How can a poll of only 1,004 Americans represent 260 million people with only a 3 percent margin of error? *Scientific American*, March 2004. URL https://perma.cc/D28G-PJ44.

Gelman, A. and Stern, H. 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.

Gemini Team Google, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., Gaffney, C., Daruki, S., Sercinoglu, O., Gleicher, Z., Love, J., Voigtlaender, P., Jain, R., Surita, G., Mohamed, K., Blevins, R., Ahn, J., Zhu, T., Kawintiranon, K., Firat, O., Gu, Y., Zhang, Y., Rahtz, M., Faruqui, M., Clay, N., Gilmer, J., Co-Reyes, J. D., Penchev, I., Zhu, R., Morioka, N., Hui, K., Haridasan, K., Campos, V., Mahdieh, M., Guo, M., Hassan, S., Kilgour, K., Vezer, A., Cheng, H.-T., Liedekerke, R. d., Goyal, S., Barham, P., Strouse, D. J., Noury, S., Adler, J., Sundararajan, M., Vikram, S., Lepikhin, D., Paganini, M., Garcia, X., Yang, F., Valter, D., Trebacz, M., Vodrahalli, K., Asawaroengchai, C., Ring, R., Kalb, N., Soares, L. B., Brahma, S., Steiner, D., Yu, T., Mentzer, F., He, A., Gonzalez, L., Xu, B., Kaufman, R. L., Shafey, L. E., Oh, J., Hennigan, T., Driessche, G. v. d., Odoom, S., Lucic, M., Roelofs, B., Lall, S., Marathe, A., Chan, B., Ontanon, S., He, L., Teplyashin, D., Lai, J., Crone, P., Damoc, B., Ho, L., Riedel, S., Lenc, K., Yeh, C.-K., Chowdhery, A., Xu, Y., Kazemi, M., Amid, E., Petrushkina, A., Swersky, K., Khodaei, A., Chen, G., Larkin, C., Pinto, M., Yan, G., Badia, A. P., Patil, P., Hansen, S., Orr, D., Arnold, S. M. R., Grimstad, J., Dai, A., Douglas, S., Sinha, R., Yadav, V., Chen, X., Gribovskaya, E., Austin, J., Zhao, J., Patel, K., Komarek, P., Austin, S., Borgeaud, S., Friso, L., Goyal, A., Caine, B., Cao, K., Chung, D.-W., Lamm, M., Barth-Maron, G., Kagohara, T., Olszewska, K., Chen, M., Shivakumar, K., Agarwal, R., Godhia, H., Rajwar, R., Snaider, J., Dotiwalla, X., Liu, Y., Barua, A., Ungureanu, V., Zhang, Y., Batsaikhan, B.-O., Wirth, M., Qin, J., Danihelka, I., Doshi, T., Chadwick, M., Chen, J., Jain, S., Le, Q., Kar, A., Gurumurthy, M., Li, C., Sang, R., Liu, F., Lamprou, L., Munoz, R., Lintz, N., Mehta, H., Howard, H., Reynolds, M., Aroyo, L., Wang, Q., Blanco, L., Cassirer, A., Griffith, J., Das, D., Lee, S., Sygnowski, J., Fisher, Z., Besley, J., Powell, R., Ahmed, Z., Paulus, D., Reitter, D., Borsos, Z., Joshi, R., Pope, A., Hand, S., Selo, V., Jain, V., Sethi, N., Goel, M., Makino, T., May, R., Yang, Z., Schalkwyk, J., Butterfield, C., Hauth, A., Goldin, A., Hawkins, W., Senter, E., Brin, S., Woodman,

O., Ritter, M., Noland, E., Giang, M., Bolina, V., Lee, L., Blyth, T., Mackinnon, I., Reid, M., Sarvana, O., Silver, D., Chen, A., Wang, L., Maggiore, L., Chang, O., Attaluri, N., Thornton, G., Chiu, C.-C., Bunyan, O., Levine, N., Chung, T., Eltyshev, E., Si, X., Lillicrap, T., Brady, D., Aggarwal, V., Wu, B., Xu, Y., McIlroy, R., Badola, K., Sandhu, P., Moreira, E., Stokowiec, W., Hemsley, R., Li, D., Tudor, A., Shyam, P., Rahimtoroghi, E., Haykal, S., Sprechmann, P., Zhou, X., Mincu, D., Li, Y., Addanki, R., Krishna, K., Wu, X., Frechette, A., Eyal, M., Dafoe, A., Lacey, D., Whang, J., Avrahami, T., Zhang, Y., Taropa, E., Lin, H., Toyama, D., Rutherford, E., Sano, M., Choe, H., Tomala, A., Safranek-Shrader, C., Kassner, N., Pajarskas, M., Harvey, M., Sechrist, S., Fortunato, M., Lyu, C., Elsayed, G., Kuang, C., Lottes, J., Chu, E., Jia, C., Chen, C.-W., Humphreys, P., Baumli, K., Tao, C., Samuel, R., Santos, C. N. d., Andreassen, A., Rakićević, N., Grewe, D., Kumar, A., Winkler, S., Caton, J., Brock, A., Dalmia, S., Sheahan, H., Barr, I., Miao, Y., Natsev, P., Devlin, J., Behbahani, F., Prost, F., Sun, Y., Myaskovsky, A., Pillai, T. S., Hurt, D., Lazaridou, A., Xiong, X., Zheng, C., Pardo, F., Li, X., Horgan, D., Stanton, J., Ambar, M., Xia, F., Lince, A., Wang, M., Mustafa, B., Webson, A., Lee, H., Anil, R., Wicke, M., Dozat, T., Sinha, A., Piqueras, E., Dabir, E., Upadhyay, S., Boral, A., Hendricks, L. A., Fry, C., Djolonga, J., Su, Y., Walker, J., Labanowski, J., Huang, R., Misra, V., Chen, J., Skerry-Ryan, R. J., Singh, A., Rijhwani, S., Yu, D., Castro-Ros, A., Changpinyo, B., Datta, R., Bagri, S., Hrafnkelsson, A. M., Maggioni, M., Zheng, D., Sulsky, Y., Hou, S., Paine, T. L., Yang, A., Riesa, J., Rogozinska, D., Marcus, D., Badawy, D. E., Zhang, Q., Wang, L., Miller, H., Greer, J., Sjos, L. L., Nova, A., Zen, H., Chaabouni, R., Rosca, M., Jiang, J., Chen, C., Liu, R., Sainath, T., Krikun, M., Polozov, A., Lespiau, J.-B., Newlan, J., Cankara, Z., Kwak, S., Xu, Y., Chen, P., Coenen, A., Meyer, C., Tsihlas, K., Ma, A., Gottweis, J., Xing, J., Gu, C., Miao, J., Frank, C., Cankara, Z., Ganapathy, S., Dasgupta, I., Hughes-Fitt, S., Chen, H., Reid, D., Rong, K., Fan, H., Amersfoort, J. v., Zhuang, V., Cohen, A., Gu, S. S., Mohananey, A., Ilic, A., Tobin, T., Wieting, J., Bortsova, A., Thacker, P., Wang, E., Caveness, E., Chiu, J., Sezener, E., Kaskasoli, A., Baker, S., Millican, K., Elhawaty, M., Aisopos, K., Lebsack, C., Byrd, N., Dai, H., Jia, W., Wiethoff, M., Davoodi, E., Weston, A., Yagati, L., Ahuja, A., Gao, I., Pundak, G., Zhang, S., Azzam, M., Sim, K. C., Caelles, S., Keeling, J., Sharma, A., Swing, A., Li, Y., Liu, C., Bostock, C. G., Bansal, Y., Nado, Z., Anand, A., Lipschultz, J., Karmarkar, A., Proleev, L., Ittycheriah, A., Yeganeh, S. H., Polovets, G., Faust, A., Sun, J., Rrustemi, A., Li, P., Shivanna, R., Liu, J., Welty, C., Lebron, F., Baddepudi, A., Krause, S., Parisotto, E., Soricut, R., Xu, Z., Bloxwich, D., Johnson, M., Neyshabur, B., Mao-Jones, J., Wang, R., Ramasesh, V., Abbas, Z., Guez, A., Segal, C., Nguyen, D. D., Svensson, J., Hou, L., York, S., Milan, K., Bridgers, S., Gworek, W., Tagliasacchi, M., Lee-Thorp, J., Chang, M., Guseynov, A., Hartman, A. J., Kwong, M., Zhao, R., Kashem, S., Cole, E., Miech, A., Tanburn, R., Phuong, M., Pavetic, F., Cevey, S., Comanescu, R., Ives, R., Yang, S., Du, C., Li, B., Zhang, Z., Iinuma, M., Hu, C. H., Roy, A., Bijwadia, S., Zhu, Z., Martins, D., Saputro, R., Gergely, A., Zheng, S., Jia, D., Antonoglou, I., Sadovsky, A., Gu, S., Bi, Y., Andreev, A., Samangooei, S., Khan, M., Kocisky, T., Filos, A., Kumar, C., Bishop, C., Yu, A., Hodkinson, S., Mittal, S., Shah, P., Moufarek, A., Cheng, Y., Bloniarz, A., Lee, J., Pejman, P., Michel, P., Spencer, S., Feinberg, V., Xiong, X., Savinov, N., Smith, C., Shakeri, S., Tran, D., Chesus, M., Bohnet, B., Tucker, G., Glehn, T. v., Muir, C., Mao, Y., Kazawa, H., Slone, A., Soparkar, K., Shrivastava, D., Cobon-Kerr, J., Sharman, M., Pavagadhi, J., Araya, C., Misiunas, K., Ghelani, N., Laskin, M., Barker, D., Li, Q., Briukhov, A., Houlsby, N., Glaese, M., Lakshminarayanan, B., Schucher, N., Tang, Y., Collins, E., Lim, H., Feng, F., Recasens, A., Lai, G., Magni, A., Cao, N. D., Siddhant, A., Ashwood, Z., Orbay, J., Dehghani, M., Brennan, J., He, Y., Xu, K., Gao, Y., Saroufim, C., Molloy, J., Wu, X., Arnold, S., Chang, S., Schrittwieser, J., Buchatskaya, E., Radpour, S., Polacek, M., Giordano, S., Bapna, A., Tokumine, S., Hellendoorn, V., Sottiaux, T., Cogan, S., Severyn, A., Saleh, M., Thakoor, S., Shefey, L., Qiao, S., Gaba, M., Chang, S.-y., Swanson, C., Zhang, B., Lee, B., Rubenstein, P. K., Song, G., Kwiatkowski, T., Koop, A., Kannan, A., Kao, D., Schuh, P., Stjerngren, A., Ghiasi, G., Gibson, G., Vilnis, L., Yuan, Y., Ferreira, F. T., Kamath, A., Klimenko, T., Franko, K., Xiao, K., Bhattacharya, I., Patel, M., Wang, R., Morris, A., Strudel, R., Sharma, V., Choy, P., Hashemi, S. H., Landon, J., Finkelstein, M., Jhakra, P., Frye, J., Barnes, M., Mauger, M., Daun, D., Baatarsukh, K., Tung, M., Farhan, W., Michalewski, H., Viola, F., Quitry, F. d. C., Lan, C. L., Hudson, T., Wang, Q., Fischer, F., Zheng, I., White, E., Dragan, A., Alayrac, J.-b., Ni, E., Pritzel, A., Iwanicki, A., Isard, M., Bulanova, A., Zilka, L., Dyer, E., Sachan, D., Srinivasan, S., Muckenhirn, H., Cai, H., Mandhane, A., Tariq, M., Rae, J. W., Wang, G., Ayoub, K., FitzGerald, N., Zhao, Y., Han, W., Alberti, C., Garrette, D., Krishnakumar, K., Gimenez, M., Levskaya, A., Sohn, D., Matak, J., Iturrate, I., Chang, M. B., Xiang, J., Cao, Y., Ranka, N., Brown, G., Hutter, A., Mirrokni, V., Chen, N., Yao, K., Egyed, Z., Galilee, F., Liechty, T., Kallakuri, P., Palmer, E., Ghemawat, S., Liu, J., Tao, D., Thornton, C., Green, T., Jasarevic, M., Lin, S., Cotruta, V., Tan, Y.-X., Fiedel, N., Yu, H., Chi, E., Neitz, A., Heitkaemper, J., Sinha, A., Zhou, D., Sun, Y., Kaed, C., Hulse, B., Mishra, S., Georgaki, M., Kudugunta, S., Farabet, C., Shafran, I., Vlasic, D., Tsitsulin, A., Ananthanarayanan, R., Carin, A., Su, G., Sun, P., V, S., Carvajal, G., Broder, J., Comsa, I., Repina,

A., Wong, W., Chen, W. W., Hawkins, P., Filonov, E., Loher, L., Hirnschall, C., Wang, W., Ye, J., Burns, A., Cate, H., Wright, D. G., Piccinini, F., Zhang, L., Lin, C.-C., Gog, I., Kulizhskaya, Y., Sreevatsa, A., Song, S., Cobo, L. C., Iyer, A., Tekur, C., Garrido, G., Xiao, Z., Kemp, R., Zheng, H. S., Li, H., Agarwal, A., Ngani, C., Goshvadi, K., Santamaria-Fernandez, R., Fica, W., Chen, X., Gorgolewski, C., Sun, S., Garg, R., Ye, X., Eslami, S. M. A., Hua, N., Simon, J., Joshi, P., Kim, Y., Tenney, I., Potluri, S., Thiet, L. N., Yuan, Q., Luisier, F., Chronopoulou, A., Scellato, S., Srinivasan, P., Chen, M., Koverkathu, V., Dalibard, V., Xu, Y., Saeta, B., Anderson, K., Sellam, T., Fernando, N., Huot, F., Jung, J., Varadarajan, M., Quinn, M., Raul, A., Le, M., Habalov, R., Clark, J., Jalan, K., Bullard, K., Singhal, A., Luong, T., Wang, B., Rajayogam, S., Eisenschlos, J., Jia, J., Finchelstein, D., Yakubovich, A., Balle, D., Fink, M., Agarwal, S., Li, J., Dvijotham, D., Pal, S., Kang, K., Konzelmann, J., Beattie, J., Dousse, O., Wu, D., Crocker, R., Elkind, C., Jonnalagadda, S. R., Lee, J., Holtmann-Rice, D., Kallarackal, K., Liu, R., Vnukov, D., Vats, N., Invernizzi, L., Jafari, M., Zhou, H., Taylor, L., Prendki, J., Wu, M., Eccles, T., Liu, T., Kopparapu, K., Beaufays, F., Angermueller, C., Marzoca, A., Sarcar, S., Dib, H., Stanway, J., Perbet, F., Trdin, N., Sterneck, R., Khorlin, A., Li, D., Wu, X., Goenka, S., Madras, D., Goldshtein, S., Gierke, W., Zhou, T., Liu, Y., Liang, Y., White, A., Li, Y., Singh, S., Bahargam, S., Epstein, M., Basu, S., Lao, L., Ozturel, A., Crous, C., Zhai, A., Lu, H., Tung, Z., Gaur, N., Walton, A., Dixon, L., Zhang, M., Globerson, A., Uy, G., Bolt, A., Wiles, O., Nasr, M., Shumailov, I., Selvi, M., Piccinno, F., Aguilar, R., McCarthy, S., Khalman, M., Shukla, M., Galic, V., Carpenter, J., Villela, K., Zhang, H., Richardson, H., Martens, J., Bosnjak, M., Belle, S. R., Seibert, J., Alnahlawi, M., McWilliams, B., Singh, S., Louis, A., Ding, W., Popovici, D., Simicich, L., Knight, L., Mehta, P., Gupta, N., Shi, C., Fatehi, S., Mitrovic, J., Grills, A., Pagadora, J., Munkhdalai, T., Petrova, D., Eisenbud, D., Zhang, Z., Yates, D., Mittal, B., Tripuraneni, N., Assael, Y., Brovelli, T., Jain, P., Velimirovic, M., Akbulut, C., Mu, J., Macherey, W., Kumar, R., Xu, J., Qureshi, H., Comanici, G., Wiesner, J., Gong, Z., Ruddock, A., Bauer, M., Felt, N., GP, A., Arnab, A., Zelle, D., Rothfuss, J., Rosgen, B., Shenoy, A., Seybold, B., Li, X., Mudigonda, J., Erdogan, G., Xia, J., Simsa, J., Michi, A., Yao, Y., Yew, C., Kan, S., Caswell, I., Radebaugh, C., Elisseeff, A., Valenzuela, P., McKinney, K., Paterson, K., Cui, A., Latorre-Chimoto, E., Kim, S., Zeng, W., Durden, K., Ponnapalli, P., Sosea, T., Choquette-Choo, C. A., Manyika, J., Robenek, B., Vashisht, H., Pereira, S., Lam, H., Velic, M., Owusu-Afriyie, D., Lee, K., Bolukbasi, T., Parrish, A., Lu, S., Park, J., Venkatraman, B., Talbert, A., Rosique, L., Cheng, Y., Sozanschi, A., Paszke, A., Kumar, P., Austin, J., Li, L., Salama, K., Perz, B., Kim, W., Dukkipati, N., Baryshnikov, A., Kaplanis, C., Sheng, X., Chervonyi, Y., Unlu, C., Casas, D. d. L., Askham, H., Tunyasuvunakool, K., Gimeno, F., Poder, S., Kwak, C., Miecnikowski, M., Mirrokni, V., Dimitriev, A., Parisi, A., Liu, D., Tsai, T., Shevlane, T., Kouridi, C., Garmon, D., Goedeckemeyer, A., Brown, A. R., Vijayakumar, A., Elqursh, A., Jazayeri, S., Huang, J., Carthy, S. M., Hoover, J., Kim, L., Kumar, S., Chen, W., Biles, C., Bingham, G., Rosen, E., Wang, L., Tan, Q., Engel, D., Pongetti, F., Cesare, D. d., Hwang, D., Yu, L., Pullman, J., Narayanan, S., Levin, K., Gopal, S., Li, M., Aharoni, A., Trinh, T., Lo, J., Casagrande, N., Vij, R., Matthey, L., Ramadhana, B., Matthews, A., Carey, C. J., Johnson, M., Goranova, K., Shah, R., Ashraf, S., Dasgupta, K., Larsen, R., Wang, Y., Vuyyuru, M. R., Jiang, C., Ijazi, J., Osawa, K., Smith, C., Boppana, R. S., Bilal, T., Koizumi, Y., Xu, Y., Altun, Y., Shabat, N., Bariach, B., Korchemniy, A., Choo, K., Ronneberger, O., Iwuanyanwu, C., Zhao, S., Soergel, D., Hsieh, C.-J., Cai, I., Iqbal, S., Sundermeyer, M., Chen, Z., Bursztein, E., Malaviya, C., Biadsy, F., Shroff, P., Dhillon, I., Latkar, T., Dyer, C., Forbes, H., Nicosia, M., Nikolaev, V., Greene, S., Georgiev, M., Wang, P., Martin, N., Sedghi, H., Zhang, J., Banzal, P., Fritz, D., Rao, V., Wang, X., Zhang, J., Patraucean, V., Du, D., Mordatch, I., Jurin, I., Liu, L., Dubey, A., Mohan, A., Nowakowski, J., Ion, V.-D., Wei, N., Tojo, R., Raad, M. A., Hudson, D. A., Keshava, V., Agrawal, S., Ramirez, K., Wu, Z., Nguyen, H., Liu, J., Sewak, M., Petrini, B., Choi, D., Philips, I., Wang, Z., Bica, I., Garg, A., Wilkiewicz, J., Agrawal, P., Li, X., Guo, D., Xue, E., Shaik, N., Leach, A., Khan, S. M., Wiesinger, J., Jerome, S., Chakladar, A., Wang, A. W., Ornduff, T., Abu, F., Ghaffarkhah, A., Wainwright, M., Cortes, M., Liu, F., Maynez, J., Terzis, A., Samangouei, P., Mansour, R., Kępa, T., Aubet, F.-X., Algymr, A., Banica, D., Weisz, A., Orban, A., Senges, A., Andrejczuk, E., Geller, M., Santo, N. D., Anklin, V., Merey, M. A., Baeuml, M., Strohman, T., Bai, J., Petrov, S., Wu, Y., Hassabis, D., Kavukcuoglu, K., Dean, J., and Vinyals, O. 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].

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., Santos, E. d. O., Järviniemi, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., Stewart, N., Grechuk, B., Grechuk, T., Enugandla, S. V., and Wildon, M. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI, December 2024. URL http://arxiv.org/abs/2411.04872. arXiv:2411.04872 [cs].

Goemans, A., Buhl, M. D., Schuett, J., Korbak, T., Wang, J., Hilton, B., and Irving, G. Safety case template for

frontier AI: A cyber inability argument, November 2024. URL http://arxiv.org/abs/2411.08088. arXiv:2411.08088 [cs].

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

Goldstein, J. A. and Sastry, G. 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.

Gong, Y., Shrestha, R., Claypoole, J., Cogswell, M., Ray, A., Kanan, C., and Divakaran, A. BloomVQA: Assessing Hierarchical Multi-modal Comprehension. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., Linde, J. v. d., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., Maaten, L. v. d., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., Oliveira, L. d., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh,

K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. AI control: improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 16295–16336, Vienna, Austria, July 2024. JMLR.org.

Grosse-Holz, F. and Jorgensen, O. Early Insights from Developing Question-Answer Evaluations for Frontier AI | AISI Work, September 2024. URL https://perma.cc/ANU7-MVZZ.

Grote, T. 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.

Groves, R. M., Jr, F. J. F., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. *Survey Methodology*. John Wiley & Sons, September 2011. ISBN 978-1-118-21134-2.

Gu, Z., Zhang, L., Zhu, X., Chen, J., Huang, W., Zhang, Y., Wang, S., Ye, Z., Gao, Y., Feng, H., and Xiao, Y. DetectBench: Can Large Language Model Detect and Piece Together Implicit Evidence? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Guo, K., Nan, B., Zhou, Y., Guo, T., Guo, Z., Surve, M., Liang, Z., Chawla, N. V., Wiest, O., and Zhang, X. 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.

Gupta, H., Verma, S., Anantheswaran, U., Scaria, K., Parmar, M., Mishra, S., and Baral, C. Polymath: A Challenging Multi-modal Mathematical Reasoning Benchmark, October 2024. URL http://arxiv.org/abs/2410.14702. arXiv:2410.14702 [cs].

Gureckis, T. M. Mechanical Turk - The Poisoned Well?, June 2021. URL https://perma.cc/P6ML-4CHD.

Gusenbauer, M. 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.

Hackenburg, K., Ibrahim, L., Tappin, B. M., and Tsakiris, M. 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.

Haddaway, N. R., Collins, A. M., Coughlin, D., and Kirk, S. 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.

Halevi, G., Moed, H., and Bar-Ilan, J. 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.

Hamotskyi, S., Levbarg, A.-I., and Hänig, C. Eval-UAtion 1.0: Benchmark for Evaluating Ukrainian (Large) Language Models. In Romanyshyn, M., Romanyshyn, N., Hlybovets, A., and Ignatenko, O. (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/.

Hardy, A., Reuel, A., Meimandi, K. J., Soder, L., Griffith, A., Asmar, D. M., Koyejo, S., Bernstein, M. S., and Kochenderfer, M. J. 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].

Hatzius, J., Briggs, J., Kodnani, D., and Pierdomenico, G. The Potentially Large Effects of Artificial Intelligence on Economic Growth. Technical report, Goldman Sachs Economics Research, March 2023. URL https://perma.cc/YM4A-N84S.

Heiding, F., Lermen, S., Kao, A., Schneier, B., and Vishwanath, A. 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].

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. In *Proceedings of the 9th International Conference on Learning Representations*, October 2020. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset. In *Advances in Neural Information Processing Systems*, volume 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Hennessy, E. A., Johnson, B. T., and Keenan, C. 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.

Hewing, M. and Leinhos, V. 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].

Hijazi, F., Alharbi, S., AlHussein, A., Shairah, H., Alzahrani, R., Alshamlan, H., Turkiyyah, G., and Knio, O. ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models. In Habash, N., Bouamor, H., Eskander, R., Tomeh, N., Abu Farha, I., Abdelali, A., Touileb, S., Hamed, I., Onaizan, Y., Alhafni, B., Antoun, W., Khalifa, S., Haddad, H., Zitouni, I., AlKhamissi, B., Almatham, R., and Mrini, K. (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/.

Hildebrandt, C., Woodlief, T., and Elbaum, S. ODDdiLLMma: 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.

Hou, G., Zhang, W., Shen, Y., Tan, Z., Shen, S., and Lu, W. 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].

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., and Rieser, V. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In Davis, B., Graham, Y., Kelleher, J., and Sripada, Y. (eds.), *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 169–182, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.23. URL https://aclanthology.org/2020.inlg-1.23/.

Hoyle, R. H. *Statistical Strategies for Small Sample Research*. SAGE, March 1999. ISBN 978-0-7619-0886-9.

Hua, W., Liu, O., Li, L., Amayuelas, A., Chen, J., Jiang, L., Jin, M., Fan, L., Sun, F., Wang, W., Wang, X., and Zhang, Y. Game-theoretic LLM: Agent Workflow for Negotiation Games, November 2024. URL http://arxiv.org/abs/2411.05990. arXiv:2411.05990 [cs].

Huang, J.-t., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z., and Lyu, M. 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.

Ibrahim, L., Huang, S., Ahmad, L., and Anderljung, M. 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://perma.cc/2BSU-J8QL.

Imbens, G. W. and Rubin, D. B. *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.

Ivanov, I. 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.

Jackson, M. and Cox, D. R. 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.

Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé, H., and Wallach, H. 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.

Jain, R., Sojitra, D., Acharya, A., Saha, S., Jatowt, A., and Dandapat, S. Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In Bouamor, H., Pino, J., and Bali, K. (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/.

Ji, A., Kojima, N., Rush, N., Suhr, A., Vong, W. K., Hawkins, R., and Artzi, Y. Abstract Visual Reasoning with Tangram Shapes. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (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/.

Jimenez, C. E., Russakovsky, O., and Narasimhan, K. CARETS: A Consistency And Robustness Evaluative Test Suite for VQA. In Muresan, S., Nakov, P., and Villavicencio, A. (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/.

Jing, Y., Jin, R., Hu, J., Qiu, H., Wang, X., Wang, P., and Xiong, D. 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].

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., and Narayanan, A. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18):eadk3452, May 2024a. doi: 10.1126/sciadv.adk3452. URL https://www.science.org/doi/10.1126/sciadv.adk3452. Publisher: American Association for the Advancement of Science.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. AI Agents That Matter, July 2024b.

URL http://arxiv.org/abs/2407.01502. arXiv:2407.01502 [cs].

Kaushik, D., Lipton, Z. C., and London, A. J. 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.

Kertzer, J. D. and Renshon, J. 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.

Kodali, P., Goel, A., Asapu, L., Bonagiri, V. K., Govil, A., Choudhury, M., Shrivastava, M., and Kumaraguru, P. 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].

Kruk, J., Marchini, M., Magu, R., Ziems, C., Muchlinski, D., and Yang, D. Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

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

Laine, R., Chughtai, B., Betley, J., Hariharan, K., Balesni, M., Scheurer, J., Hobbhahn, M., Meinke, A., and Evans, O. 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.

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

Lazar, J., Feng, J. H., and Hochheiser, H. (eds.). *Research Methods in Human Computer Interaction*. Morgan Kaufmann, Boston, January 2017. ISBN 978-0-12-805390-4. doi: 10.1016/B978-0-12-805390-4.09991-X. URL https://www.sciencedirect.com/science/article/pii/B978012805390409991X.

Lebrun, B., Temtsin, S., Vonasch, A., and Bartneck, C. 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.

Leeuw, D. d. To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21(2):233, June 2005. ISSN 0282423X. URL https://www.semanticscholar.org/paper/To-mix-or-not-to-mix-data-collection-modes-in-Leeuw/4e09d55e25393eb22b47c4b0d82d4dfa84bc24d0. Num Pages: 233 Place: Stockholm, Sweden Publisher: Statistics Sweden (SCB).

LeGris, S., Vong, W. K., Lake, B. M., and Gureckis, T. M. 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].

Lei, X., Gomez, L., Bai, H. Y., and Bashivan, P. IWISDM: Assessing instruction following in multimodal models at scale, July 2024a. URL http://arxiv.org/abs/2406.14343. arXiv:2406.14343 [cs].

Lei, X., Gomez, L., Bai, H. Y., and Bashivan, P. iWISDM: Assessing instruction following in multimodal models at scale. In Lomonaco, V., Melacci, S., Tuytelaars, T., Chandar, S., and Pascanu, R. (eds.), *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 457–480. PMLR, August 2025. URL https://proceedings.mlr.press/v274/lei25a.html.

Lei, Z., Liang, T., Hu, H., Zhang, J., Zhou, Y., Shao, Y., Li, L., Li, C., Wang, C., Yan, H., and Guo, Q. GAOKAO-Eval: Does high scores truly reflect strong capabilities in LLMs?, December 2024b. URL http://arxiv.org/abs/2412.10056. arXiv:2412.10056.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., and Batra, D. 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.

Li, B., Lin, Z., Peng, W., Nyandwi, J. d. D., Jiang, D., Ma, Z., Khanuja, S., Krishna, R., Neubig, G., and Ramanan, D. NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples, October 2024a. URL http://arxiv.org/abs/2410.14669. arXiv:2410.14669 [cs].

Li, H., Ning, Y., Liao, Z., Wang, S., Li, X. L., Lu, X., Zhao, W., Brahman, F., Choi, Y., and Ren, X. In Search of the Long-Tail: Systematic Generation of Long-Tail Inferential Knowledge via Logical Rule Guided Search. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Li, L., Lei, J., Gan, Z., Yu, L., Chen, Y.-C., Pillai, R., Cheng, Y., Zhou, L., Wang, X., Wang, W. Y., Berg, T. L., Bansal, M., Liu, J., Wang, L., and Liu, Z. 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.

Li, S., Li, L., Liu, Y., Ren, S., Liu, Y., Gao, R., Sun, X., and Hou, L. VITATECS: A Diagnostic Dataset for Temporal Concept Understanding of Video-Language Models. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (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.

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

Liao, T., Taori, R., Raji, I. D., and Schmidt, L. 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.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (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/.

List, J. A., Sadoff, S., and Wagner, M. 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.

Liu, J., Nguyen, T., Shang, M., Ding, H., Li, X., Yu, Y., Kumar, V., and Wang, Z. Learning Code Preference via Synthetic Evolution, October 2024a. URL http://arxiv.org/abs/2410.03837. arXiv:2410.03837 [cs].

Liu, P., Lin, H., Liao, M., Xiang, H., Han, X., and Sun, L. WebDP: Understanding Discourse Structures in Semi-Structured Web Documents. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (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/.

Liu, R., Geng, J., Peterson, J., Sucholutsky, I., and Griffiths, T. L. Large Language Models Assume People are More Rational than We Really are. In *Proceedings of the the 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dAeET8gxqg.

Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., and Hou, L. TempCompass: Do Video LLMs Really Understand Videos? In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Lohr, S. L. *Sampling: Design and Analysis*. Texts in Statistical Sciences. CRC Press, 3rd edition edition, 2022. URL https://www.routledge.com/Sampli

ng-Design-and-Analysis/Lohr/p/book/9780367279509.

Lu, L., Neale, N., Line, N. D., and Bonn, M. 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.

Lu, M., Cho, H. J., Shi, W., May, J., and Spangher, A. 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].

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. 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.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Muresan, S., Nakov, P., and Villavicencio, A. (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/.

Luettgau, L., Coppock, H., Dubois, M., Summerfield, C., and Ududec, C. HiBayES: A Hierarchical Bayesian Modeling Framework for AI Evaluation Statistics, May 2025. URL http://arxiv.org/abs/2505.05602. arXiv:2505.05602 [cs] version: 1.

Luo, X., Tong, S., Fang, Z., and Qu, Z. 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.

Mangalam, K., Akshkulakov, R., and Malik, J. 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.

Marie, B., Fujita, A., and Rubino, R. Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. In Zong, C., Xia, F., Li, W., and Navigli, R. (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. 7297–7306, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.566. URL https://aclanthology.org/2021.acl-long.566/.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. 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://perma.cc/ZVW4-YG9B.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, October 2024. doi: 10.1073/pnas.2322420121. URL https://www.pnas.org/doi/10.1073/pnas.2322420121. Publisher: Proceedings of the National Academy of Sciences.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Watters, P., and Halgamuge, M. N. 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.

McKee, K. R. Human Participants in AI Research: Ethics and Transparency in Practice. *IEEE Transactions on Technology and Society*, 5(3):279–288, September 2024. ISSN 2637-6415. doi: 10.1109/TTS.2024.3446183. URL https://ieeexplore.ieee.org/document/10664609. Conference Name: IEEE Transactions on Technology and Society.

McNulty, K. Power Analysis to Estimate Required Sample Sizes for Modeling. In *Handbook of Regression Modeling in People Analytics: With Examples in R and Python*. CRC Press, 2021. ISBN 978-1-003-19415-6. URL https://peopleanalytics-regression-book.org/gitbook/power-tests.html.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. 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.

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

Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., and Scialom, T. 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.

Mihalcea, R., Ignat, O., Bai, L., Borah, A., Chiruzzo, L., Jin, Z., Kwizera, C., Nwatu, J., Poria, S., and Solorio, T. 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].

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

Miller, J., Krauth, K., Recht, B., and Schmidt, L. 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.

Mirza, A., Alampara, N., Kunchapu, S., Ríos-García, M., Emoekabu, B., Krishnan, A., Gupta, T., Schilling-Wilhelmi, M., Okereke, M., Aneesh, A., Elahi, A. M., Asgari, M., Eberhardt, J., Elbeheiry, H. M., Gil, M. V., Greiner, M., Holick, C. T., Glaubitz, C., Hoffmann, T., Ibrahim, A., Klepsch, L. C., Köster, Y., Kreth, F. A., Meyer, J., Miret, S., Peschel, J. M., Ringleb, M., Roesner, N., Schreiber, J., Schubert, U. S., Stafast, L. M., Wonanke, D., Pieler, M., Schwaller, P., and Jablonka, K. M. Are large language models superhuman chemists?, November 2024. URL http://arxiv.org/abs/2404.01475. arXiv:2404.01475 [cs].

Mizrahi, M., Yardeni Seelig, S., and Shahaf, D. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In Cohn, T., He, Y., and Liu, Y. (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/.

Mohr, J. W. and Ghaziani, A. 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.

Montalan, J. R., Ngui, J. G., Leong, W. Q., Susanto, Y., Rengarajan, H., Aji, A. F., and Tjhi, W. C. Kalahi: A handcrafted, grassroots cultural LLM evaluation suite for Filipino, December 2024. URL http://arxiv.org/abs/2409.15380. arXiv:2409.15380 [cs].

Moskvichev, A. K., Odouard, V. V., and Mitchell, M. 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.

Mukhopadhyay, S., Rajgaria, A., Khatiwada, P., Gupta, V., and Roth, D. MAPWise: Evaluating Vision-Language Models for Advanced Map Queries, August 2024. URL http://arxiv.org/abs/2409.00255. arXiv:2409.00255 [cs].

Mukhopadhyay, S., Rajgaria, A., Khatiwada, P., Shrivastava, M., Roth, D., and Gupta, V. MAPWise: Evaluating Vision-Language Models for Advanced Map Queries. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9348–9378, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.473/.

Nangia, N. and Bowman, S. R. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark. In Korhonen, A., Traum, D., and Màrquez, L. (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/.

Nangia, N., Sugawara, S., Trivedi, H., Warstadt, A., Vania, C., and Bowman, S. R. What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? In Zong, C., Xia, F., Li, W., and Navigli, R. (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/.

Neuhäuser, M. and Ruxton, G. D. *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.

Norlund, T., Hagström, L., and Johansson, R. Transferring Knowledge from Vision to Language: How to Achieve it and how to Measure it? In Bastings, J., Belinkov, Y., Dupoux, E., Giulianelli, M., Hupkes, D., Pinter, Y., and Sajjad, H. (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/.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., and Vazire, S. 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.

Obeidat, R., Al-Harahsheh, Y., Al-Ayyoub, M., and Gharaibeh, M. 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, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., Bourcy, C. d., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., Lohmann, F. v., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. 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://perma.cc/VTF7-6FHX.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. 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.

Page, S. A. and Nyeboer, J. 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.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. 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.

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

Parker, C., Scott, S., and Geddes, A. *Snowball Sampling*. SAGE Publications Ltd, 2019. ISBN 978-1-5297-4761-4. doi: 10.4135/9781526421036831710. URL https://methods.sagepub.com/foundations/snowball-sampling.

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

Paskov, P., Byun, M. J., Wei, K., and Webster, T. Preliminary suggestions for rigorous GPAI model evaluations. Technical report, RAND, May 2025. URL https://www.rand.org/pubs/perspectives/PEA3971-1.html.

Patty, J. W. and Penn, E. M. 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.

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., Deletang, G., Ruoss, A., El-Sayed, S., Brown, S., Dragan, A., Shah, R., Dafoe, A., and Shevlane, T. Evaluating Frontier Models for Dangerous Capabilities, April 2024. URL http://arxiv.org/abs/2403.13793. arXiv:2403.13793 [cs].

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche Buc, F., Fox, E., and Larochelle, H. 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.

Prabhakaran, V., Mostafazadeh Davani, A., and Diaz, M. On Releasing Annotator-Level Labels and Information in Datasets. In Bonial, C. and Xue, N. (eds.), *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pp. 133–138, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.14. URL https://aclanthology.org/2021.law-1.14/.

Prolific. Representative samples FAQ, April 2025. URL https://perma.cc/VMR3-V2EB.

Qureshi, N., Edelen, M., Hilton, L., Rodriguez, A., Hays, R. D., and Herman, P. M. 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.

Reese, M. L. and Smirnova, A. 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.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. 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.

Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O'Connel, D., Arnott, P., Bloom, R., Broadley, T., Garcia, K., Goodrich, B., Hasin, M., Jawhar, S., Kinniment, M., Kwa, T., Lajko, A., Rush, N., Sato, L. J. K., Arx, S. V., West, B., Chan, L., and Barnes, E. HCAST: Human-Calibrated Autonomy Software Tasks, March 2025. URL http://arxiv.org/abs/2503.17354. arXiv:2503.17354 [cs].

Reisenzein, R. and Junge, M. 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.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. 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.

Roberts, J., Han, K., Houlsby, N., and Albanie, S. SciFIBench: Benchmarking Large Multimodal Models for Scientific Figure Interpretation, December 2024. URL http://arxiv.org/abs/2405.08807. arXiv:2405.08807 [cs].

Rosellini, A. J. and Brown, T. A. 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.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., and Grefenstette, E. 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.

Sakai, Y., Kamigaito, H., and Watanabe, T. mCSQA: Multilingual Commonsense Reasoning Dataset with Unified Creation Strategy by Language Models and Humans. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, May 2025. URL http://arxiv.org/abs/2505.10573. arXiv:2505.10573 [cs].

Sanchez, M. E. 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.

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

Sanyal, S., Xiao, T., Liu, J., Wang, W., and Ren, X. Are Machines Better at Complex Reasoning? Unveiling Human-Machine Inference Gaps in Entailment Verification. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10361–10386, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.618. URL https://aclanthology.org/2024.findings-acl.618/.

Sarrami-Foroushani, P., Travaglia, J., Debono, D., Clay-Williams, R., and Braithwaite, J. 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.

Sassoon, G. W. 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.

Saxon, M., Holtzman, A., West, P., Wang, W. Y., and Saphra, N. Benchmarks as Microscopes: A Call for Model Metrology. In *Proceedings of the First Conference on Language Modeling*, August 2024. URL https://openreview.net/forum?id=bttKwCZDkm&noteId=Yfwy2d4fiT.

Schoot, R. v. d. and Miočević, M. (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.

Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., and Kowald, D. Reproducibility in Machine Learning-based Research: Overview, Barriers and Drivers, July 2024. URL http://arxiv.org/abs/2406.14325. arXiv:2406.14325 [cs].

Shah Jahan, M., Khan, H. U., Akbar, S., Umar Farooq, M., Gul, S., and Amjad, A. 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.

Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In Webber, B., Cohn, T., He, Y., and Liu, Y. (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/.

Shaw, A. and Hargittai, E. 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.

Sheehan, K. B. 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.

Shin, E., Johnson, T. P., and Rao, K. 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.

Shrier, D., Emanuel, J., and Harris, M. Is Your Job AI Resilient? *Harvard Business Review*, October 2023. URL https://perma.cc/B7PC-3AKJ.

Si, C., Yang, D., and Hashimoto, T. 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.

Siska, C., Marazopoulou, K., Ailem, M., and Bono, J. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560/.

Solon, G., Haider, S. J., and Wooldridge, J. M. 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].

Someya, T. and Oseki, Y. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In Vlachos, A. and Augenstein, I. (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/.

Sourati, Z., Ilievski, F., Sommerauer, P., and Jiang, Y. ARN: Analogical Reasoning on Narratives, September 2024. URL http://arxiv.org/abs/2310.00996. arXiv:2310.00996 [cs].

Sprague, Z. R., Ye, X., Bostrom, K., Chaudhuri, S., and Durrett, G. 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.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro,

C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., Melo, G. d., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, S., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. 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.

Stantcheva, S. 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.

Steinbach, P., Gernhardt, F., Tanveer, M., Schmerler, S., and Starke, S. Machine Learning State-of-the-Art with Uncertainties, April 2022. URL http://arxiv.org/abs/2204.05173. arXiv:2204.05173 [cs].

Stodden, V. and Miguez, S. 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.

Strauss, M. E. and Smith, G. T. 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.

Subramonian, A., Yuan, X., Daumé III, H., and Blodgett, S. L. It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (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/.

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

Tahsin Mayeesha, T., Md Sarwar, A., and Rahman, R. M. 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.

Taktasheva, E., Fenogenova, A., Shevelev, D., Katricheva, N., Tikhonova, M., Akhmetgareeva, A., Zinkevich, O., Bashmakova, A., Iordanskaia, S., Kurenshchikova, V., Spiridonova, A., Artemova, E., Shavrina, T., and Mikhailov, V. TAPE: Assessing Few-shot Russian Language Understanding. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (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/.

Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., and Melas-Kyriazi, L. 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.

Tedeschi, S., Bos, J., Declerck, T., Hajič, J., Hershcovich, D., Hovy, E., Koller, A., Krek, S., Schockaert, S., Sennrich, R., Shutova, E., and Navigli, R. What's the Meaning of Superhuman Performance in Today's NLU? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (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/.

Thrush, T., Moore, J., Monares, M., Potts, C., and Kiela, D. I am a Strange Dataset: Metalinguistic Tests for Language Models. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., and Neubig, G. 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.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., and Hiscock, M. 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.

Traylor, F. 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://perma.cc/7G9T-B9EA.

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://perma.cc/9YZJ-HGT7.

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://perma.cc/X5YP-HLZG.

Valliant, R., Dever, J. A., and Kreuter, F. *Practical Tools for Designing and Weighting Survey Samples*. Statistics for Social and Behavioral Sciences. Springer International Publishing, Cham, 2018. ISBN 978-3-319-93631-4 978-3-319-93632-1. doi: 10.1007/978-3-319-93632-1. URL http://link.springer.com/10.1007/978-3-319-93632-1.

Valmeekam, K., Marquez, M., Sreedharan, S., and Kambhampati, S. 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.

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. 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.

Verma, V., Fleisig, E., Tomlin, N., and Klein, D. Ghostbuster: Detecting Text Ghostwritten by Large Language Models. In Duh, K., Gomez, H., and Bethard, S. (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/.

Veselovsky, V., Ribeiro, M. H., Cozzolino, P., Gordon, A., Rothschild, D., and West, R. Prevalence and prevention of large language model use in crowd work, October 2023a. URL http://arxiv.org/abs/2310.15683. arXiv:2310.15683 [cs].

Veselovsky, V., Ribeiro, M. H., and West, R. 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].

Wadhawan, R., Bansal, H., Chang, K.-W., and Peng, N. 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.

Wallach, H., Desai, M., Pangakis, N., Cooper, A. F., Wang, A., Barocas, S., Chouldechova, A., Atalla, C., Blodgett, S. L., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., Reed, S., Sheng, E., Vann, D., Vaughan, J. W., Vogel, M., Washington, H., and Jacobs, A. Z. Evaluating Generative AI Systems is a Social Science Measurement Challenge, November 2024. URL http://arxiv.org/abs/2411.10939. arXiv:2411.10939.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pp. 1–12, February 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-00986-z. URL https://www.nature.com/articles/s42256-025-00986-z. Publisher: Nature Publishing Group.

Wang, X., Jiang, L., Hernandez-Orallo, J., Stillwell, D., Sun, L., Luo, F., and Xie, X. Evaluating General-Purpose AI with Psychometrics, December 2023. URL http://arxiv.org/abs/2310.16379. arXiv:2310.16379 version: 2.

Webson, A., Loo, A., Yu, Q., and Pavlick, E. Are Language Models Worse than Humans at Following Prompts? It's Complicated. In Bouamor, H., Pino, J., and Bali, K. (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/.

Wei, K., Ezell, C., Gabrieli, N., and Deshpande, C. How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance, October 2024. URL http://arxiv.org/abs/2410.13042. arXiv:2410.13042 [cs].

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from Language Models, December 2021. URL http://arxiv.org/abs/2112.04359. arXiv:2112.04359 [cs].

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., and Isaac,

W. Sociotechnical Safety Evaluation of Generative AI Systems, October 2023. URL http://arxiv.org/abs/2310.11986. arXiv:2310.11986 [cs].

Weidinger, L., Mellor, J., Pegueroles, B. G., Marchal, N., Kumar, R., Lum, K., Akbulut, C., Diaz, M., Bergman, S., Rodriguez, M., Rieser, V., and Isaac, W. STAR: SocioTechnical Approach to Red Teaming Language Models, October 2024. URL http://arxiv.org/abs/2406.11757. arXiv:2406.11757 [cs].

Weissweiler, L., Köksal, A., and Schütze, H. 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].

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

West, B. T. Paradata in Survey Research. *Survey Practice*, 4(4), August 2011. doi: 10.29115/SP-2011-0018. URL https://www.surveypractice.org/article/3036-paradata-in-survey-research.

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Kinniment, M., Lajko, A., Nix, S., Sato, L., Saunders, W., Taran, M., West, B., and Barnes, E. 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].

Winters, B. D., Gurses, A. P., Lehmann, H., Sexton, J. B., Rampersad, C. J., and Pronovost, P. J. 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.

Wu, M.-H. and Quinn, A. 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.

Wu, X., Ding, Y., Li, B., Lu, P., Yin, D., Chang, K.-W., and Peng, N. 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].

Wu, Y., Tang, X., Mitchell, T., and Li, Y. 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.

Xiang, T., Li, L., Li, W., Bai, M., Wei, L., Wang, B., and Garcia, N. 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.

Xiao, X., Su, Y., Zhang, S., Chen, Z., Chen, Y., and Liu, T. Confidence in Large Language Model Evaluation: A Bayesian Approach to Limited-Sample Challenges, April 2025. URL http://arxiv.org/abs/2504.21303. arXiv:2504.21303 [cs].

Xiao, Z., Zhang, S., Lai, V., and Liao, Q. V. Evaluating Evaluation Metrics: A Framework for Analyzing NLG Evaluation Metrics using Measurement Theory. In Bouamor, H., Pino, J., and Bali, K. (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/.

Yao, F., Zhuang, Y., Sun, Z., Xu, S., Kumar, A., and Shang, J. Data Contamination Can Cross Language Barriers. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Yasin, A., Fatima, R., Wen, L., Afzal, W., Azhar, M., and Torkar, R. 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/.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., and Zhang, X. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *Proceedings of the Thirteenth International Conference on Learning Representations*,

October 2024. URL https://openreview.net/forum?id=3GTtZFiajM.

Yin, H., Liu, K.-H., Sun, M., Chen, Y., Zhang, B., Liu, J., Sehgal, V., Panchal, R. R., Hotaj, E., Liu, X., Guo, D., Zhang, J., Wang, Z., Jiang, S., Li, H., Chen, Z., Chen, W.-Y., Yang, J., and Wen, W. 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.

Ying, L., Collins, K. M., Wong, L., Sucholutsky, I., Liu, R., Weller, A., Shu, T., Griffiths, T. L., and Tenenbaum, J. B. On Benchmarking Human-Like Intelligence in Machines, February 2025. URL http://arxiv.org/abs/2502.20502. arXiv:2502.20502 [cs].

Yue, X., Ni, Y., Zheng, T., Zhang, K., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. 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.

Zamecnik, A., Barthakur, A., Wang, H., and Dawson, S. Mapping Employable Skills in Higher Education Curriculum Using LLMs. In Ferreira Mello, R., Rummel, N., Jivet, I., Pishtari, G., and Ruipérez Valiente, J. A. (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.

Zerroug, A., Vaishnav, M., Colin, J., Musslick, S., and Serre, T. 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.

Zhang, C., Liu, X., Ziska, K., Jeon, S., Yu, C.-L., and Xu, Y. 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.

Zhang, K., Choi, Y., Song, Z., He, T., Wang, W. Y., and Li, L. Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Zhang, S., Liu, J., and Ying, Z. 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.

Zhang, S., Xu, J., and Alvero, A. Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization. *Sociological Methods & Research*, pp. 00491241251327130, May 2025. ISSN 0049-1241. doi: 10.1177/00491241251327130. URL https://doi.org/10.1177/00491241251327130. Publisher: SAGE Publications Inc.

Zhang, Y., Lu, J., and Jaitly, N. Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games. In Ku, L.-W., Martins, A., and Srikumar, V. (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/.

Zhao, D., Andrews, J. T. A., Papakyriakopoulos, O., and Xiang, A. 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.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., and Duan, N. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Duh, K., Gomez, H., and Bethard, S. (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/.

Zhou, H. and Hong, Y. 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].

Zhou, K., Blodgett, S. L., Trischler, A., Daumé III, H., Suleman, K., and Olteanu, A. Deconstructing NLG Evaluation: Evaluation Practices, Assumptions, and Their Implications. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (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/.

Zhou, Y., Yang, J., Guo, K., Chen, P.-Y., Gao, T., Geyer, W., Moniz, N., Chawla, N. V., and Zhang, X. LabSafety Bench: Benchmarking LLMs on Safety Issues in Scientific Labs, October 2024. URL http://arxiv.org/abs/2410.14182. arXiv:2410.14182 [cs].

Zhu, H., Kapoor, R., Min, S. Y., Han, W., Li, J., Geng, K., Neubig, G., Bisk, Y., Kembhavi, A., and Weihs, L. 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.

Zhuo, T. Y., Vu, M. C., Chim, J., Hu, H., Yu, W., Widyasari, R., Yusuf, I. N. B., Zhan, H., He, J., Paul, I., Brunner, S., Gong, C., Hoang, T., Zebaze, A. R., Hong, X., Li, W.-D., Kaddour, J., Xu, M., Zhang, Z., Yadav, P., Jain, N., Gu, A., Cheng, Z., Liu, J., Liu, Q., Wang, Z., Lo, D., Hui, B., Muennighoff, N., Fried, D., Du, X., Vries, H. d., and Werra, L. V. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions, October 2024. URL http://arxiv.org/abs/2406.15877. arXiv:2406.15877 [cs].

Zickar, M. J. 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 Results from Systematic Review

This appendix contains summary statistics for select items from the checklist in Appendix B.

Please note the following limitations and caveats to the statistics below. First, summary statistics are not presented for all items; all original data and additional statistics are available on Github: `https://github.com/kevinlwei/human-baselines`. Second, some results in this section contain imputed data; the imputation scheme is detailed in Appendix B, and individual items are marked with "default" responses below. Finally, additional limitations specific to each section are also noted.

All tables below present statistics for the full set of human baselines we reviewed ($n = 115$), as well as the same statistics for a subset of human baselines from evaluations that are ($n = 7$) frequently used in industry model cards. This subset consists of: MMMU (Yue et al., 2024), GPQA (Rein et al., 2024), MATH (Hendrycks et al., 2021), DROP (Dua et al., 2019), ARC (LeGris et al., 2024), CONCEPTARC (Moskvichev et al., 2023), and EGOSCHEMA (Mangalam et al., 2023).

A.0. Paper Information

Figures 2–4 contain frequency charts for the years, publication venues, and languages of reviewed evaluations. Note that the list of "top ML/AI conferences & journals" in Figure 3 is taken from the ICML Reviewer FAQs (i.e., the list of venues for which ICML reviewers are expected to have published in); [8] we understand that conference and journal "rankings" are inherently subjective and do not take a position on which publication venues are objectively "better" or "worse." Figure 3 also contains workshops and non-archival venues.

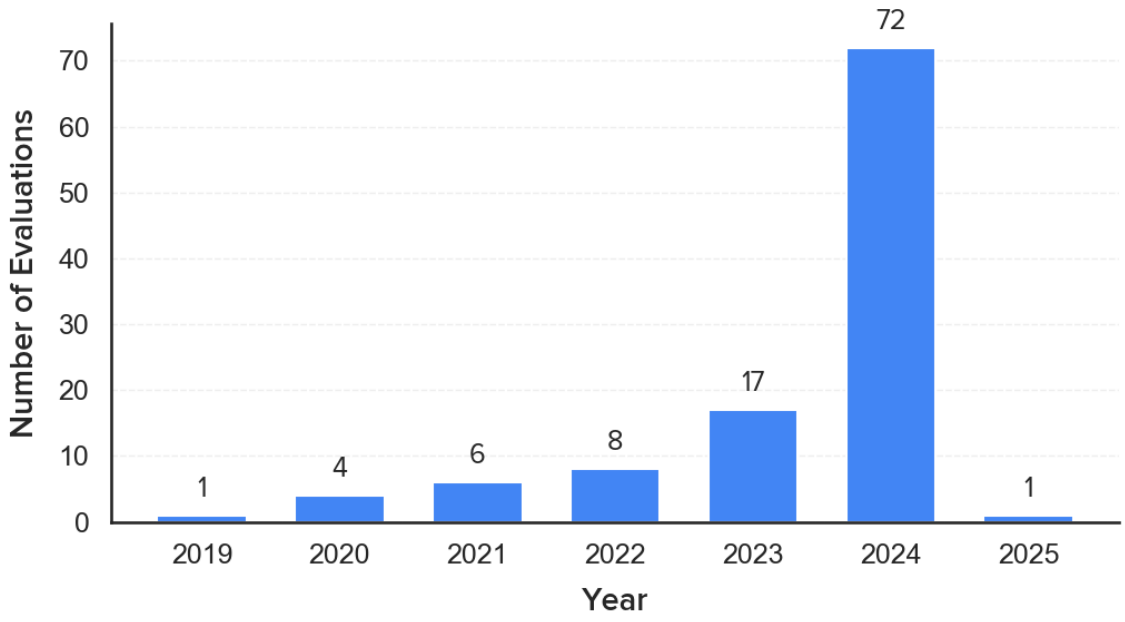

Figure 2: Frequency of years in which reviewed evaluations were published.

[8] `https://perma.cc/D6TK-YQFJ`.

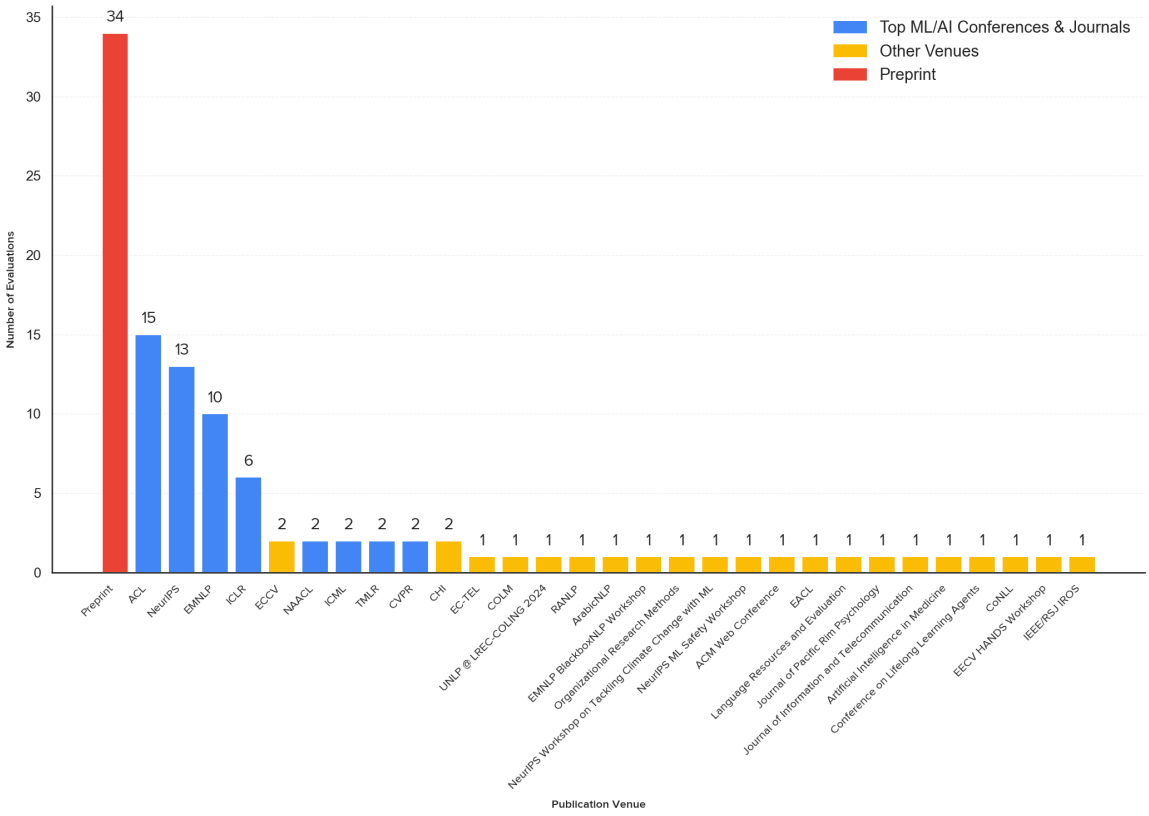

Figure 3: Frequency of publication venues of reviewed evaluations, in descending order. "Top ML/AI conferences & journals" are: ICML, NeurIPS, ICLR, UAI, AISTATS, COLT, ALT, JMLR, TMLR, CVPR, ICCV, ACL, NAACL, EMNLP, and SIMODS.

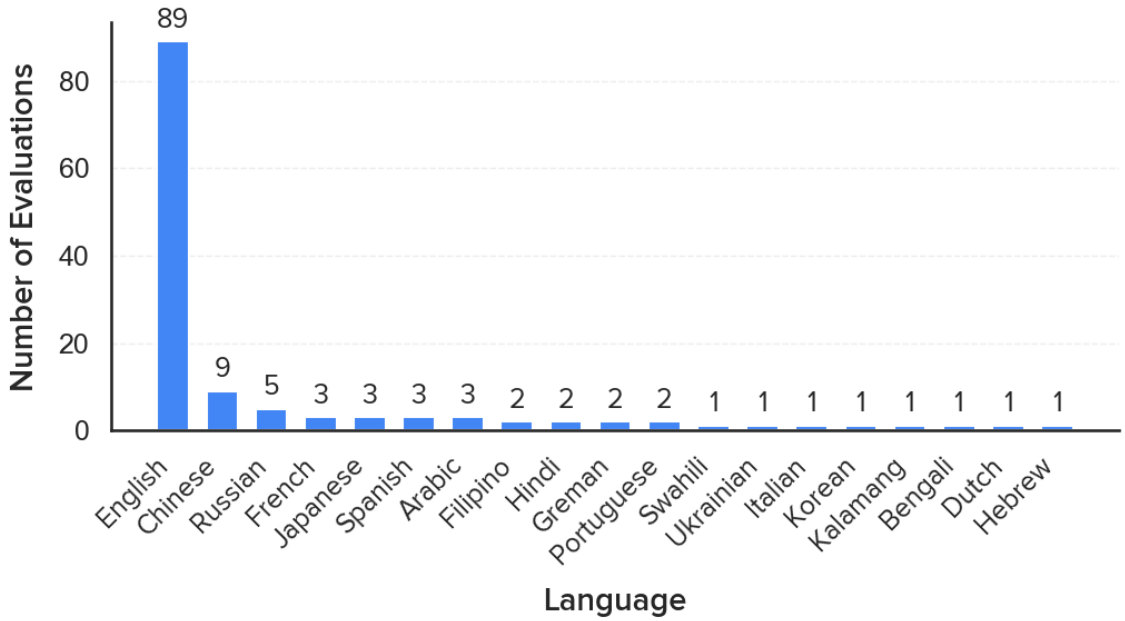

Figure 4: Frequency of languages in which reviewed evaluations' items were written, in descending order. Note that individual items may contain items in multiple languages.

A.1. Baseline Design & Implementation

NB: the "test set equivalence" row in Table 1 is not a separate item in our checklist but rather is imputed from the number of evaluation items in the AI test set and the human test set, as well as an item (not reported below but available on our Github[9]) about whether comparisons between human and AI performance are made on the same test set.

Question	**All Baselines** ($n = 115$)		**Model Card Baselines** ($n = 7$)	
	Yes	No	Yes	No
Test Set Equivalence: Were human and AI test sets identical? (Default: No)	59.13% 68	40.87% 47	57.14% 4	42.86% 3
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)? (Default: No)	16.52% 19	83.48 96	28.57% 2	71.43% 5
1.8 **Power Analysis**: Did the authors conduct power analysis in order to determine baseline size? (Default: No)	1.74% 2	98.26% 113	0.00% 0	100.00% 7
1.10 **Pre-Registration**: Was the baseline/eval design pre-registered? (Default: No)	1.74% 2	98.26% 113	0.00% 0	100.00% 7

Table 1: Summary statistics for baseline design & implementation items (with imputation)

Question	**All Baselines** ($n = 115$)			**Model Card Baselines** ($n = 7$)		
	Yes	No	Unknown	Yes	No	Unknown
1.6 **Iterative Design**: Was the experimental setup of the baseline iteratively designed with participatory methods?	34.78% 40	12.17% 14	63.04% 61	42.86% 3	28.57% 2	28.57% 2
1.7 **Amount of Effort**: Does the baseline control for the amount of effort by human baseliners and AIs	13.91% 16	35.65% 41	50.43% 58	28.57% 2	57.14% 4	14.29 1
1.9 **Ethics Review**: Was the study approved or exempted by an IRB, or did it undergo other ethics review?	13.91% 16	2.61% 3	83.48% 96	0.00% 0	14.29% 1	85.71% 6

Table 2: Summary statistics for baseline design & implementation items (no)

[9] https://github.com/kevinlwei/human-baselines.

A.2. Baseliner Recruitment

NB: note that our statistics for Q2.1 are likely significant over-estimates, as we erred on the side of more generous annotations. Most papers did not explicitly specify populations of interest; papers that gestured at baseline demographics (e.g., "our baseliners were 18–25 years old") were assumed to have specified a population of interest defined by those demographics (e.g., the population in the previous example would be all adults 18–25 years of age).

Question	**All Baselines** ($n = 115$)		**Model Card Baselines** ($n = 7$)	
	Yes	No	Yes	No
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? (Default: No)	42.61% 49	57.39% 66	57.14% 4	42.86% 3
2.3 **Quality Control in Recruitment**: Were human baseliners pre-qualified or excluded during the recruitment process for any reason? (Default: Yes)	28.70% 33	71.30% 82	28.57% 2	71.43% 5
2.4 **Author Baseliners**: Did the authors or members of the research team also serve as human baseliners? (Default: No)	9.33% 14	91.67% 101	28.57% 2	71.43% 5
2.5 **Baseliner Train/Test Contamination**: Did the recruitment process exclude baseliners who had been exposed to the eval questions previously? (Default: No)	7.83% 9	92.17% 106	14.29% 1	85.71% 6
2.6 **Baseliner Training**: Did the human baseliners receive training for the baseline? (Default: No)	22.61% 26	77.39% 89	42.86% 3	57.14% 4

Table 3: Summary statistics for baseliner recruitment items (with imputation)

Question	**All Baselines** ($n = 115$)			**Model Card Baselines** ($n = 7$)		
	Convenience	Crowdsource	Unknown	Convenience	Crowdsource	Unknown
2.2 **Baseliner Sampling Strategy**: How were the human baseliners recruited?	31.30% 36	32.17% 37	36.52% 42	28.57% 2	42.86% 3	28.57% 2

Table 4: Summary statistics for Q2.2 Baseliner Sampling Strategy (no imputation)

Question	**All Baselines** ($n = 115$)			**Model Card Baselines** ($n = 7$)		
	Yes	No	Unknown	Yes	No	Unknown
2.7 **Baseliner Testing Compensation**: Were the human baseliners compensated for completing the baseline? (Default: No)	41.74% 48	11.03% 13	46.96% 54	57.14% 4	0.00% 0	42.86% 3

Table 5: Summary statistics for Q2.7 Baseliner Testing Compensation (no imputation)

A.3. Baseline Execution

Question	**All Baselines** ($n = 115$)		**Model Card Baselines** ($n = 7$)	
	Yes	No	Yes	No
3.2 **Quality Control in Execution**: Were quality checks implemented or data cleaned/excluded during the data collection process (i.e., after baseliners were recruited)? (Default: No)	23.48% 27	76.52% 88	28.57% 2	71.43% 5
3.3 **UI Equivalence**: Did the human baseliners and AIs have access to the same UI for each item? (Default: No)	12.17% 14	87.83% 101	14.29% 1	85.71% 6
3.4 **Instruction Equivalence**: Did the human baseliners and AIs have access to the same instructions/prompt/question for each item? (Default: No)	24.35% 28	75.65% 87	14.29% 1	85.71% 6
3.5 **Tool Access Equivalence**: Did the human baseliners and AIs have access to the same (technical) tools for each item? (Default: Yes)	89.57% 103	10.43% 12	71.43% 5	28.57% 2
3.6 **Explanations**: Did the eval/baseline collect explanations from the human baseliners, after the evaluation was conducted? (Default: No)	11.30% 13	97.39% 112	42.86% 3	57.14% 4

Table 6: Summary statistics for baseline execution items (with imputation)

A.4. Baseline Analysis

Question	**All Baselines** ($n = 115$) Yes	**All Baselines** ($n = 115$) No	**Model Card Baselines** ($n = 7$) Yes	**Model Card Baselines** ($n = 7$) No
4.1 **Statistical Significance**: Did the eval test for statistically significant differences between AI and human performance? (Default: No)	8.70% 10	91.30% 105	0.00% 0	100.00% 7
4.2 **Uncertainty Estimate**: Did the paper present a measure of uncertainty for the AI and human baseline results? (Default: No)	33.04% 38	66.96% 77	14.29% 1	85.71% 7
4.3 **Evaluation Metric Equivalence**: Was the same evaluation metric measured/compared for both humans and AIs? (Default: Yes)	93.91% 108	6.09% 7	100.00% 7	0.00% 0
4.4 **Evaluation Scoring Criteria Equivalence**: Was the same scoring rubric used for both AI and human results? (Default: Yes)	98.26% 113	1.74% 2	100.00% 7	0.00% 0
4.5 **Evaluation Scoring Method Equivalence**: Was the same scoring method used for both AI and human results? (Default: Yes)	95.65% 110	4.35% 5	100.00% 7	0.00% 0

Table 7: Summary statistics for baseline analysis items (with imputation)

Question	**All Baselines** ($n = 115$) Point	Interval	Distribution	**Model Card Baselines** ($n = 7$) Point	Interval	Distribution
4.2.1 **Estimate Type**: Is the reported baseline a point estimate, an interval estimate, or a distribution? (Default: Point Estimate)	63.48% 73	32.17% 37	4.35% 5	71.43% 5	14.29% 1	14.29% 1

Table 8: Summary statistics for Q4.2.1 Estimate Type (with imputation)

A.5. Baseline Documentation

Question	**All Baselines** ($n = 115$)		**Model Card Baselines** ($n = 7$)	
	Yes	No	Yes	No
5.1.1 **Reporting Sample Demographics**: Were demographics for human baseliners, e.g., race, gender, etc. reported? (Default: No)	22.61% 26	77.39% 89	28.57% 2	71.43% 5
5.1.2 **Reporting Baseline Instructions**: Were instructions/guidelines given to human baseliners reported? (Default: No)	40.00% 46	60.00% 69	42.86% 3	57.14% 4
5.1.3 **Reporting Time to Completion**: Was time to completion for eval items reported? (Default: No)	20.87% 24	79.13% 91	28.57% 2	71.43% 5
5.2 **Baseline Data Availability**: Is the (anonymized) human baseline data publicly available? (Default: No)	21.74% 25	78.26% 90	28.57% 2	71.43% 5
5.3 **Experimental Materials Availability**: Are experimental materials used to implement the eval/baseline publicly available? (Default: No)	46.97% 54	55.65% 64	57.14% 4	42.86% 3
5.4 **Analysis Code Availability**: Is the code used to analyze the eval/baseline publicly available? (Default: No)	40.87% 47	59.13% 56	57.14% 4	42.86% 3

Table 9: Summary statistics for baseline documentation items (with imputation)

B. 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 imputed—i.e., 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 is no consensus in the literature on common practices.
  - For items without default responses, "Unknown/Unreported" labels were not adjusted.

B.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)

B.1. Baseline Design & Implementation

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"

B.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" (Default), "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", "Partial", "No" (Default), "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), "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?

B.3. Baseline Execution

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 Execution**: 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" (Default), "Unknown/Unreported", or "N/A"

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

3.2.2 **Execution 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 (Unchecked by default)

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 external tools; respond yes if the human had internet access and the AI did not (but was trained on the internet)
Select one of: "Yes" (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"

B.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" (Default), "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: "Yes" (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: "Yes" (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: "Yes" (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"

B.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"

C. Appendix: Case Studies

This appendix contains both positive and negative examples of human baselines. Positive examples of (high quality and highly transparent) human baselines include Wijk et al. (2024); LeGris et al. (2024); Rein et al. (2025); Brodeur et al. (2025), and we discuss the first two of these below.[10] We do not claim that positive examples are entirely positive or that they follow all our recommendations in Section 4, merely that they are substantially more robust and transparent than other literature.

For additional negative case studies, see Tedeschi et al. (2023).

C.1. Positive Example: Wijk et al. (2024)

RE-BENCH Wijk et al. (2024) is an evaluation of AI research engineering capabilities, and it includes an expert human baseline. We discuss each baseline lifecycle stage below.

Design & Implementation. The benchmark consists of 7 tasks, on which both humans and AI systems were evaluated. It is unclear if the tasks were iteratively designed, but the authors report validating the tasks on human performance, and they also report significant qualitative results (task trajectories). The sample size was only 61, and no power analysis was conducted, but given the small size of the population of interest[11], it's possible that this sample size is sufficient to ensure reasonable statistical power. The authors do not report whether the study underwent ethics review.

Recruitment. The population of interest is specified as human experts with AI research engineering expertise, i.e., as defined by years of experience, research output, hiring screens, and graduate education. The baseliner sample was recruited using convenience sampling, and the authors used well-defined criteria and rigorous screening methods (including some who completed a CodeSignal screen) as quality controls during the recruitment process. Baseliners were compensated at competitive rates ($1855 per expert on average).

Execution. It is unclear if quality controls were used post-data collection. The authors somewhat control for method effects by controlling the baseliners' coding environment, and both humans and AI systems were permitted internet access. Level of effort is specified by comparing performance in a specified time interval; all humans were given 8 hours per task, and the authors also report performance when both humans and AI systems are given 2 hours per task. The authors collect logs and other qualitative data.

Analysis. The authors report performance intervals over time, and evaluation metrics, rubrics, and scoring methods are consistent.

Documentation. Significant detail about the baselining methods and the baseliner sample are reported (e.g., professional backgrounds). All task environments are released on Github at `https://github.com/METR/ai-rd-tasks/tree/main`, and agent trajectories are also provided at `https://transcripts.metr.org/`. The manuscript notes that "[a]nalysis code and anonymized human expert data [are] coming soon," though we were unable to find these materials as of the time of this writing (June 2025).

C.2. Positive Example: LeGris et al. (2024)

H-ARC ("Human-ARC") (LeGris et al., 2024)[12] is the generalist (non-expert) human baseline for ARC, a visual program synthesis benchmark. We discuss each baseline lifecycle stage below.

Design & Implementation. Human results are collected on the entirety of the ARC evaluation and test sets, ensuring that the test set is consistent across humans and AI systems. It is unclear whether baseline instruments were iterated on. No power analysis was conducted, but the sample size consisted of 1768 baseliners, which is higher than the rule-of-thumb 1000 needed to represent the U.S. adult population (though it is unclear what the exact population of interest is) (Gelman,

[10] METR has produced a number of high-quality baselines, most recently Rein et al. (2025) (which was published after our review and therefore not included in our results above). Brodeur et al. (2025) was similarly not included as it was not caught by our search terms in the literature review. Our choice to discuss only two examples is not based on the quality of the underlying baselines but due to the fact that we believe the chosen examples to be somewhat more well-known in the ML research community.

[11] In fact, the exact size of the population is likely not known with substantial precision.

[12] Note that 3 of 4 authors are affiliated with the Department of Psychology of New York University. We notice generally that many high-quality human baselines are created by experts with interdisciplinary backgrounds beyond pure ML research, likely due to more robust scientific research norms in other disciplines. Cf. discussion in (Burden et al., 2025).

2004).[13] No IRB approval was reported.

Recruitment. No population of interest is explicitly specified, though we can assume the intended population is a large population of human adults (participants were specified to be 18–77 years old). Baseliners are selected via a crowdsourced sample from Amazon Mechanical Turk and CloudResearch. No detailed quality controls are specified, but the CloudResearch service contains built-in tools to improve data quality. Baseliners were compensated $10 and also awarded a performance bonus of $1.

Execution. The authors report that some baseliner data is incomplete and conduct a robustness check when excluding and imputing missing data. Baseliners were compensated and also given three tries at each task, which is some measure of effort. The authors collected explanations from baseliners.

Analysis. The authors report interval estimates; no statistical tests were conducted that compared human vs. AI performance, though the authors reported a number of within-sample statistical tests (i.e., comparing different segments of baseliners). Evaluation rubrics and scoring methods are consistent.

Documentation. Significant detail about the baselining methods and the baseliner sample are reported (e.g., age, gender, and other demographic information).[14] Data and code are released on Github at `https://github.com/Le-Gris/h-arc`.

C.3. Positive Examples: Limiting Baseline Interpretations

Not all human baselines are able to maximize scientific robustness, e.g., due to cost considerations. In these cases, researchers can consider scaling back interpretations of human baselines and clearly outlining methodological shortcomings.

One example of baselines in this vein are those contained in Laine et al. (2024), which is a benchmark of AI systems' levels of situational awareness.[15] The authors are careful to conduct the baseline only for relevant tasks, and the baseline is interpreted as an upper-bound on performance.[16]

C.4. Negative Example: Sourati et al. (2024)

ARN Sourati et al. (2024) is a benchmark on "Analogical Reasoning on Narratives" and contains a human baseline. We discuss each baseline lifecycle stage below.

Design & Implementation. Baseliners complete only 120 of 1,096 items, and performance metrics on that subset are then compared to AI performance on the entire evaluation dataset. It is unknown whether baseline instruments were iteratively developed. The population of interest is not specified, but the baseline seems to be intended as a generalist (non-expert) baseline, so a sample size of 2 is clearly insufficient to robustly estimate performance metrics of a broad human population. No IRB approval is reported.

Recruitment No population of interest is specified, and the human baseline was conducted by two research assistants (it is unclear if the baseliners had already been exposed to the evaluation items). No quality controls in recruitment are specified.

Execution Almost nothing about the baseline execution is specified, though the authors do provide the instructions given to baseliners.

Analysis Only point estimates are reported. Furthermore, the language used to report human performance somewhat overstates baseline results. For instance, it is difficult to make claims that "models are not as good as humans at distinguishing analogies from distractors (57 vs 96%)" when there are only two baseliners.[17]

Documentation The authors do not report most methodological details, and all that is known about the baseliner sample is that it consists of two research assistants. As of this writing, we are unable to find publicly available code or baseline results (though the benchmark items are available online).

[13] Note, though, that the baseliners are unevenly distributed across 800 tasks.

[14] In fact, H-ARC is released as a separate and independent document from the original benchmark.

[15] In fact, the human baselines in Laine et al. (2024) are quite thoughtfully designed, and our choice to discuss this baseline is primarily due to the well-justified and well-discussed *interpretations* of the baselines in this paper (not because we believe it to fail any particular robustness criteria).

[16] More precisely, "[t]he intended interpretation is that the upper baseline is achievable and represents a high level of situational awareness (roughly comparable to the role-playing humans)."

[17] A more accurate statement might be along the lines of "models are not as good as two undergraduate students at [university] . . . ."

C.5. Negative Examples: Non-Transparent Reporting

A number of studies report almost nothing about their human baselines except for the results. Some examples include: Chiu et al. (2024); Jing et al. (2023); Mukhopadhyay et al. (2024); Yue et al. (2024); Blinov et al. (2022); Gong et al. (2024); Yin et al. (2024). For instance, Gong et al. (2024) contains only two sentences about its human baseline:

> "We perform human baseline on proposed dataset (1200 core VQA samples) with a small group of adult reviewers. The human baseline reports an average accuracy of 89% with 2% standard deviation."

It is nearly impossible to interpret this result: how many "adult reviewers" participated in the baseline? How were these baseliners selected? What instructions were given to these baseliners? Without substantially more detail about study methodology and about the baseliner sample, readers are entirely unable to determine the quality of the baseline and the extent to which the baseline is generalizable to other populations or settings.

D. Appendix: Discussion on Expert Human Baselines

Given the rapid advancement of AI systems, researchers, policymakers, and the general public have an interest in benchmarking AI capabilities against those of the highest performing humans in a given domain. As such, many human baselines are now expert baselines. Some preliminary discussion of these baselines follows:

- Expert baselines are well-suited for estimating the *maximum* possible human performance—i.e., the best that humans can currently perform on an evaluation item (as in, e.g., Obeidat et al. 2024; Laine et al. 2024).[18]
- For any expert baselines, evaluators should develop and report clear standards for what constitutes expertise. Note that not all expertise is tied to professional experience or educational credentials. See Diaz & Smith (2024) for a discussion on constructions of "expertise" in machine learning research.
- Expert populations are often small, and smaller baseliner sample sizes are increasingly more acceptable as evaluation items become more specialized.
- Given the nature and often small population of human experts within a given domain, evaluators may collect convenience samples while maintaining clear standards on expertise. See Wijk et al. (2024) for an example of rigorous recruitment criteria in a convenience sample. Evaluators can also consider snowball sampling—i.e., sampling by asking study participants to recruit other participants from their networks into the study; see Parker et al. (2019) for an overview of snowball sampling.
- Another example of an expert human baseline is Asiedu et al. (2025).
- When calculating results, taking the maximum of all scores per item is acceptable for estimating maximum performance. Note, however, that sample maxima have extreme distributions, and measures of uncertainty should be calculated differently compared to other sample statistics.
- As foundation models are trained on data from across the internet (and thus have "seen" relevant information already), comparing expert performance with internet access with AI system performance is likely a fair comparison, even where the AI system does not have internet access.

[18] Researchers could also attempt to estimate the mean of expert performance, but this estimation would be more difficult as expert populations are often unknown size, smaller than, and less accessible than non-expert populations.

E. 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 E.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 B.

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

E.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 10, then filtered according to the criteria in Table 10. 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.

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 10: Inclusion criteria for meta-review articles.

Our search process yielded a total of 29 articles to be included in our meta-review (listed in Table 11). 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, execution, 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.

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)

Table 11: A complete list of the 29 articles included in our meta-review.

E.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 similar to that of Zhao et al. (2025).

First, we conducted a systematic search for relevant literature. To begin, we queried Google Scholar in December 2024 for articles containing the keywords in Table 12.[19] 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.

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

[19]Note on terminology: the literature currently has no unified, standard terminology for referring to human baselines. Terms such as "human baseline," "expert baseline," and "human performance baseline" (the terms contained in Table 12) are commonly used; however, some authors use "human benchmark" or even "human evaluation." We discovered usage of the term "human benchmark" after our search was concluded; although a qualitative analysis indicated that this term was less frequently used than those in "human benchmark," and a spot check leads us to believe that our results would not differ significantly even with inclusion of this term, the exclusion of this search term is nevertheless a limitation to the coverage of our systematic search. We deliberately excluded "human evaluation" from our search terms, as this term is normally used to describe human annotations, grading, or scoring of AI outputs rather than human baselines (e.g., as used in Howcroft et al. 2020), though a number of our included articles nevertheless contained this term.

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"

Table 12: Search terms for systematic literature review of human baselines

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[20] for articles containing human baselines, and MB identified evaluation datasets with human baselines used in industry evaluations by scanning the model cards/system 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).[21] 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 13 and 14.[22] As Google Scholar does not always index the most authoritative 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.[23] 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 15.

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 all authors, with results stored in a Google Sheet. Questions that arose during the final coding process were adjudicated via discussion. All codes were then validated by KW, PP, SD, and MB (with no coder validating their own codes). After

[20]Elicit is an AI search and analysis for researchers (Elicit).

[21]The date filter in Table 14 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).

[22]Note that although one of our exclusion criteria is for articles published before 2020, we make one exception and nevertheless include one article from 2019 (Dua et al., 2019) due to its prevalence in industry model cards.

[23]Since the time of this writing, published versions of four articles we reviewed have become available. We cite to the preprints that we reviewed below (de Haan et al., 2024; Bai et al., 2024; Mukhopadhyay et al., 2024; Lei et al., 2024a), but the published versions are available at: de Haan et al. 2025; Bai et al. 2025; Mukhopadhyay et al. 2025; Lei et al. 2025.

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 multimodal) 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

Table 13: Inclusion criteria for systematic review of human baselines.

validation was complete, 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 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 six 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; Laine et al., 2024; Suvarna et al., 2024), bringing the total number of human baselines up to 115.

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 observational/real-world data or a human baseline from a pre-existing dataset)	• Articles using observational/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

Table 14: Exclusion criteria for systematic review of human baselines.

	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); 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); de Haan et al. (2024); Hackenburg et al. (2023); Hamotskyi et al. (2024); Heiding et al. (2024); Hendrycks et al. (2021); 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); 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)

Table 15: A complete list of the 109 articles included in our systematic review of human baselines. Note that we analyze 115 individual baselines from these articles, as a single article may contain multiple baselines (see explanation in text).

F. Appendix: Additional Resources

This appendix contains a non-exhaustive list of further resources for researchers interested in running a human baseline. Most items in this list are cited in the relevant sections above, and we provide this list purely for convenience.

For practical guidance on designing human baselines (and surveys):

- Aside from this paper, your first stop for highly practical guidance on baseline design should be the appendix in Stantcheva (2023),[24] which contains detailed information on sampling methods, crowdsourcing, writing survey questions, response bias, and other useful resources.
- For guidance on validity in AI evaluations, see Salaudeen et al. (2025).
- For guidance on power analysis, see Card et al. (2020).
- For best practices on ensuring data quality on Amazon Mechanical Turk, see Lu et al. (2022a).
- For methods to prevent AI usage in human baselines, see Veselovsky et al. (2023a).
- For additional best practices in AI evaluation, see Reuel et al. (2024); Paskov et al. (2024); Biderman et al. (2024); Grosse-Holz & Jorgensen (2024).
- For a summary on uncertainty estimation and other statistical methods, see Miller (2024) (with the caveat of Bowyer et al. 2025).
- For early work on dealing with small sample sizes in evaluations, see Luettgau et al. (2025); Xiao et al. (2025).
- For best practices on reproducibility and transparency in ML research, see Kapoor et al. (2024a); Semmelrock et al. (2024).

For more comprehensive reference texts:

- For an introduction to measurement theory, see Bandalos (2018).
- For an introduction to (human) survey research, see Groves et al. (2011) or Valliant et al. (2018).
- For an introduction to survey sampling (likely more complex than is necessary for most evaluations), see Lohr (2022).
- For an introduction to survey weights, see Valliant et al. (2018).
- For an introduction to hierarchical modelling and sample size estimation, see chapters 8 and 11 of McNulty (2021) (chosen as it provides examples in Python).
- For a discussion of measurement equivalence or measurement invariance (ensuring that measurement instruments capture the same concept across different populations), see Davidov et al. (2014).[25]
- For discussions on using and interpreting p-values in statistics, see McShane et al. (2019); Gelman & Stern (2006).

[24] Available directly at: `https://www.annualreviews.org/content/journals/10.1146/annurev-economics-091622-010157#supplementary_data` (archived at `https://perma.cc/EZ9X-XK6A`).

[25] Although this discussion is in the context of measurements in different populations of *humans*, many measurement equivalence questions also arise between populations of humans and AI models.

G. 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`.