DICE: A Framework for Dimensional and Contextual Evaluation of Language Models

Aryan Shrivastava
aashrivastava@uchicago.edu
University of Chicago

Paula Akemi Aoyagui
paula.aoyagui@mail.utoronto.ca
University of Toronto

Abstract

Language models (LMs) are increasingly being integrated into a wide range of applications, yet the modern evaluation paradigm does not sufficiently reflect how they are actually being used. Current evaluations rely on benchmarks that often lack direct applicability to the real-world contexts in which LMs are being deployed. To address this gap, we propose **Di**mensional and **C**ontextual **E**valuation (DICE), an approach that evaluates LMs on granular, context-dependent dimensions. In this position paper, we begin by examining the insufficiency of existing LM benchmarks, highlighting their limited applicability to real-world use cases. Next, we propose a set of granular evaluation parameters that capture dimensions of LM behavior that are more meaningful to stakeholders across a variety of application domains. Specifically, we introduce the concept of ***context-agnostic*** parameters—such as robustness, coherence, and epistemic honesty—and ***context-specific*** parameters that must be tailored to the specific contextual constraints and demands of stakeholders choosing to deploy LMs into a particular setting. We then discuss potential approaches to operationalize this evaluation framework, finishing with the opportunities and challenges DICE presents to the LM evaluation landscape. Ultimately, this work serves as a practical and approachable starting point for context-specific and stakeholder-relevant evaluation of LMs.

Keywords

contextual evaluation, language models, evaluation framework

1 Introduction

Language models (LMs) have significantly evolved from the early n-gram models first proposed in 1948 [74] to powerful neural models that are highly capable and general. Furthermore, new training paradigms have enabled models to follow user instructions [13, 59], making them more approachable to broader society. This has led to the widespread adoption of LMs across a range of real-world domains. For example, LMs are now applied in healthcare [e.g., 53, 70, 76, 82, 89], education [e.g., 15, 20, 32, 85], law [e.g., 3, 23, 35, 39], finance [e.g., 43, 49, 55, 79], and human resources [e.g., 48, 60, 86, 87]. The proliferation of LMs into diverse domains emphasizes the need for effective, meaningful, and relevant evaluation methods to highlight their context-specific capabilities and limitations. Particularly, stakeholders, including those that are non-technical, rely on LM evaluations to determine, whether, how, and which models to deploy into their specific context.

The standard approach to evaluating modern LMs is via benchmarks. They serve as a standardized set of tasks that assess key capabilities and limitations of LMs, facilitating consistent evaluation which ultimately enables comparisons between models under various settings [52]. As such, benchmark results are often the primary factor stakeholders consider when deciding whether to deploy an LM in a specific context and, if so, which model best meets their needs. However, many works have noted the limitations of the benchmarking paradigm in LM evaluation. For one, benchmarks suffer from construct invalidity, meaning they often do not sufficiently measure what they are intended to measure [66]. This issue is exacerbated largely due to overgeneralized claims of LM performance made by model developers [67]. Additionally, most benchmarks measure performance on tasks that are not relevant to how LMs are being used in the real-world [52]. This highlights the insufficiency of relying solely on benchmarks to guide deployment decisions across contexts.

Thus, in this position paper, we propose **DICE** (**Di**mensional & **C**ontextual **E**valuation), a framework that granularizes and contextualizes LM evaluation to better support domain-specific stakeholders. We begin by reviewing the current landscape of LM evaluation, considering both the utility and limitations of benchmarks (§2). Next, we outline DICE (§3). As a part of this framework, we introduce the concept of *context-agnostic* dimensions, which are relevant across all domains, and *context-specific* dimensions, which must be designed to the particular needs and constraints of a given domain. We also discuss how DICE can be operationalized. Finally, we explore the opportunities this framework presents for improving LM evaluation alongside the challenges that may arise with regards to its adoption (§5). Ultimately, we contribute a novel perspective to LM evaluation with the aim to make LM evaluation more contextual and stakeholder-relevant. This work serves as a precursor to a larger, empirical study and we hope to foster further discussion on the development of more adaptive, context-aware evaluation methodologies that better reflect real-world requirements.

2 Current Evaluations of Language Models

2.1 The Modern Benchmarking Paradigm

Currently, LMs are typically evaluated via *benchmarks*: "a dataset . . . and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods" [66]. It has become standard practice to use these benchmarks to track progress, identify weaknesses, and facilitate comparative analysis of LMs [67]. For example, model developers use benchmarks

HEAL @ CHI, Yokohama, Japan

to 1) guide the development of models by identifying areas for improvement and validating progress and 2) encourage the adoption of their models by reporting state-of-the-art results. Furthermore, model users often rely on benchmark performance to determine which LM best suits their specific use case—or whether to use one at all. Overall, benchmarking offers a standardized approach to evaluating LMs, serving the needs of developers and users alike.

Over time, numerous benchmarks have emerged, evaluating different aspects of LM performance. While there are many valid categorizations, we broadly categorize benchmarks into those of:

General-Purpose Factual Knowledge. These benchmarks consist of question datasets with verifiable factual answers, typically formatted as multiple-choice or short-answer tasks. These encompass some of the most widely reported benchmarks in LM evaluation. Prominent examples include GLUE [84], MMLU [29], BIG-bench [80], and Humanity's Last Exam [63].

Reasoning Ability. These benchmarks assess an LM's ability for logical deduction and problem-solving. Mathematical reasoning is a commonly tested area, as shown by MATH [30] and GSM8K [14]. Additionally, DROP [18] measures discrete reasoning over multiple paragraphs, while ARC-AGI [12] measures abstract-reasoning and pattern-recognition. The recent emergence of "reasoning" models such as OpenAI's o-series [56] and the DeepSeek models [16] underscores the importance of evaluating reasoning ability.

Human-Preference Alignment. These benchmarks assess how well an LM's outputs align with human preferences, extending beyond evaluations of objective correctness. Typically performance is measured via human assessment or by leveraging LLMs-as-a-judge. Chatbot Arena [11], MT-bench [97], and AlpacaEval [42] are prominent benchmarks in this category.

Task-Specific Performance. While the previous types of benchmarks often seek the measure broad and *general* capabilities of LMs, task-specific benchmarks constrain evaluation into a specified domain. Coding benchmarks such as HumanEval [10] and SWE-bench [33] are a key focus of the community. But there still exist benchmarks spanning a diverse array of domains including healthcare [e.g., 34, 61] and law [e.g., 21, 27, 40].

2.2 Limitations of the Benchmarking Paradigm

While benchmarking has become the dominant framework for evaluating LMs, in its current form, it comes with several limitations that hinder its ability to fully capture model capabilities as they pertain to real-world environments. Here, we identify and examine *four core limitations* while focusing our discussion on their implications for stakeholders seeking to deploy LMs effectively.

Diminishing Reliability. Benchmarks often suffer from *saturation*, when most models reach close-to-perfect levels of performance, rendering them ineffective for comparison [50]. Furthermore, benchmarks often leak into LMs' training sets, inflating their reported performance [17]. Thus, benchmark results become less reliable for stakeholders seeking to accurately assess LMs.

Limited and Overgeneralized Reporting. Model developers typically report on a narrow set of benchmarks, often emphasizing those measuring general-purpose factual knowledge or reasoning [e.g., 2, 4, 19, 47, 57, 83]. This limited scope constrains model evaluation, which may not accurately reflect how an LM would perform in diverse real-world contexts. Furthermore, prioritizing evaluations on a model's general capabilities fails to provide stakeholders with meaningful insights into how a model will behave in specialized environments that require contextual knowledge.

Construct Invalidity. Benchmarks, especially those that aim to measure general-purpose capabilities, suffer from *construct invalidity* [66]. That is, benchmarks often heralded as those of "general" performance often inappropriately measure the capabilities they are evaluating for. This results in misleading assessments of a model's overall ability, obfuscating how a model would actually behave in specified real-world scenarios.

Misalignment with Real-World Use-Cases. The way in which LMs are being used by individuals "in the wild" often concerns topics that do not align with how LMs are evaluated [94, 96]. Furthermore, even benchmarks that seek to measure task-specific capabilities do not accurately reflect similar tasks faced in the real world. For example, many medical benchmarks rely on board-style exams that fail to capture the ambiguities, complexities, and subjectivity of real medical scenarios [37].

2.3 Positive Steps in LM Evaluation

Despite the limitations of the benchmarking paradigm, recent efforts in LM evaluation have introduced new approaches that aim to provide more robust, comprehensive, interpretable, and context-aware assessments of model performance. HELM aims to improve the transparency of language models through a multi-metric evaluation approach that goes beyond evaluating simple accuracy on benchmarks [44]. CheckEval also takes a multi-metric approach by decomposing tasks into human-defined dimensions, demonstrating that this methodology has a strong correlation with human judgments [38]. Other studies have evaluated LMs on human-centric tasks, such as Chatbot Arena, which collects tasks through crowdsourcing [11], and WildBench, which leverages real-world user queries [46]. Lastly, there has been a rise in context-specific benchmarks designed with the input of domain experts. For example, LegalBench, a legal reasoning benchmark, was collaboratively constructed through an interdisciplinary process involving subject matter experts to capture practically useful and interesting capabilities [27]. And MENTAT, a mental healthcare decision-making benchmark, was created by psychiatrists to capture the ambiguities faced by mental healthcare practitioners [37]. We use many ideas put forth by these positive examples of LM benchmarking to inform our evaluation framework DICE, which we will outline in §3.

3 Dimensional and Contextual Evaluation

Motivated by the positive strides in LM evaluation that aim to address the limitations of traditional benchmarking, DICE is a multi-metric evaluation framework that aims to assess context-aware granular dimensions of LM behavior. We argue that by dicing up LM evaluations into such dimensions, we make them more interpretable, and thus more actionable, to stakeholders. Decomposing evaluation criteria into smaller, well-defined units can enable a more focused evaluation methodology, improving construct validity and providing stakeholders with clear, explicit assessment criteria that facilitate more straightforward interpretations of model performance [38]. In contrast to traditional benchmarks that often measure performance on broadly defined tasks using a single metric (typically accuracy) [44], evaluating models along granular dimensions allows for a more precise characterization of LM capabilities. Even when multiple benchmarks are considered, they largely measure coarsely defined abilities, making it difficult to disentangle specific factors contributing to model performance [66]. By considering a structured decomposition of model behavior, we can more effectively identify specific LM strengths and weaknesses, reducing ambiguity in evaluation. This enhances the diagnostic power of LM assessments, providing stakeholders with a clearer understanding of where models excel and where they fall short, revealing explicit trade-offs between models [44]. Ultimately this enables clearer and better-informed decision-making regarding model selection and deployment as it pertains to specific use-cases.

We classify our proposed dimensions into: ***context-agnostic*** (§3.1) and ***context-specific*** (§3.2) dimensions. Context-agnostic dimensions apply broadly across application domains and serve as a tractable starting point. Context-specific dimensions incorporate stakeholders in the evaluation process, ensuring that assessments align with the contextual requirements of the stakeholder-defined deployment environment. DICE represents a call to shift towards a more contextualized approach to LM evaluation while maintaining some of the tractability and systematic benefits of benchmarking.

We note that we will not make any claims about *how* to evaluate all of these dimensions, as doing so would be infeasibly complex. For instance, entire studies are dedicated to evaluating individual dimensions such as consistency in various contexts [e.g., 71, 75, 92]. Instead, DICE focuses on addressing key limitations of the current benchmarking paradigm, providing a structured approach that can be unified and operationalized (§4) to better support stakeholders in making informed decisions about LM use in specific contexts.

3.1 Context-Agnostic Dimensions

Context-agnostic dimensions refer to granular dimensions of LM behavior that are relevant across diverse use cases. By introducing context-agnostic dimensions, we maintain a sense of standardization within DICE by establishing a shared foundation for evaluation that is applicable across diverse contexts. We note that although these dimensions are phrased as context-agnostic, the required behavior along these dimensions certainly still vary depending on stakeholder needs and specific deployment contexts. We discuss how context-agnostic dimensions can be operationalized in §4.

Faithfulness: *Are LMs faithful to user instructions?* Despite modern LMs being tuned to follow user instructions [59], they may not perfectly do so [65, 88, 98]. A model is faithful if it accurately follows user instructions without distorting information. Failures in faithfulness can manifest in different ways, such as responding in the wrong format or generating outputs that include irrelevant details. Evaluating faithfulness provides stakeholders with how reliably a model can follow user instructions, a critical requirement of LM ability across all contexts.

Coherence: *Do LM outputs make sense?* For a model to be useful in a particular context, it should speak about relevant topics in a manner that makes sense. Coherence is not just limited to the grammatical soundness of text - it should also ensure that LM outputs are logically consistent and relevant to the context. Assessing coherence allows stakeholders to determine whether a model can articulate ideas in a structured and comprehensible manner that aligns with the needs of the given use case. Since clear and logically sound communication is essential across diverse applications, coherence serves as a foundational dimension for evaluating LM performance in any context.

Robustness: *Are LM outputs invariant to prompt perturbations?* In real-world contexts, LMs are frequently asked to complete similar tasks despite being faced with variations in prompting [44]. Previous work has shown that outputs tend to vary greatly under these circumstances, despite prompts calling for the same behavior from the LM [72, 75, 92, 95]. This is particularly important in domains that require reliability and reproducibility, such as legal analysis or scientific research. Conversely, in more creative contexts such as writing, some degree of variability may be desirable. Nonetheless, analyzing robustness provides valuable insights into a model's suitability for specific use cases. We note that this definition does not refer to other forms of robustness such as adversarial robustness [91] that may not be fundamental concerns to many contexts.

Epistemic Honesty: *Are LMs honest about their knowledge limitations?* Models across contexts are certainly going to face knowledge gaps. Epistemic honesty concerns whether a model reliably acknowledges these gaps - reliably stating "I don't know" rather than generating misleading responses [90]. Evaluating epistemic honesty is crucial to understanding a model's trustworthiness and usability - one that fails to acknowledge its limitations is misleading while one that is overly hesitant is impractical to use. Evaluating this dimension enables stakeholders to define and find the correct balance of model transparency specific to the demands of the application domain.

Efficiency: *What will it cost to deploy a model?* Deploying LMs into any context always comes with associated costs. For one, it costs money to deploy and use an LM. Furthermore, if an LM needs to be fine-tuned, one will likely need to pay for API calls, compute, and/or data. In addition to monetary costs, efficiency also encompasses time-related factors, such as inference speed and latency. Recent advancements scale test-time compute, which uses more computational resources during model inference for better results. For example, OpenAI's deep research model may take tens of minutes to run [58], but provide much more comprehensive results than other models that respond almost instantly. While different contexts certainly have different requirements and desires for LM efficiency, all contexts have certain constraints regarding how much it should cost to deploy and use a model.

For full transparency, we also acknowledge other dimensions that could have been included in this set but were deliberately excluded for various reasons. We chose to exclude *accuracy* as a context-agnostic dimension as there are many contexts where there is no well-defined notion of ground-truth. For example, there is no notion of correctness when LMs are tasked with subjective decision-making, which is a common real-world application [94]. Furthermore, domains such as creative writing or open-ended brainstorming often do not have well-defined notions of correctness. Thus, accuracy is not a universally applicable measure of LM behavior and so we do not consider it a core context-agnostic dimension. In the LM alignment literature, the *Helpful, Honest, and Harmless* (HHH) principle is a framework for aligning AI systems with human values [31]. However, we choose to exclude these from the context-agnostic set. While these broad categories are useful for other purposes, we believe they lack the granularity necessary for the precise behavioral evaluation DICE proposes. For a similar reason, we choose to exclude *bias and fairness*. While these are undeniably important to measure the adverse effects that LMs can have on society [1, 8, 44], their definitions lack granularity across different contexts. Instead, we encourage stakeholders to define more specific aspects of bias and fairness such as demographic performance gaps or anti-stereotype reinforcement as context-specific dimensions. Lastly, *uncertainty calibration*, which refers to the fact that if LMs are x% confident in their answer, they should be correct x% of the time. However, this again ties back into notions of correctness, which many contexts do not carry a well-defined notion of. By excluding these, we ensure that context-agnostic dimensions remain granular and truly relevant across *all* contexts.

While these dimensions provide an in-depth starting point, we do not claim that this list is exhaustive. There are likely many additional relevant context-agnostic dimensions, and we encourage future work to address this.

3.2 Context-Specific Dimensions

Context-specific dimensions refer to aspects of LM behavior that are relevant to the particular environment in which an LM is deployed. Because these dimensions are necessarily stakeholder-defined, it is neither possible nor practical to provide an exhaustive list. Instead, we explore how different stakeholders may define dimensions through case studies that illustrate how they can vary across diverse contexts.

Specifically, we consider deployment environments of mental healthcare and education. We choose to study these two contexts because they encompass distinct requirements illustrating the necessity of context-specific evaluation. It is important to emphasize that the dimensions we define are not intended to be a definitive or exhaustive standard but rather as illustrative examples. Our goal is to encourage stakeholders, both within these domains and beyond, to critically evaluate and refine dimensions that align with their specific needs and constraints rather than establish exhaustive lists.

3.2.1 *Case Study I: Mental Healthcare.*

Trigger Warning: This section contains and discusses mentions of sensitive mental health topics.

Many countries, including the United States, face national-level mental health crises [9, 69], while access to mental healthcare remains limited and insufficient [68, 81]. In an effort to make mental healthcare more accessible to those that would otherwise go untreated, many mental health practitioners are turning to AI-enabled digital mental health tools, with a particular focus on LMs, to enable personalized, real-time support for patients [28, 64].

Consider a psychiatrist who seeks to reduce wait times at their clinic by introducing LMs to assist in preliminary psychiatric evaluations. Following the requirements of task-autonomous AI in mental healthcare proposed by [26], we discuss dimensions of LM behavior the psychiatrist may evaluate in addition to the above-mentioned context-agnostic dimensions in order to make a well-informed decision regarding whether, or which, models to deploy.

For one, an LM should prioritize the prevention of harm to the user (e.g., in cases of suicide or self-harm) or others. For example, a patient with schizophrenia (a chronic brain disorder with symptoms that include delusions and hallucinations) may ask an LM to advise them on how to remove a chip from their brain [26]. Thus, the psychiatrist may define the dimension of ***harm prevention***, which asks whether the LM actively discourages and prevents harm through their responses. To complement this dimension of evaluation, the psychiatrist may additionally be concerned with ***sycophancy***, where an LM tends to affirm a user's thoughts, even if harmful. Sycophancy can lead to exacerbating distress, reinforcing negative self-affirmations, and validating delusions [26]. By evaluating these dimensions, the psychiatrist can gain information about whether LMs are sufficiently safe to deploy.

Furthermore, a psychiatrist may be concerned with an LM's ***diagnostic accuracy***, ***corrigibility***, and ***interruptability***. An LM should be able to assess individuals accurately without misleading human psychiatrists or users with incorrect diagnoses. Corrigibility ensures that if an LM provides incorrect information, it can recognize corrections and adjust its responses accordingly to maintain accuracy and reliability. Additionally, the psychiatrist may want to ensure that a human practitioner can intervene when necessary or that a user can halt the interaction at any time. Evaluating these three dimensions helps determine whether LMs are sufficiently knowledgeable, adaptable, and responsive for integration into mental healthcare applications.

Of course, it is up to the psychiatrist to determine how to operationalize the evaluations along these dimensions. Especially in the context of mental healthcare, there is a lack of pragmatic datasets that represent real-world tasks where most benchmarks simply take the form of standardized, multiple-choice tests. One potential choice could be MENTAT, a dataset designed by psychiatrists that captures the real-world ambiguities of mental healthcare [37].

3.2.2 *Case Study II: Education.*

Since the introduction of tools such as ChatGPT, LMs have increasingly been integrated into educational contexts. Not only are they used by students to solve homework, but they can also be used as study assistants, teaching assistants and adaptive learning tools [85]. In this case study, we consider a high school looking to adopt an LM to serve as a teaching assistant tasked with duties such as question generation, curriculum design, and automatic grading across all subjects.

Because the school is seeking to use one LM across all subjects, here is a case where it would be beneficial to deploy a more generally capable LM. Thus, the school can evaluate the dimension of ***accuracy*** on general high school knowledge. To operationalize this dimension, the school can evaluate on a subset of MMLU [29], restricting their analysis to only the tasks labeled at high school level difficulty, such as high school math, history, or psychology. Performing this analysis enables a school to determine whether an LM meets the necessary knowledge requirements to function effectively as a general teaching assistant. For example, it provides information regarding whether an LM will be able to generate factually correct statements or grade students accurately. Furthermore, a school may be concerned with ***pedagogical alignment*** [77]. An LM should be able to generate questions, design curricula, and provide feedback that aligns with the school's learning objectives for students and teaching values. For example, if the school's pedagogy emphasizes constructive learning, when evaluating a student's short-answer response, an LM should not merely point out the flaws in the student's reasoning but also provide alternative reasoning that builds from what the student already wrote.

Furthermore, the school may desire LM behaviors such as ***adherence,*** and ***explainability***. The first can be exemplified in adherence to teacher-provided rubrics or documents to grade students or design curricula rather than use its own preconceptions. For example, the rubric may specify ignoring grammatical errors in a student's short answer to place stronger emphasis on the student's conceptual understanding of the content. Additionally, a teacher may want an exam that assesses content only within the scope of a textbook. Evaluating adherence allows a school to deploy an LM that is flexible to the requirements of the school rather than rely on the model's preconceptions that may lead to misaligned grading practices or curriculum design. Regarding explainability, the teacher may want to be able to understand why an LM graded the way it did or decided to include certain topics in a lesson plan while omitting others. An explainable LM ensures that grading or curriculum design remains transparent, allowing teachers to verify its reasoning, identify potential biases, and make well-informed adjustments when required.

Again, our goal with these case studies is not to provide exhaustive lists of dimensions. Rather, we illustrate the types of considerations stakeholders may make in different contexts in order to define relevant context-specific dimensions. We show that distinct behaviors are desired out of LMs just within the contexts of mental healthcare and education. This points to the insufficiency of using broad-scoped evaluation methodologies and the necessity to contextualize evaluations to make them more relevant to stakeholders.

4 Operationalizing Evaluation

As discussed, DICE assesses LMs on context-aware dimensions, making evaluations more interpretable by reducing ambiguity and highlighting explicit trade-offs based on stakeholder-defined preferences. However, for DICE to be actionable to stakeholders, it must be ***operationalized***. Here, we outline key considerations and approaches for implementing DICE in a systematic manner.

While the above discussion primarily concerned the identification of dimensions on which to evaluate LMs on, we did not discuss *how* to actually measure them. Each dimension, whether context-agnostic or context-specific, needs to have clearly defined evaluation protocols and metrics. The construction of such evaluation protocols should involve stakeholders to ensure the chosen metric accurately reflects the requirements of the context [67]. This applies to both context-agnostic and context-specific dimensions. For example, when measuring robustness, a stakeholder in creative writing may use n-gram metrics such as BLEU [62], valuing fine-grained variations in diction and syntax and preferring lower levels of consistency for more diverse and creative outputs. On the other hand, a stakeholder in healthcare may use metrics such as BERTScore [93] to de-prioritize structural differences in texts by focusing on semantic similarity and desire higher levels of consistency, ensuring straightforward, predictable outputs. For many dimensions, human evaluation is likely the most suitable method. For example, the dimension of pedagogical alignment explained in the Education case study likely requires teacher input to ensure that LM outputs correspond with relevant teaching styles and promote instructor-specified learning objectives for students.

Once stakeholders determine how to measure individual dimensions of interest, the natural next step is to determine how one can aggregate measurements to ensure that results meaningfully inform decisions. Not all dimensions are of equal importance, and not every dimension is equally important across contexts. As an example, consider the dimension of epistemic honesty, again in the domains of healthcare and creative writing. A medical practitioner may value a model aware of its knowledge limitations much more than a novelist. Thus, it is not sufficiently meaningful to take a simple average of performance along the dimensions — a common approach in existing LM evaluations [e.g., 29, 38]. Rather, stakeholders should define a context-specific weighting system in order to ensure that the most critical dimensions are prioritized while those that are less relevant do not dominate the decision-making process. One approach that stakeholders can take is to implement a priority order to help balance trade-offs among dimensions [31]. The flexibility to weight different dimensions of LM behavior is a key advantage of DICE. By considering granular dimensions, stakeholders can more clearly identify the types of behavior they seek in a model specific to their use case and explicitly examine trade-offs to ensure their final decision aligns with their contextual needs.

While defining dimensions of LM behavior and determining meaningful measurements are crucial steps for stakeholder-relevant evaluations, these alone are not enough. For LM evaluations to be truly context-aware, we must also ensure that they are evaluated on datasets that accurately reflect their intended deployment contexts. Evaluating on general-purpose benchmarks or benchmarks misaligned with the intended deployment context does not provide sufficiently representative information, thus misguiding decision-making regardless of how well-defined the dimensions or measurements are. This underscores the urgent need for evaluating on context-specific datasets that capture the complexities of a diverse array of real-world use cases of LMs. As we move towards a more context-aware paradigm of LM evaluation, it is crucial to address both the opportunities that arise from DICE and the challenges that must be addressed to ensure its successful adoption.

5 Opportunities and Challenges

5.1 Opportunities

This section presents further opportunities for DICE in LM evaluation. While most benchmarks assume the existence of an objectively correct answer, many decisions in human-AI collaboration are inherently context-dependent and open to interpretation. Thus, a more pluralistic approach [78] is required. For instance, in bias detection for content moderation, there is often disagreement about what is and what is not considered offensive material [25] and multiple assessments might be possible depending on cultural context due to the need for subjective interpretation [5, 22]. In such cases, DICE's context-dependent dimensions for LM evaluation could include well-known metrics such as *toxicity level*, but also less explored ones, such as *cultural alignment* [51] or *perspective-taking* to better evaluate if an individual or population are being harmed [6]. Other open-to-interpretation scenarios include HR decisions for hiring [24], conflict resolution [73] and group decision-making [11] where it becomes important to assess which perspectives the LM is emphasizing or suppressing to mitigate biases to ensure fairness.

DICE also facilitates stakeholders to take a more active role in LM evaluation, aligning with the emphasis on carefully considering stakeholder needs within their specific domains to inform the design of evaluation criteria [45]. Stakeholder knowledge provides valuable contextual insights into the real-world settings where the models will be deployed, including domain-specific expertise and more detailed business requirements and constraints. Therefore, understanding and catering to stakeholder needs will improve DICE's relevance and applicability. The target audience of stakeholders for our framework includes business decision-makers who do not possess technical knowledge in machine learning but are imbued with recommending or selecting an LM for an application. This audience requires clear performance metrics suitable to their business needs that allow for cross-comparison of models, which we aim to enable with DICE. Another potential application would be to support stakeholders in choosing between existing benchmarks. Since DICE's dimensions define the desired LM behavior parameters, it would be possible to align dimensions against existing benchmark metrics and map out the most suitable option. For example, *LegalBench* identifies six categories of legal reasoning [27]. If a legal stakeholder's dimensions correspond with any of these dimensions, this benchmark becomes a good candidate for evaluation. Thus, our framework does not aim to replace benchmarks, but rather, extract the most meaningful interpretations of benchmarks to stakeholders.

5.2 Challenges

To ensure the longevity of our proposed evaluation framework, DICE, we anticipate challenges related to scalability, flexibility and trade-offs. The first challenge, scalability, relates to the demands for datasets to power the framework, with special concerns for the scarcity of domain-specific datasets. The second challenge, flexibility, includes similar issues, also inherent to a data-hungry system, as adaptability to new use cases and domains will require more data to prevent the framework from becoming static, outdated and irrelevant to stakeholders. The last challenge is about negotiating trade-offs between dimensions, as it is known there are tensions between certain criteria, for example balancing robustness, fairness and accuracy in deep neural networks [41]. While DICE's construction enables this type of analysis that would otherwise be diffcult to do, this presents a multi-objective optimization problem that may be difficult to resolve [7]. Furthermore, multiple stakeholders having competing needs and goals presents another challenge to the implementation of DICE.

While many of these challenges remain open problems in LM evaluation, we explore potential approaches for addressing them within the research community. Pertaining to the need for domain-specific datasets, we acknowledge that while there are good examples, they are often still misaligned with the context or scarce. Thus, we encourage the broader Human-Computer Interaction and Computer Science community to collaborate with domain experts in order to create appropriate, well-aligned datasets. This will enable contextually relevant evaluation across contexts, which DICE relies on for maximum utility. Pertaining to the need for an adaptable framework, future work may explore strategies to ensure DICE remains responsive to evolving use cases and stakeholder requirements. By design, DICE allows stakeholders to add, remove, or reweight evaluation dimensions, enabling a flexible assessment process. However, maintaining adaptability also requires access to dynamic, context-specific datasets, which remains an open challenge. Addressing this issue will be critical to ensuring that DICE continues to provide relevant and meaningful evaluations across diverse applications. Pertaining to negotiating trade-offs between dimensions, there exist many approaches to resolving difficulties in aligning and resolving multi-objective optimization [e.g., 36, 54]. We encourage future work to assess how these approaches align with human preferences, thus guiding the choice of optimization frameworks that best capture stakeholder preferences, which can then be integrated into DICE.

As stated, this work serves as a precursor to a larger empirical study in which we will aim to address these challenges as well. Specifically, we will conduct a series of case studies each focusing on a particular domain. For each, we aim to determine both context-specific dimensions and how to evaluate context-specific dimensions through a literature search (e.g., to identify applicable benchmarks) and domain expert guidance. After obtaining LM performance across all dimensions, we will examine various aggregation methods (e.g., weighted average or lexicographic pareto dominance) and compare the resulting rankings with the rankings of human experts to validate the utility and reliability of DICE.

6 Conclusion

As LMs become integrated into diverse facets of society, evaluations must evolve beyond the current benchmarking paradigm to better serve real-world applications. Thus, we introduced DICE as a framework to enable a more granular, context-aware evaluation of LM behavior. We argued that DICE makes evaluations more meaningful, interpretable, and actionable to stakeholders. Furthermore, by specifying context-agnostic and context-specific dimensions, we provided a structured approach for context-aware evaluation that remains tractable across diverse domains. We then explored how DICE could be operationalized by discussing how to evaluate and aggregate dimensions.

While DICE faces challenges pertaining to the need for context-specific datasets, evolving use cases, and resolving trade-offs in multi-objective optimization, it ultimately presents many opportunities for contextualizing the current LM evaluation landscape, fostering more adaptive, transparent, and stakeholder-driven assessments.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 298–306. https://doi.org/10.1145/3461702.3462624

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Robin AI. 2025. Large Language Models in the Legal Industry. https://www.robinai.com/post/large-language-models-legal-industry

[4] Anthropic. 2024. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family Accessed: 2025-02-10.

[5] Paula Akemi Aoyagui, Sharon Ferguson, and Anastasia Kuzminykh. 2024. Exploring Subjectivity for more Human-Centric Assessment of Social Biases in Large Language Models. *arXiv preprint arXiv:2405.11048* (2024).

[6] Paula Akemi Aoyagui, Kelsey Stemmler, Sharon Ferguson, Young-Ho Kim, and Anastasia Kuzminykh. 2025. A Matter of Perspective(s): Contrasting Human and LLM Argumentation in Subjective Decision-Making on Subtle Sexism. *arXiv preprint arXiv:2502.14052* (2025). https://arxiv.org/abs/2502.14052

[7] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR* abs/2204.05862 (2022). https://doi.org/10.48550/arXiv.2204.05862

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[9] CDC. 2024. https://www.cdc.gov/mental-health/about/what-cdc-is-doing.html Accessed: 02-17-2025.

[10] 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. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 331, 30 pages.

[12] François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019).

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25, 70 (2024), 1–53.

[14] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[15] Education Copilot. 2024. Education Copilot. https://educationcopilot.com/

[16] DeepSeek. [n. d.]. deepseek: Into the unknown.

[17] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12039–12050. https://doi.org/10.18653/v1/2024.findings-acl.716

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

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[20] Andy Extance. 2023. ChatGPT has entered the classroom: how LLMs could transform education. Nature. https://www.nature.com/articles/d41586-023-03507-3

[21] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289* (2023).

[22] Sharon Ferguson, Paula Akemi Aoyagui, Rimsha Rizvi, Young-Ho Kim, and Anastasia Kuzminykh. 2024. The Explanation That Hits Home: The Characteristics of Verbal Explanations That Affect Human Perception in Subjective Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–37.

[23] Robin Gainer and Kosta Starostin. 2024. How LLMs Can Boost Legal Productivity (with Accuracy and Privacy). https://cohere.com/blog/how-llms-can-boost-legal-productivity-with-accuracy-and-privacy

[24] Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315* (2024).

[25] Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *Comput. Surveys* 55, 13s (2023), 1–32.

[26] Declan Grabb, Max Lamparth, and Nina Vasan. 2024. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation. In *First Conference on Language Modeling*. https://openreview.net/forum?id=1pgfvZj0Rx

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

[28] Salah Hamdoun, Rebecca Monteleone, Terri Bookman, and Katina Michael. 2023. AI-based and digital mental health apps: Balancing need and risk. *IEEE Technology and Society Magazine* 42, 1 (2023), 25–36.

[29] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*. https://openreview.net/forum?id=d7KBjmI3GmQ

[30] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *Advances in Neural Information Processing Systems* (2021).

[31] Yue Huang, Chujie Gao, Yujun Zhou, Kehan Guo, Xiangqi Wang, Or Cohen-Sasson, Max Lamparth, and Xiangliang Zhang. 2025. Position: We Need An Adaptive Interpretation of Helpful, Honest, and Harmless Principles. *arXiv preprint arXiv:2502.06059* (2025).

[32] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Shahbaz Khan, and Ibrahim Haleem Khan. 2023. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 2 (2023), 100115. https://doi.org/10.1016/j.tbench.2023.100115

[33] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=VTF8yNQM66

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

[35] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and S Yu Philip. 2024. Large language models in law: A survey. *AI Open* (2024).

[36] Leonardo Lai, Lorenzo Fiaschi, and Marco Cococcioni. 2020. Solving mixed Pareto-Lexicographic multi-objective optimization problems: The case of priority chains. *Swarm and Evolutionary Computation* 55 (2020), 100687. https://doi.org/10.1016/j.swevo.2020.100687

[37] Max Lamparth, Declan Grabb, Amy Franks, Scott Gershan, Kaitlyn N. Kunstman, Aaron Lulla, Monika Drummond Roots, Manu Sharma, Aryan Shrivastava, Nina Vasan, and Colleen Waickman. 2025. Moving Beyond Medical Exam Questions: A Clinician-Annotated Dataset of Real-World Tasks and Ambiguity in Mental Healthcare. arXiv:2502.16051 [cs.CL] https://arxiv.org/abs/2502.16051

[38] Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. CheckEval: Robust Evaluation Framework using Large Language Model via Checklist. *arXiv preprint arXiv:2403.18771* (2024).

[39] LexisNexis. 2024. New Survey Data from LexisNexis Points to Seismic Shifts in Law Firm Business Models and Corporate Legal Expectations Due to Generative AI. https://www.lexisnexis.com/community/pressroom/b/news/posts/new-survey-data-from-lexisnexis-points-to-seismic-shifts-in-law-firm-business-models-and-corporate-legal-expectations-due-to-generative-ai.

[40] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2024. LegalAgentBench: Evaluating LLM Agents in Legal Domain. *arXiv preprint arXiv:2412.17259* (2024).

[41] Jingyang Li and Guoqiang Li. 2024. The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey. *Comput. Surveys*

(2024).

[42] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

[43] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '23)*. Association for Computing Machinery, New York, NY, USA, 374–382. https://doi.org/10.1145/3604237.3626869

[44] 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. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW Featured Certification, Expert Certification.

[45] Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100* (2023).

[46] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=MKEHCx25xp

[47] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[48] Xiaoye Ma, Weiheng Liu, Changyi Zhao, and Liliya R Tukhvatulina. 2024. Can Large Language Model Predict Employee Atrition?. In *Proceeding of the 2024 5th International Conference on Computer Science and Management Technology*. 1164–1172.

[49] Carsten Maple, Alpay Sabuncuoglu, Lukasz Szpruch, Andrew Elliott, and Tony Zemaitis Gesine Reinert. 2024. The Impact of Large Language Models in Finance: Towards Trustworthy Adoption. *The Alan Turing Institute* (2024). https://www.turing.ac.uk/news/publications/impact-large-language-models-finance-towards-trustworthy-adoption

[50] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2024. Artificial Intelligence Index Report 2024. *Institute for Human-Centered AI* (2024).

[51] Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. *arXiv preprint arXiv:2309.12342* (2023).

[52] Timothy R McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *arXiv preprint arXiv:2402.09880* (2024).

[53] Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, et al. 2024. The application of large language models in medicine: A scoping review. *Iscience* 27, 5 (2024).

[54] Subhojyoti Mukherjee, Anusha Lalitha, Sailik Sengupta, Aniket Deshmukh, and Branislav Kveton. 2024. Multi-Objective Alignment of Large Language Models Through Hypervolume Maximization. *arXiv preprint arXiv:2412.05469* (2024).

[55] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *arXiv preprint arXiv:2406.11903* (2024).

[56] OpenAI. [n. d.]. Reasoning best practices.

[57] OpenAI. 2024. ChatGPT. https://chatgpt.com/

[58] OpenAI. 2025. Introducing deep research.

[59] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[60] Hugh O'Neal. 2024. How Large Language Models (LLMs) are Reshaping HR Management. https://www.metadialog.com/blog/large-language-models-for-hr-how-to-use-it-in-human-resource/

[61] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*. PMLR, 248–260.

[62] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL '02)*. Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[63] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Daron Anderson, Tung Nguyen, Mobeen Mahmood, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Jessica P. Wang, Pawan Kumar, Oleksandr Pokutnyi, Robert Gerbicz, Serguei Popov, John-Clark Levin, Mstyslav Kazakov, Johannes Schmitt, Geoff Galgon, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Zachary Giboney, Gashaw M. Goshu, Joan of Arc Xavier, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, John Wydallis, Mark Nandor, Ankit Singh, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Darling Duclosel, Jungbae Nam, Jennifer Zampese, Ryan G. Hoerr, Aras Bacho, Gautier Abou Loume, Abdallah Galal, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Lianghui Li, Sumeet Motwani, Christian Schroeder de Witt, Edwin Taylor, Johannes Veith, Eric Singer, Taylor D. Hartman, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Joshua Robinson, Aleksandar Mikov, Ameya Prabhu, Longke Tang, Xavier Alapont, Justine Leon Uro, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Julien Guillod, Yuqi Li, Joshua Vendrow, Vladyslav Kuchkin, Ng Ze-An, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy, Dakotah Martinez, Ben Pageler, Nick Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Subrata Mishra, Ariel Ghislain Kemogne Kamdoum, Tobias Kreiman, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Tim Tarver, Haline Heidinger, Rafael Sayous, Stefan Ivanov, Joseph M Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Ali Dehghan, Andres Algaba, Brecht Verbeken, David Noever, Ragavendran P V, Lisa Schut, Ilia Sucholutsky, Evgenii Zheltonozhskii, Derek Lim, Richard Stanley, Shankar Sivarajan, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Jennifer Sandlin, Anmol Sahu, Yuzheng Hu, Sara Fish, Nasser Heydari, Archimedes Apronti, Kaivalya Rawal, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Alan Goldfarb, Sergey Ivanov, Rafał Poświata, Chenguang Wang, Daofeng Li, Donato Crisostomi, Andrea Achilleos, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Elliott Thornley, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Keith Schneider, Zakayo Kazibwe, Nate Stambaugh, Mukhwinder Singh, Ilias Magoulas, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Veit Elser, Kanu Priya Agarwal, Victor Efren Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, Václav Rozhoň, Ziqiao Ma, Christian Stump, Mohammed Berkani, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccardo, Ferenc Jeanplong, Niv Cohen, Josef Tkadlec, Paul Rosu, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, John Arnold Ambay, Abhishek Shukla, Yan Carlos Leyva Labrador, Alan Givré, Hew Wolff, Vivien Rossbach, Muhammad Fayez Aziz, Younesse Kaddar, Yanxu Chen, Robin Zhang, Jiayi Pan, Antonio Terpin, Niklas Muennighoff, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Adam Jones, Jainam Shah, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor Kretov, Kaustubh Sridhar, Zienab EL-Wasif, Anji Zhang, Daniel Pyda, Joanna Tam, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, Daniel Bugas, David Aldous, Jesyin Lai, Shannon Coleman, Mohsen Bahaloo, Jiangnan Xu, Sangwon Lee, Sandy Zhao, Ning Tang, Michael K. Cohen, Micah Carroll, Orr Paradise, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Benedito Alves de Oliveira Junior, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Sztyber-Betley, Priti Shukla, Jonathan Crozier, Antonella Pinto, Shreyas Verma, Prashant Joshi, Zheng-Xin Yong, Allison Tee, Jérémy Andréoletti, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Khánh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Jonathan Roberts, William Alley, Kunyang Sun, Ryan Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Sreenivas Goud Raparthi, Pablo Hernández-Cámara, Freddie Martin, Dmitry Malishev, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ziye Chen,

Biró Bálint, M Saiful Bari, Peyman Kassani, Zihao Wang, Behzad Ansarinejad, Laxman Prasad Goswami, Yewen Sun, Hossam Elgnainy, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro José Moyano, Rajat Maheshwari, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Javier Gimenez, Innocent Enyekwe, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Tim Santens, Khalida Meer, Harrison K Wang, Kalyan Ramakrishnan, Evan Chen, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mündler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Carter Harris, Jason Gross, Ilya Gusev, Asankhaya Sharma, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayezi, Sergei Bogdanov, Alexander Piperski, Marc Carauleanu, David K. Zhang, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Wiktor Morak, Wenjie Ma, William Held, Tran Duc Huy, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Emilien Duc, Bita Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, Mátyás Vincze, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Jiang Muzhen, Fredrik Ekström, Angela Hammon, Oam Patel, Nicolas Remy, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Mike Battaglia, Mohammad Maghsoudimehrabani, Hieu Hoang, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Stephen Mensah, Nathan Andre, Anton Peristyy, Chris Harjadi, Himanshu Gupta, Stephen Malina, Samuel Albanie, Will Cai, Mustafa Mehkary, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Shreen Gul, Gunjan Chhablani, Zhehang Du, Adrian Cosma, Colin White, Robin Riblet, Prajvi Saxena, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Shiv Halasyamani, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Renas Bacho, Vincent Ginis, Aleksandr Maksapetyan, Florencia de la Rosa, Xiuyu Li, Guillaume Malod, Leon Lang, Julien Laurendeau, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Yiğit Yalın, Gbenga Daniel Obikoya, Luca Arnaboldi, Rai, Filippo Bigi, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Alesia Yakimchyk, Huanxu, Liu, Olle Häggström, Emil Verkama, Himanshu Narayan, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Łucki, Wen-Ding Li, Justin Xu, Kevin Joseph Scaria, Freddie Vargus, Farzad Habibi, Long, Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Declan Grabb, Ida Bosio, Tony Fruhauff, Ido Akov, Eve J. Y. Lo, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Yibo Jiang, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Muhammad Rehan Siddiqi, Alon Ragoler, Justin Tan, Deepakkumar Patil, Rebeka Plecnik, Aaron Kirtland, Roselynn Grace Montecillo, Stephane Durand, Omer Faruk Bodur, Zahra Adoul, Mohamed Zekry, Guillaume Douville, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Gözdenur Demir, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Hsiaoyun Milliron, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Abdelkader Dendane, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Russell Campbell, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I. M. J. McInnis, Alex Hoover, Sören Möller, Song Bian, John Lai, Tejal Patwardhan, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249* (2025).

[64] Yue Qi. 2024. Pilot Quasi-Experimental Research on the Effectiveness of the Woebot AI Chatbot for Reducing Mild Depression Symptoms among Athletes. *International Journal of Human–Computer Interaction* (2024), 1–8.

[65] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13025–13048. https://doi.org/10.18653/v1/2024.findings-acl.772

[66] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. 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)*. https://openreview.net/forum?id=j6NxpQbREA1

[67] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. 2024. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=hcOq2buakM

[68] Andis Robeznieks. 2022. Doctor shortages are here—and they'll get worse if we don't act fast. https://www.ama-assn.org/practice-management/sustainability/doctor-shortages-are-here-and-they-ll-get-worse-if-we-don-t-act Accessed: 02-17-2025.

[69] Ian RH Rockett, Eric D Caine, Aniruddha Banerjee, Bina Ali, Ted Miller, Hilary S Connery, Vijay O Lulla, Kurt B Nolte, G Luke Larkin, Steven Stack, et al. 2021. Fatal self-injury in the United States, 1999–2018: Unmasking a national mental health crisis. *EClinicalMedicine* 32 (2021).

[70] Ashish Sarraju, Dennis Bruemmer, Erik Van Iterson, Leslie Cho, Fatima Rodriguez, and Luke Laffin. 2023. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *Jama* 329, 10 (2023), 842–844.

[71] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems* 36 (2024).

[72] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

[73] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[74] C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[75] Aryan Shrivastava, Jessica Hullman, and Max Lamparth. 2024. Measuring Free-Form Decision-Making Inconsistency of Language Models in Military Crisis Simulations. *arXiv preprint arXiv:2410.13204* (2024).

[76] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[77] Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard Baraniuk. 2024. Pedagogical Alignment of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13641–13650. https://doi.org/10.18653/v1/2024.findings-emnlp.797

[78] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070* (2024).

[79] Alex Spyrou and Brian Pisaneschi. 2024. Practical Guide for LLMs in the Financial Industry. https://rpc.cfainstitute.org/research/the-automation-ahead-content-series/practical-guide-for-llms-in-the-financial-industry

[80] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret,

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

[81] Ching-Fang Sun, Christoph U Correll, Robert L Trestman, Yezhe Lin, Hui Xie, Maria Stack Hankey, Raymond Paglinawan Uymatiao, Riya T Patel, Vemmy L Metsutnan, Erin Corinne McDaid, et al. 2023. Low availability, long wait times, and high geographic disparity of psychiatric outpatient care in the US. *General Hospital Psychiatry* 84 (2023), 12–17.

[82] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[83] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[84] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (Eds.). Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[85] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105* (2024).

[86] Weijie Xu, Jay Desai, Fanyou Wu, Josef Valvoda, and Srinivasan H Sengamedu. 2024. HR-Agent: A Task-Oriented Dialogue (TOD) LLM Agent Tailored for HR Applications. *arXiv preprint arXiv:2410.11239* (2024).

[87] Weijie Xu, Zicheng Huang, Wenxiang Hu, Xi Fang, Rajesh Cherukuri, Naumaan Nayyar, Lorenzo Malandri, and Srinivasan Sengamedu. 2024. HR-MultiWOZ: A Task Oriented Dialogue (TOD) Dataset for HR LLM Agent. In *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, Estevam Hruschka, Thom Lake, Naoki Otani, and Tom Mitchell (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 59–72. https://aclanthology.org/2024.nlp4hr-1.5/

[88] Jianhao Yan, Yun Luo, and Yue Zhang. 2024. RefuteBench: Evaluating Refuting Instruction-Following for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13775–13791. https://doi.org/10.18653/v1/2024.findings-acl.818

[89] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine* 5, 1 (2022), 194.

[90] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for Honesty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems.* https://openreview.net/forum?id=67K3Xlvw8L

[91] Zeyu Yang, Zhao Meng, Xiaochen Zheng, and Roger Wattenhofer. 2024. Assessing Adversarial Robustness of Large Language Models: An Empirical Study. *arXiv preprint arXiv:2405.02764* (2024).

[92] Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, Junbo Zhao, et al. 2023. Assessing hidden risks of LLMs: an empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235* (2023).

[93] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations.* https://openreview.net/forum?id=SkeHuCVFDr

[94] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=Bl8u7ZRlbM

[95] Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Improving the Robustness of Large Language Models via Consistency Alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 8931–8941. https://aclanthology.org/2024.lrec-main.782/

[96] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=BOfDKxfwt0

[97] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

[98] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911* (2023).