Dataset Debt in Biomedical Language Modeling

Jason Alan Fries[*1,2] **Natasha Seelam**[*3] **Gabriel Altay**[*4] **Leon Weber**[*5,12]

Myungsun Kang[*6] **Debajyoti Datta**[*7] **Ruisi Su**[*8] **Samuele Garda**[*5]

Bo Wang[9] **Simon Ott**[10] **Matthias Samwald**[10] **Wojciech Kusa**[11]

[1] Stanford University [2] Snorkel AI [3] Sherlock Biosciences [4] Tempus Labs, Inc.
[5] Humboldt-Universität zu Berlin [6] Immuneering Corporation [7] University of Virginia
[8] Sway AI [9] Massachusetts General Hospital [10] Medical University of Vienna [11] TU Wien
[12] Max Delbrück Center for Molecular Medicine [*] Equal Contribution

Abstract

Large-scale language modeling and natural language prompting have demonstrated exciting capabilities for few and zero shot learning in NLP. However, translating these successes to specialized domains such as biomedicine remains challenging, due in part to biomedical NLP's significant *dataset debt* – the technical costs associated with data that are not consistently documented or easily incorporated into popular machine learning frameworks at scale. To assess this debt, we crowdsourced curation of datasheets for 167 biomedical datasets. We find that only 13% of datasets are available via programmatic access and 30% lack any documentation on licensing and permitted reuse. Our dataset catalog is available at: https://tinyurl.com/bigbio22.

1 Introduction

Natural language prompting has recently demonstrated significant benefits for language model pretraining, including unifying task inputs for large-scale multi-task supervision (Raffel et al., 2019) and improving zero-shot classification via explicit, multi-task prompted training data (Wei et al., 2022; Sanh et al., 2022). With performance gains reported when scaling to thousands of prompted training tasks (Xu et al., 2022), tools that enable large-scale integration of expert-labeled datasets hold great promise for improving zero-shot learning.

However, translating these successes to specialized domains such as biomedicine face strong headwinds due in part to the current state of dataset accessibility in biomedical NLP. Recently *data cascades* was proposed as a term-of-art for the costs of undervaluing data in machine learning (Sambasivan et al., 2021). We propose a similar term, *dataset debt*, to capture the technical costs (Sculley et al., 2015) of using datasets which are largely open and findable, but inconsistently documented, structured, and otherwise inaccessible via a consistent, programmatic interface. This type of debt creates significant practical challenges when integrating complex domain-specific corpora into popular machine learning frameworks.

We claim that biomedical NLP suffers from significant dataset debt. For example, while HuggingFace's popular Datasets library (Lhoest et al., 2021) contains over 3,000 datasets, biomedical data are underrepresented and favor tasks with general domain appeal such as question answering or semantic similarity (PubmedQA, SciTail, BIOSSES). To assess the state of biomedical dataset debt, we built, to our knowledge, the largest catalog of metadata for publicly available biomedical datasets. We document provenance, licensing, and other key attributes per (Gebru et al., 2021) to help guide future efforts for improving dataset access and machine learning reproducibility.

Our effort found low overall support for programmatic access, with only 13% (22/167) of our datasets present in the Datasets hub. Despite a proliferation of schemas designed to standardize dataset loading and harmonize task semantics. there remains no consistent, API interface for easily incorporating biomedical data into language model training at scale.

2 Data-Centric Machine Learning

Deep learning models are increasingly moving to commodified architectures. *Data-centric machine learning* (vs. model-centric) is inspired by the observation that the performance gains provided by novel architectures are often smaller than gains obtained using better training data. We outline some key challenges and opportunities in data-centric language modeling. These are broadly applicable to NLP, but have strong relevance to biomedicine

Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, pages 137 - 145

and the current state of dataset debt.

2.1 Curating and Cleaning Training Data

Popular language models such as GPT-3 (Brown et al., 2020) do not incorporate scientific or medical corpora in their training mixture, contributing to their lower performance when used in biomedical domains and few-shot tasks (Moradi et al., 2021). Additionally, simply training the language model on in-domain data might lead to non-trivial risks associated with the recapitulated biases from the training corpora (Zhang et al., 2020; Gururangan et al., 2022).

In scientific literature, discounting source provenance could manifest as language models parroting conflicting or inaccurate scientific findings. Zhao et al.(Zhao et al., 2022) curated scientific corpora to identify patient-specific information (e.g., mining PubMed Central to identify case reports that respect licensing for re-use and re-distribution). With sufficient metadata and dataset provenance, this level of curation could be extended to the entire training corpus for a biomedical language model.

Data cleaning has a large impact on language model performance. Deduplicating data leads to more accurate, more generalizable models requiring fewer training steps (Cohen et al., 2013; Lee et al., 2021). Cleaning up the consistency of answer response strings was reported to improve biomedical question answering (Yoon et al., 2021). Duplication contamination is a serious risk in biomedical datasets, which often iteratively build or extend prior annotations, introducing risk of test leakage in evaluation (Elangovan et al., 2021).

2.2 Programmatic Labeling

Biomedical domains require specialized knowledge, making expert-labeled datasets time-consuming and expensive to generate. In limited-data settings, distant and weakly supervised methods (Craven and Kumlien, 1999) are often used to combine curated, structured resources (e.g., knowledge bases, ontologies) with expert rules to programmatically label data. These approaches have demonstrated success across NER, relation extraction, and other biomedical applications (Kuleshov et al., 2019; Fries et al., 2021). However these approaches typically are applied to real, albeit unlabeled data, creating challenges when modeling rare classes. A recent trend is transforming structured resources directly into realistic-looking, but synthetic training examples. KELM (Agarwal et al., 2021) converts Wiki knowledge graph triplets into synthesized natural language text for language model pretraining.

Natural language prompting has emerged as a powerful technique for zero/few shot learning, where task guidance from prompts reduces sample complexity (Le Scao and Rush, 2021). Cross-lingual prompting (English prompts, non-English examples) has demonstrated competitive classification performance (Lin et al., 2021). Training language models directly on prompts has resulted in large gains in zero-shot performance over GPT-3 as well as producing models with fewer trained parameters (Sanh et al., 2022; Wei et al., 2022).

PromptSource (Bach et al., 2022) is a recent software platform for creating prompts and applying them to existing labeled datasets to build training data. These developments highlight a promising trend toward defining programmatic transformations on top of existing datasets, enabling them to be configured into new tasks. However, leveraging large-scale prompting remains challenging in biomedicine due to the lack of programmatic access to a large, diverse collections of biomedical datasets and tasks.

2.3 Diverse Evaluation and Benchmarking

Inspired by standardized benchmarks in general domain NLP research (Wang et al., 2018, 2019), BioNLP takes similar initiatives by establishing a benchmark of 10 datasets spanning 5 tasks (Peng et al., 2019, BLUE), an improved benchmark on BLUE with 13 datasets in 6 tasks (Gu et al., 2022, BLURB), and a benchmark of 9 different tasks for Chinese biomedical NLP (Zhang et al., 2021, CBLUE). While these benchmarks provide tools for consistent evaluation, only BLURB supports a leaderboard and none directly provide dataset access. Evaluation frameworks that provide programmatic access are often restricted to single and well-established tasks and impose pre-processing choices that can make inconsistent performance comparisons (Crichton et al., 2017; Weber et al., 2021).

To the best of our knowledge, there are currently no zero-shot evaluation frameworks for biomedical data similar to BIG-Bench[1], which currently contains little-to-no biomedical tasks.

Evaluation frameworks must also allow probing the trained language models' intrinsic properties,

[1]https://github.com/google/BIG-bench

rather than only measure downstream classification performance. Following (Petroni et al., 2019) in the general NLP domain, (Sung et al., 2021) introduce BioLAMA, a benchmark making available 49K biomedical knowledge triplets to probe the relational knowledge present in pre-trained language models.

3 Datasets Summary

3.1 Metadata/Datasheet Curation

Our inclusion criteria targeted expert-annotated datasets designated as public, reusable research benchmarks for one or more NLP tasks. We excluded: (1) multimodal datasets where removing the non-text modality undermines the task, e.g., visual question answering, audio transcription, image-to-text generation; (2) general resource datasets, e.g, the PMC Open Access Subset, MIMIC-III (Johnson et al., 2016); (3) derived resources, e.g., knowledge bases constructed via text mining; and (4) modeling artifacts, e.g., static embeddings or pretrained language models.

We recruited 8 volunteers to identify datasets and crowdsource their metadata curation for an open, community dataset catalog. Participants reviewed dataset publications and websites which described the curation process, and then completed the metadata schema outlined in Table 1 This schema loosely assesses compliance with FAIR data principles (Wilkinson et al., 2016).

Our initial effort identified 101 datasets. We combined this list with a contemporaneously curated catalog of biomedical datasets, identified via systematic literature review (Blagec et al., 2022). Since the catalog described in Blagec et al. (2022) was generated using broader inclusion criteria (e.g., non-public data, imaging and video datasets) we identified 104/475 entries that met our criteria. After merging, we conducted a second round of crowdsourcing to annotate metadata, resulting in our current catalog of 167 biomedical datasets. We did not conduct a formal assessment of inter-annotator agreement.

Field	Description
Name	Dataset name
Task Types	NER, NED, QA, NLI, coreference resolution, etc.
Domain	Corpora domain: biomedical or clinical/health
File Format	BioC, JSON, etc.
Annotations	Expert label provenance
API Access	Available via HuggingFace Datasets?
Splits	Canonical definitions for training/validation/testing splits
License	Provided license type
Languages	Included languages
Multilingual	Parallel corpora
Publication	Manuscript describing dataset
Year	Publication year
Citations	Google Scholar counts
Homepage	Website describing dataset
Public URL	Open URL (no authentication)
Dead Link	Dataset no longer accessible

Table 1: Metadata collected for all biomedical datasets. See Appendix A for more details on each category.

4 Results

4.1 Dataset Access

Only 22/167 (13%) of biomedical datasets are available via the Datasets API, despite 123/167 (74%) being openly hosted on public websites. The remaining datasets require authentication to access (21%) or were dead links (5%).

Format	Name	Count	Total
Structured	BioC	5	3%
Structured	BRAT	16	10%
Structured	CoNLL	11	7%
Structured	PubTator	4	2%
Semi-structured	XML	26	16%
Semi-structured	JSON	43	26%
Semi-structured	TSV/CSV	15	9%
Semi-structured	TMX	1	1%
Plain Text	Standoff	13	8%
Plain Text	Text	25	15%
Plain Text	ARFF	1	1%
Binary	Word	1	1%
Binary	Excel	2	1%
Unknown	Unknown	4	2%

Table 2: Distribution of file formats for biomedical datasets.

Table 2 outlines the diversity of commonly used biomedical file formats. Most datasets are provided in semi-structured form (51%), followed by structured (22%), and non-standard plain text files

(17%). There are several structured formats which propose a data model for parsing and standardizing task semantics (e.g., BRAT (Stenetorp et al., 2012), BioC (Comeau et al., 2013)). However, for information extraction tasks which could use these formats, only 31/86 (36%) actually do.

Table 2 outlines dataset licensing, broken down into six categories, largely based on commercial vs. non-commercial restrictions. These cover broad classes of licensing, ranging from permissive Creative Commons Share-Alike licenses to dataset-specific data-use agreements (DUA). Nearly 30% of datasets are publicly available online yet do not include any licensing information. A further 16.8% have DUA requirements, but include unclear language on what restrictions are placed on dataset usage.

License	Restrictions	Count	Percent
Public	C/NC	56	33.5%
Public	NC	13	7.8%
DUA	C/NC	12	7.2%
DUA	NC	8	4.8%
DUA	?	28	16.8%
Unknown	?	50	29.9%

Table 3: Dataset licenses. Restrictions are commercial (C), non-commercial (NC) and unknown (?).

4.2 Dataset and Task Diversity

Biomedical datasets (i.e., tasks built from scientific publications) made up 68% of available datasets while clinical datasets (patient notes, health news, clinical trial reports) made up 32%.

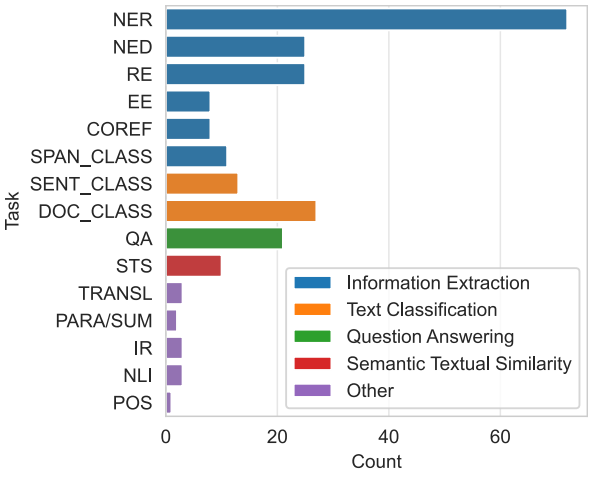

Figure 1: All NLP tasks, broken down into 5 categories (see legend). Note datasets often support multiple tasks.

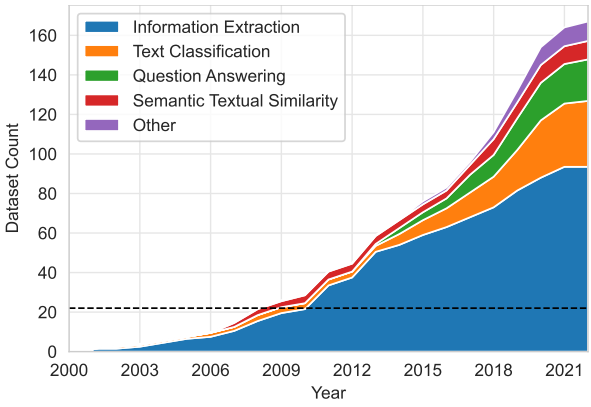

Figure 2: Cumulative count of datasets by task, ordered by year of dataset release. The black dashed line indicates the total number available via the Datasets API.

Fig.2 shows the overall homogeneity of public biomedical datasets as of 2022. Information extraction tasks (e.g., NER, NED, releation extraction, coreference resolution) comprise 56%, followed by 20% text classification (e.g, document labeling, sentiment analysis), 13% question answering, and 6% semantic similarity.

Task Category	Eng.	Non-Eng.
Information Extraction	128	34
Text Classification	33	10
Question Answering	21	0
Semantic Textual Similarity	10	0
Other	12	6

Table 4: Task category counts by English (Eng.) and Non-English (Non-Eng.) languages.

Given all tasks, 14 languages are covered. Five languages make up 95% of all datasets. English is the majority (80%), followed by Spanish (7.5%), German (2.4%), French (2.4%), and Chinese (2.4%). Table 4 contains counts of task categories binned into English and Non-English . Question answering and semantic similarity have zero non-English datasets.

5 Conclusion

In this work, we outlined several challenges in training biomedical language models. With increasingly large biomedical language models (Yang et al., 2022), limitations in the quality and properties of training data grow more stark. We argue that biomedical NLP suffers from significant dataset debt, with only 13% of datasets accessible via API

access and readily usable in state-of-the-art NLP tools. Current biomedical datasets are homogeneous, largely focusing on NER and relation extraction tasks, and predominantly English language. These limitations highlight opportunities presented by recent data-centric machine learning methods such as prompting, which enables experts to inject task guidance into training and more easily reconfigure existing datasets into new training tasks.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. 2022. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *arXiv preprint arXiv:2201.07040*.

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.

Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2013. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC bioinformatics*, 14(1):1–15.

Donald C Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, et al. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, August 6-10, 1999, Heidelberg, Germany*, pages 77–86. AAAI.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):1–14.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1):1–11.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Thorsten Gruber. 2014. Academic sell-out: how an obsession with metrics and rankings is damaging academia. *Journal of Marketing for Higher Education*, 24(2):165–177.

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 Trans. Comput. Heal.*, 3(1):2:1–2:23.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *CoRR*, abs/2201.10474.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Volodymyr Kuleshov, Jialin Ding, Christopher Vo, Braden Hancock, Alexander Ratner, Yang Li, Christopher Ré, Serafim Batzoglou, and Michael Snyder. 2019. A machine-compiled database of genome-wide association studies. *Nature communications*, 10(1):1–8.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*.

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. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M.

Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*.

Xi Yang, Nima Pour Nejatian, Hoo Chang Shin, Kaleb Smith, Christopher Parisien, Colin Compas, Mona Flores, Ying Zhang, Tanja Magoc, Christopher Harle, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *medRxiv*.

Wonjin Yoon, Jaehyo Yoo, Sumin Seo, Mujeen Sung, Minbyul Jeong, Gangwoo Kim, and Jaewoo Kang. 2021. Ku-dmis at bioasq 9: Data-centric and model-centric approaches for biomedical question answering. In *CEUR Workshop Proceedings*, volume 2936, pages 351–359. CEUR-WS.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Lei Li, Xiang Chen, Shumin Deng, Luoqiu Li, Xin Xie, Hongbin Ye, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Mosha Chen, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Huajun Chen, Buzhou Tang, and Qingcai Chen. 2021. CBLUE: A chinese biomedical language understanding evaluation benchmark. *CoRR*, abs/2106.08087.

Zhengyun Zhao, Qiao Jin, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient notes and relations extracted from case reports in pubmed central. *arXiv preprint arXiv:2202.13876*.

A Appendix

A.1 Metadata Overview

This section contains detailed descriptions of each metadata field collected for the dataset catalog.

A.1.1 Name

The dataset name, preferring short forms (BC5CDR) as typically used on homepages or scientific publications over verbose ones (“BioCreative 5 Chemical Disease Relation Task").

A.1.2 Task Types

Datasets contain labels for one or more tasks. Tables 5 and 6 outline the tasks we consider in this work.

Name	Abbreviation
Named Entity Recognition	NER
Named Entity Disambiguation	NED
Relation Extraction	RE
Event Extraction	EE
Coreference Resolution	COREF
Span Classification	SPAN
Document Classification	DOC
Sentence Classification	SENT
Semantic Textual Similarity	STS
Question Answering	QA
Translation	TRANSL
Paraphrasing	PARA
Summarization	SUM
Natural Language Inference	NLI
Part-of-Speech Tagging	POS
Information Retreival	IR

Table 5: All task types.

A.1.3 Domain

Source domain of the dataset.

- *Biomedical*: Tasks are defined for scientific literature (e.g., PubMed abstacts, full-text publications from the PMC Open Access Subset).
- *Clinical*: Tasks are defined for clinical notes from patient electronic health records, health-related questions from social media or news websites, clinical trial reports, etc.

A.1.4 File format

File formats provided by the original dataset creators.

Category	Abbreviation
Information Extraction	NER
Information Extraction	NED
Information Extraction	RE
Information Extraction	EE
Information Extraction	COREF
Information Extraction	SPAN
Text Classification	DOC
Text Classification	SENT
Semantic Textual Similarity	STS
Question Answering	QA
Other	TRANSL
Other	PARA
Other	SUM
Other	NLI
Other	POS
Other	IR

Table 6: Task categories.

A.1.5 Annotations

Provenance of labels used to create a dataset.

- *Manual*: Expert annotators directly label data instances. This may include multiple rounds of adjudication.
- *Model-assisted Manual*: Experts verify, correct, or augment the output of a model (e.g., pre-annotated entities are used by annotators to define relations).
- *Crowdsourced*: Labels are the result of a voting process over multiple annotator's labels.
- *Rules*: Heuristics developed by experts and applied to unlabeled text to create annotations. This includes a wide range of weak/distant supervision techniques.
- *Found*: Generated from "in-the-wild" data, such as aligned pairs of translated text mined from web pages.
- *Unlabeled*: no human-generated labels (e.g., the PMC Open Subset).

A.1.6 API Access

URL of HuggingFace's Datasets implementation, otherwise "no".

A.1.7 Splits

Are canonical train, validation, and test sets defined by the dataset creators? If so, which sets are provided. `value` ∈ { `NONE, train, valid, test` }.

A.1.8 License

License information accompanying the dataset. Unknown licenses means the annotator could not find any information or formal legal documents on the homepage, software repository (e.g, GitHub, Google Code), or README with the data itself.

- *Public*: Creative Commons (CC BY 3.0/4.0, CC BY-SA 3.0/4.0), Public Domain, GNU Free Documentation License, GNU Common Public License v3.0, MIT License, Apache License 2.0
- *Public Non-commercial*: Creative Commons (CC BY NC 2.0/3.0/4.0, CC BY-NC-SA 4.0), CSIRO Data License (Non-commercial), Public for Research
- *DUA-NC*: DUA for non-commercial use only.
- *DUA-C/NC*: DUA for commercial and non-commercial uses.
- *DUA-UNK*: DUA with unknown restrictions.
- *Unknown*: Public-Unknown, Public w/ Registration

A.1.9 Languages

Languages used in the labeled dataset.

A.1.10 Multilingual

Dataset contains aligned pairs for two or more languages.

A.1.11 Publication, Year

URL to the manuscript, DOI, and year of publication.

A.1.12 Citations

Current citation count from Google Scholar, as of 02-22-2022. This measure was collected to provide a weak measure of dataset visibility. We note that citation count is a problematic measure of valuation and subject to many criticisms (Gruber, 2014).

A.1.13 Homepage, Public URL

URL of website describing and hosting the dataset. If the dataset has a direct download link, denote if it is public or only available after authentication.

A.1.14 Dead Link

URL of dataset homepage, as documented in the source publication, is no longer active.

A.2 Domain-specific

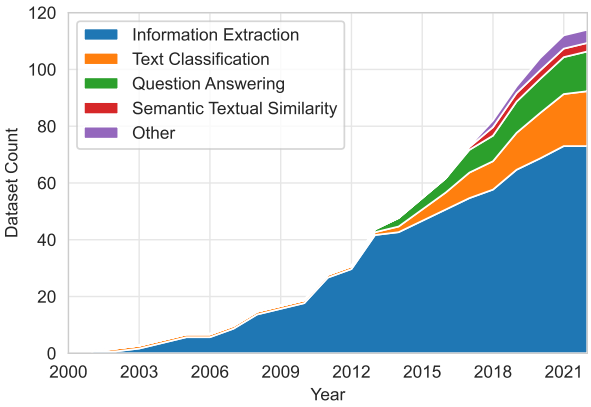

Figure 3: Scientific/biomedical domain (e.g., PubMed abstracts) cumulative distribution of available tasks, ordered by year of dataset release.

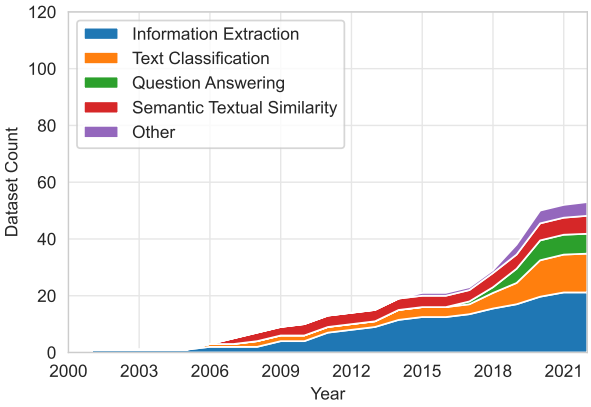

Figure 4: Clinical domain (e.g., patient notes) cumulative distribution of available tasks, ordered by year of dataset release.

A.3 Languages

Task	English	Non-English
NER	60	18
NED	21	9
RE	22	3
EE	8	0
COREF	8	0
SPAN_CLASS	9	4
SENT_CLASS	12	2
DOC_CLASS	21	8
QA	21	0
STS	10	0
TRANSL	3	5
PARA/SUM	2	0
IR	3	0
NLI	3	0
POS	1	1

Table 7: Tasks by language