2/21/2026 5:58:05 PM

Compare Results

Old File:		New File:
2024.emnlp-main.293.pdf	versus	**2024_emnlp-main_293.pdf**
11 pages (180 KB)		**9 pages (184 KB)**
10/31/2024 10:32:46 PM		2/21/2026 7:34:13 AM

Total Changes

860

Content

41 Replacements
55 Insertions
156 Deletions

Styling and Annotations

312 Styling
296 Annotations

Go to First Change (page 2)

Academics Can Contribute to Domain-Specialized Language Models

Mark Dredze Genta Indra Winata Prabhanjan Kambadur Shijie Wu
Ozan Irsoy Steven Lu Vadim Dabravolski David S Rosenberg
Sebastian Gehrmann

Abstract

Large language models (LLMs) have demonstrated impressive performance across a wide range of tasks, but training state-of-the-art models requires substantial computational and financial resources, placing them largely out of reach for academic institutions. This has led to concerns that academic researchers may be excluded from meaningful participation in the development of cutting-edge LLMs. In this paper, we argue that academics can still play a critical role by focusing on domain-specialized LLMs. We discuss the motivations for domain-specific models, outline strategies for developing and evaluating them within academic constraints, and highlight opportunities for impactful research that does not require industrial-scale resources.

1 Introduction

Large language models (LLMs) have rapidly transformed natural language processing (NLP), achieving remarkable performance across a broad range of tasks. Models such as GPT-3 and its successors have demonstrated strong few-shot and zero-shot capabilities, enabling applications that were previously out of reach. These advances have been driven by scaling model size, training data, and computational resources.

However, the increasing scale of state-of-the-art LLMs has also created a widening gap between industry and academia. Training frontier models requires massive computational infrastructure and financial investment, often available only to large technology companies. As a result, academic researchers may feel excluded from contributing to the development of cutting-edge LLMs.

Despite these challenges, we argue that academia can continue to play a vital and distinctive role in LLM research. Rather than competing directly with industry on general-purpose foundation models, academic researchers can focus on domain-specialized LLMs tailored to specific areas of expertise. These models can provide high value while remaining feasible within academic resource constraints.

In this paper, we outline the case for domain-specific LLMs, discuss how they can be developed and evaluated in academic settings, and highlight research directions where academics can make meaningful contributions.

2 LLMs: A Brief History

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 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 (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 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.

[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 Med-PaLM 2 (Nori et al., 2023).

3 The Need for Domain-Specific LLMs

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; García-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). The interaction between in-context learning and parameter updates remains an open question, particularly for specialized domains where data may be scarce.

How should we evaluate domain-specific LLMs? Evaluation in specialized domains requires collaboration with subject matter experts and the development of high-quality benchmarks that reflect real-world use. Many existing benchmarks were not designed with domain-specific applications in mind, and new evaluation paradigms may be needed to measure utility, safety, and robustness in context.

4 Evaluation of Domain-Specific Models

Evaluation of Domain-Specific Models

Developing domain-specific LLMs requires rigorous evaluation that reflects the needs of the target domain. While general-purpose benchmarks provide useful comparisons across models, they often fail to capture the nuances and practical requirements of specialized applications. Domain-specific evaluation must therefore go beyond leaderboard performance and consider task relevance, reliability, and alignment with domain standards.

Collaboration with domain experts is essential in constructing meaningful evaluation datasets and metrics. Experts can help identify realistic use cases, define acceptable error rates, and assess qualitative aspects such as reasoning quality and factual consistency. Without such collaboration, models risk optimizing for superficial benchmark gains rather than practical utility.

Another challenge lies in data contamination and benchmark leakage. As commercial LLMs are trained on increasingly large and opaque datasets, it becomes difficult to ensure that evaluation benchmarks have not been included in training data. This concern is especially pronounced for widely used academic benchmarks. For domain-specific models trained within academic settings, transparent data documentation and careful curation can help mitigate these issues.

Finally, evaluation should account for deployment considerations, including robustness, safety, and fairness within the domain context. A model that performs well on isolated test sets may

still fail under real-world conditions. Ongoing monitoring and iterative evaluation are therefore necessary components of responsible domain-specific model development.

5 The Role of Academics

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 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.

6 Limitations

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

Ethics Statement

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

8 References

References

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.

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.

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.

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.

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.

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.

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, 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.

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

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.

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.

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

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.

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.

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.

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.

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, and et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. CoRR, abs/2112.02721.

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.

Alexander V Eriksen, Sören Möller, and Jesper Ryg. 2023. Use of gpt-4 to diagnose complex clinical cases.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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

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.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

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? 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.

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, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, 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.

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.

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.

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.

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 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.

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 Yüksekgönül, 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.

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.

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.

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.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [ILLEGIBLE]