Academics Can Contribute to Domain-Specialized Language Models

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

February 21, 2026

Abstract

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

1 Introduction

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

While the trend to develop a single, general- purpose generative model is a net positive change that has resulted in impressive results, it has also slowed down progress in other areas of NLP. First, we are less focused on problems that cannot be

solved with a chat- like interface. Second, the best- performing LLMs are often commercial systems, which are sometimes opaque about training data, system architecture, and training details. Third, frequent model updates hinder reproducibility.

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

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

2 LLMs: A Brief History

While modern LMs date back to 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 finetuning 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?

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). MedPALM has shown that adapting even giant LLMs to a specific domain leads to vastly increased performance (Singhal et al., 2022, 2023). Furthermore, the release of LLaMA (Touvron et al., 2023a) led quickly to Alpaca (Taori et al.,

2023) and a wave of new fine- tuned versions of LLaMA for specific tasks. This trend strongly indicates that domain- specific models, especially for constrained sizes, are still highly relevant.

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

3 The Need for Domain-Specific LLMs

In general, web data does not reflect the needs of all NLP systems. Historically, the community has developed systems for specialized domains such as finance, law, bio- medicine, and science. Accordingly, there have been efforts to build LLMs for these domains (Wu et al., 2023; Taylor et al., 2022; Singhal et al., 2022; 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.

4.1 Depth-first Evaluation

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

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

4.2 Sound Metrics

For convenience, most benchmark tasks are formulated as multiple choice question answering or classification. This is not how LLMs are often used. For much more common generation tasks, researchers have been ringing alarms about broken evaluations (Gehrmann et al., 2023). It is dubious whether we gain insights into non- task- specific generation through NLU benchmarks. If we are performing the depth- first evaluation of a generation task, a remaining hurdle - and why researchers fall back to NLU tasks - is the lack of robust metrics. While there is much recent work on

better metrics (Celikyilmaz et al., 2020; Gehrmann et al., 2023), a troubling trend is the use of LLMs as evaluators (e.g., Sellam et al., 2020; Chiang et al., 2023). This approach poses many risks, including the implicit assumption that the evaluating model has access to the ground truth judgment. While there are some promising results, using an LLM out of the box should be avoided (e.g., Wang et al., 2023a,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.

4.3 Products are not Baselines

If we really do want to evaluate 100+ tasks, there are many issues with the soundness of evaluation setups. At this scope, it is impossible to run careful ablation studies or to assess the effect of changes to methodology in a causal manner. Moreover, different LLMs respond differently to prompts. The BLOOM evaluation averaged over multiple prompts and found significant variance (Workshop et al., 2022). This variance leads to a lack of reproducibility: LLaMA (Touvron et al., 2023a) claimed high MMLU (Hendrycks et al., 2021) performance but didn't release the prompts that led to them.[1] Similarly, the evaluation scheme makes a

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

difference (Liang et al., 2022, Fig. 33). High evaluation costs mean benchmarks pick a small number of setups (sometimes only one) for each task, which introduces further bias, making it hard to construct fair benchmarks on many tasks.

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

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

5 The Role of Academics

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

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

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & Zoph, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[2] Agostinelli, F., McAleer, S., Shmakov, A., & Baldi, P. (2019). Solving the rubik's cube with deep reinforcement learning and search. Nature Machine Intelligence, 1(8), 356–363.

[3] Aji, A. F., Winata, G. I., Mahendra, R., Koto, F., Romadhony, A., & Prasojo, R. E. (2022). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 815–834).

[4] Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). Palm 2 technical report. arXiv preprint arXiv:2305.10403.

[5] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., ... & Nobles, A. L. (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] Blair-Stanek, A., Holzenberger, N., & Van Durme, B. (2023a). Can gpt-3 perform statutory reasoning? arXiv preprint arXiv:2302.06100.

[7] Blair-Stanek, A., Holzenberger, N., & Van Durme, B. (2023b). LexGPT 0.1: Pre-trained gpt-3 for legal NLP. arXiv preprint arXiv:2302.05703.

[8] Bohnet, B., Tran, V. Q., Verga, P., Aharoni, R., & Berant, J. (2022). Attributed question answering: Evaluation and modeling for attributed large language models. arXiv preprint arXiv:2212.08009.

[9] Bolton, E., Hall, D., Yasunaga, M., Lee, T., Manning, C., & Liang, P. (2023). Biomedlm: A domain-specific large language model for biomedical text. arXiv preprint arXiv:2303.00915.

[10] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[11] Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799.

[12] Chen, J. H., Himmelstein, D. S., & Zittrain, J. L. (2024). Evaluating large language models on medical exam questions. JAMA, 331(5), 431–433.

[13] Chiang, C. H., Lee, H. Y., & Chen, Y. N. (2023). Can large language models be an alternative to human evaluation? arXiv preprint arXiv:2305.01955.

[14] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240), 1–113.

[15] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2022). Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

[16] Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., ... & Vinyals, O. (2021). The benchmark lottery. arXiv preprint arXiv:2107.07002.

[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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 (pp. 4171–4186).

[18] Dhole, K. D., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., ... & et al. (2021). Nl-augmenter: A framework for task-sensitive natural language augmentation. arXiv preprint arXiv:2112.02721.

[19] Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Gardner, M. (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 (pp. 1286–1305).

[20] Eriksen, A. V., Moller, S., & Ryg, J. (2023). Use of gpt-4 to diagnose complex clinical cases. [ILLEGIBLE].

[21] Ethayarajh, K., & Jurafsky, D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4846–4853).

[22] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391–409.

[23] Garcia-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., ... & et al. (2024). Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. arXiv preprint arXiv:2404.07613.

[24] Gehrmann, S., Clark, E., & Sellam, T. (2023). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. Journal of Artificial Intelligence Research, 77, 103–166.

[25] Goyal, T., Li, J. J., & Durrett, G. (2022). News summarization and evaluation in the era of GPT-3. arXiv preprint arXiv:2209.12356.

[26] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., ... & et al. (2024). Olmo: Accelerating the science of language models. arXiv preprint arXiv:2402.00838.

[27] Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (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 (pp. 8342–8360).

[28] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). Retrieval augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929–3938).

[29] Han, X., Zhao, W., Ding, N., Liu, Z., & Sun, M. (2022). PTR: Prompt tuning with rules for text classification. AI Open, 3, 182–192.

[30] Han, T., Kumar, A., Agarwal, C., & Lakkaraju, H. (2024). Towards safe large language models for medicine. In ICML 2024 Workshop on Models of Human Feedback for AI Alignment.

[31] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In International Conference on Learning Representations.

[32] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & et al. (2022). An empirical analysis of compute-optimal large language model training. Advances in Neural Information Processing Systems, 35, 30016–30030.

[33] Hooker, S. (2021). The hardware lottery. Communications of the ACM, 64(12), 58–65.

[34] Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

[35] Ignat, O., Jin, Z., Abzaliev, A., Biester, L., Castro, S., Deng, N., ... & 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) (pp. 8050–8094).

[36] Jelinek, F. (1976). Continuous speech recognition by statistical methods. Proceedings of the IEEE, 64(4), 532–556.

[37] Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858.

[38] Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., ... & Williams, A. (2021). Dynabench: Rethinking benchmarking in NLP. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4110–4124).

[39] Kim, E.-A., Pan, H., Mudur, N., Taranto, W., Venugopalan, S., Bahri, Y., & Brenner, M. (2024). Performing Hartree-Fock many-body physics calculations with large language models. Bulletin of the American Physical Society.

[40] Lai, V., Nguyen, C., Ngo, N., Nguyen, T., Dernoncourt, F., Rossi, R., & Nguyen, T. (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 (pp. 318–327).

[41] Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., ... & Alsentzer, E. (2023). Do we still need clinical language models? In Conference on health, inference, and learning (pp. 578–597).

[42] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (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 (pp. 7871–7880).

[43] Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020b). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems.

[44] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

[45] Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., ... & et al. (2021). Few-shot learning with multilingual language models. arXiv preprint arXiv:2112.10668.

[46] Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., ... & 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.

[47] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1–35.

[48] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726–742.

[49] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[50] Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., ... & Wei, J. (2023). The FLAN collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688.

[51] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022). Biogpt: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics, 23(6), bbac409.

[52] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). Hallucination-free? assessing the reliability of leading ai legal research tools. arXiv preprint arXiv:2405.20362.

[53] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[54] Mille, S., Dhole, K. D., Mahamood, S., Perez-Beltrachini, L., Gangal, V., Kale, M. S., ... & Gehrmann, S. (2021). Automatic construction of evaluation suites for natural language generation datasets. In Advances in Neural Information Processing Systems.

[55] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

[56] Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... & et al. (2021). Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

[57] Nicholas, G., & Bhatia, A. (2023). Lost in translation: Large language models in non-english content analysis. arXiv preprint arXiv:2306.07377.

[58] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[59] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[60] Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., & Riedel, S. (2020). How context affects language models' factual predictions. In Automated Knowledge Base Construction.

[61] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., ... & Riedel, S. (2021). KILT: a benchmark for knowledge intensive language tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2523–2544).

[62] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. [MISSING].

[63] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[64] Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446.

[65] Rae, J. W., Potapenko, A., Jayakumar, S. M., & Lillicrap, T. P. (2019). Compressive transformers for long-range sequence modelling. arXiv preprint arXiv:1911.05507.

[66] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485–5551.

[67] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., ... & et al. (2020). Mastering Atari, Go, Chess and Shogi by planning with a learned model. Nature, 588(7839), 604–609.

[68] Schumacher, E., & Dredze, M. (2019). Learning unsupervised contextual representations for medical synonym discovery. JAMA Open, 2(4), 538–546.

[69] Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7881–7892).

[70] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

[71] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2022). Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

[72] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., ... & Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

[73] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.

[74] Strong, E., DiGiammarino, A., Weng, Y., Kumar, A., Hosamani, P., Hom, J., & Chen, J. H. (2023). Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Internal Medicine, 183(9), 1028–1030.

[75] Suzgun, M., Scales, N., Scharli, N., Gehrmann, S., Tay, Y., Chung, H. W., ... & et al. (2023). Challenging big-bench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13003–13051).

[76] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

[77] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., ... & Stojnic, R. (2022). Galactica: A large language model for science. arXiv preprint arXiv:2211.09085.

[78] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Lample, G. (2023a). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[79] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[80] Tricot, A., & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. Educational Psychology Review, 26, 265–283.

[81] Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Zidek, A., ... & et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature, 596(7873), 590–596.

[82] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

[83] Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., & Rajani, N. (2021). BERTology meets biology: Interpreting attention in protein language models. In International Conference on Learning Representations.

[84] Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., ... & Sui, Z. (2023a). Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926.

[85] Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., ... & et al. (2023b). How far can camels go? Exploring the state of instruction tuning on open resources. arXiv preprint arXiv:2306.04751.

[86] Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In International Conference on Learning Representations.

[87] Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., ... & Fung, P. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 815–834).

[88] Winata, G. I., Zhao, H., Das, A., Tang, W., Yao, D. D., Zhang, S.-X., & Sahu, S. (2024). Preference tuning with human feedback on language, speech, and vision tasks: A survey. arXiv preprint arXiv:2409.11564.

[89] BigScience Workshop, Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilic, S., ... & et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

[90] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

[91] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 483–498).

[92] Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., ... & Abdulnour, R.-E. E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. The Lancet Digital Health, 6(1), e12–e22.

[93] Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., ... & et al. (2021). The AI index 2021 annual report. arXiv preprint arXiv:2103.06312.

[94] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & et al. (2022). Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

[95] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K. R., & Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. arXiv preprint arXiv:2301.13848.

[96] Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., ... & Levy, O. (2023). LIMA: Less is more for alignment. arXiv preprint arXiv:2305.11206.