The Data Provenance Project

Shayne Longpre [1] **Robert Mahari** [1] **Niklas Muennighoff** [2] **Anthony Chen** [3] **Kartik Perisetla** [4]
William Brannon [1] **Jad Kabbara** [1] **Luis Villa** [5] **Sara Hooker** [6]

Abstract

A wave of recent language models have been powered by large collections of natural language datasets. The sudden race to train models on these disparate collections of incorrectly, ambiguously, or under-documented datasets has left practitioners unsure of the legal and qualitative characteristics of the models they train. To remedy this crisis in data transparency and understanding, in a joint effort between experts in machine learning and the law, we've compiled the most detailed and reliable metadata available for data licenses, sources, and provenance, as well as fine-grained characteristics like language, text domains, topics, usage, collection time, and task compositions. Beginning with nearly 40 popular instruction (or "alignment") tuning collections, we release a suite of open source tools for downloading, filtering, and examining this training data. Our analysis sheds light on the fractured state of data transparency, particularly with data licensing, and we hope our tools will empower more informed and responsible data-centric development of future language models.

1. Introduction

The latest wave of language models, both public (Chung et al., 2022; Taori et al., 2023; Geng et al., 2023) and proprietary (including Bard Anil et al., 2023, ChatGPT Ouyang et al., 2022, GPT-4 OpenAI, 2023), are capable of a range of general reasoning abilities (Wei et al., 2022). This is attributed in large part to the diversity and richness of their training data, including pre-training corpora, and finetuning datasets paired with instructions (Wei et al.; Sanh et al., 2021) or human feedback (Ouyang et al., 2022). Natural language training data is comprised of hundreds of data sources, both for pre-training (Gao et al., 2020), and for instruction tuning, as compiled by academics (Wang et al., 2022b; Longpre et al., 2023a; Muennighoff et al., 2022), synthetically generated by models (Taori et al., 2023; Wang et al., 2022a), or aggregated by platforms like Hugging Face (Lhoest et al., 2021).

A Crisis in Data Transparency A central challenge to model developers is crowd-sourcing quality collections of data, with reliable information on their contents and limitations (Longpre et al., 2023b). Recent trends have seen massive collections with sparser documentation and attribution (Wang et al., 2022c), a lack of Dataset Cards or Datasheets (Gebru et al., 2021; Pushkarna et al., 2022), even non-disclosure of training sources (OpenAI, 2023; Anil et al., 2023), and ultimately a decline in understanding the raw training data mixtures (Dodge et al., 2021). This lack of understanding can lead to data leakages between training and test data, or exposing personally identifiable information (PII) (Bubeck et al., 2023), poor quality models than anticipated, and unintended biases or behaviours (Welbl et al., 2021; Xu et al., 2021). The Data Provenance effort aims to remedy this deterioration in data documentation by compiling relevant metadata for thousands of popular datasets (a large undertaking), and expanding their metadata with a much richer taxonomy than Hugging Face, Papers with Code, or other aggregators. To empower better data documentation and understanding, we provide tools to (a) filter, download, and explore data collections, but also (b) auto-generate a Data Provenance Card, as a supplement to Datasheets (Gebru et al., 2021) for sections on data sourcing and composition.

Unreliable Data Provenance & Licensing. Our annotation collection process revealed systemic problems particularly data licenses, where existing aggregations are a mix of sparse, ambiguous, and incorrect. For instance, 513 of Super-Natural Instruction's 1556 tasks having "Unknown" licenses (Wang et al., 2022b), leaving a substantial information gap. As a result, much of this data is unusable for risk-averse or non-academic practitioners. Second, the annotations that are collected are often incorrect—we randomly sampled 84 datasets with HuggingFace URLs, and found 49% were incorrectly labeled, usually as more per-

*Equal contribution [1]MIT [2]Hugging Face [3]University of California, Irvine [4]Apple [5]Tidelift [6]Cohere For AI. Correspondence to: Shayne Longpre <slongpre@media.mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

missive. Third, ambiguously licensed datasets, such as ShareGPT (Vercel, 2023) and Alpaca (Taori et al., 2023), which leverage generations from proprietary models, have been commonly adopted, with developers only realizing the ramifications after their models are fully trained and even deployed. These informational gaps, ambiguities, and license mis-transcriptions have led to misapprehensions (eg. LMSYS-Org (2023) claiming models are "compatible with commercial usage" despite Flan-T5 finetuning on datasets licensed as Non-Commercial), license revisions post-public release (with MPT-StoryTeller (Frankle, 2023)), and even lawsuits (*eg.* Stability AI (Arstechnica, 2023)).

Uncertainty in legal interpretations (Section 3) means practitioners rely on many signals beyond licenses to decide their risk tolerance in using a training dataset. Discussion with practitioners revealed these signals include the dataset creators, the original data sources, as well as precedence of use (downloads and citations). With legal experts, we design a pipeline for tracing dataset provenance, including their original sources, their licenses, and subsequent use.

The legal treatment of training data has significant consequences on AI development. Records in training data are normally copyrighted, and it remains unclear if the process of converting training data into model weights violates the copyrights of the original authors (Quang, 2021; Epstein et al., 2023). Using copyrighted training data may be considered "fair use" because the use of copyrighted material is far from the original purpose of the material (Sobel, 2017). Independent of these open legal questions, a transparent chain of provenance for data contributes to transparent and responsible AI development. To this end, the Data Provenance Project helps developers clearly attribute the data they use. We enumerate our contributions:

1. **Designing Large-scale Data Documentation** A combined human-machine data documentation framework, designed by legal and AI experts, to collect reliable information on dataset licenses, characteristics and provenance.
2. **The First Empirical Analysis of Natual Language Dataset Licenses** This work comprises the first large scale, empirical comparison of data licensing practices for natural language datasets.
3. **Tools for Data Provenance** An open source repository for downloading, filtering, and exploring detailed properties of thousands of text datasets. This includes tools to auto-generate *Data Provenance Cards* for future documentation best practices.

2. The Data Provenance Project

The Data Provenance Project remedies the described challenges with a large-scale expert-guided annotation of popular text datasets. We tailor our annotations, described in Section 2.1, to characteristics relevant for designing a well-informed training corpus. To begin with we target nearly 40 popular instruction or "alignment" finetuning data collections, for which we list a few examples in Table 1.

2.1. Data on Data

Our information collection spans (I) *identifier information*, bridging metadata from several aggregators, including Hugging Face, Papers with Code, Semantic Scholar, and ArXiv, (II) detailed *dataset characteristics* for a richer understanding of training set composition, and (III) *dataset provenance* for licensing and attribution. Our repository of tools then allow practitioners to filter data on any of the listed criteria and visualize the other characteristics of the resulting data. For their selected criteria they may then generate a human readable, markdown summary, or "Data Provenance Card" of the used datasets, and their compositional properties for languages, tasks, and licenses.

Identifier Information

1. **Dataset Identifiers**: The dataset's name, associated paper title, and description of the dataset.
2. **Dataset Aggregator Links**: A link each major aggregator, including Hugging Face, Papers with Code, Semantic Scholar, and ArXiv allows us to incorporate and compare their crowdsourced metadata.
3. **Collection**: The name and URL to the data collection of which this dataset is apart.

Dataset Characteristics

1. **Languages**: Each of the languages represented in the dataset. We use automated methods to augment language identification, as aggregators are often sparse.
2. **Task Categories**: The 20+ task categories represented in the instructions, such as Question Answering, Translation, Program Synthesis, Toxicity Identifcation, Creative Writing, and Roleplaying.
3. **Text Topics**: An automated annotation of the topics discussed in the datasets.
4. **Text Length Metrics**: The minimum, maximum, and mean number of dialog turns per conversation, of characters per user and per assistant.
5. **Format**: The format and intended use of the data. The options are zero-shot prompts, few-shot prompts, chain-of-thought prompts, multi-turn dialog, and response ranking.
6. **Time of Collection**: The time as which the work was published, which acts as an upper bound estimate on the age of the text.

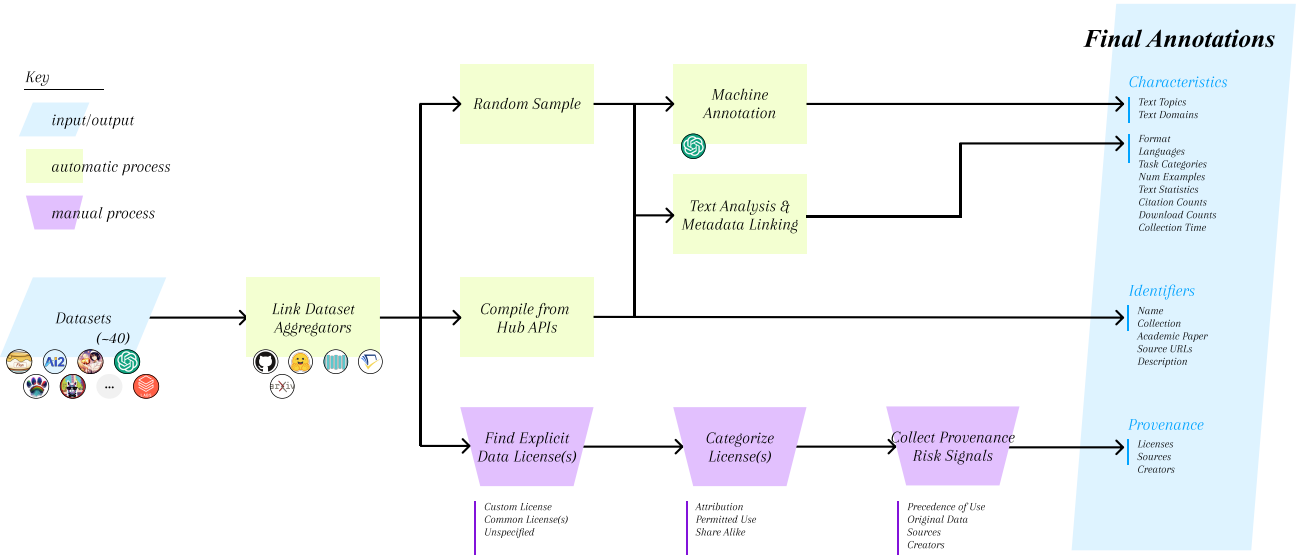

Figure 1. The metadata collection pipeline, combining manual and automatic processes.

Dataset Provenance

1. **Licenses**: The license name and URLs associated with the data, using the process described in Section 2.2
2. **Text Source**: The original sources of the text, often Wikipedia, Reddit, or other scraped online/offline sources.
3. **Creators**: The institutions of the dataset authors, including universities, corporations, and other organizations.
4. **Attribution**: The attribution information for the authors of the paper associated with the dataset.
5. **Citation & Download Counts**: The citation and Hugging Face download count for the paper and dataset, dated June 2023. This acts as an estimate or community use, and is commonly used as precedence to decide on the risk-level for using these datasets when the license is Unspecified.

2.2. Data Provenance Collection

This data was collected with a mix of manual and automated techniques. Connecting disparate dataset information hubs, like Github, Hugging Face and Semantic Scholar, required manual search. Annotating and verifying license information, in particular, required a carefully guided manual workflow, designed with legal practitioners. Once these information hubs were connected, it was possible to synthesize or scrape additional metadata, such as dataset languages, task categories, and time of collection. And for richer details on each dataset, like text topics and source, we used carefully tuned prompts on language models inspecting each dataset.

Automated Annotation Methods Based on the manually retrieved pages, we automatically extract Licenses from HuggingFace configurations and GitHub pages. We leverage the Semantic Scholar public API (Kin, 2023) to retrieve the released date and current citation counts associated to academic publications. Additionally, we compute a series of other helpful, but often overlooked data properties such as text metrics, and dialog turns. We elected to measure sequence length in characters rather than word tokens, for fairer treatment by language and script.

API Annotation Methods While task categories have been the established measurement of data diversity in recent instruction tuning work (Sanh et al., 2021), they omit other perspectives on data diversity and representation. To augment this, we use OpenAI's ChatGPT API to help annotate for richer text features. We randomly sampled 100 examples per dataset and carefully curated model prompts (with examples) to suggest up to 3 topics discussed in the text.

To annotate for the original data sources, we embedded the text from the dataset's academic paper into a prompt for ChatGPT's API. Our manual verification of these prompts and their answers showed reliable and accurate results for simple tasks like topic classification, language identification, and keyword extraction (data sources).

License Annotation Workflow Natural language datasets are often combined into large collections which can make some of the requirements of open-source licenses challenging to operationalize. Machine learning practitioners usually

want to segment these datasets into categories, representing whether they are allowed to train on the data, evaluate on it, modify it, and/or re-distribute it. Our license annotation workflow follows these steps:

1. **Compile all License Information** We aggregate all licensing information from Github, ArXiv, Hugging Face, Papers with Code, and the collection itself (e.g. Super-Natural Instructions).
2. **Search for explicit Data Licenses** The human annotator searches for a license specifically given to the dataset (not the code) by the authors. A license is found if (a) the Github repository mentions or links a license for the data, (b) the Hugging Face Dataset Card mentions or links a license for the data, (c) the Hugging Face license label was uploaded by the dataset creator themselves, (d) a dataset-specific license is linked from Papers with Code, or the paper itself.
3. **Identify a License Type** A license may fall into a set of common categories (e.g. MIT, Apache 2, CC BY SA, etc.), be a "Custom" license, a permission Request Form, or if none was found for the data, "Unspecified". If a dataset is comprised of multiple parts, the annotator can label a list of license types. The license type is accompanied by a license URL and license notes from the annotator, for posterity and categorizing "Custom" licenses (the next step).
4. **Categorize Licenses** Based on discussions with industry experts, we categorize licenses based on four important features: is data usage limited to academic or non-commercial purposes (**Permitted Use**), does the data source need to be attributed (**Attribution**), and do derivatives of the data need to licensed under the the same terms as the original (**Share-Alike**).
5. **Additional Provenance** In practice, legal teams may wish to balance their risk tolerance with more nuanced criteria. For instance, they may be satisfied with using (more permissive) Github licenses, even when it is ambiguous whether these apply to the code or the data. They may also wish to include or exclude datasets based on whether these are widely used in practice, where the original data was sourced from, and if the creator is a competitor. To supplement the above license categories, we also collect all this metadata for filtering.

Even these simple steps raise challenges: for example, the attribution clauses of many licenses require a copy of the original license to be included with the licensed material. For a large collection, this could mean that thousands of different licenses would need to be attached to a trained model. Our project aims to comply with the terms of license agreements while also focusing on pragmatic usability. For example, to resolve the issue of attribution, we assemble a master list of licenses for each user's request and also create a more concise data card that attributes datasets more concisely and conveniently.

3. Legal Discussion

The legal landscape surrounding the use of alignment data is complex. It remains unclear how license terms should be interpreted for data usage and many relevant laws, including copyright and fair use, are ill-defined in this context.

Most open source licenses were designed for software, but we find them being attached to alignment data. These licenses were not originally conceived with data utilization in mind, rendering their terms ambiguous when applied to this new context. This issue is further exacerbated when multiple datasets, each potentially governed by a different open-source license, are amalgamated into collections. A primary concern arises in determining whether machine learning models trained on these data collections should be classified as "derivative" works. If so, such models would, for example, be subject to copyleft requirements, which may preclude developers from utilizing multiple datasets released under incompatible copyleft licenses or require them to assemble hundreds or thousands of licenses to satisfy attribution requirements.

If license terms are deemed to be inapplicable to data, datasets may nevertheless be protected under copyright laws. This situation offers some potential legal avenues for data utilization, such as statutory exceptions - exemplified by the European Union's Data Mining exception - or fair use standards. Yet again, the application of these exceptions and standards to machine learning remains ill defined, further complicated by jurisdictional differences in legal standards.

In the face of these pervasive legal uncertainties, practitioners' decisions regarding data usage are ultimately guided by a blend of factors including the specific licensing terms, the origin of datasets, and the degree of usage of a given dataset by others. Navigating this landscape requires striking a delicate balance between risk mitigation and the need for sufficient resources. This equation, however, varies across regions, applications, and corporate environments, influenced by factors such as competition, risk, and regional legislation.

In creating a repository of all this information, we hope to give practitioners the ability to make informed choices, while encouraging dataset creators with the ability to be more thoughtful about the licenses that they select. Ultimately, data licenses could be leveraged to promote more responsible, inclusive, and transparent machine learning practices.

References

The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Arstechnica. Stable diffusion copyright lawsuits could be a legal earthquake for ai, 2023. URL https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Conover, M., Hayes, M., Mathur, A., Meng, X., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm. https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm, 2023.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. 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, 2021.

Epstein, Z., Hertzmann, A., Herman, L., Mahari, R., Frank, M. R., Groh, M., Schroeder, H., Smith, A., Akten, M., Fjeld, J., et al. Art and the science of generative ai. *Science*, 380(6650):1110–1111, 2023.

Ethayarajh, K., Zhang, H., Wang, Y., and Jurafsky, D. Stanford human preferences dataset, 2023. URL https://huggingface.co/datasets/stanfordnlp/SHP.

Frankle, J. Tweet by mosaic ml. https://twitter.com/jefrankle/status/1654848529834078208, 2023.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., and Song, D. Koala: A dialogue model for academic research. Blog post, April 2023. URL https://bair.berkeley.edu/blog/2023/04/03/koala/.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

LMSYS-Org. Lm-sys: Fastchatt5. https://github.com/lm-sys/FastChat#FastChat-T5, 2023.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023a.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., and Ippolito, D. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, toxicity, 2023b.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Nguyen, H., Suri, S., Tsui, K., and Schuhmann, C. The open instruction generalist (oig) dataset. https://laion.ai/blog/oig-dataset/, 2023.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1776–1826, 2022.

Quang, J. Does training ai violate copyright law? *Berkeley Tech. LJ*, 36:1407, 2021.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. *ICLR 2022*, 2021. URL https://arxiv.org/abs/2110.08207.

Sileo, D. tasksource: A dataset harmonization framework for streamlined nlp multi-task learning and evaluation. 2023.

Sobel, B. L. Artificial intelligence's fair use crisis. *Columbia Journal of Law & the Arts*, 41:45, 2017.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *NeurIPS*, 2020.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Vercel. Sharegpt, 2023. URL https://sharegpt.com/.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language model with self generated instructions, 2022a. URL https://arxiv.org/abs/2212.10560.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A. S., Naik, A., Stap, D., et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022b. URL https://arxiv.org/abs/2204.07705.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, 2022c.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *TMLR*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., and Huang, P.-S. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2447–2469, 2021.

Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., and Klein, D. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2390–2397, 2021.

A. Data Collections

In the Data Provenance Project we compile nearly 40 collections of instruction and alignment tuning datasets. A subset of these are listed in Table 1 to illustrate the diversity in source, characteristics, languages, and licenses. It also illustrates the focus on popular and commonly used datasets in the community.

COLLECTION	LICENSE	LANGS	NOTABLE CHARACTERISTICS
Anthropic HH	MIT	1	Response helpfulness scores (RLHF) (Bai et al., 2022; Ganguli et al.)
Dolly 15k	CC BY-SA	1	15k human generated instructions and responses (Conover et al., 2023)
OpenAssistant	Apache 2.0	35	161k human written multi-turn conversations (Köpf et al., 2023)
Flan Collection	Various	38	Largest public academic task collection (Longpre et al., 2023a)
xP3x	Various	277	Largest multilingual academic task collection (Muennighoff et al., 2022)
Tasksource	Various	1	485 Wide-ranging English recasted classification tasks (Sileo, 2023)
LAION OIG	Various	$\sim$ 50	43M instruction-outputs from varied data sources (Nguyen et al., 2023)
SHP	Reddit API	1	385k Ranked Reddit thread responses (RLHF) (Ethayarajh et al., 2023)
ShareGPT	OAI NC	1	OpenAI responses crowdsourced by a browser extension (Vercel, 2023)
Self-Instruct	OAI NC	1	Synthetic instruction-outputs generated from GPT-3 (Wang et al., 2022a)
WebGPT	OAI NC	1	20k RLHF question-answer-human ratings (Nakano et al., 2021)
OpenAI Summ.	OAI NC	1	93k model summary, human rating pairs (Stiennon et al., 2020)

Table 1. **Alignment Tuning Collections and their characteristics.** For data taken from proprietary sources, we annoate with OpenAI's Non-compete (OAI NC) clause, or with Reddit's API terms.