

Academics Can Contribute to Domain-Specialized Language Models

Mark Dredze^{1,2}, Genta Indra Winata^{3,*}, Prabhanjan Kambadur¹, Shijie Wu^{4,*}, Ozan Irsoy¹, Steven Lu¹, Vadim Dabrowski¹, David S Rosenberg¹, and Sebastian Gehrmann¹

¹Bloomberg

²Johns Hopkins University

³Capital One

⁴Anthropic

*The project was completed during work at Bloomberg.

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 [10, 14, 88, 93, 78]. By training on massive amounts of data from many sources, these models can do well on extremely broad professional and linguistic examinations [1, 4], college-level knowledge questions [28, 38], and collections of reasoning tasks [74].

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: [92] found that roughly 30

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 [34], we summarize very recent history to describe the current environment. In the wake of the popularization of neural word embeddings by word2vec [51], contextualized representations of language as features for supervised systems were realized by ELMo [58] followed by BERT [17, 45]. BERT and subsequent models became the base models for supervised systems utilizing task-specific fine-tuning and continued pre-training for new domains [25], e.g., for clinical tasks ELMO [67] and clinicalBERT [32].

Parallel work utilized transformers for autoregressive LLMs, resulting in GPT [61], GPT-2 [62], BART [40, 46], CTRL [35], T5 [65, 90], and XGLM [43]. 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 [10] 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 [14, 4, 93, 77, 64].

While only a few could train large models, many studied how best to use them: prompt engineering [47], prompt tuning [26, 85], evaluation [42], among many other topics. Commercial LLM APIs, and eventually open source models [93, 88, 77, 78, 23], facilitated this work. [33] noted the massive research shift to LLMs reflected in Google Scholar citations.

Subsequent work in instruction tuning [57] and fine-tuning [85, 15, 48] 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?¹

Despite this view, multiple domain-specific LLMs have demonstrated that domain-specific data leads to models that outperform much larger models [89, 76]. Med-PaLM has shown that adapting even giant LLMs to a specific domain leads to vastly increased performance [70, 71].² Furthermore, the release of LLAMA [77] led quickly to Alpaca [75] 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 [29, 63, 89] and others providing none [1]. 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 [?]. Furthermore, we rely on benchmarks in place of deep engagement with an application and its stakeholders.

¹Distillation for task-specific models remains popular if smaller models are desired [31].

²We acknowledge that the biomedical domain is a rapidly developing area, and GPT-4 without fine-tuning was reported to surpass MedPaLM 2 [56].

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 [89, 76, 70, 9, 49, 39, 20]. 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.

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 [76, 9], mixing general and domain-specific data [89], and fine-tuning existing models [70, 71]. 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?

What is the role of in-context learning and fine-tuning? Both LIMA [95] and Med-PaLM [70] use a small number of examples to tune a model. With expanding context size, we may soon rely entirely on in-context learning [59]. 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?

How can LLMs be integrated with domain-specific knowledge? Specialized knowledge is key in many domains. RAG [41, 24] and KILT-derived works [60] focus on knowledge-intensive tasks by including retrieval steps. Work on attributed QA [8] takes a similar approach, as do search LLMs that require interaction with retrieved data [54]. 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 [53, 36?]. The primary evaluation approach for LLMs has been to evaluate on a broad set of these narrow benchmarks [42, 72]. 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.

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, [19, 22]). 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 [52, 18]. 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 [91? , 5, 27, 12, 73], law [7, 6, 50], and physics [37], among other areas. We encourage more work on evaluation practices for specific tasks that can handle various model setups and yield informative insights [94, 42].

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 [21]. 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 [11, 21], a troubling trend is the use of LLMs as evaluators (e.g., [68, 13]). 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., [83, 84]). 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.

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 [88]. This variance leads to a lack of reproducibility: LLAMA [77] claimed high MMLU [28] performance but didn't release the prompts that led to them.³ Similarly, the evaluation scheme makes a difference [42]. 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 [33] and the academic and industry relationship is changing [44]. 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.

³There was model evaluation: [<https://huggingface.co/blog/open-llm-leaderboard-mmlu>] (<https://huggingface.co/blog/open-llm-leaderboard-mmlu>)

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 [87], 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 [80, 82] and games as proxies for reasoning [69, 2, 66]. 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 [3, 86, 55].

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 [79]. 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.

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.

Ethics Statement

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

References

- [1] 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*.
- [2] 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.
- [3] 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.

- [4] 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*.
- [5] 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.
- [6] 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.
- [7] 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*.
- [8] Bernd Bohnet, Vinh 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/2212.08037.
- [9] Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2023. BioMedLM. [<https://github.com/stanford-crfm/BioMedLM>] (<https://github.com/stanford-crfm/BioMedLM>).
- [10] 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.
- [11] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv*, 2006.14799.
- [12] 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*.
- [13] 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 See <https://vicuna.lmsys.org/> (<https://www.google.com/search?q=https://vicuna.lmsys.org/>) (accessed 14 April 2023), 2(3):6.
- [14] 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 Sep. 2022. [MISSING TITLE] [MISSING JOURNAL]

- [15] 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*, 2210.11416.
- [16] 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.
- [17] 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.
- [18] 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, et al. 2021. [MISSING TITLE] [MISSING JOURNAL]
- [19] 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.
- [20] Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, 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*.
- [21] 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.
- [22] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356.
- [23] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. [MISSING TITLE] [MISSING JOURNAL]
- [24] [MISSING REFERENCE: Guu et al. 2020]
- [25] [MISSING REFERENCE: Gururangan et al. 2020]
- [26] [MISSING REFERENCE: Han et al. 2022]
- [27] [MISSING REFERENCE: Han et al. 2024]
- [28] [MISSING REFERENCE: Hendrycks et al. 2021]
- [29] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lissa 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.
- [30] Sara Hooker. 2021. The hardware lottery. *Commun. ACM*, 64(12):58–65.

- [31] 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.
- [32] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.
- [33] Oana Ignat, Zhijing 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? [MISSING DETAILS]
- [34] [MISSING REFERENCE: Jelinek 1976]
- [35] [MISSING REFERENCE: Keskar et al. 2019]
- [36] [MISSING REFERENCE: Kiela et al. 2021]
- [37] [MISSING REFERENCE: Kim et al. 2024]
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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. [MISSING TITLE] [MISSING JOURNAL]
- [42] [MISSING REFERENCE: Liang et al. 2022]
- [43] [MISSING REFERENCE: Lin et al. 2021]
- [44] 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*.
- [45] Yinhan Liu, Ji... [MISSING AUTHORS] 2019. [MISSING TITLE] [MISSING JOURNAL]
- [46] [MISSING REFERENCE: Liu et al. 2020]
- [47] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- [48] [MISSING REFERENCE: Longpre et al. 2023]

- [49] [MISSING REFERENCE: Luo et al. 2022]
- [50] [MISSING REFERENCE: Magesh et al. 2024]
- [51] [MISSING REFERENCE: Mikolov et al. 2013]
- [52] [MISSING REFERENCE: Mille et al. 2021]
- [53] [MISSING REFERENCE: Mitchell et al. 2019]
- [54] [MISSING REFERENCE: Nakano et al. 2021]
- [55] [MISSING REFERENCE: Nicholas and Bhatia 2023]
- [56] [MISSING REFERENCE: Nori et al. 2023]
- [57] [MISSING REFERENCE: Ouyang et al. 2022]
- [58] [MISSING REFERENCE: Peters et al. 2018]
- [59] [MISSING REFERENCE: Petroni et al. 2020]
- [60] [MISSING REFERENCE: Petroni et al. 2021]
- [61] [MISSING REFERENCE: Radford et al. 2018]
- [62] [MISSING REFERENCE: Radford et al. 2019]
- [63] [MISSING REFERENCE: Rae et al. 2019]
- [64] [MISSING REFERENCE: Rae et al. 2021]
- [65] [MISSING REFERENCE: Raffel et al. 2020]
- [66] [MISSING REFERENCE: Schrittwieser et al. 2020]
- [67] [MISSING REFERENCE: Schumacher and Dredze 2019]
- [68] [MISSING REFERENCE: Sellam et al. 2020]
- [69] [MISSING REFERENCE: Silver et al. 2016]
- [70] [MISSING REFERENCE: Singhal et al. 2022]
- [71] [MISSING REFERENCE: Singhal et al. 2023]
- [72] [MISSING REFERENCE: Srivastava et al. 2022]
- [73] 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.
- [74] Mirac Suzgun, Nathan Scales, [MISSING AUTHORS] 2023. [MISSING TITLE] [MISSING JOURNAL]
- [75] [MISSING REFERENCE: Taori et al. 2023]
- [76] [MISSING REFERENCE: Taylor et al. 2022]
- [77] [MISSING REFERENCE: Touvron et al. 2023a]
- [78] [MISSING REFERENCE: Touvron et al. 2023b]

- [79] [MISSING REFERENCE: Tricot and Sweller 2014]
- [80] [MISSING REFERENCE: Tunyasuvunakool et al. 2021]
- [81] [MISSING REFERENCE: Vaswani et al. 2017]
- [82] [MISSING REFERENCE: Vig et al. 2021]
- [83] [MISSING REFERENCE: Wang et al. 2023a]
- [84] [MISSING REFERENCE: Wang et al. 2023b]
- [85] [MISSING REFERENCE: Wei et al. 2022]
- [86] [MISSING REFERENCE: Winata et al. 2023]
- [87] [MISSING REFERENCE: Winata et al. 2024]
- [88] tao Bai, Zachary Seid, Zhao Xinran, Zhuoye 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*, abs/2206.04615.
- [89] [MISSING REFERENCE: Wu et al. 2023]
- [90] [MISSING REFERENCE: Xue et al. 2021]
- [91] 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 healthcare: A model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.
- [92] 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*.
- [93] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui 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*.
- [94] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language... [MISSING TITLE] [MISSING JOURNAL]
- [95] [MISSING REFERENCE: Zhou et al. 2023]