Compare Results

Old File:		New File:
2024.emnlp-main.293.pdf	versus	**2024_emnlp-main_293.pdf**
11 pages (180 KB)		**17 pages (304 KB)**
10/31/2024 10:32:46 PM		2/21/2026 8:00:47 AM

Total Changes

967

Content

69 Replacements
72 Insertions
89 Deletions

Styling and Annotations

414 Styling
323 Annotations

Go to First Change (page 2)

Academics Can Contribute to Domain-Specialized Language Models

Mark Dredze Genta Indra Winata Prabhanjan Kambadur
Shilie Wu Ozan Irsoy Steven Lu Yadim Dabravolski
David S. Rosenberg Sebastian Gehrmann

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.

[ILLEGIBLE] 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: Zhang et al. (2021) found that roughly 30

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: Zhang et al. (2021) found that roughly

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 Jelinek (1976), 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 (Peters 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 ELMo (Schumacher and Dredze, 2019) 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,b; Groeneveld et al., 2024), facilitated this work. Ignat et al. (2024) 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 could 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 (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, 2023).[2] 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

[1] Distillation for task-specific models remains popular if smaller models are desired (Hsieh et al., 2023).

[2] We acknowledge that the biomedical domain is a rapidly developing area, and GPT-4 without fine-tuning was reported to surpass MedPaLM 2 (Nori et al., 2023).

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; 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.

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, 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?*

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?*

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 Jurafsky, 2020). The primary evaluation approach for LLMs has been to evaluate on a broad set of these narrow benchmarks (Liang et al., 2022, HELM) (Srivastava et al., 2022, BIG-Bench). 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, 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,b; 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).

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 evaluation 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,b). 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 (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.[3] 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

[3] Similarly, the evaluation scheme makes a difference (Liang et al., 2022, Fig. 33).

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?[4]](https://huggingface.co/blog/open-llm-leaderboard-mmlu)

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 Bhatia, 2023).

General-purpose LLMs are unlikely to solve problems in many important domains, with many open research problems that can only be solved by

[4]There was significant confusion surrounding model evaluation: https://huggingface.co/blog/open-llm-leaderboard-mmlu

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 Sweller, 2014). 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 APl-based LLMs.

Ethics Statement

We confirm that all ethical concerns are addressed.

References

Sy s tems, 33:181 1 -1901.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Asli Celikyilmaz,Elizabeth Clark, and Jianfeng Gao.

2020. Evahntion of text generation: A survey. arXlv,

Diogo Almeida, Janko Altenschmidt, Sam Altman,

2006.14799.

Shyamal Anadkat, et al. 2023. Gpt-4 technical report.

arXiv p re p rint arXiv : 2 3 0 3.0877 4. Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark

Forest Agostinelli, Stephen McAleer, Alexander Dredze. 2024. Benchmarking large language models

on answering and explaining challenging medical

Shmakov, and Pierre Baldi. 2019. Solving the Rubik's Cube with deep reinforcement learning and questions. arXiv preprint arXiv : 2402. I 8060.

search. N ature M achine Int e ll i g enc e, 1 (8) : 3 56-3 63. Wei-Lin Chiang, Zhuohan L| Zi Lin, Ying Sheng,
Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan
Zhrrang, Yonghao Zhtang, Joseph E Gonzalez, et al.
Cahyawijaya, Ade Romadhony, Rahmad Mahendra, 2023. Yicrlna: An open-source chatbot impressing
Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, gpt-4 with 90Vo* many languages: Building the indonesian nlp benchmark dataset and model. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7265-7279.
Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert L Johnson, Cesar Cibeira, Mahmoud Mostafa, et al. 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.
John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Jennifer J Zhu, Jessica B Kelley, Rachna Reddy, Joseph A Ha, C Yiu Cho, and Davey M Smith. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum.
ternal Medicine*JAMA Internal Medicine*, 183(6):589–596.
Iz Beltagy, Arman Cohan, and Kyle Lo. 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.
Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023a. The humans pretend and the ai acts: A simulation of legal education with generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.
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.
Bernd Bohnet, Mnh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, and Michael Collins. 2022. Attributed question answering: Eval-

uating the attribution of answers to their sources. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4621–4632.

Elliot Bolton, Shubham Gupta, and Christopher D. Manning. 2023. Sapiens: A system for fine-tuning language models for reasoning in science. arXiv preprint arXiv:2306.11470.

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

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. arXiv, 2006.14799.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and Ion Stoica. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and others. 2022. PaLM: Scaling language modeling with pathways. arXiv

preprint arXiv:2204.02311.
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.
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.
Mostafa Dehghani, Alexey Gritsenko, Mario N. Belk, and others. 2021. Scaling vision with sparse mixture of experts. arXiv preprint arXiv:2106.05974.
Bhavya Dhole, Gaurav Kumar, and Romain Paulus. 2021. Robustness and domain transferability of neural summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10358–10375.
Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2021. Documenting large webtext corpora: A case study on the colosal clean crawled corpus. arXiv preprint arXiv:2104.08758.
Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. arXiv preprint arXiv:2009.13888.
Emma Eriksen, Michael W. Vowels, and others. 2023. GPT-4 on medical licensing exams. *[ILLEGIBLE]*
Alexander R. Fabbri, Wojciech Kryscinski, 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.
Iker Garcia-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:2402.[ILLEGIBLE].
Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the broken evaluation of text generation. *Journal of Machine Learning Research*, 24(168):1–76.
Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. NewsSumm: A dataset for summarizing news. *[ILLEGIBLE]*
Dirk Groeneveld, Kyle Lo, and others. 2024. OLMo: Accelerating the science of language models. *[ILLEGIBLE]*
Suchin Gururangan, Ana Marasovic, 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.
Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938.
Naman Goyal, and others. 2022. *[ILLEGIBLE]*
Xiao Han, Weilin Xia, and others. 2022. Prompt tuning is competitive. *[ILLEGIBLE]*
Eric Han, and others. 2024. *[ILLEGIBLE]*
Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*.
Leo L. Hooker. 2021. The hardware lottery. *Communications of the ACM*, 64(12):58–65.
Jordan Hoffmann, Sebastian Borgeaud, and others. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.
Kuan-Hao Huang, Jiaan-Der Chen, and others. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *[ILLEGIBLE]*
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. *[ILLEGIBLE]*

Nicolas Ignat, and others. 2024. *[ILLEGIBLE]*

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

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.

Douwe Kiela, and others. 2021. Dynabench: Rethinking benchmark. arXiv preprint arXiv:2104.14337.

Jungwoo Kim, and others. 2024. *[ILLEGIBLE]*

Steven Lu, and others. *[ILLEGIBLE]*

Qing Lai, and others. 2023. *[ILLEGIBLE]*

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.

Patrick Lewis, and others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *[ILLEGIBLE]*

Fangzhao Liu, and others. 2023. *[ILLEGIBLE]*

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.

Yinhan Liu, and others. 2020. *[ILLEGIBLE]*

Percy Liang, and others. 2022. *[ILLEGIBLE]*

Xi Victoria Lin, and others. 2021. XGLM: A multilingual autoregressive language model. arXiv preprint arXiv:2112.10668.

Littman, and others. 2022. *[ILLEGIBLE]*

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.

Yi Luo, and others. 2022. *[ILLEGIBLE]*
Magesh, and others. 2024. *[ILLEGIBLE]*
Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
Rodolfo Mille, and others. 2021. *[ILLEGIBLE]*
Margaret Mitchell, and others. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.
Nicholas and Bhatia. 2023. *[ILLEGIBLE]*
Nori, and others. 2023. *[ILLEGIBLE]*
OpenAI. 2023. *[ILLEGIBLE]*
Long Ouyang, and others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 35.
Fabio Petroni, Tim Rockt"aschel, and others. 2020. How contextual are contextualized word representations? *[ILLEGIBLE]*
Fabio Petroni, and others. 2021. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of NAACL*.
Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *[ILLEGIBLE]*
Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *[ILLEGIBLE]*
Jack Rae, Sebastian Borgeaud, and others. 2021. Scaling language models: Methods, analysis insights. arXiv preprint arXiv:2112.10684.
Jack Rae, and others. 2019. *[ILLEGIBLE]*
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. *Journal of Machine Learning Research*, 21(140):1–67.
Schrittwieser, and others. 2020. *[ILLEGIBLE]*
Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of ACL*.

Singhal, and others. 2022. Large language models encode clinical knowledge.
[ILLEGIBLE]
Singhal, and others. 2023. Towards expert-level medical question answering with large language models.
[ILLEGIBLE]
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,
Srivastava, and others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.
arXiv preprint arXiv:2206.04615.
Strong, and others. 2023. *[ILLEGIBLE]*
Mirac Suzgun, Nathan Scales, Nathanael Sch"arli, 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: EMNLP 2023, pages 9699–9717.
Taylor, and others. 2022. Galactica: A large language model for science.
arXiv preprint arXiv:2211.09085.
Taori, and others. 2023. Alpaca: A strong, replicable instruction-following model.
[ILLEGIBLE]
Tricot and Sweller. 2014. Domain-specific knowledge and learning.
[ILLEGIBLE]
Touvron, and others. 2023a. LLaMA: Open and efficient foundation language models.
arXiv preprint arXiv:2302.13971.
Touvron, and others. 2023b. Llama 2: Open foundation and fine-tuned chat models.
arXiv preprint arXiv:2307.09288.
Tunyasuvunakool, and others. 2021. Highly accurate protein structure prediction with AlphaFold.
Nature, 596:583–589.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 30.
Vig, and others. 2021. *[ILLEGIBLE]*
Wang, and others. 2023a. *[ILLEGIBLE]*
Wang, and others. 2023b. *[ILLEGIBLE]*
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*.
Workshop, and others. 2022. BLOOM: A 176B-parameter open-access multilingual language model.
arXiv preprint arXiv:2211.05100.
Shilie Wu, and others. 2023. BloombergGPT: A large language model for finance.
arXiv preprint arXiv:2303.17564.
Xue, and others. 2021. mT5: A massively multilingual pre-trained text-to-text transformer.
In *Proceedings of NAACL*.
Zack, and others. 2024. *[ILLEGIBLE]*
Tianyi Zhang, and others. 2021. *[ILLEGIBLE]*
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/2301.13949.
Winata, and others. 2023. *[ILLEGIBLE]*
Winata, and others. 2024. *[ILLEGIBLE]*
Zhou, and others. 2023. LIMA: Less is more for alignment.
arXiv preprint arXiv:2305.11206.