DOI: 10.1111/nyas.15007

CONCISE REPORT

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

THE SCIENCES

Holistic Evaluation of Language Models

Rishi Bommasani | **Percy Liang** | **Tony Lee**

Center for Research on Foundation Models, Stanford University, Stanford, California, USA

Correspondence
Rishi Bommasani, Center for Research on Foundation Models, Stanford University, Stanford, California.
Email: nlprishi@stanford.edu

Funding information
Google and Schmidt Futures AI2050 Initiative

Abstract

Language models (LMs) like GPT-3, PaLM, and ChatGPT are the foundation for almost all major language technologies, but their capabilities, limitations, and risks are not well understood. We present Holistic Evaluation of Language Models (HELM) to improve the transparency of LMs. LMs can serve many purposes and their behavior should satisfy many desiderata. To navigate the vast space of potential scenarios and metrics, we taxonomize the space and select representative subsets. We evaluate models on 16 core scenarios and 7 metrics, exposing important trade-offs. We supplement our core evaluation with seven targeted evaluations to deeply analyze specific aspects (including world knowledge, reasoning, regurgitation of copyrighted content, and generation of disinformation). We benchmark 30 LMs, from OpenAI, Microsoft, Google, Meta, Cohere, AI21 Labs, and others. Prior to HELM, models were evaluated on just 17.9% of the core HELM scenarios, with some prominent models not sharing a single scenario in common. We improve this to 96.0%: all 30 models are now benchmarked under the same standardized conditions. Our evaluation surfaces 25 top-level findings. For full transparency, we release all raw model prompts and completions publicly. HELM is a living benchmark for the community, continuously updated with new scenarios, metrics, and models https://crfm.stanford.edu/helm/latest/.

KEYWORDS
artificial intelligence, evaluation, foundation models, language models, natural language processing, transparency

INTRODUCTION

Language models (LMs) are everywhere. Models like OpenAI's ChatGPT have taken the world by storm: they power products across almost every sector like Microsoft's Bing search engine, Instacart's AI shopping assistant, and Spotify's AI DJ. Kids use these models to help them write homework essays on Hamlet; software engineers use these models to help them write code. When you write an email and get prompted with a suggestion for how to end your sentence, that is an LM making that suggestion. These models have already been used to coauthor *Economist* articles and award-winning essays, cocreate screenplays, and coconstruct testimonies before the U.S. Senate. Meanwhile, there has been extensive discussion of their risks: they can be toxic, dishonest, possibly used to spread disinformation, and the practices surrounding their data and their deployment raise serious legal and ethical issues.

LMs are emblematic of a paradigm shift in the field of artificial intelligence (AI) toward foundation models.[1] At its core, an LM is a box that takes in text and generates text (Figure 1). Despite their simplicity, when these models are trained on broad data at an immense scale, they can be specialized to myriad downstream scenarios. The same model could be used to summarize legal documents, retrieve information from medical records, or answer questions about your favorite baseball player. Yet, the immense surface of model capabilities,

limitations, and risks remains poorly understood. The rapid development, rising impact, and inadequate understanding demand that we make LMs transparent.

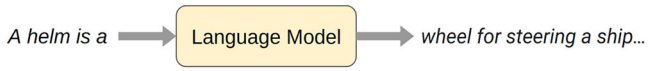

FIGURE 1 Language model. A language model takes text (a prompt) and generates text (a completion). Despite the simple interface, LMs can perform many tasks like answering question or summarizing documents.

Transparency and evaluation

The societal impact of AI is ever-growing with LMs leading the charge. In February 2023, just 4 months after ChatGPT was released, Reuters reported that ChatGPT was the fastest-growing consumer application in history (surpassing Google, Facebook, Instagram, TikTok, and more). If this technology is increasingly important, how can we make it **transparent** to understand it better?

Evaluation is the most established approach for understanding AI systems. By building benchmarks that concretely measure different aspects of an LM, we can characterize its strengths and weaknesses. In fact, benchmarks orient AI. They encode values and priorities[2,3] that specify directions for the AI community to improve upon.[4–8] When implemented and interpreted appropriately, they enable the broader community to influence AI's trajectory.

But what does it mean to evaluate an LM? How does one evaluate an LM? LMs are general-purpose text interfaces that could be applied across a vast expanse of scenarios. How do we navigate this space of use cases? And for each scenario, we should have a broad set of desiderata: models should be accurate, fair, robust, efficient, and so on. How do models fare across these many fronts?

Unfortunately, while LMs are increasingly salient, transparency lags behind: models from Google, Microsoft, Meta, OpenAI, and more had not been evaluated in the same way to enable clear comparison. In fact, prior to our work, no *standard* existed for LM evaluation. We have developed a new benchmarking approach, Holistic Evaluation of Language Models (HELM), which provides transparency through standardized evaluation.

Holistic evaluation requires:

1. **Broad coverage and recognition of incompleteness**. Given the vast surface of capabilities and risks, we need to evaluate LMs across many scenarios and metrics. However, it is impossible to fully cover the space. Therefore, holistic evaluation simultaneously should provide breadth and foreground what is currently missing.
2. **Multi-metric measurement**. Societally beneficial systems reflect many values, not just accuracy. Holistic evaluation should represent these plural desiderata. In other words, we want to understand *trade-offs*: is one model more accurate but less fair than another? Are less biased models also less toxic?
3. **Standardization**. Models should be evaluated under standardized conditions: the evaluation procedure should be controlled to not deliberately favor some models over others. Once the evaluation is defined, all LMs (including those developed in the future) should be evaluated against the same standard.

Overall, holistic evaluation builds transparency by characterizing LMs in their totality.

HELM

HELM has two levels: (i) an abstract taxonomy of *scenarios* and *metrics* to define the design space for LM evaluation and (ii) a concrete set of implemented scenarios and metrics that were selected to prioritize coverage (e.g., different English varieties), value (e.g., user-facing applications), and feasibility (e.g., limited engineering resources).

Recognition of incompleteness

To grapple with the vast evaluation surface, we first taxonomize the space of *scenarios* (where LMs can be applied) and *metrics* (what we want them to do). We visualize our approach in Figure 2. A scenario consists of a task, a domain (consisting of what genre the text is, who wrote it, and when it was written), and the language. We then prioritize a subset of scenarios and metrics based on societal relevance (e.g., user-facing applications), coverage (e.g., different English dialects/varieties), and feasibility (i.e., we have limited compute). In contrast to prior benchmarks (e.g., SuperGLUE,[9] EleutherAI LM Harness,[10] and BIG-Bench[11]), which enumerate a set of scenarios and metrics, we situate our scenarios in a larger taxonomy to make explicit what is currently missing. Examples for what we miss in the first version of HELM include: languages beyond English, nontraditional tasks, such as copywriting, and metrics that capture human–LM interaction.

Multi-metric measurement

Most existing benchmarks consider scenarios with a single main metric (usually accuracy), relegating the evaluation of other desiderata (e.g., toxicity) to separate scenarios (e.g., RealToxicityPrompts[12]). This is insufficient: we believe it is integral that all desiderata be evaluated in the same contexts where we expect to deploy models. For each of our 16 core scenarios, we measure seven metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency). The multi-metric approach (Figure S1) makes explicit potential trade-offs and helps ensure the nonaccuracy desiderata are not treated as second-class citizens to accuracy (see Ref. 13).

Targeted evaluations

In addition, we perform targeted evaluations: 26 finer-grained scenarios that isolate specific skills (e.g., mathematical reasoning,

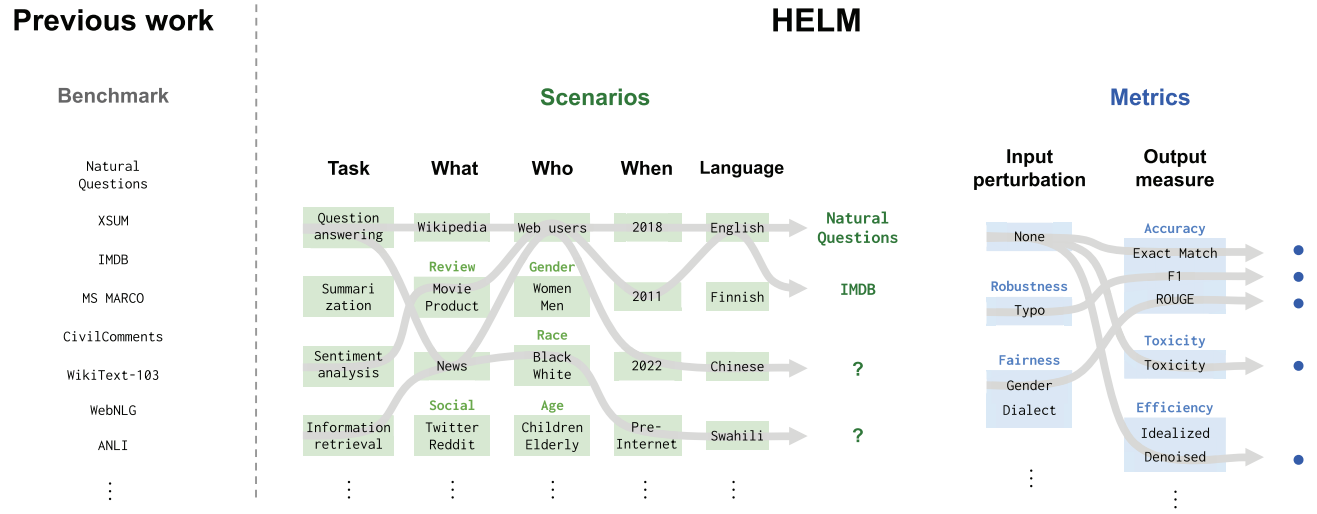

FIGURE 2 The HELM approach. Previous language model benchmarks are collections of datasets with canonical metric(s), usually accuracy (*left*). In comparison, HELM adopts a top-down approach: we explicitly state what we want to evaluate (i.e., scenarios and metrics). Given this taxonomy, we deliberately implement and evaluate a subset, articulating what we currently miss (e.g., coverage of languages beyond English). Abbreviation: HELM, Holistic Evaluation of Language Models.

commonsense knowledge) and risks (e.g., disinformation generation, copyrighted content). This includes 21 scenarios that are either entirely new in this work or that previously were not used in mainstream LM evaluation.

Standardization

As LMs become the substrate for language technologies, the absence of an evaluation standard compromises the community's ability to see the full landscape of LMs. As an example, of the 405 datasets evaluated across all major language modeling works at the time of writing, the extent to which models evaluate on these datasets is uneven. Different models are often evaluated on different scenarios: models such as Google's T5 (11B) and Anthropic-LM v4-s3 (52B) were not evaluated on a single dataset in common in their original works. Several models (e.g., AI21 Labs' J1-Grande v1 (17B), Cohere's Cohere xlarge v20220609 (52.4B), and Yandex's YaLM (100B)) essentially do not report public results (to our knowledge).

To rectify this status quo, we benchmark 30 prominent LMs on HELM. The models come from 12 organizations: AI21 Labs, Anthropic, BigScience, Cohere, EleutherAI, Google, Meta, Microsoft, NVIDIA, OpenAI, Tsinghua University, and Yandex. Critically, the models differ in their public *access* (see Ref. 14): some are open (e.g., BigScience's BLOOM (176B)), others are limited access via Application Programming Interface (API) (e.g., OpenAI's davinci (175B)), and still others are closed (e.g., Microsoft/NVIDIA's TNLG v2 (530B)). For our 16 core scenarios, models were previously evaluated on 17.9% of our scenarios (even after compiling evaluations dispersed across different prior works), which we improve to 96.0% (Figure S2).

The importance of adaptation

To benchmark these models, we must specify an *adaptation* procedure that uses the general-purpose LM to tackle a given scenario. In this work, we adapt all LMs through few-shot prompting, as pioneered by GPT-3.[15] We chose simple and generic prompts to encourage the development of generic language interfaces that do not require model-specific incantations.

EMPIRICAL FINDINGS

We ran more than 4900 evaluations of different models on different scenarios. This amounts to over 12 billion tokens of model inputs and outputs, spanning 17 million model calls, which costs $38K for the commercial models (under current pricing schemes) and almost 20K GPU hours for the open models, which were run on the Together Research Computer. Through this, we identify 25 top-level findings, from which we extract five salient points:

1. **Instruction tuning**, the practice of fine-tuning LMs with human feedback, pioneered by OpenAI and Anthropic, is highly effective in terms of accuracy, robustness, and fairness (Figure S3). Instruction tuning allows smaller models (e.g., Anthropic-LM v4-s3 (52B)) to compete with models 10x the size (Microsoft/NVIDIA's TNLG v2 (530B)), though within a model family, scaling improves accuracy. Unfortunately, how the instruction tuning was performed for these models is not public knowledge.
2. Currently, **open** models (e.g., Meta's OPT (175B), Big Science's BLOOM (176B), and Tsinghua University's GLM (130B)) underperform relative to the nonopen models (e.g., OpenAI's

text-davinci-002, Microsoft/NVIDIA's TNLG v2 (530B), and Anthropic-LM v4-s3 (52B)) as seen in Figure S4, Open models have improved dramatically over the last year, but it will remain to be seen how these dynamics unfold, and what this says about power in the language modeling space.

3. In Figure S5, we find that (average) **accuracy** is correlated with **robustness** (e.g., models are not misled by typos) and fairness (e.g., models perform well across English dialects), though there are some scenarios and models where there are large drops in robustness and fairness. Our multi-metric approach allows us to monitor these deviations and ensure that we do not lose sight of considerations beyond accuracy.
4. The **adaptation** strategy (e.g., prompting) has a large effect on the results: there is no universally best strategy as the best choice depends on both the model and use case (Figure S6). Sometimes even the qualitative trends themselves change, such as the relationship between accuracy and calibration (which captures whether the model knows what it does not know). This shows the importance of standardized, controlled evaluations, so that we can attribute performance to the model versus the adaptation strategy. This result also shows that models are not yet *interoperable*, an important property for building a robust ecosystem of natural language interfaces.
5. We found **human evaluation** essential in some cases. For **summarizing documents**, we find that LMs produce effective summaries (as measured via human evaluation), but the reference summaries in standard summarization datasets (i.e., **CNN/DailyMail** and **XSUM**) are actually worse (under the same human evaluations). Models fine-tuned on these datasets appear to do well according to automatic metrics, such as ROUGE-L, but they also underperform few-shot prompting of LMs. This suggests that better summarization datasets are desperately needed. For **disinformation generation**, we find that OpenAI's text-davinci-002 and Anthropic-LM v4-s3 (52B) are effective at generating realistic headlines that support a given thesis, but results are more mixed when prompting models to generate text encouraging people to perform certain actions. While using LMs for disinformation is not yet a slam dunk, this could change as models become more powerful. Thus, periodic benchmarking is crucial for tracking risks and potential misuse.

RELATED WORK AND DISCUSSION

The rise of LMs

Language modeling has a long-standing tradition of study across human language processing and computational language processing.[16–27,15,28] Language modeling has also been seen as a grand challenge for AI, most notably in the Hutter Prize and the associated enwiki8 benchmark on data compression.[a] However, in contrast to these prior framings, where LMs were viewed as standalone generative models, the models we study in this work instead are better understood by situating LMs in two broader contexts. First, given the models function as adaptable foundations for the myriad scenarios they are tested on, we view LMs as foundation models in service of building performant systems for these downstream use cases.[1] Second, as we demonstrate in our agnosticism on how the models are constructed, we view LMs as natural language interfaces (see Ref. 29).

As Bommasani et al. (Ref. 1, section 1.1) describe, the rise of LMs in Natural Language Processing (NLP) initiated the foundation model paradigm. Specifically, ELMo,[30] GPT,[26] and BERT[27] demonstrated that pretraining using language modeling objectives could produce powerful general-purpose representations for many downstream use cases, building on prior evidence of the successes of pretraining.[31,32] Further, these works, especially GPT and later GPT-2,[33] produced models with qualitatively better generative capabilities than what had been seen previously.

Together, these formative works ushered in a significant change in the status of language modeling in NLP: LMs rapidly became the substrate for almost all modeling work, especially with the advent of open infrastructure through Hugging Face Transformers[34] and models developed for languages beyond English (e.g., multilingual-BERT and XLM[27,35]). Since then, we have seen a proliferation of different organizations building LMs, often through conceptually similar means, with a rapid growth in scale and resource intensity. Notably, some of the models (e.g., TNLG v2 (530B)) we benchmark 1000× larger than ELMo and BERT. These models can cost millions of dollars to train, requiring extensive systems-level optimizations and dedicated large-scale compute.[36] These changes have also translated from research to deployment: LMs are directly exposed as commercial APIs or are integrated into ubiquitous products as part of an emerging commercial ecosystem.[37]

Benchmarks in NLP

Similar to language modeling, benchmarking has a long history in NLP. As Spärck Jones put it in her ACL Lifetime Achievement Award speech, "proper evaluation is a complex and challenging business."[5] To address this challenge, the practice of benchmarking rose to prominence as the core methodology in the 1980s and, especially, the 1990s (see Refs. 38 and 4). This transition was well-demonstrated by initiatives, such as the Message Understanding Conference (MUC[39]) and the Text Retrieval Conference (TREC[40]). And it coincided with a broader shift in the field toward statistical and data-driven methods with large datasets (e.g., the Penn Treebank[41]) and new venues like the Conference on Empirical Methods for Natural Language Processing.[42]

More than a decade later, with the rise of deep learning in the 2010s,[43–48,31,32,49–52] larger benchmarks, such as SNLI[53] and SQuAD,[54] were developed to provide both adequate data for training systems in addition to evaluating systems. This parallels concurrent developments in other areas of AI: most notably, the ImageNet benchmark[55] that shaped modern computer vision. Like their

[a] https://en.wikipedia.org/wiki/Hutter_Prize

predecessors, these benchmarks assign each model a single score (e.g., the SQuAD F1 score) to measure the accuracy for a single task.

As more general-purpose approaches to NLP grew, often displacing more bespoke task-specific approaches, new benchmarks, such as SentEval,[56] DecaNLP,[57] GLUE,[58] and SuperGLUE,[9] coevolved to evaluate their capabilities. In contrast to the previous class of benchmarks, these benchmarks assign each model a vector of scores to measure the accuracy for a suite of scenarios. In some cases, these benchmarks also provide an aggregate score (e.g., the GLUE score, which is the average of the accuracies for each of the constituent scenarios).

More recently, this theme of meta-benchmarks that assess model accuracy across a range of tasks has continued (see Ref. 1, section 4.4.3): for example, GEM[59] provides a suite for natural language generation tasks, XTREME[60] provides a suite for tasks spanning numerous languages, and GEMv2[61] provides a suite for generation across languages. This approach is also the dominant approach to LM evaluation,[b] often with even broader collections: Brown et al.[15] popularized the approach in their work on GPT-3, where they evaluated on 42 datasets. Indeed, this is the approach used in all the works that introduced models we evaluate in this work. Efforts like the EleutherAI Language Model Evaluation Harness,[10] HuggingFace's Evaluate library,[62] and Big-Bench[11] have centralized and expanded these evaluations into systematic repositories.

Situated against this landscape, what differentiates our work is our holistic approach, which manifests in both our benchmark design process and our concrete benchmark. HELM is the byproduct of an explicit two-step process: we taxonomize the space for LM evaluation, structured around use cases (scenarios) and desiderata (metrics), and then systematically select points in a way that reflects our priorities. This makes explicit the aspiration, the concrete benchmark, and, consequently, what our benchmark lacks that we should aspire to evaluate. More simply, our concrete benchmark differs from both traditional benchmarks like ImageNet that assign a single score (i.e., the ImageNet accuracy) and meta-benchmarks like GLUE that assign a score vector (i.e., the accuracies on the GLUE datasets) to each model. Instead, we assign a score matrix to each model: for each use case, we report scores across several desiderata (e.g., accuracy, calibration, robustness, fairness, and efficiency).

Independent of the fact we measure holistically, one may wonder what the relationship is between the scenarios we select and those evaluated in prior works. To help understand this relationship, we document the scenarios that were evaluated for in past work (e.g., the scenarios evaluated by Chowdhery et al.[28] in the PaLM paper or by Gao et al.[10] in the EleutherAI Language Model Evaluation Harness) as well as past results for the models we evaluate on our scenarios (e.g., the **HellaSwag** accuracy reported by Brown et al.[15] in the GPT-3 paper). Further, to build on BIG-Bench specifically, we highlight that our codebase integrates all BIG-Bench scenarios, augmented with metrics beyond accuracy and the ability to evaluate all models we support. We emphasize that, currently, no common standard exists for language modeling evaluation, especially as the capabilities, harms, and limitations of these models are still being understood through the ongoing design of evaluations. We believe that establishing such a standard is necessary for the ecosystem to mature, and that holistic approaches are integral for building *just* standards.

[b] We note Efrat et al.[64] as a very recent counterexample to this trend that takes a much more minimalistic and succinct "unit-testing" perspective.

CONCLUSION

LMs have transformed AI, ushering in the paradigm of foundation models. The reach of modern LMs extends well beyond research, with LMs being rapidly productionized into consequential and ubiquitous language technologies,[37] which we expect to only increase in the near-term future. We lack transparency on LMs at present, which is especially concerning given their rapid growth and burgeoning impact: as a community, we do not understand LMs in their totality. For this reason, we have pushed for holistic evaluation in this effort, as we believe holistic evaluation is a critical means for providing the necessary transparency for LMs.

Transparency begets trust and standards. Viewing benchmarks as models for social change, given that they orient the development of AI systems, our broader objective is to transform foundation models from immature emerging technologies to reliable tools that support human flourishing. With this objective in mind, we recognize the history and trajectory of AI benchmarking aligns with institutional privilege.[63] Benchmarks set the agenda and orient progress: we should aspire for holistic, pluralistic, and democratic benchmarks.[3] Given the understated but significant power of benchmarks to drive change, which in turn indicates that benchmark design confers power, we foreground our objectives for HELM along with its limitations. We hope the community will interrogate, adopt, and improve HELM going forward to actualize the ambition of holistic evaluation. In this way, we hope that holistic evaluations for LMs and for other classes of foundation models will give rise to useful, responsible, and societally beneficial technology.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the study.

ACKNOWLEDGMENTS

HELM was the year-long effort of a team of 50 people. Many others also contributed valuable feedback and guidance; see the paper for the full list of contributors and acknowledgments. We would like to especially thank AI21 Labs, Cohere, and OpenAI for providing credits to run experiments on their limited-access models, as well as Anthropic and Microsoft for providing API access to their closed models. We are grateful to BigScience, EleutherAI, Google, Meta, Tsinghua University, and Yandex for releasing their open models, and to Together for providing the infrastructure to run all the open models. Finally, we would also like to thank Google for providing financial support through a Stanford HAI–Google collaboration as well as the AI2050 program at Schmidt Futures (Grant G-22-63429).

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
2. Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 4846–4853.
3. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. New York: Association for Computing Machinery.
4. Spärck Jones, K., & Galliers, J. R. (1995). Evaluating natural language processing systems: An analysis and review. *Number 1083 in Lecture Notes in Computer Science*. Springer Verlag.
5. Spärck Jones, K. (2005). ACL lifetime achievement award: Some points in a time. *Computational Linguistics*, *31*(1), 1–14.
6. Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
7. Bowman, S. R., & Dahl, G. (2021). What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
8. Raji, I. D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (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)*.
9. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*.
10. Gao, L., Tow, J., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., McDonell, K., Muennighoff, N., Phang, J., Reynolds, L., Tang, E., Thite, A., Wang, B., Wang, K., & Zou, A. (2021). A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*.
11. Srivastava, A., Rastogi, A., Rao, A. B., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.
12. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
13. Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.
14. Liang, P., Bommasani, R., Creel, K. A., & Reich, R. (2022). The time is now to develop community norms for the release of foundation models.
15. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
16. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.
17. Lounsburg, F. G. (1954). Transitional probability, linguistic structure and systems of habit-family hierarchies. Psycholinguistics: A Survey of Theory and Research.
18. Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96–106.
19. Baker, J. K. (1975). *Stochastic modeling for automatic speech understanding*. Morgan Kaufmann Publishers Inc.
20. Baker, J. K. (1975). The dragon system – An overview. IEEE Transactions on Acoustic Speech Signal Processing.
21. Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*, 532–556.
22. Jelinek, F. (1990). *Self-organized language modeling for speech recognition*. Morgan Kaufmann Publishers Inc.
23. Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
24. Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.
25. Merity, S., Keskar, N. S., & Socher, R. (2018). An analysis of neural language modeling at multiple scales. *ArXiv*, abs/1803.08240.
26. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Technical report, OpenAI.
27. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*.
28. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N. M., Prabhakaran, V., … Fiedel, N. (2022). PaLM: Scaling language modeling with pathways.
29. Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., Gerard-Ursin, I., Li, X. L., Ladhak, F., Rong, F., Wang, R. E., Kwon, M., Park, J. S., Cao, H., Lee, T., Bommasani, R., Bernstein, M., & Liang, P. (2022). Evaluating human–language model interaction. *ArXiv*, abs/2212.09746.
30. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *North American Association for Computational Linguistics (NAACL)*.
31. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
32. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 129–136.
34. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
35. Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*.
36. Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., & Zaharia M. (2021). Efficient large-scale language model training on GPU clusters using megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*.

37. Bommasani, R., Soylu, D., Liao, T., Creel, K. A., & Liang, P. (2023). Ecosystem graphs: The social footprint of foundation models. arXiv preprint arXiv:2303.15772.
38. Liberman, M. (2010). Obituary: Fred jelinek. *Computational Linguistics*, *36*(4), 595–599.
39. Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.
40. Voorhees, E. M., & Harman, D. (1998). The Text REtrieval Conferences (TRECs). In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13–15, 1998*. Baltimore, MD: Association for Computational Linguistics.
41. Marcus, M., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3*.
42. EMNLP. (1996). Conference on empirical methods in natural language processing.
43. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning (ICML)*.
44. Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Association for Computational Linguistics (ACL)*.
45. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.
46. Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems (NeurIPS)*.
47. Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *International Conference on Machine Learning (ICML)*.
48. Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *International Conference on Machine Learning (ICML)*.
49. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
50. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.
51. Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.
53. Bowman, S., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Empirical Methods in Natural Language Processing (EMNLP)*.
54. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*.
55. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*.
56. Conneau, A., & Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.
57. McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
58. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*.
59. Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., … Zhou, J. (2021). 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). Association for Computational Linguistics.
60. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.
61. Gehrmann, S., Bhattacharjee, A., Mahendiran, A., Wang, A., Papangelis, A., Madaan, A., McMillan-Major, A., Shvets, A. V., Upadhyay, A., Yao, B., Wilie, B., Bhagavatula, C., You, C., Thomson, C., Garbacea, C., Wang, D., Deutsch, D., Xiong, D., Jin, D., … Hou, Y. (2022). Gemv2: Multilingual nlg benchmarking in a single line of code. *ArXiv*, abs/2206.11249.
62. von Werra, L., Tunstall, L., Thakur, A., Luccioni, A. S., Thrush, T., Piktus, A., Marty, F., Rajani, N., Mustar, V., Ngo, H., Sanseviero, O., vSavsko, M., Villanova, A., Lhoest, Q., Chaumond, J., Mitchell, M., Rush, A. M., Wolf, T., & Kiela, D. (2022). Evaluate&evaluation on the hub: Better best practices for data and model measurements.
63. Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
64. Efrat, A., Honovich, O., & Levy, O. (2022). Lmentry: A language model benchmark of elementary language tasks.
65. Tay, Y., Dehghani, M., Tran, V. Q., García, X., Bahri, D., Schuster, T., Zheng, H., Houlsby, N., & Metzler, D. (2022). Unifying language learning paradigms. *ArXiv*, abs/2205.05131.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Bommasani, R., Liang, P., & Lee, T. (2023). Holistic Evaluation of Language Models. *Ann NY Acad Sci.*, *1525*, 140–146. https://doi.org/10.1111/nyas.15007