Pre-training data selection for biomedical domain adaptation using journal impact metrics

Mathieu Laï-king and **Patrick Paroubek**
Université Paris-Saclay, CNRS,
Laboratoire Interdisciplinaire des Sciences du Numérique,
91400, Orsay, France
{mathieu.laiking,patrick.paroubek}@lisn.upsaclay.fr

Abstract

Domain adaptation is a widely used method in natural language processing (NLP) to improve the performance of a language model within a specific domain. This method is particularly common in the biomedical domain, which sees regular publication of numerous scientific articles. PubMed, a significant corpus of text, is frequently used in the biomedical domain. The primary objective of this study is to explore whether refining a pre-training dataset using specific quality metrics for scientific papers can enhance the performance of the resulting model. To accomplish this, we employ two straightforward journal impact metrics and conduct experiments by continually pre-training BERT on various subsets of the complete PubMed training set, we then evaluate the resulting models on biomedical language understanding tasks from the BLURB benchmark. Our results show that pruning using journal impact metrics is not efficient. But we also show that pre-training using fewer abstracts (but with the same number of training steps) does not necessarily decrease the resulting model's performance.

1 Introduction

Advances in deep learning for natural language processing (NLP) in recent years have enabled transfer learning to develop (Ruder et al., 2019), particularly since the creation of *Transformers* (Vaswani et al., 2017).

One type of transfer learning aims to start with a pre-training phase where the model learns the general language structure and then a second phase where the model can be fine-tuned for a specific task. In the context of deep learning for NLP, this method avoids re-training a model from scratch for each new task, starting with a model that already has general language knowledge. These pre-trained models generally use a large corpus of text.

A specialized domain, such as finance or the biomedical domain, may contain numerous tasks. In the case of language, a specialized domain has a specific vocabulary containing terms more rarely found in general texts. We can observe this phenomenon when looking at tokens produced by a biomedical tokenizer against a general tokenizer (Boukkouri et al., 2022). Moreover, tasks may require domain-specific knowledge not found in general sources. So, to improve the performance of a model previously trained on a general domain to a specific domain, it is interesting to use a corpus specific to the domain to which we wish to adapt our model.

Most of the data used for pre-training in the biomedical field are research articles and papers that can be either abstracts, full texts, or a combination of both. This data generally originates from large public databases such as PubMed or PubMedCentral (for full-text articles). However, to our knowledge, no study has examined the selecting subsets of these large databases for pre-training using metrics specific to scientific papers. That leads us to our research question: Can a language model be adapted to the biomedical domain by efficiently selecting scientific documents in the pre-training data while maintaining or improving its performance?

This paper presents our experiments on adapting the pretrained BERT-base model to the biomedical domain. We get the PubMed January 2024 baseline corpus and define different subset configurations using journal impact metrics: h-index (Hirsch, 2005) and Scimago Journal Rank or SJR (Guerrero-Bote and Moya-Anegón, 2012). We then perform continual pre-training from the BERT-base model (Devlin et al., 2019) and evaluate it on several tasks from the BLURB benchmark (Gu et al., 2022).

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 363–369
August 16, 2024. ©2024 Association for Computational Linguistics

2 Related work

2.1 Domain-adaptive and domain-specific pre-training for the biomedical domain

The adaptation of neural models to the biomedical domain has been extensively studied in recent years, focusing on BERT-type models and, more recently, large generative language models. We distinguish two main categories regarding the pre-training data:

- *Mixed-domain pre-training*, where the model has seen data from different domains during the pre-training: it can either be a model that has been pre-trained on a general corpus and then trained on in-domain data or a model trained simultaneously on data from multiple domains, such as biomedical and clinical for example (Lee et al., 2019; Beltagy et al., 2019; Peng et al., 2019).

- *Domain-specific pre-training*, where the model only sees data from a single domain during pre-training. The hypotheses are that by using a domain-specific vocabulary, the models learn more accurate representations of specific in-domain terms (that would be divided by the sub-word tokenization with a general corpus) and that it reduces noise introduced by text completely unrelated to the domain (Beltagy et al., 2019; Boukkouri et al., 2022; Lewis et al., 2020; Gu et al., 2022).

2.2 Pre-training data quality for large language models

Several works focus on selecting sequences using quality metrics for pre-training Transformer models in the general domain, particularly with the advent of large language models and the evolution of the size of pre-training datasets for these models (Zhou et al., 2023; Attendu and Corbeil, 2023; Marion et al., 2023; Das and Khetan, 2023).

The adaptation of large language models using scientific articles has been largely studied. However, only a few have emphasized the quality of scientific articles used. For the Galactica model (Taylor et al., 2022), they only mention applying *"several quality filters, including excluding papers from journals with certain keywords and also excluding papers with a low journal impact factor"*. Most other models that used PubMed or PubMed-Central for pre-training do not mention any specific selection of data at the document level; most focus on preprocessing steps at the content level (bibliography references, authors, figures and tables, etc.) when dealing with full-text articles (Luo et al., 2022; Wu et al., 2023; Luo et al., 2023; Chen et al., 2023).

3 Methods

We use the same methodology as Marion et al. (2023), with some small modifications :

Let D be a large dataset containing documents and ξ a metric assigning a score to a document. We build a subset $P_{c\xi}$ by adding instances that fit our selection criteria c :

$$P_{c\xi} = \{d_i \in D | c_{0\xi} \leq \xi(d_i)) \leq c_{1\xi}\} \qquad (1)$$

Where $c_{0\xi}$ and $c_{1\xi}$ are the lower and upper bound for the criteria c and the metric ξ. For each metric, we consider two selection criteria: keeping top or middle percentiles[1] of D as the data to be kept. This serves as verifying if the model learns better with high quality documents (defined by the metric, for our metrics, higher is better). We keep either 25% or 50% of the documents in D. So for instance, if we take the 25 $\%$ middle for the metric ξ, we should compute the 37.5 $\%$ and 62.5 $\%$ percentiles with respect to metric ξ, which corresponds to $c_{0\xi}$ and $c_{1\xi}$, and keep the documents between these two percentiles.

Then, we tokenize each document in the subset, and we concatenate them into sequences of length equal to the model's context length. This differs from Marion et al. (2023) as we do the filtering before tokenization (because our metrics are applied on a document, not on a sequence of tokens). These sequences are then used to pre-train a model. The goal is then to pre-train a model on a subset of the whole training set while retaining or improving the model's performance.

3.1 Pre-training corpus

We use the PubMed Baseline corpus comprising all article abstracts deposited on the PubMed database until January 2024. Using PubMed metadata, we filter out abstracts that are not in English, abstracts whose text is not available, and abstracts whose ISSN journal identifier is not present (we filter this to have enough abstracts with a score as our pruning metrics are based on journal impact). After

[1] we do not use the bottom percentiles because in our case, for the SJR metric, more than 25% of the dataset had the same value : 0

filtering, the total corpus is comprised of 15.9B tokens.

We did not perform a pre-training experiment using the non-filtered PubMed set because we did not have enough articles with journal identifiers to obtain convenient metric percentiles. Still, we expect this filtering to already impact the overall quality of the corpus.

3.2 Quality metrics

The nature of the datasets used for general model training (by which we mean models that are not domain-specific) differs from those used in the biomedical field. They are generally huge datasets comprising texts extracted from the Internet on various sites. In our case, these are research articles from the same database. This presupposes a text quality that is adequate in certain respects (generally correct syntax and formal language, unlike texts found on the Internet).

We wanted to use metrics specific to scientific articles that have meaning for scientific article readers. So, we decided to use journal impact metrics. We used the metadata available on PubMed. This type of metric can provide insight into the probable impact that a paper can have but does not necessarily ensure scientific quality. However, we believe filtering with impact metrics in a large corpus can help reduce the noise, help the model learn biomedical language, and learn biomedical knowledge more efficiently. We use the h-index (Hirsch, 2005) and the SJR (Guerrero-Bote and Moya-Anegón, 2012) as the data is publicly available on the Scimago website[2]. For comparison, we also perform a random score assignation on all papers from the dataset; we do not perform multiple random assignations to limit the compute cost.

We computed the percentiles for SJR and h-index and, as there were zero values for the SJR index (for the 12.5% and 25% percentiles), we did not perform all the pre-trainings for the *mid* criteria, we only considered the *25 %* subset. This is also why we did not consider the bottom percentiles. We also did not perform the pre-training on the complete set because of time and resource constraints, but we plan to do it in future work.

3.3 Pre-processing

We define

We tokenize the whole dataset and concatenate the text of the different abstracts into sequences of length 512 tokens (maximum sequence length for the model we use: BERT (Devlin et al., 2019)). . We keep 5 % of this set as validation data.

3.4 Model and pre-training

We use the original *BERT-base* model (Devlin et al., 2019), continue pre-training on the defined datasets with masked language modeling, and compare the resulting models. For each pre-training (on each subset), we fix a shared global number of steps so that each model sees the same quantity of tokens: we select the number of steps as the total number needed for one epoch on the entire PubMed corpus. For the runs with the subsets, the model will run multiple epochs until it reaches the total number of steps, with data shuffling between epochs (for example, two epochs for the run where we take the top 50% of PubMed abstracts with respect to h-index).

We train with a sequence length of 512 and a batch size of 8192[3], which gives us a total of 3598 steps. We use a linear schedule with 10 % warmup and a peak learning rate of $1e-4$. For the other hyperparameters, we follow the original BERT paper. We train our different models on 2 NVIDIA A100 GPUs.

3.5 Evaluation and fine-tuning

We evaluate the produced pre-trained models on some of the datasets from the BLURB benchmark (Gu et al., 2022). We also re-evaluate the BERT-based model to ensure a consistent evaluation with our fine-tuning scripts. We excluded the PICO and Sentence Similarity tasks (EBM-PICO (Nye et al., 2018) and BIOSSES (Soğancıoğlu et al., 2017)), for which we had trouble reproducing similar and consistent results across runs to those obtained in the BLURB paper, as they did not share any code to perform the fine-tuning and evaluation. So, we are left with the following evaluation tasks :

- Named entity recognition (NER) : BC5-chem & BC5-disease (Li et al., 2016), BC2GM (Smith et al., 2008), JNLPBA (Collier and Kim, 2004) and NCBI-disease (Doğan et al., 2014). We evaluate the models for NER tasks using the *entity-level F1 score*. We model the entities using BIO tags.

[2] https://www.scimagojr.com/journalrank.php

[3] We perform gradient accumulation and data parallelism to get this batch size.

	base	random		h-index				sjr		
				mid		top		mid	top	
	0%	25%	50%	25%	50%	25%	50%	25%	25%	50%
BC5-chem	87.31	90.03	**90.24**	89.40	89.93	89.51	89.52	89.72	89.61	89.89
BC5-disease	77.09	**81.09**	80.72	81.05	80.73	80.38	80.68	81.00	80.76	80.60
BC2GM	75.32	79.17	79.01	79.51	79.51	79.52	79.41	78.74	79.01	**79.87**
JNLPBA	76.77	78.02	77.85	77.51	77.95	78.13	**78.41**	78.13	78.28	77.90
NCBI-disease	81.59	84.89	84.45	**85.09**	84.84	84.63	84.71	84.97	84.30	84.98
HoC	79.22	84.41	84.74	84.72	84.56	84.83	84.71	84.54	**85.07**	84.76
ChemProt	77.07	79.25	78.83	78.94	79.72	79.00	**79.92**	78.77	79.62	78.96
DDI	**89.11**	87.54	87.70	87.91	88.27	86.46	86.80	87.05	85.92	87.76
GAD	76.82	78.09	78.24	78.31	77.38	77.34	**78.39**	77.42	78.35	77.00
BioASQ	72.19	75.93	75.63	75.63	75.24	74.84	**76.07**	75.85	75.50	75.22
PubMed QA	**55.24**	55.20	55.20	55.16	55.12	54.78	55.16	55.20	55.22	55.20
Micro avg.	77.07	79.42	79.33	79.38	79.39	79.04	**79.43**	79.22	79.24	79.29
Macro avg.	75.89	78.56	78.55	78.59	78.53	78.25	**78.64**	78.41	78.53	78.46

Table 1: Comparison of the performance of our pretrained models on the different evaluation tasks from the BLURB benchmark (Gu et al., 2022). *'base'* model is the BERT$_{BASE}$ model (Devlin et al., 2019) from which we continue the pre-training. For the macro average, we average the datasets from the same task and then average the performance on each task. For each task or average, the **best performance is in bold** and the second best performance is underlined.

- Relation extraction : ChemProt(M. et al., 2017), DDI (Herrero-Zazo et al., 2013), GAD (Bravo et al., 2015). We evaluate the models for relation extraction using the *micro F1 score*. We use entity dummyfication with start and end tags and use the [CLS] token to classify relations.

- Document classification : HoC (Baker et al., 2016), for which we measure the *micro F1 score*.

- Question answering : PubMedQA (Jin et al., 2019) and BioASQ Task 7b (Nentidis et al., 2020). We evaluate these tasks using *accuracy*.

4 Results and Discussion

To limit random effects, we perform the fine-tuning multiple times with different random seeds, as described in the BLURB paper: using five seeds for all datasets except for BioASQ and PubMedQA, for which we use ten seeds (because they are smaller in size). We then report the average performance across the different seeds for each dataset in the table 1.

4.1 Improvement against non biomedical model

All models trained on biomedical data perform better than the base model trained only on general-domain data. However, for a fair comparison, we should train it for the same amount of steps on non-biomedical data.

4.2 Are journal impact metrics important for the model ?

We obtain the best results in micro and macro averages for the model trained on the top 50% of the entire set with respect to the h-index of the journal in which abstracts have been published. Overall, the h-index metric performs better than SJR, which may be because the SJR percentile values are very close to each other, so the quality differences are less important.

However, the performance differences are low when we compare to the SJR metric or even when selecting abstracts randomly, regardless of the proportion of abstracts we keep. So, journal impact metrics do not seem important when selecting pretraining data from a corpus of scientific articles. We then should find more appropriate metrics to define the quality of a single abstract or test it on a full-text article corpus (so that the impact of a single document is higher).

4.3 Is it better to pre-train a model using more abstracts ?

If we compare the performance difference when training with 25% of the data against 50%, we globally have better performances (except for the random selection), but these differences are not significant. So, it would be interesting to perform further pre-training experiments using different subset sizes to investigate which number of documents is optimal for the domain adaptation.

5 Conclusion

This paper presents our early experiments on selecting the pre-training data for the biomedical domain. We show that the journal impact metrics are not better than the random selection at a fixed number of training steps. We also observe that reducing the number of abstracts in the training set does not necessarily decrease the final model performance and show the need to investigate how many documents we need to pre-train a model without losing performance.

Further directions include finding better metrics (or combinations of metrics) to assess the quality of a document in the pre-training corpus, investigating metrics at a different level (at the corpus level using various mixtures of biomedical domains), and using a corpus of full-text articles.

6 Acknowledgments

This project was provided with computer and storage resources by GENCI at IDRIS thanks to the grant 20XX-AD011014707 on the supercomputer Jean-Zay's A100 partition .

References

Jean-michel Attendu and Jean-philippe Corbeil. 2023. NLU on Data Diets: Dynamic Data Subset Selection for NLP Classification Tasks. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 129–146, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? Comparing pre-training strategies of BERT in the medical domain. In *LREC 2022 - Language Resources and Evaluation Conference*, page 2626.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinformatics*, 16(1):55.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint*.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

Devleena Das and Vivek Khetan. 2023. DEFT: Data Efficient Fine-Tuning for Large Language Models via Unsupervised Core-Set Selection. *arXiv preprint*.

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.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Vicente P. Guerrero-Bote and Félix Moya-Anegón. 2012. A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, 6(4):674–688.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

J. E. Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016:baw068.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Shenmin Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *undefined*.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. *arXiv preprint*.

Krallinger M., Rabal O., and Lourenço A. 2017. Overview of the biocreative vi chemical-protein interaction track. *Proceedings of the BioCreative VI Workshop,*, 141-146.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale. *arXiv preprint*.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. Results of the seventh edition of the BioASQ Challenge. volume 1168, pages 553–568.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Larry Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W. John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(2):S2.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. *arXiv preprint*.

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*, volume 30. Curran Associates, Inc.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers. *arXiv preprint*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. *arXiv preprint*.