Academics Can Contribute to Domain-Specialized Language Models

Mark Dredze[1,2] Genta Indra Winata[3]*Equal contribution Prabhanjan Kambadur[1] Shiqi Wu[4]*Equal contrib
Ozan Irsoy[1] Steven Lu[1] Vadim Dabravolski[1] David S. Rosenberg[1] Sebastian Gehrmann[1]
[1]Bloomberg [2]Johns Hopkins University [3]Capital One [4]Anthropic
mdredze@bloomberg.net

Abstract

Commercially available models dominate academic leaderboards. While impressive, this has concentrated research on creating and adapting general-purpose models to improve NLP leaderboard standings for large language models. However, leaderboards collect many individual tasks and general-purpose models often underperform in specialized domains; domain-specific or adapted models yield superior results. This focus on large general-purpose models excludes many academics and draws attention away from areas where they can make important contributions. We advocate for a renewed focus on developing and evaluating domain- and task-specific models, and highlight the unique role of academics in this endeavor.

1 Introduction

Natural language processing (NLP) research has historically produced domain- and task-specific supervised models. The field has shifted course in the past few years, with a singular focus on general-purpose generative large language models (LLMs) that, rather than focusing on a single task or domain, do well across many tasks [Brown et al.2020, Chowdhery et al.2022, Workshop et al.2022, Zhang et al.2022, Touvron et al.2023b]. By training on massive amounts of data from many sources, these models can do well on extremely broad professional and linguistic examinations [Achiam et al.2023, Anil et al.2023], college-level knowledge questions [Hendrycks et al.2021, Lai et al.2023], and collections of reasoning tasks [Suzgun et al.2023].

While the trend to develop a single, general-purpose generative model is a net positive change that has resulted in impressive results, it has also slowed down progress in other areas of NLP. First, we are less focused on problems that cannot be solved with a chat-like interface. Second, the best-performing LLMs are often commercial systems, which are sometimes opaque about training data, system architecture, and training details. Third, frequent model updates hinder reproducibility.

The resources required to train large general language models naturally constrain research to large organizations, and researchers (or academics) outside of these organizations have become dependent on closed commercial systems, or open systems with limited transparency regarding their training data. This is partly reflected in broader AI trends: zhang2021ai found that roughly 30% of papers at AI conferences (including *CL) have a Fortune 500 tech affiliation. Increased resources contribute to the success of transformer-based LLMs [Vaswani et al.2017], with available hardware [Hooker2021] and benchmarks [Dehghani et al.2021] both playing a deciding role in what models end up being developed. By optimizing the average score across hundreds of shallow tasks, we are smoothing out any signal that would be gained from deeply engaging with individual tasks. Developing domain-specific models can help identify model and training choices that yield improvements on tasks within those domains.

In this paper, we argue for renewed attention to domain-specific models with rigorous and domain-expert informed evaluations. Because many academics are excluded from LLM development due to resource constraints, attention has been drawn away from research areas where academics can make the greatest contributions: deep dives on specific challenging problems. Thus, we propose several research questions to reorient the research community towards developing domain-specific models and applications, where academics are uniquely suited to lead.

2 LLMs: A Brief History

While modern LMs date back to jelinek1976continuous, we summarize very recent history to describe the current environment. In the wake of the popularization of neural word embeddings by word2vec [Mikolov et al.2013], contextualized representations of language as features for supervised systems were realized by ELMo [Peterson et al.2018] followed by BERT [Devlin et al.2019, Liu et al.2019]. BERT and subsequent models became the base models for supervised systems utilizing task-specific fine-tuning and continued pre-training for new domains [Gururangan et al.2020], e.g., for clinical tasks clinicalELMo [Schumacher and Dredze20 and clinicalBERT [Huang et al.2019].

Parallel work utilized transformers for autoregressive LLMs, resulting in GPT [Radford et al.2018], GPT-2 [Radford et al.2019], BART [Lewis et al.2020a, Liu et al.2020], CTRL [Keskar et al.2019], T5 [Raffel et al.2020, Xue et al.2021], and XGLM [Lin et al.2021]. These models had some few-shot capabilities, but they could each be adapted (fine-tuned) for a specific task of interest. Some models were available to academics, though training a new model was beyond reach for many.

GPT-3 [Brown et al.2020] greatly increased model size and changed our understanding of LLMs. Impressive in-context (few-shot) learning pushed the idea that a single large model could solve a wide range of tasks. While the cost of resources meant training was restricted to a few groups, work focused on training bigger models [Chowdhery et al.2022, Anil et al.2023, Zhang et al.2022, Touvron et al.2023a, Rae et al.2021]. While only a few could train large models, many studied how best to use them: prompt engineering [Liu et al.2023], prompt tuning [Han et al.2022, Wei et al.2022], evaluation [Liang et al.2022], among many other topics. Commercial LLM APIs, and eventually open source models [Zhang et al.2022, Workshop et al.2022, Touvron et al.2023a, Touvron et al.2023b, Groeneveld et al.2024], facilitated this work. ignat2024has noted the massive research shift to LLMs reflected in Google Scholar citations. Subsequent work in instruction tuning [Ouyang et al.2022] and fine-tuning [Wei et al.2022, Chung et al.2022, Longpre et al.2023] have further centralized research around general-purpose models. Many consider fine-tuning for specific applications to be obsolete: why would you tune a model for a specific task when you can tune a single model to do well on all tasks?

Despite this view, multiple domain-specific LLMs have demonstrated that domain-specific data leads to models that outperform much larger models [Wu et al.2023, Taylor et al.2022]. Med-PaLM has shown that adapting even giant LLMs to a specific domain leads to vastly increased performance [Singhal et al.2022, Singhal et al.2023]. Furthermore, the release of LLaMA [Touvron et al.2023a] led quickly to Alpaca [Taori et al.2023] and a wave of new fine-tuned versions of LLaMA for specific tasks. This trend strongly indicates that domain-specific models, especially for constrained sizes, are still highly relevant.

To be clear, our concern is not with closed models, which play an important role in the model ecosystem. Models range from full to limited to no access, with some closed models providing incredibly detailed information [Hoffmann et al.2022, Rae et al.2019, Wu et al.2023] and others providing none [Achiam et al.2023]. Our lament over this focus on general models, either open or closed, is that it draws attention away from work on task- and domain-specific models and evaluations. Academics have become product testers, instead of focusing on tasks where they can play a unique role. Moreover, existing academic benchmarks increasingly serve a reduced purpose for commercial models; we are hill-climbing on benchmarks without a way to ensure existing LLMs have not been trained to excel on these benchmarks [Dodge et al.2021]. Furthermore, we rely on benchmarks in place of deep engagement with an application and its stakeholders.

3 The Need for Domain-Specific LLMs

In general, web data does not reflect the needs of all NLP systems. Historically, the community has developed systems for specialized domains such as finance, law, bio-medicine, and science. Accordingly, there have been efforts to build LLMs for these domains [Wu et al.2023, Taylor et al.2022, Singhal et al.2022, Singhal et al.2023, Bolton et al.2023, Luo et al.2022, Lehman et al.2023, Garcia-Ferrero et al.2024]. We need a deep investment in how best to develop and evaluate these models in partnership with domain experts. How should we best integrate insights gained from the development of general-purpose models with these efforts? We propose several research directions.

3.1 How can general-purpose models inform domain-specific models?

Building domain-specific models should benefit from insights and investments into general-purpose models. There are several strategies: training domain-specific models from scratch [Taylor et al.2022, Bolton et al.2023], mixing general and domain-specific data [Wu et al.2023], and fine-tuning existing models [Singhal et al.2022, Singhal et al.2023]. Focusing on domain-specific needs, applications, and knowledge with guidance from topic experts will benefit us in acquiring a better model for specific NLP tasks. Which approach yields the best results for task performance and overall cost?

3.2 What is the role of in-context learning and fine-tuning?

Both LIMA [Zhou et al.2023] and Med-PaLM [Singhal et al.2022] use a small number of examples to tune a model. With expanding context size, we may soon rely entirely on in-context learning [Petroni et al.2020]. This blurs the lines between changing model parameters and conditioning during inference. Beyond inference speed tradeoffs between the two, there may be value in tuning on tens of thousands (or more) of examples. Which domain-specific examples are the most effective to include and in what manner?

3.3 How can LLMs be integrated with domain-specific knowledge?

Specialized knowledge is key in many domains. RAG [Lewis et al.2020b, Guu et al.2020] and KILT-derived works [Petroni et al.2021] focus on knowledge-intensive tasks by including retrieval steps. Work on attributed QA [Bohnet et al.2022] takes a similar approach, as do search LLMs that require interaction with retrieved data [Nakano et al.2021]. Rich updated knowledge sources will always exist beyond the model, especially in environments like medicine, finance, and many academic disciplines.

4 Evaluation of Domain-Specific Models

The evaluation of NLP systems is at a crossroads, and the downstream usage of LLMs and evaluation approaches have diverged. Benchmarks assume that their results translate to insights into similar tasks and usefulness for commercial applications. But benchmarks have become increasingly narrow in scope, oftentimes assessing one metric on a single, often flawed, dataset [Mitchell et al.2019, Kiela et al.2021, Ethayarajh and Jurafsky2020]. The primary evaluation approach for LLMs has been to evaluate on a broad set of these narrow benchmarks [Liang et al.2022, Srivastava et al.2022]. High average performance argues for a broad range of capabilities; however, one size may not fit all. Since specific uses of LLMs are typically much more narrow, we identify three major issues and associated research opportunities with this approach.

4.1 Depth-first Evaluation

Current approaches focus on a single model doing everything well on average instead of being useful in a single domain. However, it is widely acknowledged that the standard benchmarks for most tasks are insufficient (e.g., for summarization, [Fabbri et al.2021, Goyal et al.2022]). Task-specific evaluations have thus adopted additional protocols that measure how well models transfer to different domains, how robust they are, and whether they stand up to concept drift [Mille et al.2021, Dhole et al.2021]. These details disappear when benchmarking on 100+ tasks. Yet, a model's usefulness is not solely defined by doing okay on everything but rather by how well it performs in specific and narrow tasks that provide value. This value is only realized if the model does not suffer from catastrophic failures.

Exemplar studies that perform deep dives on LLMs for specific tasks exist in healthcare [Zack et al.2024, Eriksen et al.2023, Ayers et al.2023, Han et al.2024, Chen et al.2024, Strong et al.2023], law [Blair-Stanek et al.2023a, Blair-Stanek et al.2023b, Magesh et al.2024], and physics [Kim et al.2024], among other areas. We encourage more work on evaluation practices for specific tasks that can handle various model setups and yield informative insights [Zhang et al.2023, Liang et al.2022].

4.2 Sound Metrics

For convenience, most benchmark tasks are formulated as multiple choice question answering or classification. This is not how LLMs are often used. For much more common generation tasks, researchers have been ringing alarms about broken evaluations [Gehrmann et al.2023]. It is dubious whether we gain insights into non-task-specific generation through NLU benchmarks. If we are performing the depth-first evaluation of a generation task, a remaining hurdle—and why researchers fall back to NLU tasks—is the lack of robust metrics. While there is much recent work on better metrics [Celikyilmaz et al.2020, Gehrmann et al.2023], a troubling trend is the use of LLMs as evaluators (e.g., [Sellam et al.2020, Chiang et al.2023]). This approach poses many risks, including the implicit assumption that the evaluating model has access to the ground truth judgment. While there are some promising results, using an LLM out of the box should be avoided (e.g., [Wang et al.2023a, Wang et al.2023b]). Moreover, it is unclear how to evaluate the evaluator when it is a non-deterministic API, or how to scale the development of learned metrics and quantify the strength of a metric.

4.3 Products are not Baselines

If we really do want to evaluate 100+ tasks, there are many issues with the soundness of evaluation setups. At this scope, it is impossible to run careful ablation studies or to assess the effect of changes to methodology in a causal manner. Moreover, different LLMs respond differently to prompts. The BLOOM evaluation averaged over multiple prompts and found significant variance [Workshop et al.2022]. This variance leads to a lack of reproducibility: LLaMA [Touvron et al.2023a] claimed high MMLU [Hendrycks et al.2021] performance but didn't release the prompts that led to them. Similarly, the evaluation scheme makes a difference [Liang et al.2022]. High evaluation costs mean benchmarks pick a small number of setups (sometimes only one) for each task, which introduces further bias, making it hard to construct fair benchmarks on many tasks.

An additional issue with the current benchmarking approach is that the best-performing models are often commercial APIs. With limited transparency regarding data and training, we cannot fairly evaluate these models (e.g., data leakage). Furthermore, task-specific tuning may have been selected based on these specific benchmarks. Moreover, the underlying models change frequently, so it is unclear whether a result will hold for long.

These evaluation issues prompt significant open questions: 1) How do we develop consistent evaluation setups across models that give true measures of performance? 2) How do we develop evaluation setups and metrics more closely aligned with downstream usage? 3) How do we develop evaluation suites that support depth-first evaluation and not breadth-first benchmarking?

5 The Role of Academics

A focus on general-purpose LLMs has forced academics to work with large base models and perhaps, shifted the focus to solve problems of immediate industrial interest. Many academics feel excluded from current research trends [Ignat et al.2024] and the academic and industry relationship is changing [Littman et al.2022]. Shifting attention back to domain-specific applications emphasizes areas where academics hold an advantage: partnerships with domain experts to invest in specific tasks, and consideration of broader societal needs.

Developing domain-specific models requires domain expertise and universities are diverse academic environments that house experts in many domains. Collaborations with these experts can identify data sources, tasks, and challenges important within each domain. Furthermore, these collaborations are the best avenues for better alignment of evaluations with use cases [Winata et al.2024], and can support the development of proper metrics. These collaborations are necessary to explore wide open interdisciplinary topics, such as models for protein structure prediction [Tunyasuvunakool et al.2021, Vig et al.2021] and games as proxies for reasoning [Silver et al.2016, Agostinelli et al.2019, Schrittwieser et al.2020]. This includes developing domain-specific resources, which require domain experts to properly design and construct the datasets. Further, areas where industry underinvests are those where academics could focus attention. For example, low-resource languages are not served by a general-purpose multilingual LLM, nor will we reasonably have enough data to support current LLM training methods. Dialects and variations in languages are still wide open topics [Aji et al.2022, Winata et al.2023, Nicholas and Bhatia2023].

General-purpose LLMs are unlikely to solve problems in many important domains, with many open research problems that can only be solved by domain-specific approaches. Focusing on domain-specific knowledge will benefit us in acquiring a better model and developing application strategies more aligned with how humans learn domain-specific knowledge [Tricot and Sweller2014]. For many interdisciplinary areas, subject matter experts are essential, and the problems must be defined clearly. The first pass from an LLM is often impressive, but it hides the trenches and areas where things are most interesting. We need a renewed focus on developing and evaluating domain-specific models and applications, an area where academics can play a leading role. Let us not be distracted by claims that a single model solves all tasks, and instead deeply explore and understand the needs and challenges of specific domains.

6 Limitations

The literature that we explored in this opinion paper is limited to the area of LLMs. We study the history of LLMs from the literature on word embeddings, encoder-only, and generative transformers to the latest advancement of API-based LLMs.

7 Ethics Statement

Our work does not include any experiments or use of data. No potential ethical issues in this work.

References

[Achiam et al.2023] 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*.

[Agostinelli et al.2019] Forest Agostinelli, Stephen McAleer, Alexander Shmakov, and Pierre Baldi. 2019. Solving the Rubik's Cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8):356–363.

[Aji et al.2022] Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

[Anil et al.2023] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

[Ayers et al.2023] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596.

[Blair-Stanek et al.2023a] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023a. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

[Blair-Stanek et al.2023b] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023b. Openai cribbed our tax example, but can gpt-4 really do tax? *arXiv preprint arXiv:2309.09992*.

[Bohnet et al.2022] Bernd Bohnet, Mnh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *CoRR*, abs/2212.08039.

[Bolton et al.2023] Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2023. BioMedLM. `https://github.com/stanford-crfm/BioMedLM`.

[Brown et al.2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[Celikyilmaz et al.2020] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

[Chen et al.2024] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*.

[Chiang et al.2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. `https://vicuna.lmsys.org`.

[Chowdhery et al.2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

[Chung et al.2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

[Dehghani et al.2021] Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. 2021. The benchmark lottery. *CoRR*, abs/2107.07002.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Dhole et al.2021] Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang

Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard H. Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, et al. 2021. NL-Augmenter: A framework for task-sensitive natural language augmentation. *CoRR*, abs/2112.02721.

[Dodge et al.2021] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305.

[Eriksen et al.2023] Alexander V Eriksen, Soren Møller, and Jesper Ryg. 2023. Use of gpt-4 to diagnose complex clinical cases.

[Ethayarajh and Jurafsky2020] Kawin Ethayarajh and Dan Jurafsky. 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)*, pages 4846–4853, Online. Association for Computational Linguistics.

[Fabbri et al.2021] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

[Garcia-Ferrero et al.2024] Iker Garcia-Ferrero, Rodrigo Agerri, Aitziber Atutxa, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, et al. 2024. Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. *arXiv preprint arXiv:2404.07613*.

[Gehrmann et al.2023] Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

[Goyal et al.2022] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356.

[Groeneveld et al.2024] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

[Gururangan et al.2020] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

[Guu et al.2020] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

[Han et al.2022] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. PTR: Prompt tuning with rules for text classification. *AI Open*, 3:182–192.

[Han et al.2024] Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Towards safe large language models for medicine. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

[Hendrycks et al.2021] 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*.

[Hoffmann et al.2022] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al.

2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

[Hooker2021] Sara Hooker. 2021. The hardware lottery. *Commun. ACM*, 64(12):58–65.

[Hsieh et al.2023] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

[Huang et al.2019] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

[Ignat et al.2024] Oana Ignat, Zhengjin Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, et al. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094.

[Jelinek1976] Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

[Keskar et al.2019] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

[Kiela et al.2021] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringhia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 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*, pages 4110–4124, Online. Association for Computational Linguistics.

[Kim et al.2024] Eun-Ah Kim, Haining Pan, Nayantara Mudur, William Taranto, Subhashini Venugopalan, Yasaman Bahri, and Michael Brenner. 2024. Performing Hartree-Fock many-body physics calculations with large language models. *Bulletin of the American Physical Society*.

[Lai et al.2023] Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

[Lehman et al.2023] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR.

[Lewis et al.2020a] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

[Lewis et al.2020b] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

[Liang et al.2022] 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 Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan 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. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

[Lin et al.2021] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

[Littman et al.2022] Michael L Littman, Ifeoma Ajunwa, Guy Berger, Craig Boutilier, Morgan Currie, Finale Doshi-Velez, Gillian Hadfield, Michael C Horowitz, Charles Isbell, Hiroaki Kitano, et al. 2022. Gathering strength, gathering storms: The one hundred year study on artificial intelligence (ai100) 2021 study panel report. *arXiv preprint arXiv:2210.15767*.

[Liu et al.2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

[Liu et al.2020] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

[Liu et al.2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Longpre et al.2023] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The FLAN collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

[Luo et al.2022] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

[Magesh et al.2024] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mille et al.2021] Simon Mille, Kaustubh D. Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Sanjay Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Advances in Neural Information Processing Systems*.

[Mitchell et al.2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29–31, 2019*, pages 220–229. ACM.

[Nakano et al.2021] Reiichiro Nakano, Jacob Hilton, Suchir Bhatia, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

[Nicholas and Bhatia2023] Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

[Nori et al.2023] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

[Ouyang et al.2022] 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*, 35:27730–27744.

[Peterson et al.2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[Petroni et al.2020] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

[Petroni et al.2021] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim

Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

[Radford et al.2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

[Radford et al.2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Rae et al.2021] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

[Rae et al.2019] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.

[Raffel et al.2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

[Schrittwieser et al.2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering Atari, Go, Chess and Shogi by planning with a learned model. *Nature*, 588(7839):604–609.

[Schumacher and Dredze2019] Elliot Schumacher and Mark Dredze. 2019. Learning unsupervised contextual representations for medical synonym discovery. *JAMIA Open*, 2(4):538–546.

[Sellam et al.2020] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

[Singhal et al.2022] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schiürli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

[Singhal et al.2023] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

[Silver et al.2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[Srivastava et al.2022] 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, Ameet Annasaheb Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew D. 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 Mullokan-dov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, Bridget R. Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Ozyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin,

Blake Stephen Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Tatiana Ramirez, Clara Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Daniel H Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel González, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, D. 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, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth P. Donaway, Ellie Pavlick, Emanuele Rodolì, Emma FC Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fan Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Han Sol Kim, Hannah Rashkin, Hanna Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hubert Wong, Ian Aik-Soon Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, John K. Miller, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, J. Brooker Simon, James Koppel, James Zheng, James Zou, Jan Kočoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Narain Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jenni Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Oluwadara Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Jane Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Ochieng 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, Luca Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lutfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Madotto Andrea, Maheen Saleem Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Márcio Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melissa Arnaud, Melvin Andrew McElrath, Michael A. Yee, Michael Cohen, Mi Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Monica Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, N. Mukund Varma, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas S. Roberts, Nicholas Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha 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, Pi Bei Hwang, P. Milkowski, Piyush S. Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, QING LYU, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphael 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, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib J. Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Sam Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi S. Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soohwan Lee, Spencer Bradley Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Rose Biderman, Stephanie C. Lin, S. Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq A. Ali, Tatsuo Hashimoto, Te-Lin Wu, Theo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, T. N. Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler O'Brien 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, W. Vossen, Xiang Ren, Xiaoyu Tong, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yang Song, Yasaman Bahri, Ye Ji Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yun-tao Bai, Zachary Seid, Zhao Xinran, Zhuyue Zhao, Zi Fu Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu, Sahib Singh, and Uri Shaham. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

[Strong et al.2023] Eric Strong, Alicia DiGiammarino, Yingjie Weng, Andre Kumar, Poonam Hosamani, Jason Hom, and Jonathan H Chen. 2023. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Internal Medicine*, 183(9):1028–1030.

[Suzgun et al.2023] Mirac Suzgun, Nathan Scales, Nathanael Schiürli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

[Taori et al.2023] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

[Taylor et al.2022] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

[Touvron et al.2023a] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

[Touvron et al.2023b] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[Tricot and Sweller2014] André Tricot and John Sweller. 2014. Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, 26:265–283.

[Tunyasuvunakool et al.2021] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Zidek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

[Vig et al.2021] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Rajani. 2021. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*.

[Wang et al.2023a] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

[Wang et al.2023b] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023b. How far can camels go? Exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

[Wei et al.2022] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

[Winata et al.2023] Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, et al. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834.

[Winata et al.2024] Genta Indra Winata, Hanyang Zhao, Anirban Das, Wen-pin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. 2024. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *arXiv preprint arXiv:2409.11564*.

[Workshop et al.2022] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

[Wu et al.2023] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

[Xue et al.2021] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

[Zack et al.2024] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

[Zhang et al.2021] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. 2021. The AI index 2021 annual report. *arXiv preprint arXiv:2103.06312*.

[Zhang et al.2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

[Zhang et al.2023] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. *CoRR*, abs/2301.13848.

[Zhou et al.2023] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Faeze Brahma, Vishrav Chaudhary, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. *CoRR*, abs/2305.11206.