Benchmarks as Microscopes: A Call for Model Metrology

Michael Saxon **Ari Holtzman** **Peter West**
William Yang Wang **Naomi Saphra**
University of California, Santa Barbara University of Chicago
University of Washington Kempner Institute at Harvard University

Abstract

Modern language models (LMs) pose a new challenge in capability assessment. Static benchmarks inevitably saturate without providing confidence in the deployment tolerances of LM-based systems, but developers nonetheless claim that their models have generalized traits such as reasoning or open-domain language understanding based on these flawed metrics. The science and practice of LMs requires a new approach to benchmarking which measures specific capabilities with dynamic assessments. To be confident in our metrics, we need a new discipline of *model metrology*—one which focuses on how to generate benchmarks that predict performance under deployment. Motivated by our evaluation criteria, we outline how building a community of model metrology practitioners—one focused on building tools and studying how to measure system capabilities—is the best way to meet these needs to and add clarity to the AI discussion.

1 Introduction

Just how good are our current language models? It's hard to say. Even the latest benchmarks are not scalable, relevant, or durable enough to predict performance in real-world settings.

Although engineering (Saka et al., 2024; Goyal et al., 2024), research (Bommasani et al., 2021; Bai et al., 2022), and policy (Cihon, 2019; NIST, 2023) decisions are grounded in the supposed capabilities of AI systems—particularly language models (Tolan et al., 2021; NIST, 2022)—even experts disagree about the nature (Li et al., 2023; Morris et al., 2023) and extent (Jumelet & Hupkes, 2018; Bender & Koller, 2020; Park et al., 2022) of these capabilities.

Most LM breakthroughs are judged either through *aggregate performance on narrow benchmarks* (Achiam et al., 2023) or *one-off manual analyses* (Bubeck et al., 2023). These assessments then inform conversations around scaling (Hoffmann et al., 2022), risk (Falco et al., 2021), and deployability. Popular static benchmarks inevitably **saturate** (Beyer et al., 2020; Ott et al., 2022) as each consecutive generation of models over-optimizes for performance on its evaluation sets—a process exacerbated by those datasets contaminating future training data—without resolving fundamental impasses over the nature of LMs or usefully informing their deployment. *We need benchmark practices that yield meaningful observations.*

The emergence of a new discipline: from homemade microscopes to optical metrology

In 1609, Galileo Galilei built one of the first optical telescopes. When he turned it around, Galileo found that he could also observe very small objects close up—and so microscopy was born (Singer, 1914). For centuries, these tools were built by the same scientists using them to make fundamental discoveries (La Berge, 1999).

Eventually, the expertise required to design the precise tools to advance science outstripped scientists' glassworking skills. By the 20th century, large teams of specialists built orbital space telescopes (Leverington, 2012) and microscopes became mass-manufactured commodities (Davidson & Abramowitz, 2002). The science and engineering of measurement tools have coalesced into specialized disciplines.

Corresponding authors: saxon@ucsb.edu, nsaphra@fas.harvard.edu

At present, our LM evaluation practices resemble the state of astronomy and microbiology in the early 17th century—the same community analyzing the object of study (models) is also building the tools for that analysis (benchmarks). We believe that LMs require an independent professional and scientific community dedicated to building analytic tools just as microscope and telescope building have. Just as metrology—the science of measurement—coalesced from a useful skill for natural scientists into an independent discipline, we call for formalizing model evaluation research into a new field, **model metrology**.

First, we enumerate fundamental problems in language model assessment (**Section 2**). Popular benchmarks wrongheadedly attempt to measure *generalized capabilities*, a poorly defined goal (§2.1), distracting us from capturing real-world utility (§2.2). These issues and others, such as static benchmark saturation, are widely recognized, but have gone unsolved due to cultural disconnects between model builders, evaluators, and real-world users (§2.3).

From there we identify critical qualities that useful and concrete benchmarks *should* have (**Section 3**)—**constrained** settings, **dynamic** examples, and **plug-and-play** deployability—to motivate our proposals for rectifying the aforementioned issues, leading into our discussion of how a dedicated model metrology community is the way to provide them (**Section 4**).

Model metrologists can serve as a bridge between scientists, practitioners, and users (§4.1) and build benchmarks that meet our desiderata (§4.2). Within their field they will share knowledge, techniques, tooling, and theory (§4.3) to enable rigorous critique and auditing of other metrologists' work (§4.4), and advance the overall rigor of AI science. But how can we build the model metrology field?

Section 5 introduces possible first steps, identifying existing disparate communities of domain-specific *proto-metrologists* and discussing how they may be organized into a unified community (§5.1). By soliciting real-world measurement needs (§5.2) and engaging with AI subfields that need better measurement practices (§5.3), the field can naturally grow.

Section 6 concludes by noting the role metrology can play in high-level discussions around the fundamental nature of AI (§6.1), and how AI researchers—like the astronomers, phyicists, and biologists who came before them—might transition model metrology from an exploratory science, to a formal field, to finally, a mature engineering discipline (§6.2).

2 Problems with current benchmarks for LMs

While benchmarks have long driven AI progress, they are now used to support increasingly grandiose claims. When research communities believe that "solving" a benchmark represents core progress toward generalized intelligence, interest and investment naturally follow. Raji et al. (2021) document how the *common task framework*—public contests between systems assessed on common train and test sets (Donoho, 2017)—enabled advancements in concrete and tightly-scoped problems such as automatic speech recognition and machine translation, but has since been inappropriately extended to claim generalized capabilities in pretrained vision (Russakovsky et al., 2015) and language (Wang et al., 2019) models.

Current language modeling practices have shifted from training and testing on specific benchmark datasets to testing models in a zero-shot setting. Consequently, many researchers assume that *because* LMs aren't deliberately trained on task-specific train sets (Ge et al., 2023; Bai et al., 2024), performance on these benchmarks is *stronger evidence* of general capability than for fine-tuned models (Piantadosi & Hill, 2022; Mitchell & Krakauer, 2023). Though these assumptions are controversial—evidenced by the remarkably divergent perspectives of NLP and AI researchers (Michael et al., 2023)—the glamour of the promise of general intelligence has carried claims of generalized intelligence to a credulous public (Neri & Cozman, 2020), driving anxieties over AI risk (Ambartsoumean & Yampolskiy, 2023).

These attempts to assess general capabilities address a legitimate need: evaluation is important for guiding advances and comparing models (Phillips et al., 2000). Because modern LMs are used as *everything systems*, we may naturally wish to characterize their general capabilities (Morris et al., 2023). However, *when attempting to assess general capabilities, evaluations neither capture general competency nor predict performance on many downstream applications*.

On the wrasse fish & the pitfalls of generalized capabilities

As the apocryphal Einstein quote goes, "if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid" (O'Toole, 2017).

Ironically, there is at least one example of fish intelligence outpacing primates, namely the economic puzzles solved in labs by the cleaner wrasse, a symbiotic species that lives in coral reefs and feeds on the parasites of larger fish. The fish are given meal options which should be eaten in a particular order due to variable reliability, and they find the optimal solution faster than capuchins, chimpanzees, orangutans, and even one researcher's four-year-old daughter (Salwiczek et al., 2012).

The wrasse evolved to preferentially treat regular customers over reef visitors, tracking clientele across thousands of daily parasite cleanings (Gibson & Barnes, 2000). An artificial "wrassebot" would be ready to deploy only when it exhibits game theoretically-optimal strategies and "machiavellian" (Bshary, 2011) manipulations of clientele—but there's no need to solve grade school math problems or translate text.

2.1 Generalized capabilities are hard to define and contentious.

Narrow LM capability benchmarks are often derived either from tests for humans—e.g., GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021)—or from *existing constrained benchmarks* for specific engineering problems in natural language processing such as question answering (Ho et al., 2023) or entailment recognition (Bowman et al., 2015). Performance on these benchmarks often predicts performance on similar test sets on the same task, to the degree that a violation of this expectation can be evidence of test-set contamination (Paster, 2023; Jain et al., 2024, fig. 5). However, claims regarding the real-world reliability of these systems are often unsupported (Liao et al., 2021). We believe these benchmarks cannot characterize broader capabilities, even when aggregated.

Critics of generalized capabilities benchmarks note that they lack **construct validity**—strong evidence that any evaluation represents a capability (O'Leary-Kelly & Vokurka, 1998) that they claim to measure (Davis, 2023). This discordance between metrics and resulting claims was already present for fine-tuned model evaluation (Raji et al., 2021) but has since worsened, as LM developers and their allies claim generalized intelligence (Bubeck et al., 2023), often based on a huge set of limited benchmark scores (Fei et al., 2022; Achiam et al., 2023). Similar claims are made for more specific abstract capabilities when researchers attribute benchmark performance to faculties like abstract reasoning (Yasunaga et al., 2021), language understanding (Moore, 2022), or common sense knowledge (Zhao et al., 2023b).

We are at an impasse. Though these claims of generality are contested (Murty et al., 2023), they are hard to conclusively disprove. Bender & Koller (2020) argue axiomatically that understanding cannot be acquired through the LM objective. Poor generalization across time (Lazaridou et al., 2021), tasks (Yang et al., 2022), and heuristics (Singhal et al., 2023) have also been provided as empirical counterarguments to LM capability claims. Given that humans struggle to even evaluate intelligence in other animals (De Waal, 2016), how can we assess slippery abstract notions like reasoning (Manning, 2022) in AI systems?

2.2 Benchmarks can aim for generality—or they can be valid and useful.

When benchmarks claim to test abstract capabilities, critics often question whether that capability is *necessary* or just *sufficient* for their solution (Potts, 2020). To substantiate a capability claim, a task must require said capability, but it may be impossible to prove that relationship. After all, both a studious human scholar and an answer key achieve 100% accuracy on an exam, but a piece of paper clearly does not possess the scholar's intelligence.

Consider a developer of a real-world application based on an LM, which we dub *builder-consumers*. For a builder-consumer, a useful benchmark must simply test if a system—regardless of abstract capabilities—is performant on their task. Meaningfully representing the deployment setting makes a benchmark **ecologically valid** (De Vries et al., 2020).

Are generality and ecological validity fundamentally in tension? Recent efforts to unify these goals such as HELM (Liang et al., 2023) provide a large collection of scores by harvesting existing benchmarks divided into specific scenarios (e.g., news domain tasks). By contextualizing these tasks within categories—and providing disaggregated scores over them—they aim to preserve task-level construct validity (Liao & Xiao, 2023) while capturing a "holistic" view of an LM's capabilities. Though many of these tasks may have construct validity, these holistic evaluation attempts do not capture generality—each benchmark represents a tiny view of a broader "task universe" (Liao et al., 2021). Ecological validity is particularly problematic when discussing "AGI-level" capabilities (Morris et al., 2023), though discussing the credibility of those notions is not central here.

Most model consumers are developing applications that rely on predictable LM behavior and therefore need to evaluate consistent, constrained capabilities for their domain. However, developers such as OpenAI (Achiam et al., 2023), Anthropic (Anthropic, 2023), and Google (Team et al., 2023; 2024) instead continue to focus on the same MMLU, GSM8K, and HumanEval test suites rather than on application-specific tests. These platforms provide API access to these models as a paid service, so *why aren't they benchmarking customer-relevant capabilities?* Maybe they are convinced that general intelligence is quantifiable by these benchmarks. Perhaps the deeper issue is that building bespoke evaluations is hard, and the domains are innumerable—will a collection of constrained tests ever be large enough to placate critics? Do scientists have a role to play in producing these gap-closing evaluations?

2.3 We know existing benchmarks are flawed. Why do we keep using them?

These criticisms are not novel—indeed, they're commonly expressed sentiments. Regardless, these flawed benchmarks remain dominant. Saturation is broadly acknowledged as a problem whose mechanisms present a fundamental challenge to benchmark validity. GSM8K, long used to assess mathematical reasoning, is fully saturated for "frontier" models—GPT-4 achieves near-100% accuracy with prompting and decoding tricks (Zhou et al., 2024a). We know models overfit even to hidden (but static) test sets (Gorman & Bedrick, 2019). We know that the long tail of incomprehensibly large pretraining datasets (Mitchell et al., 2022; Elazar et al., 2023) inevitably enables answer memorization (Alzahrani et al., 2024) or heuristic learning (Poliak et al., 2018; McCoy et al., 2019; Wang et al., 2021; Saxon et al., 2023) through similar examples (Peng et al., 2023; Kandpal et al., 2023). Poor construct validity is widely noted (Jacobs & Wallach, 2021), as is the futility of measuring generalized capabilities (Casares et al., 2022). So why do we still rely on these benchmarks? Perhaps:

1. Misalignment of interest/incentives for researchers and needs of users.
2. Fundamental difficulties in building benchmarks that meet our desiderata.
3. The allure of general intelligence attracts public, media, and investor attention.

As long as the LM community is the primary benchmark building community, these problems will persist. The incentives for academic researchers building benchmarks is 'impact,' as measured through citations and public use (Kang et al., 2023), and primary incentive for industrial actors is to demonstrate the superiority of their latest product. Unsurprisingly, 'top benchmarks' from a small number of elite institutions have become primary measures of AI progress (Koch et al., 2021). In the short term, a carefully-scoped and rigorously designed benchmark is as impactful as a well-hyped but soon-to-saturate one—but the former is much harder to make than the latter. The development of best practices for benchmark building cannot rely on the incentives of scientific machine learning research.

3 Qualities of useful, concrete benchmarks

The problems above can only be solved if we *abandon general metrics when making deployment decisions*. **Why study the wrasse fish's intelligence outside of the reef?**

We need a benchmarking culture and practice that empowers consumers to *specify their desired constraints* and *generate their own benchmarks*. Model producers should track progress by these scenario-driven evaluations. Good LM capability evaluations are:

Constrained. Benchmarks that characterize model behavior on a concrete task where domain experts can describe the boundaries a good system should stay within are *constrained*. Often, concrete problems can be defined in terms of verifiable rules. Given these boundaries as an objective, issues of scenario innumerability and ecological validity become much more tractable, as we can rely on domain expert understanding of real-world inputs.

Dynamic. A fixed dataset is easily memorized. To avoid stale metrics, we prefer dynamically generated test simulations where performance is measured by task-specific constraints.

Plug-and-play. Benchmark generating processes must be plug-and-play, i.e., run easily using customer-specific models or constraints. Expensive, one-off benchmarks are unlikely to foster open academic discourse—or provide value to downstream developers—and human preferences are expensive to gather and difficult to replicate. Accessible evaluation setups will entice domain experts to apply and even publish their constraints.

These desiderata are mutually reinforcing and enabled by *abandoning generality* and centering the needs of real-world builder-consumers and end users. Because model developers and AI researchers usually aim to improve general capabilities, focused real-world evaluation must instead rely on dedicated practitioners, i.e. metrologists.

4 The promise of a model metrology community

In this section, we explain how and why a dedicated discipline of model metrology can resolve the aforementioned issues, producing benchmarks that meet our desiderata and improve the state of LM production, use, and analysis. We lay out the benefits a dedicated metrology community would bring, the types of problems it would enable better solutions to, the research and engineering activities metrologists might undertake, and the cultural changes in broader AI science it would effect, discussing existing relevant work throughout.

4.1 A dedicated community can better connect researchers, developers, and users.

Given their role in spurring investment, flawed benchmarks and metrics lead researchers to waste time and effort developing methods that are ill-suited to any real-world application (Hoyle et al., 2021). Currently, LM benchmarking is disengaged from builder-consumers (Yang et al., 2023) and model-based service end-users (Xiao et al., 2024). Model metrology requires social change in both scientific and product communities to mitigate this disengagement. Even if tools and methods existed that could convert user constraints into quality dynamic benchmarks (see §4.2), enabling prospective LM-based application developers to successfully capture their use case in specifications may still be a challenge.

For example, consider the *Air Canada chatbot incident*. A Canadian court found Air Canada liable for paying a customer a nonexistent, off-policy bereavement discount promised by an LM-based customer service agent (Melnick, 2024). This agent, presumably using a GPT-based model, failed to adhere to policy—thereby failing as a customer service agent—despite GPT-3.5's near state-of-the-art performance on a massive suite of generalized benchmarks.

This incident might be described as the underlying LM lacking several different abstract capabilities. Perhaps the agent failed at *rule-following*, if the rules for bereavement discounts was specified somewhere in the system prompt context. Perhaps the agent failed in *common-sense reasoning*, if this specific policy wasn't explicitly provided but the list of all allowable discounts was. Regardless of the source of the error, it is unlikely that this specific failure case could have been predicted through generalized benchmark results.

However, a builder-consumer developing customer service agents could have tested for failures like this, but currently lacks the tooling to do so. Within the customer service agent domain, *common sense* entails not making promises that violate policy. A developer could manually enumerate every line of the policy, probing the agent for examples where it would fail. Model developers aren't thinking what abstract capability failures mean in diverse domains, *but domain experts who will use the models know what their needs are*. Model metrologist-developed methods will enable builder-consumers to plug their own constraints in to generate ecologically valid benchmarks for their specific task.

4.2 Metrologists will produce targeted dynamic benchmarks for complex problems.

How might a dedicated community build evaluations that meet our desiderata? Let's consider the Air Canada incident. Suppose a developer had a concrete list of rules for a customer service agent to follow, including not promising off-policy transactions. How do we use such rules to dynamically evaluate an LM for this constrained domain?

One technique could be to leverage an adversarial LM as a source of variation, generating many test cases attacking the task-domain constraints. In our airline customer service example, an LM could be prompted to generate role-play scenarios of various challenging customers: a child pranking the system, a client who struggles with technology, a jailbreaker looking for a big discount, or a panicked and angry stranded traveler.

Model outputs conditioned on these adversarial test dialogs could be judged deterministically against policy constraints, detecting issues like diversion from company policy. Even though the evaluation is dynamic, no human evaluator needs to manually enumerate all edge cases, as an expert has already fixed the deterministic rules. This style of evaluation exists—within the silo of LM security research (Shayegani et al., 2024; Zhu et al., 2023)—a metrology community would further its development, dissemination, and deployment.

Evaluation scenarios like this one are possible with current technology, as LM adversaries have been already been used for stress testing and assessment (Chan et al., 2024). There is mounting evidence that, using reversal, LMs can generate exemplars that they can't correctly respond to (Berglund et al., 2023; West et al., 2023). Given well-scoped constraints, model outputs can be evaluated deterministically, e.g., using variation between minimal pairs (Ribeiro et al., 2020) or fulfillment of a set of requirements (Hu et al., 2023), rather than using arbitrary and opaque LM-judgements of dubious reliability (Oh et al., 2024).

Though we have proposed a dynamic evaluation pipeline for one constrained setting, we are not claiming to have described the best way to produce a strong benchmark-generating process for all settings. Concerted research is necessary to develop best practices for metrology. Professional model metrologists will have to be competent at developing, formalizing, and sharing insights from disparate benchmark efforts for the benefit of the field.

4.3 Model metrologists will establish shared knowledge & techniques.

Even as we develop increasingly sophisticated evaluation methods, our community lacks consensus on their validity and best practices for their use. Metrics that use automatic scores from reference-similarity (Kocmi & Federmann, 2023) or correctness (Wang et al., 2023a; Mizumoto & Eguchi, 2023) are hotly debated (Chiang & Lee, 2023). For metrologists to use a technique with confidence, they need community consensus on its efficacy. We need:

Shared framings of abstract capabilities across concrete settings. Although abstract capabilities like "reasoning," "understanding," or "rule-following" are ill-defined and unquantifiable in general, they can be used to frame desired behavior and edge cases to avoid in constrained settings. For example, the apparent lack of rule-following exhibited in the Air Canada incident suggests that constraint-based adversarial agent testing may uncover failure modes in a customer service chat bot setting. Techniques developed in pursuit of that evaluation could probably be leveraged for many other rule-following problems in other constrained settings. Transfer of this knowledge would be facilitated by having dedicated metrologists—rather than customer service chat bot developers—building and promoting these constrained evaluations. By comparing the results of a method deployed across disparate settings, metrologists will guide further tool development.

Evidence Centered Benchmark Design (ECBD) (Liu et al., 2024) is an example of proto-metrology work toward this direction. They lay out a framework for assessing whether a benchmark actually captures a desired (often abstract) capability through analysis of specific "test items" within a target application context. Effort in using a standardized framework to describe a capability in one domain may transfer to another, and through the accumulation of evidence on how these framings perform in the wild, metrologists will develop a more sophisticated vocabulary to develop the practice of model measurement and assessment.

A shift from observations to theories and science. The AI community is sharing observations about LMs too fast for any one researcher to follow. Without replication and meta-analyses, scientists cannot determine which observations expose meaningful patterns and which represent random idiosyncrasies. For example, the evidence connecting benchmarks on various domains is weak (Fergusson et al., 2023). Although different math tests are usefully correlated (Paster, 2023) and small test sets of under 100 samples estimate on large static multi-task benchmarks (Polo et al., 2024), the future of these approaches remains uncertain. Can we predict the limits of this generalization? Will it hold for new classes of models? Will it hold for dynamic benchmarking or constraint-based benchmarks? A dedicated discipline could synthesize the growing evaluation literature by replicating, analyzing, and eventually canonizing findings into useful *theories of metrology*.

The luminiferous aether & advances in science driven by surprising measurements

When 17th century physicists first proposed that light is a wave, they posited that it must travel through a physical medium, which they dubbed the *luminiferous aether*. As the earth moves through space, the theory predicted that an *aether wind* would be observed, making the speed of light on earth different in different directions.

However, the aether remained unobservable until the late 19th century invention of the *interferometer*, an instrument to measure light interference patterns. The Michelson-Morley experiment, intended to demonstrate the direction of the aether wind by comparing the speed of light in orthogonal directions, failed to find any differences. **Enabled by advances in measurement technology**, this "most famous failed experiment" (Blum & Lototsky, 2006) revolutionized 20th century physics, ultimately giving rise to relativity and quantum theory (Shankland, 1964).

By producing new measurement tools, model metrologists can not only validate existing models but drive LM science as a whole. As in other sciences, better measurements can precipitate questions that our current scientific paradigms are not yet capable of asking (Kuhn, 1962). For example, improved metrics can enable more sophisticated testing of scaling laws (Schaeffer et al., 2024) and discover associations between specific capabilities and error types. While it is impossible to predict what future paradigm shifts will look like, we are confident that surprising yet high-confidence observations—which metrology is intended to enable—will have an important role to play.

Quality benchmark-building tools. In the long term, metrologists should aim to develop tooling for automated benchmark generation by domain experts who are not necessarily evaluation experts, as discussed above. This will require technical innovations such as methods to expand a high-level task description into a set of exemplars, or prompting an LM to behave adversarially against a task-specific LM system. Among other tools, metrologists must develop prompting techniques that test the boundaries of rule-following in one setting (e.g., customer service) and generalize better to other settings (e.g., planning navigation). These concretely motivated—but generalized—techniques would be more appropriate than those explicitly designed for reasoning assessment. Perhaps the best way to test rule-following is to monitor agents interacting with simulated situations. Proper evaluation, however, requires both creativity and broad knowledge. Metrologists could stress-test chat bots by eliciting interactive personae by prompting (Cheng et al., 2023).

Task Me Anything (Zhang et al., 2024) is a recent example of work toward automated evaluation that is accessible out-of-the-box to non-experts. The authors propose a technique to build a multimodal LM evaluation set to answer specific queries about language model capabilities such as "which model is best at counting objects?" by selecting samples from existing static benchmarks. Techniques such as this coupled with sample generation strategies could be shared between metrologists across application domains in combination with other tools such as automated sample generation to build better dynamic benchmarks.

While some tasks are well-suited for constraint-based evaluation, for others (such as general purpose chat agents) the target is human preference. The gold standard for human preference evaluation in interactive LM applications are *competitive interactive evaluations* such as Chatbot Arena (Chiang et al., 2024) where multiple systems are compared head-to-head on genuine human-generated queries. Although "chat agents" are not a particularly constrained domain, these evaluations do capture genuine human preference dynamically. Unfortunately, they are expensive to run and inconsistent, as any new system will be run head-to-head against all other models on new human interactions that must be collected over time. In light of these shortcomings *LM-as-a-judge* techniques have been proposed, where an evaluator language model is used as a proxy for human preference feedback (Zheng et al., 2023). Metrologists might expand this technique beyond chat bot evaluation.

4.4 Metrology culture prioritizes data work, methodological rigor, and proactive criticism.

Despite its importance, data work is devalued in AI research communities compared to more glamorous theoretical, empirical, and modeling work (Sambasivan et al., 2021). Even without a cultural shift in AI research, a metrology-focused community should center data and benchmarking work as first-class contributions.

Recent existing work has promoted rigor in benchmarking by identifying model cheating on benchmarks (Chen et al., 2024), finding errors in dynamic benchmark generators (Saxon et al., 2024a), and producing meta-metrics to find benchmark failure modes (Saxon et al., 2024b). However, these efforts are *reactively* responding to flawed work, rather than *proactively* identifying best practices for capability measurement. A metrology community should aspire to be the latter.

5 How do we build the model metrology discipline?

Having established the motivation, purpose, and necessity of the dedicated discipline of model metrology, we now discuss ways we might build the field. Because an evaluation discipline is not part of the existing language modeling community's culture—including among model consumers—metrology can only be built alongside substantial social change.

5.1 Uniting proto-metrology communities

Many ML, NLP, and AI conferences already hold benchmarks and evaluation tracks. Work on metric development and benchmarking best practices has regularly appears at ICLR (Lu et al., 2024), *ACL (Maynez et al., 2023), CVPR (Xu et al., 2022), and NeurIPS (Zhang et al., 2023). These researchers are effectively *proto-metrologists* establishing foundations for this field. As a starting point, current and aspiring metrology researchers should be familiar with these pioneering works. Workshops and evaluation-focused venues could facilitate cross-engagement between aspiring metrologists and ultimately provide an intellectual home. For inspiration, we look to existing communities of proto-metrologists concentrated in domain-specific venues (Saphra et al., 2024).

The machine translation (MT) community has invested considerably in rigorous benchmarks and metrics. The Conference on Machine Translation (WMT) (Kocmi et al., 2023) has run shared tasks that *simultaneously benchmark translation systems and translation quality metrics* since 2006 (Bojar et al., 2016). With buy-in from both academic and industrial researchers developing MT systems, the annual WMT shared task evaluates MT systems and metrics on new language pairs and domains every year. In so doing, the MT community continually refreshes their measurement practices as the field advances. Those MT researchers focused primarily on building quality evaluation metrics are effectively *MT metrologists* already.

Similarly, the Generation, Evaluation & Metrics (GEM) Workshops (Gehrmann et al., 2023), focused on advancing and evaluating text generation systems for specific tasks like data-to-text and summarization, are another good example of a proto-metrology community. Alongside shared tasks on building these text generation systems, they emphasize research on metrics for evaluating generated text against gold references to better compare competing submissions. Their GEM benchmark was an early attempt at building a protocol for living text generation assessment where new tasks and metrics could be slotted into a comprehensive benchmark over time (Gehrmann et al., 2021).

Strong proto-metrology work is continuously being published—both within these communities and at larger venues—yet knowledge transfer between these proto-metrologists is scattershot. Topic modeling researchers have assessed how different topic model quality metrics vary considerably across corpus domains (Hoyle et al., 2021), yet despite its general transferability, this finding is confined to the topic modeling community. Text-to-image researchers have invented intricate ways to generate directed graphs of requirements to check prompt-image faithfulness (Cho et al., 2023), yet outside this community, these techniques have not disseminated. The move from annual static competitions (Harman, 1993) to dynamic, modular packages (Thakur et al., 2021) has greatly advanced the practical state-of-the-art in information retrieval—but researchers outside IR may be unaware.

Right now we effectively rely on happy accidents to spread evaluation knowledge across these disciplinary barriers. By organizing this work not as "`[domain]`'s evaluation research" but as *core model metrology research applied to* `[domain]`, we can render this knowledge transfer routine. Model metrology venues should bring the culture of WMT and GEM to evaluation writ large, so lessons from all these domains can be shared—to start we should collect and promote these disparate threads of capability measurement research as a cohesive whole.

5.2 Soliciting novel constraints and edge cases to benchmark

A model metrology community must be built by engaging with domain experts and model consumers. Metrology practitioners can use case studies such as the Air Canada incident as starting points to experiment with constrained benchmark-generating systems, but ultimately *useful constraints can only be provided by domain experts* who know the boundaries of their tasks. Academic metrologists and metrology venues should actively solicit specifications for new tasks. These expert-designed settings could be framed as shared tasks or even as concrete evaluation bounties. While we expect that industry and nonprofit customers developing LM-based applications will happily contribute their constraint specifications, as they stand to benefit most when their needs are prioritized by model training institutions, incentivizing academic researchers to invest in this work may prove challenging.

After all, data-focused conferences such as LREC already publish specialized training and test sets for constrained-domain tasks—but our academic incentive structures based on 'prestige' and technical 'novelty' devalue these venues compared to general AI venues like NeurIPS. A metrology community may effect sociological change toward valuing this grounded and concrete work by providing opportunities for genuine technical novelty (eg., benchmark generating processes) alongside crucial data work on builder-consumer-relevant constraints. Practical model metrology will probably emerge as a profitable industry where consultants operate similarly to penetration testing teams in information security, customizing stress tests for each scenario based on their broad knowledge of evaluation.

5.3 Engaging with related fields

At its start, model metrology draws on the evaluation work already taking place in many domains of machine learning, artificial intelligence, and natural language processing research. Though organizing this work in a cohesive community has many benefits, model metrology will only succeed if it continues to closely engage with its parent communities—but those communities stand to benefit from a stronger metrology culture as well.

Empirical research requires measurable dependent variables. Beyond building better benchmarks, model metrologists will be well-positioned to build *ecologically valid observational tools* to inject rigor into capabilities-focused empirical research. In particular, we anticipate black-box model analysis (Belinkov et al., 2023), human-computer interaction (Liu et al., 2023), robustness (Zhong & Wang, 2023), and interpretability (Räuker et al., 2023) research will benefit from engagement with metrologists.

While it is widely accepted that model metrics improve at larger scales, the metrics chosen are often critiqued. Scaling laws have been claimed on LM loss (Hoffmann et al., 2022), static benchmark performance (Srivastava et al., 2023), or—in the extreme case—a completely decontextualized y-axis simply labeled "intelligence" (Anthropic, 2023), yet these metrics cannot fully support popular claims about scaling laws in general intelligence or practical utility. A different test set and metric can elicit inverse (McKenzie et al., 2023) and U-shaped scaling curves (Wei et al., 2023) with respect to model size, demonstrating a need for serious discussion of measurement methodology led by metrologists.

6 Conclusions

The time has come for disparate threads of research in LM benchmarking and evaluation scattered across many domains to coalesce into a cohesive model metrology community.

Investments into techniques to build constrained, dynamic, and plug-and-play evaluations to end user or builder-consumer specifications will enable more granular and concrete model evaluation. Adoption of these granular evaluations at scale will enable model makers more holistically argue for the utility of their LMs over competitors.

A community focused on these best practices for measurement and assessment will in turn be able to leverage this knowledge to identify new experimental directions into deeper model capabilities. If successful, the agenda will operationalize many areas of interest within AI, including alignment, fairness, common sense, knowledge, and reasoning. These nebulous objectives can be freed from the innumerability and construct validity issues they acquire when treated as generalized open-domain capabilities.

Metrology will produce real-world applicable evaluation techniques to help LM users make informed decisions and help model developers track incremental progress. We believe grounded progress benchmarks will also improve public discourse around AI.

When benchmarks attempt to simultaneously track core scientific progress and product effectiveness, they fail to achieve either. A dedicated evaluation discipline will be able to make great evaluations for each goal with shared methodology.

6.1 Model metrology & the artificial general intelligence (AGI) discussion

Present benchmarking culture suffers from conflicting efforts to practically compare models and to produce evidence in debates about AGI. Despite many high-profile cases of proven benchmark contamination, AGI optimists continue to equate improvement on static benchmarks with increases in robust intelligence. In their framing, critics are "moving the goalposts" by dismissing newly saturated benchmarks. Meanwhile, skeptics spin underperformance on a specific benchmark as evidence that a target capability is impossible for LMs. As each benchmark saturates, the cycle of hype and deflation continues and little is learned.

We believe our vision for metrology is useful as it directly gets at a core reason most people actually care about the AGI: the promise of making drop-in replacements for humans in specific jobs. If this really is what we care about, why not measure it directly? Constrained and ecologically valid benchmarking can finally decouple practical evaluation from ideological arguments about AGI.

A culture of model metrology will hopefully drive everyone to make *weaker statements* about intrinsic model capabilities grounded in quantifiable real-world capacities. These evaluation experts can promote a healthier, calmer public-facing discourse around AI.

6.2 The end game: model assessment *without* a model metrologist

In the best-case scenario, the model metrology agenda succeeds by building a mature engineering discipline and making standardized off-the-shelf benchmark-generating processes a commodity. Our proposal is modeled on the history of microscope manufacture. For most applications requiring a microscope, the exact desired instrument is already mass-produced. For truly niche applications (e.g., assessing semiconductor deformation) custom-built metrology solutions are still needed (Houghton et al., 2016). One day, LM consumers should likewise meet their measurement needs out-of-the-box without hiring a metrologist. This independence is the ultimate goal of a model metrology discipline.

Acknowledgements

Thank you to the countless colleagues who have discussed the problems of evaluation with us and who have made public statements on them, inspiring and informing this work.

Thank you Greg Durrett, Tejas Srinivasan, Leif Hancox-Li, Prajjwal Bhargava, John Kevin Cava, and our anonymous reviewers for reading and providing feedback on earlier drafts.

This work was supported in part by the National Science Foundation Graduate Research Fellowship under Grant No. 1650114, and CAREER Award under Grant No. 2048122. This work was enabled in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Miltiadis Allamanis, Sheena Panthaplackel, and Pengcheng Yin. Unsupervised evaluation of code llms with round-trip correctness. *ArXiv preprint*, abs/2402.08699, 2024. URL https://arxiv.org/abs/2402.08699.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *ArXiv preprint*, abs/2402.01781, 2024. URL https://arxiv.org/abs/2402.01781.

Vemir Michael Ambartsoumean and Roman V Yampolskiy. Ai risk skepticism, a comprehensive survey. *ArXiv preprint*, abs/2303.03885, 2023. URL https://arxiv.org/abs/2303.03885.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. Online technical report., 2023. URL https://api.semanticscholar.org/CorpusID:268232499.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv preprint*, abs/2212.08073, 2022. URL https://arxiv.org/abs/2212.08073.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36, 2024.

Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.). *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.blackboxnlp-1.0.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *ArXiv preprint*, abs/2309.12288, 2023. URL https://arxiv.org/abs/2309.12288.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *ArXiv preprint*, abs/2006.07159, 2020. URL https://arxiv.org/abs/2006.07159.

Edward K Blum and Sergey V Lototsky. *Mathematics of physics and engineering*. World Scientific Publishing Company, 2006.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. Ten years of wmt evaluation campaigns: Lessons learnt. In *LREC 2016 Workshop "Translation Evaluation–From Fragmented Tools and Data Sets to an Integrated Ecosystem"*, pp. 27–34, 2016.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Redouan Bshary. Machiavellian intelligence in fishes. *Fish cognition and behavior*, 2:240–257, 2011.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712, 2023. URL https://arxiv.org/abs/2303.12712.

Pablo Antonio Moreno Casares, Bao Sheng Loe, John Burden, Seán Ó hÉigeartaigh, and José Hernández-Orallo. How general-purpose is a language model? usefulness and safety with human prompters in the wild. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 5295–5303. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20466.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *ArXiv preprint*, abs/2403.20330, 2024. URL https://arxiv.org/abs/2403.20330.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. Compost: Characterizing and evaluating caricature in llm simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, 2023.

Cheng-Han Chiang and Hung-Yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=3MW8GKNyzI.

Jaemin Cho, Yushi Hu, Jason Michael Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

Peter Cihon. Standards for ai governance: international standards to enable global coordination in ai research & development. *Future of Humanity Institute. University of Oxford*, pp. 340–342, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Michael W Davidson and Mortimer Abramowitz. Optical microscopy. *Encyclopedia of imaging science and technology*, 2(1106-1141):120, 2002.

Ernest Davis. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.*, 56(4), 2023. ISSN 0360-0300. doi: 10.1145/3615355. URL https://doi.org/10.1145/3615355.

Harm De Vries, Dzmitry Bahdanau, and Christopher Manning. Towards ecologically valid research on language user interfaces. *ArXiv preprint*, abs/2007.14435, 2020. URL https://arxiv.org/abs/2007.14435.

Frans De Waal. *Are we smart enough to know how smart animals are?* WW Norton & Company, 2016.

David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26 (4):745–766, 2017.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *ArXiv preprint*, abs/2310.20707, 2023. URL https://arxiv.org/abs/2310.20707.

Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571, 2021.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 7216–7262. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/16b14e3f288f076e0ca73bdad6405f77-Paper-Datasets_and_Benchmarks.pdf.

Grant Fergusson, Caitriona Fitzgerald, Chris Frascella, Megan Iorio, Tom McBrien, Calli Schroeder, Ben Winters, and Enid Zhou. Generating harms: Generative ai's impact & paths forward. *Electronic Privacy Information Center*, 2023.

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5539–5568. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 96–120, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.10. URL https://aclanthology.org/2021.gem-1.10.

Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz (eds.). *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, Singapore, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.0.

R Gibson and Margaret Barnes. Evolution and ecology of cleaning symbioses in the sea. *Oceanography and marine biology: an annual review*, 38:311, 2000.

Kyle Gorman and Steven Bedrick. We need to talk about standard splits. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2786–2791, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1267. URL https://aclanthology.org/P19-1267.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pp. 1167–1168, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635739. URL https://doi.org/10.1145/3616855.3635739.

Donna Harman. Overview of the first trec conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 36–47, 1993.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*, 2023.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *ArXiv preprint*, abs/2203.15556, 2022. URL https://arxiv.org/abs/2203.15556.

Todd Houghton, Michael Saxon, Zeming Song, Hoa Nyugen, Hanqing Jiang, and Hongbin Yu. 2d grating pitch mapping of a through silicon via (tsv) and solder ball interconnect region using laser diffraction: Ieee electronic components and technology conference, 2016. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pp. 2222–2227. IEEE, 2016.

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan L. Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? the incoherence of coherence. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2018–2033, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0f83556a305d789b1d71815e8ea4f4b0-Abstract.html.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *ArXiv preprint*, abs/2403.07974, 2024. URL https://arxiv.org/abs/2403.07974.

Jaap Jumelet and Dieuwke Hupkes. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 222–231, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5424. URL https://aclanthology.org/W18-5424.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023. URL https://proceedings.mlr.press/v202/kandpal23a.html.

Donghyun Kang, TaeYoung Kang, and Junkyu Jang. Papers with code or without code? impact of github repository usability on the diffusion of machine learning research. *Information Processing & Management*, 60(6):103477, 2023. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2023.103477. URL https://www.sciencedirect.com/science/article/pii/S0306457323002145.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203, 2023.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1962.

Ann Elizabeth Fowler La Berge. The history of science and the history of microscopy. *Perspectives on Science*, 7(1):111–142, 1999.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 29348–29363, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f5bf0ba0a17ef18f9607774722f5698c-Abstract.html.

David Leverington. *A History of Astronomy: from 1890 to the Present*. Springer Science & Business Media, 2012.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. Can language models understand physical concepts? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11843–11861, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.726. URL https://aclanthology.org/2023.emnlp-main.726.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Q Vera Liao and Ziang Xiao. Rethinking model evaluation as narrowing the socio-technical gap. *ArXiv preprint*, abs/2306.03100, 2023. URL https://arxiv.org/abs/2306.03100.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Bill Yuchen Lin, Khyathi Chandu, Faeze Brahman, Yuntian Deng, Abhilasha Ravichander, Valentina Pyatkin, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL https://huggingface.co/spaces/allenai/WildBench.

Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. "what it wants me to say": Bridging the abstraction gap between end-user programmers and code-generating large language models. In

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580817. URL https://doi.org/10.1145/3544548.3580817.

Yu Lu Liu, Su Lin Blodgett, Jackie Chi Kit Cheung, Q Vera Liao, Alexandra Olteanu, and Ziang Xiao. Ecbd: Evidence-centered benchmark design for nlp. *ArXiv preprint*, abs/2406.08723, 2024. URL https://arxiv.org/abs/2406.08723.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Christopher D Manning. Human language understanding & reasoning. *Daedalus*, 151(2): 127–138, 2022.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. Benchmarking large language model capabilities for conditional generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9194–9213, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.511. URL https://aclanthology.org/2023.acl-long.511.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=DwgRm72GQF. Featured Certification.

Kyle Melnick. Air canada chatbot promised a discount. now the airline has to pay it. *The Washington Post*, 2024. URL https://www.washingtonpost.com/travel/2024/02/18/air-canada-airline-chatbot-ruling/.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, et al. What do nlp researchers believe? results of the nlp community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16334–16368, 2023.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data. *arXiv preprint arXiv:2212.05129*, 2022.

Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

Atsushi Mizumoto and Masaki Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050, 2023. ISSN 2772-7661. doi: https://doi.org/10.1016/j.rmal.2023.100050. URL https://www.sciencedirect.com/science/article/pii/S2772766123000101.

Jared Moore. Language models understand us, poorly. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 214–222, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.16.

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi: Operationalizing progress on the path to agi. *ArXiv preprint*, abs/2311.02462, 2023. URL https://arxiv.org/abs/2311.02462.

Shikhar Murty, Orr Paradise, and Pratyusha Sharma. Pseudointelligence: A unifying lens on language model evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7284–7290, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.485. URL https://aclanthology.org/2023.findings-emnlp.485.

Hugo Neri and Fabio Cozman. The role of experts in the public perception of risk of artificial intelligence. *AI & society*, 35:663–673, 2020.

NIST. AI Risk Management Framework: Second Draft. Technical report, National Institute for Standards and Technology, 2022. URL https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

NIST. ARTIFICIAL INTELLIGENCE SAFETY INSTITUTE CONSORTIUM COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT. Technical report, National Institue for Standards and Technology, 2023. URL https://www.nist.gov/system/files/documents/2023/12/15/AISIC%20FINAL%20APPROVED%20TEMPLATE_FINAL%20FINAL%2012152023%20reference%20copy.pdf.

Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. The generative ai paradox on evaluation: What it can solve, it may not evaluate. *ArXiv preprint*, abs/2402.06204, 2024. URL https://arxiv.org/abs/2402.06204.

Scott W O'Leary-Kelly and Robert J Vokurka. The empirical assessment of construct validity. *Journal of operations management*, 16(4):387–405, 1998.

Garson O'Toole. *Hemingway didn't say that: The truth behind familiar quotations*. Brilliance Audio, 2017.

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

Sungjin Park, Seungwoo Ryu, and Edward Choi. Do language models understand measurements? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1782–1792, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.128.

Keiran Paster. Testing language models on a held-out high school national finals exam. https://huggingface.co/datasets/keirp/hungarian_national_hs_finals_exam, 2023.

Zhencan Peng, Zhizhi Wang, and Dong Deng. Near-duplicate sequence search at scale for large language model memorization evaluation. *Proc. ACM Manag. Data*, 1(2), 2023. doi: 10.1145/3589324. URL https://doi.org/10.1145/3589324.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000.

Steven Piantadosi and Felix Hill. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022. URL https://openreview.net/forum?id=nRkJEwmZnM.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *ArXiv preprint*, abs/2402.14992, 2024. URL https://arxiv.org/abs/2402.14992.

Christopher Potts. Is it possible for language models to achieve language understanding? Medium post, 2020. URL https://chrisgpotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483. IEEE, 2023.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Abdullahi Saka, Ridwan Taiwo, Nurudeen Saka, Babatunde Abiodun Salami, Saheed Ajayi, Kabiru Akande, and Hadi Kazemi. Gpt models in construction industry: Opportunities, limitations, and a use case validation. *Developments in the Built Environment*, 17:100300, 2024. ISSN 2666-1659. doi: https://doi.org/10.1016/j.dibe.2023.100300. URL https://www.sciencedirect.com/science/article/pii/S2666165923001825.

Lucie H Salwiczek, Laurent Prétôt, Lanila Demarta, Darby Proctor, Jennifer Essler, Ana I Pinto, Sharon Wismer, Tara Stoinski, Sarah F Brosnan, and Redouan Bshary. Adult cleaner wrasse outperform capuchin monkeys, chimpanzees and orang-utans in a complex foraging task derived from cleaner–client reef fish cooperation. *PLoS One*, 7(11):e49068, 2012.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

Naomi Saphra, Eve Fleisig, Kyunghyun Cho, and Adam Lopez. First tragedy, then parse: History repeats itself in the new era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2310–2326, 2024.

Michael Saxon, Xinyi Wang, Wenda Xu, and William Yang Wang. PECO: Examining single sentence label leakage in natural language inference datasets through progressive evaluation of cluster outliers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3061–3074, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.223.

Michael Saxon, Yiran Luo, Sharon Levy, Chitta Baral, Yezhou Yang, and William Yang Wang. Lost in translation? translation errors and challenges for fair assessment of text-to-image models on multilingual concepts. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 572–582, Mexico City, Mexico, 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-short.48.

Michael Stephen Saxon, Fatima Jahara, Mahsa Khoshnoodi, Yujie Lu, Aditya Sharma, and William Yang Wang. Who evaluates the evaluations? objectively scoring text-to-image prompt coherence metrics with t2iscorescore (ts2). *ArXiv preprint*, abs/2404.04251, 2024b. URL https://arxiv.org/abs/2404.04251.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Robert S Shankland. Michelson-morley experiment. *American Journal of Physics*, 32(1):16–35, 1964.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael B. Abu-Ghazaleh. Vulnerabilities of large language models to adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*. Association for Computational Linguistics, 2024.

Charles Singer. Notes on the early history of microscopy. *Proceedings of the Royal Society of Medicine*, 7(Sect_Hist_Med):247–279, 1914.

Prasann Singhal, Jarad Forristal, Xi Ye, and Greg Durrett. Assessing out-of-domain language model performance from few examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2385–2397, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.175.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista

Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swkedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. *ArXiv preprint*, abs/2402.08115, 2024. URL https://arxiv.org/abs/2402.08115.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805, 2023. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *ArXiv preprint*, abs/2403.08295, 2024. URL https://arxiv.org/abs/2403.08295.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Ken Thompson. Reflections on trusting trust. *Communications of the ACM*, 27(8):761–763, 1984.

Songül Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernández-Macías, José Hernández-Orallo, and Emilia Gómez. Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks. *Journal of Artificial Intelligence Research*, 71: 191–236, 2021.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. URL https://openreview.net/forum?id=wUU-7XTL5XO.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3261–3275, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926, 2023a. URL https://arxiv.org/abs/2305.17926.

Xinyi Wang, Wenhu Chen, Michael Saxon, and William Yang Wang. Counterfactual maximum likelihood estimation for training deep networks. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 25072–25085, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d30d0f522a86b3665d8e3a9a91472e28-Abstract.html.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *ArXiv preprint*, abs/2311.09000, 2023b. URL https://arxiv.org/abs/2311.09000.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V Le. Inverse scaling can become u-shaped. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=19sGqVUxQw.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox:"what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*, 2023.

Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. Human-centered evaluation and auditing of language models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3636302. URL https://doi.org/10.1145/3613905.3636302.

Xixi Xu, Zhongang Qi, Jianqi Ma, Honglun Zhang, Ying Shan, and Xiaohu Qie. Bts: A bi-lingual benchmark for text segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19152–19162, 2022.

Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Bo Qiao, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Empower large language model to perform better on industrial domain-specific question answering. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 294–312, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.29. URL https://aclanthology.org/2023.emnlp-industry.29.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xingxu Xie, and Yue Zhang. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *ArXiv preprint*, abs/2211.08073, 2022. URL https://arxiv.org/abs/2211.08073.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 535–546, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.45. URL https://aclanthology.org/2021.naacl-main.45.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. *ArXiv preprint*, abs/2406.11775, 2024. URL https://arxiv.org/abs/2406.11775.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5484–5505. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. (inthe) wildchat: 570k chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2023a.

Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31967–31987. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/65a39213d7d0e1eb5d192aa77e77eeb7-Paper-Conference.pdf.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Li Zhong and Zilong Wang. Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:261048682.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=c8McWs4Av0.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms, 2024b.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. Promptbench: A unified library for evaluation of large language models. *ArXiv preprint*, abs/2312.07910, 2023. URL https://arxiv.org/abs/2312.07910.

Limitations *(and rebuttals)*

You overlooked existing dynamic/constrained/construct valid benchmark examples! Current automatically-generated benchmarks are largely confined to toy problems of limited interest to practitioners. For games like chess (Feng et al., 2023) and reasoning and planning problems from block worlds (Valmeekam et al., 2022; Stechly et al., 2024), test examples can be generated and evaluated deterministically. Programming is a notable exception—it is commercially relevant, deterministically evaluable, and test examples are relatively easy to dynamically generate (Allamanis et al., 2024). However, even for programming, benchmarks over fixed problem sets (e.g., HumanEval (Chen et al., 2021)) reign supreme. Clearly further emphasis on the necesity of dynamic benchmarking is needed here. Dynamic evaluations of LM coding capabilities such as LiveCodeBench (Jain et al., 2024) do source real coding problems from the internet, but this approach is difficult to transfer to other domains. Metrologists will create more dynamic benchmarks reflecting real applications.

As for constrained and ecologically valid benchmark examples we didn't discuss, WildBench (Lin et al., 2024)—a static benchmark built atop exemplars collected from WildChat (Zhao et al., 2023a)—is one rare example of a benchmark that truly is representative of its target distribution of real user dialogue. However, it still is static, and chat agents writ large are not a very constrained domain. This is the reason that the LMSys benchmarks fail to meet all our desiderata (Chiang et al., 2024). Furthermore, the actual users testing these systems are largely LM enthusiasts or researchers themselves—this strains the ecological validity of most arena-style benchmarks.

LMs evaluating LMs? How can a system measure capabilities we don't know it has? There are risks to relying on the target of analysis to self-verify. After all, how can we use a model to measure its own capabilities (or those of similar models)? One solution is to expand out a set of objectively verifiable characteristics with an LM, checked externally.

Preliminary evidence suggests that even GPT-4 fails to match human annotator performance in open-domain claim verification (Wang et al., 2023b). If we prompt a model to generate sentences, then ask GPT-4 to evaluate the generated text, how can we trust that GPT-4's judgements capture anything meaningful? These LM judgements are also problematic for evaluation because they are not even comparable, as different models produce substantially different outcomes (Zhou et al., 2024b).

Reflections on trusting trust

Ken Thompson's Turing award acceptance speech, "Reflections on Trusting Trust" (Thompson, 1984), details how he hid a backdoor Trojan horse in early source versions of a C compiler. Because the compiler was bootstrapped, i.e., new versions of a compiler were compiled by the previous version, this backdoor was nearly impossible to detect or remove without being aware of its introduction.

The backdoor was included even in versions of the compiler binary built from source code without the backdoor, as long as that source was compiled using a binary descended from Thompson's modified code. His discovery, shocking in 1984, seems almost mundane today: *If any part of a complex system is compromised, the entire system is compromised.* For modern automated metrics employing blackbox language models for their own evaluation, verification, and even training, we must view each element as potentially compromised by the flaws in proprietary models.

Even dynamic benchmarks will go stale Living benchmarks based on arbitrary rules can be gamed by exploiting the idiosyncracies of the supervising model and discrepancies in the simulation environment. We have no visibility into the decision processes and therefore no real guarantees of its validity (Oh et al., 2024). Therefore, developing a living metrology community is crucial. Researchers and practitioners will need to refresh their benchmark generators with new methods. Where generative techniques become obviated, breakthroughs in measurement can occur.