# Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation

## A Case Study of the HuggingFace and GEM Data and Model Cards

**Angelina McMillan-Major**
University of Washington
Hugging Face
angie@huggingface.co

**Salomey Osei**
KNUST
Masakhane
sosei@aimsammi.org

**Juan Diego Rodriguez**
UT Austin
juand-r@utexas.edu

**Pawan Sasanka Ammanamanchi**
IIIT Hyderabad
pass2pawan@gmail.com

**Sebastian Gehrmann**
Google Research
gehrmann@google.com

**Yacine Jernite**
Hugging Face
yacine@huggingface.co

## Abstract

Developing documentation guidelines and easy-to-use templates for datasets and models is a challenging task, especially given the variety of backgrounds, skills, and incentives of the people involved in the building of natural language processing (NLP) tools. Nevertheless, the adoption of standard documentation practices across the field of NLP promotes more accessible and detailed descriptions of NLP datasets and models, while supporting researchers and developers in reflecting on their work. To help with the standardization of documentation, we present two case studies of efforts that aim to develop reusable documentation templates – the HuggingFace data card, a general purpose card for datasets in NLP, and the GEM benchmark data and model cards with a focus on natural language generation. We describe our process for developing these templates, including the identification of relevant stakeholder groups, the definition of a set of guiding principles, the use of existing templates as our foundation, and iterative revisions based on feedback.

## 1 Introduction

Dataset and model documentation is a necessary step in identifying potential issues with machine learning (ML) systems and addressing their broader impacts (Gebru et al., 2018, 2020; Bender and Friedman, 2018, among others). In their overview of data collection and use in ML, Paullada et al. (2020) identify issues that have frequently arisen such as considerations for how subjects are represented in datasets, spurious cues that may be exploited by ML model, and concerns about the content in datasets collected through crawling methodologies. They advocate for careful documentation of datasets and their collection processes in order to surface these problems. However, best practices for documentation have seen no widespread adoption even for the most popular datasets and models. Indeed, writing such detailed documentation requires additional effort from researchers who may lack the required resources or familiarity with the process. Providing dataset and model creators with guidelines and several examples in a single place to inspire and inform prospective writers could thus drive widespread adoption of documentation.

Research efforts that involve a large number of models or datasets are particularly well positioned to develop and maintain specific guidelines and best practices by making documentation a required component for submitting contributions. By bringing together the domain expertise of participants and the experience of researchers who have a greater familiarity with documenting data and models, these efforts provide an opportunity to develop and refine templates that balance generality and informativeness. In addition, by requiring appropriate documentation for any involved model or dataset, these efforts can set a precedent that informs future endeavours. We encourage organizations to consider their role in the successful uptake of documentation practices, such as providing their members with adequate resources to understand the goals and motivations of documentation and measured steps towards integrating documentation into current research norms.

In this paper, we present two case studies of creating documentation templates and guides in natural language processing (NLP): the Hugging Face (HF) dataset hub[1] and the benchmark for Gen-

[1] https://hf.co/datasets/card-guide



*Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 121–135

eration and its Evaluation and Metrics (GEM).[2] We use the term *data card* to refer to documentation for datasets in both cases and the term *model card* to refer to documentation for models in the GEM workshop, following Mitchell et al. (2019). Focusing on these settings allows us to ground what constitutes 'good' documentation in these contexts, namely technical user-oriented information, scientific reproducibility, and social contextualization of data and data-driven systems.

## 2 Related Work

In the U.S., research involving direct interventions on human subjects is subject to the Federal Policy for the Protection of Human Subjects (or Common Rule).[3] This policy tasks institutional review boards (IRBs) with certifying that such research follows established ethical standards and regulations. While this review is sometimes decried as cumbersome (Grady, 2015), the process ensures that researchers both reflect and communicate ahead of time how the data will be collected and used, why it is necessary to answer the question at hand, and how protected information will be handled to prevent harm to the human subjects. Whereas much of the data that supports current methods in ML and especially NLP is created by, gives information about, and is used to train models that will likely affect these same human subjects, this relationship does not constitute a direct intervention as defined in the Common Rule. However, Metcalf and Crawford argue that the definition is too narrow when one considers the similarity in potential harms and advise that data-driven methods should be subject to a similar from of ethical review, which includes clear communication about the goals and mechanisms for collecting and safeguarding the data.

Despite existing literature on database documentation in HCI and related fields (Cheney et al., 2009; Bhardwaj et al., 2014), documentation in ML has only recently gained traction. In a survey of ML projects across India, East and West African countries, and the USA, Sambasivan et al. (2021) analyze compounding events causing negative, downstream effects from data issues, resulting in technical debt[4] over time, and identify insufficient documentation to be a trigger of these events in 20% of the 53 cases. They report impacts such as time and effort lost in using incorrect data, barriers to completing models, and data that had to be abandoned due to it no longer being usable.

In response to issues like the ones found by Sambasivan et al., many research groups have proposed documentation schemata for different parts of the ML pipeline. Arnold et al. (2019) introduce FactSheets to document the use, performance and security aspects of an AI product. Mitchell et al. (2019) put forward Model Cards focusing on documenting the evaluation and use of a specific model, while Gebru et al. (2018), Holland et al. (2018), and Pushkarna et al. (2021) propose Datasheets for Datasets, Dataset Nutrition Labels, and the Data Cards Playbook, respectively, as documentation schemata and processes for documenting the data used in ML and AI systems. Hutchinson et al. (2021) frame datasets as technical infrastructure and propose documentation for several stages of the development process, including the design, creation, and maintenance of the dataset.

To address the distinct challenges of working with language data, such as those summarized by Bender et al. (2021), other researchers have proposed specialized documentation for work in NLP. The first such example is by Bender and Friedman (2018) who propose a version of Data Statements for documenting aspects of the data from a linguistic perspective. In addition to documenting the data that a model sees during training, there is also the need to document experiments, especially those involving humans. Thus, Shimorina and Belz (2021) develop a Human Evaluation Datasheet, with the goal of describing human evaluation experiments rather than naturally occurring language data. Starting from the templates by Bender and Friedman (2018) and Mitchell et al. (2019), we iteratively extend and adapt our templates for the HF and GEM contexts.

## 3 Methods

To guide the development of our templates, we draw from the methods of value sensitive design (eg., Friedman and Hendry, 2019) by identifying stakeholders and assessing their values, which Friedman and Hendry define as "what is important to people in their lives, with a focus on ethics and morality" (pg. 24). The values of the various stakeholders, including developers themselves,

[2] https://gem-benchmark.com/data_cards/

[3] https://www.nidcr.nih.gov/research/human-subjects-research

[4] Cunningham (1992) employs this term to describe the accumulation of flaws in technical systems over time, using an analogy to financial debt.

may influence the development of a technology in a number of ways (Friedman and Kahn, 2003). In this section, we conduct a stakeholder analysis and describe the principles we follow in the development of our templates. In Section 7, we explore the potential social impact of the templates and our positionality in the design process.

### 3.1 Documentation Stakeholders

This work presents a documentation strategy adopted by two **organizations** (the HF dataset hub and the GEM benchmark) for two categories of **resources** (language datasets and NLG models). We identify three groups of direct stakeholders as well as indirect stakeholders whose needs we consider in designing this documentation strategy.

**The organizations.** The managing organizations play a central role in gathering, presenting, and enabling the use of the resources. The organizations are responsible for *establishing documentation standards* that need to be met for any resource.

**The resource creators.** When submitting their resources to an organization, dataset curators and model developers are required to *write documentation* to meet the organization's stated standards.

**The resource users.** The resources distributed by the organizations may further be utilized by downstream users. These users *read the provided documentation* to determine whether the resources may or may not be appropriate for their needs.

**The indirect stakeholders.** Indirect stakeholders include any person impacted by the dissemination of a resource, such as dataset publication, model deployment, and deployment of a model trained on a dataset. While indirect stakeholders might not have direct control over the way the resources are used, they may *refer to the documentation* to analyze these impacts.

### 3.2 General Approach

As discussed by Waseem et al. (2021), datasets are the result of subjective choices made by the dataset creators. We therefore approach documentation as a way for authors to communicate this subjectivity by explicitly stating the decisions that led to the resource creation and the contexts in which those decisions were made. In addition to providing developers with the opportunity to reflect on their choices in creating a resource, the documentation gives users insight as to how and why the resource was developed, which may help the user assess how appropriate the resource is for their use case, and may even surface previously unconsidered issues. In addition to the documentation formats surveyed in Section 2, the practice of reflecting on the impact of one's work is being standardized by academic conferences such as NeurIPS's broader impacts statement[5] and NAACL's ethics review.[6] We aim to encourage this practice in our local contexts through our free text templates that emphasize reflection on the topics we see as important in understanding the development and potential uses of datasets and models in NLP.

While guidance around these processes is still being standardized, we recognize that there are risks that may result from the proliferation of documentation formats and the suggestion of documentation as a way to mitigate the harms caused by ML systems. For example, documentation that is not updated to reflect changes to the documented resource may result in harms due to decisions made on inaccurate information. In other cases, the standardization of documentation may add to the difficulties experienced by inexperienced or underfunded researchers in publishing their work. Finally, authors may try to justify work that causes known harms by documenting the potential for harm without attempting to address the harms themselves. Organizations that institute documentation standards need to consider these risks when integrating documentation into their local contexts and be attentive to the varied impacts to researchers, developers, and community members.

### 3.3 Data Cards Principles

Language conveys information about not only the individual producing the language, but also about the social groups that individual is a member of and social context that the individual is producing the language in (Eckert and Rickford, 2001). For example, an accent may indicate the geographical region that a person grew up in and that person's use of a local phrase may indicate that they believe the person that they're talking to is also from that region. As such, the values of our indirect stakeholders, the people whose sociolinguistic information are embodied in the resources, are of high priority in designing the data card templates. In order for documentation to be accessible to in-

[5] https://neurips.cc/Conferences/2020/CallForPapers
[6] https://2021.naacl.org/ethics/faq/

direct stakeholders, who may not be familiar with ML terminology or academic writing, the information needs to be clearly presented and easy to find within the document. This consideration led to several revisions in our template content and presentation.

### 3.4 Model Cards Principles

The recent push by academic conferences for social impact statements in publications provides a clear way for resource creators to consider the impact to their own direct and indirect stakeholders. Furthermore, academic ML conferences such as NeurIPS have instituted reproducibility checklists for paper submissions (Pineau et al., 2020). Building off this checklist, Dodge et al. (2019) argue for greater reproducibility in ML publications by proposing their own reproducibility checklist as well as a metric for reporting performance on the validation set as a function of the model training time. Mitchell et al. (2019) focus on evaluation in their own schema for models, arguing for disaggregated results to be reported over different population subgroups in data used to evaluate the model. These three considerations - social impact, reproducibility, and evaluation - form the main aspects of our template.

## 4 Case Study I: HuggingFace Data Cards

As a first case study of data card development, we present the template developed for the HuggingFace open source NLP libraries. The full template is available in Table 1. The completed data card for the ELI5 dataset (Fan et al., 2019) is available in Appendix A.

**HF Libraries** The HuggingFace Transformers library (Wolf et al., 2020) has evolved as a centralized platform providing a common API for easily loading the weights of nearly 10,000 (at the time of writing) transformer-based models with different frameworks (PyTorch, TensorFlow, JAX) and original code bases. The Datasets library[7] takes a similar approach to providing a hub that allows users to easily discover and reuse the datasets that were used to trained those models. To that end, the library focuses on the following three features:

- A unified API for downloading and iterating through a wide variety of datasets
- A backend supported by memory-mapped Arrow arrays[8] to enable use of even large datasets in resource-constrained settings
- A documentation structure that gives users a clear overview of the available datasets and the information required to use them

To meet the latter need, we designed a data card template that provides a unified way to present this information for all of the proposed NLP datasets.

**HF data cards stakeholders** The direct stakeholders of the HF card templates include: the organization, whose goals for the library and its documentation standards are stated above; the team and HF community members, who add new datasets to the library and are encouraged to fill out as much of the cards as they can; and the library users, who may examine or train models on the provided datasets. We note that, in our case, the people who add the datasets to the library may not be the original dataset curators, and so may not have direct access to all the required information.

**Initial version** The first version of the data card consisted of 8 sections. The first three aimed to answer the question *"What is this dataset used for?"* and asked the writers to fill out information about the tasks supported, the original purpose for creating the dataset, and the languages represented within. The template then asked about the *people involved in making the dataset*, including the dataset creators, the language creators, and the annotators, if relevant. This was followed by a section titled "Data Characteristics," which covered all of the *data selection and processing steps*. In particular, considering that users might want to use several datasets developed to address similar tasks together to train a model, we wanted to surface any domain shifts or differences in text normalization in either of these last two sections.

The template then continued to a *dataset structure* section covering information about the default train/test split if provided, size of the dataset, description of the features, examples of data points, and suggested metrics to use. Broadly, the purpose of this section was to give a user any technical information they might need to train a model, such as how the dataset size might influence what size of model or regularization technique to use. Seeing

[7] https://hf.co/datasets

[8] https://arrow.apache.org/

the features and an explicit example of a data instance can also clarify the input and output features and texts.

The second to last section covered *known limitations* of the dataset, and prompted the writer to consider specifically social biases in a first subsection and any other limitations, such as common surface correlations a model might take advantage of, in another. The reasoning behind this choice is that a user might do as much harm by deploying a model exhibiting harmful biases as by deploying a model that had a high score for the chosen metric but did not actually perform the described high-level task. Finally, for the benefit of users who might want to share derivatives of the dataset or use them for commercial purposes, the last section contained the *licensing information* for the dataset. Using this schema, we wrote an initial draft card for the SNLI dataset (Bowman et al., 2015) and shared it with the HF team and the SNLI curators for comments.

**Revised version** We then expanded and reordered the template based on the initial comments. The first section became the *dataset description* consisting of: a list of links to relevant information about the dataset available elsewhere including the dataset paper, leaderboards, and the contact information of at least one person in case of further questions; a free text summary; a description of supported tasks, suggested metrics (moved here for this updated version), and leaderboards; and the languages represented. The *dataset structure* was moved just after the languages, with examples of data instances and information about the fields and splits.

The largest change was in the people and dataset characteristics sections. We restructured these into a single *dataset creation* section which now starts with a curation rationale to properly contextualize all of the choices described in the rest of the card. The template then requests information about the people involved in producing the source language and annotations and the normalization and processing steps for that data. A section was then added to specifically describe the status of Personal Identifying (PI) data in the dataset in order to both help protect the data subjects' privacy and to help the dataset users comply with existing regulations. Dataset creators were renamed dataset curators to emphasize the difference between the people making the curation choices and the people producing the source language data, and their description was moved to the very end of the data card. Finally we expanded the limitations section to a broader section on *considerations for using the data*, adding a prompt for prospective social impacts of using the dataset, both positive and negative.

**Supporting documentation writers** We made these changes to improve card readers' ability to navigate the document and find necessary information about the dataset and to assist card authors when writing their cards by clarifying the desired information for each section. To further aid authors, we developed a guide formatted with desired content and instructions for each section.[9] We intend for the template and guide to support authors of datasets both with and without existing documentation. Authors of datasets without publications can use the card as a starting point for building documentation and visibility for the dataset in the HF library, but also as an overview of what information should be included in a publication. For datasets with publications, the HF card provides authors with a more widely accessible format for documenting their dataset that does not have the length limitations of paper submissions and can be revised as needed to reflect any updates to the dataset.

Publications also have the property of being static and cannot be updated to reflect changes in the dataset. To address this, we designed the HF data cards to function as living documents. First, hosting them on GitHub allows community members at large to easily add new information or modify existing sections to reflect new findings. We see this as particularly important for the section on using the dataset as new considerations are reported as a result of novel use cases and research. We also made the decision to publish the template with the sections pre-populated with placeholder text (specifically, "More Information Needed") in order to encourage authors and community members to fill in the section when the information is available. The ability to update information helps to address the harms caused by out-of-date documentation. By integrating the data cards into the HF library, we are able to see a more complete characterization of the available datasets that is similarly up to date. This allows us to point out where fewer datasets are available for tasks and languages and make progress towards a more diverse library.

[9] https://hf.co/datasets/card-guide

## 5 Case Study II: GEM Data and Model Cards

Our second case study of data card development is the template we developed for the datasets and model submissions of the Generation, its Evaluation, and Metrics (GEM) workshop. We present the full template in Table 1. The completed data card for the ASSET dataset (Alva-Manchego et al., 2020) is available in Appendix B.

### 5.1 GEM Benchmark

The benchmark for GEM aims to standardize how research in natural language generation (NLG) is conducted with a particular focus on in-depth evaluations (Gehrmann et al., 2021). To this end, newly developed NLG models should be documented and evaluated on a set of established tasks over a range of reproducible and robust metrics. This goal can only be achieved if the infrastructure provided by the benchmark supports the creation of such documentation.

Since a benchmark may comprise multiple datasets and provide a centralized way to interact with them, we can focus on two groups of stakeholders following the descriptions in Section 3.1.

**Benchmark curators.** The curators need to *ensure that all datasets are documented* according to the requirements.

**Benchmark participants.** The participants need to *write model documentation* according to the requirements, but may be novice ML practitioners or inexperienced in writing documentation.

### 5.2 Data Cards

None of the datasets included in GEM had existing data cards. To address this issue, we develop a data card template and use it to document all the datasets involved in the benchmark. Moreover, to be able to quickly add new datasets and to help the broader NLG community construct their own data cards, we release the template and associated guide. The data card closely follows the HF data card template introduced in Section 4, with changes to target NLG-specific issues. We made these changes to address feedback after testing an initial version of the data card on the CommonGen (Lin et al., 2020) and ASSET (Alva-Manchego et al., 2020) datasets. An overview of the differences between the two templates is presented in Table 1.

The major difference between the general HF template and the NLG-specific template is that NLG datasets may contain natural language both in the input and the output. Inputs and outputs may have different sources and thus require documentation for both. In addition, the input-output pairs may be constructed in ways that are challenging to describe in the HF template. For example, output text may be crawled and undergo revisions while the input text remains the same. This difference did not lead to different sections in the data card itself, but it did lead to changes in the guidelines on how to write them.

Moreover, NLG tasks have an underlying communicative goal which differentiates them from classification and other structured tasks. It is imperative to surface the communicative goal, since it heavily influences how generated text for a particular task should be evaluated. Another category of changes concerns the context of GEM compared to general purpose data cards. For example, since GEM itself is a benchmark, information about leaderboards does not have to be prominently featured, whereas it should give credit to the original data creators early on.

We also added three GEM-specific sections: (1) Why is this dataset part of GEM, (2) Changes to the original dataset for GEM, and (3) Getting started with in-depth research on the task. The first aims to tie the collections of data cards together by situating a dataset and task within the larger goal of the benchmark. The second section is of crucial importance for any data card for a benchmark, since the benchmark may change the purpose of a dataset and the organizers could modify the underlying data by cleaning it, adding more data, or releasing a reformatted version. The final question encourages participants to engage with the data in order to develop a deeper understanding of the task formulation. Therefore, to help participants gain insights into the data, we included a section with helpful pointers to relevant papers and tutorials.

Finally, GEM is designed to be a multilingual benchmark. Since we expect to include languages with fewer resources than may be found for languages like English, we aim to consider the communities that speak those languages and the impacts that technology built with these datasets could have on them. For example, a dataset for a language with few other resources may only capture the language of a few speakers in a certain context, like

| **HuggingFace data card** | **GEM data card** |
|---|---|
| Dataset Description | Dataset and Task Description |
| • Dataset Summary | • Dataset and Task Summary |
| • Supported Tasks and Leaderboards | *See below* |
| – | • Why is this dataset part of GEM? |
| • Languages | • Languages |
| – | Meta Information |
| *See below* | • Dataset Curators |
| *See below* | • Licensing Information |
| *See below* | • Citation Information |
| *See above* | • Leaderboard |
| Dataset Structure | Dataset Structure |
| • Data Instances | • Data Instances |
| • Data Fields | • Data Fields |
| • Data Splits | • Data Statistics |
| Dataset Creation | Dataset Creation |
| • Curation Rationale | • Curation Rationale |
| – | • Communicative Goal |
| • Source Data | • Source Data |
| •• Initial Data Collection and Normalization | •• Initial Data Collection and Normalization |
| •• Who are the source language producers? | •• Who are the source language producers? |
| • Annotations | • Annotations |
| •• Annotation process | •• Annotation process |
| •• Who are the annotators? | •• Who are the annotators? |
| • Personal and Sensitive Information | • Personal and Sensitive Information |
| – | Changes to the Original Dataset for GEM |
| Considerations for Using the Data | Considerations for Using the Data |
| • Social Impact of the Dataset | • Social Impact of the Dataset |
| – | • Impact on Underserved Communities |
| • Discussion of Biases | • Discussion of Biases |
| • Other Known Limitations | • Other Known Limitations |
| – | Getting started with in-depth research on the task |
| Additional Information | – |
| • Dataset Curators | *See above* |
| • Licensing Information | *See above* |
| • Citation Information | *See above* |
| • Contributions | – |
| – | Credits for Data Cards and this Template |

Table 1: Side by side comparison of the HF and GEM data card templates. Each section is denoted by horizontal lines, subsections are denoted with •, subsubsections with ••.

| GEM Model Card |
|---|
| Social Impact |
| • Additional Data |
| • Training Process |
| • Real-World Use |
| • Measuring Impact |
| Reproducibility |
| • Model Description |
| • Model Details |
| • Model Hyperparameters |
| • Hyperparameter Specification |
| • Number of Training and Evaluation Runs |
| • Dataset Details |
| • Dependencies and External Libraries |
| • Link to Downloadable Source Code |
| • Computing Infrastructure Used |
| Evaluation details |

Table 2: An overview of the GEM model card template.

university students in the context of an experiment. Models trained on this dataset would not have access to the variation in the language that comes from speakers of other ages and in other contexts, but because there may not be other available tools, that model may become widely used and misrepresented as a general model of the language. We thus added a specific section to address potential concerns involving communities that speak those language varieties, such as the implications for model generality and privacy when speakers from small communities are easily identifiable.

### 5.3 Model Cards

Following the guiding principles outlined in Section 3.4, the model cards have three sections: social impact, reproducibility, and evaluation. A detailed overview is shown in Table 2.

**Social impact** In the first section, we invite submission authors to consider the impacts their models may have on users if they were deployed. We recognize that without further guidance, the open-ended nature of this request may make it prohibitively difficult to address. Indeed, trying to foresee all the ways in which data or modeling choices may affect all direct and indirect stakeholders is overwhelming, if not impossible. Instead, we narrow the scope to help users practice reflecting on causal relationships between design and deployment effects. We do this by providing guiding examples of models and their potential impacts. In one scenario, we consider a summarization model trained on a English Wikipedia, which is known to have various dimensions of gender bias (Wagner et al., 2015). We present two possible impacts on the output summaries based on this gender bias and suggest tests to measure the effect. We then encourage model creators to follow a similar line of thought. We ask about additional data used (and to a link to documentation if it exists), about the training process, and about a possible real-world use. We then request that the documentation author choose one aspect of one of the steps outlined, contemplate a way in which this aspect may negatively impact direct or indirect users, and propose a way to measure this impact. In particular, we believe that the latter requirement may help steer authors toward considering more plausible impacts.

**Reproducibility** The reproducibility section of the card combines elements of Mitchell et al. (2019)'s model cards and Dodge et al. (2019)'s reproducibility checklist. The sections ask the minimal number of questions which are key to reproducing the model submission. We request a model description, which includes the model type, version, the environment (i.e., versions of required software), and training algorithm used, with available space for further details. We also request a specification of dependencies and external libraries used to build the model. Authors have the option to link to their source code. Finally, authors are asked to describe the compute infrastructure used (e.g., the number of GPUs, the GPU type, and vRAM) and the training time for the final model.

Several questions concern the model hyperparameters, including the optimizer, training steps, learning rate. In addition, we elicit information about potential hyperparameter searches conducted as part of the model development. The hyperparameter search section requests information on the bounds for the hyperparameters, the number of search trials, and the method for choosing the hyperparameter values. The hyperparameter specifications for the best performing models are also requested but not required. Finally, the section ends with an optional space to list the number of training and evaluation runs and a required subsection detailing the utilized training dataset(s), including any processing on the data.

**Evaluation** The final section consists of the evaluation description. We suggest summarizing the evaluation process by including the metric details, the splits for the training, validation, and test data, and by providing the model performance on the test and validation data.

## 6 Conclusion

We detail the processes and principles we followed to produce the documentation templates used in the HF library and the GEM benchmark. We ground this work in the current discussion of documentation as a way to communicate the impacts of ML systems. As touched on in Section 3.2, extensive documentation is only a tool to support the communication of decisions that led to the creation of datasets and models and the positionality of their creators; it is not a direct solution to the harms caused by ML systems. We present our templates to encourage others to consider important questions that may be asked of their own work. The templates from both case studies are open source and we welcome contributions and feedback from authors and users to continually revise and improve them. Moreover, while the templates described in this work are designed for specific contexts and may not be fully applicable to others, they can be used as starting points for adaption to other settings.

## 7 Social Impact and Positionality Statement

**Social impact statement** Our goal is to promote the standardization of specialized documentation for NLP datasets and models. Institutional adoption and promotion may see its greatest effect in the widening of community engagement. The infrastructure used to host and maintain the documentation also facilitates revisions and smaller contributions from the involved communities. However, we are also aware of the risks that requiring this level of documentation for participation in either of our organizations may produce, such as raising the barrier to entry for those without experience in writing such documentation for language data as well as for people with fewer mentoring resources or platforms for engaging the community. Finally, while documenting the limitations of a resource is an important first step towards incrementally addressing issues, there is a risk that the act of documenting may allow creators to abdicate responsibility for these limitations in some cases, without taking any further steps to minimize negative social impacts of the systems they develop.

**Positionality statement** We are researchers at academic and industrial institutions with backgrounds in linguistics, NLP, ML, and HCI. Our guiding principles are discussed in Section 3. We aim to adapt available schemata to our specialized contexts, namely the HF library and the GEM benchmark and to present our development process as part of the general progress towards accountable and practical documentation for language datasets in ML systems. As NLP practitioners, we developed these card templates to directly support other members of the HF and NLG communities in writing documentation that answers questions that we ourselves would ask about the data and models. The completed templates will support users, researchers, and members of the public who may be impacted by these resources in understanding their contents and context.

## Acknowledgments

Thank you to our anonymous reviewers and colleagues for their thoughtful comments on this work.

## References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM*

*Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Anant Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J. Elmore, Samuel Madden, and Aditya G. Parameswaran. 2014. DataHub: Collaborative data science & dataset version management at scale.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2009. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474.

Ward Cunningham. 1992. The wycash portfolio management system. In *Addendum to the Proceedings on Object-Oriented Programming Systems, Languages, and Applications (Addendum)*, OOPSLA '92, page 29–30, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Penelope Eckert and John R Rickford. 2001. *Style and sociolinguistic variation*. Cambridge University Press, Cambridge, UK ; New York, NY.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Batya Friedman and Peter H. Kahn. 2003. Human values, ethics, and design. In Julie A. Jacko and Andrew Sears, editors, *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, Human factors and ergonomics, page 1177–1201. Lawrence Erlbaum Associates, USA.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for datasets.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Sebastian Gehrmann, Tosin P. Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondrej Dusek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Rubungo Andre Niyongabo, Salomey Osei, Ankur P. Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. *CoRR*, abs/2102.01672.

Christine Grady. 2015. Institutional review boards: purpose and challenges. *Chest*, 148(5):1148–1155.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

J. Metcalf and K. Crawford. 2016. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In

*Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program).

Mahima Pushkarna, Andrew Zaldivar, Dan Nanas, Emily Brouillet, Reena Jana, Oddur Kjartansson, Danielle Smalls, and Vivian Tsai. 2021. Data cards playbook. https://pair-code.github.io/datacardsplaybook/.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI.

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied machine learning: On the illusion of objectivity in NLP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

# A Example of a Hugging Face Data Card: ELI5

## A.1 Dataset Description

ELI5 homepage; ELI5 repository; paper: ELI5: Long Form Question Answering; contact: Yacine Jernite

**Dataset Summary** The ELI5 dataset is an English-language dataset of questions and answers gathered from three subreddits where users ask factual questions requiring paragraph-length or longer answers. The dataset was created to support the task of open-domain long form abstractive question answering (QA), and covers questions about general topics in its r/explainlikeimfive subset, science in it r/askscience subset, and history in its r/AskHistorians subset.

**Supported Tasks and Leaderboards** The dataset can be used to train a model for Open Domain Long Form QA. An LFQA model is presented with a non-factoid and asked to retrieve relevant information from a knowledge source (such as Wikipedia), then use it to generate a multi-sentence answer. The model performance is measured by how high its ROUGE score to the reference is. A BART-based model with a dense retriever trained to draw information from Wikipedia passages achieves a ROUGE-L of 0.149.

**Languages** The dataset is in English (BCP-47 code: `en`), as spoken by users of the target subreddits.

## A.2 Dataset Structure

**Data Instances** A typical data point comprises a question, with a `title` containing the main question and a `selftext` which sometimes elaborates on it, and a list of answers from the forum sorted by the number of upvotes they obtained. The URLs in each of the text fields have been extracted to respective lists and replaced by generic tokens in the text. Examples are available here.

**Data Fields** `q_id`: a unique string question ID, corresponding to its ID in the source submission dumps; `subreddit`: the source subreddit- 'explainlikeimfive', 'askscience', or 'AskHistorians'; `title`: title of the question, with URLs extracted and replaced by tokens in the form `URL_n`; `title_urls`: list of the extracted URLs, the nth element of the list was replaced by `URL_n`; `selftext`: either an empty string or an elaboration of the question; `selftext_urls`: similar to `title_urls`, but for `self_text`; `answers`: a list of answers, each answer has: `a_id` (a unique string answer ID, corresponding to its ID in the source comments dumps), `text` (the answer text with the URLs normalized), and `score` (the number of upvotes the answer had received when the dumps were created); `answers_urls`: a list of the extracted URLs (All answers use the same list, the numbering of the token continues across answer texts).

**Data Splits** The data is split into a training, validation and test set for each of the three subreddits. In order to avoid having duplicate questions in across sets, the `title` field of each of the questions were ranked by their tf-idf match to their nearest neighbor and the ones with the smallest value were used in the test and validation sets. The number of training, validation, and test examples for each subreddit are: 272,634, 9,812, and 24,512 for r/explainlikeimfive; 131,778, 2,281, and 4,462 for r/askscience; and 98,525, 4,901, and 9,764 for r/AskHistorians.

## A.3 Dataset Creation

**Curation Rationale** ELI5 was built to provide a testbed for machines to learn how to answer more complex questions, which requires them to find and combine information in a coherent manner. The dataset consists of questions that were asked by community members of three subreddits, including r/explainlikeimfive, and the answers provided by other users. The rules of the subreddit make this data well-suited for abstractive QA: the questions need to seek an objective explanation about well established facts, and the answers provided need to be understandable without any particular domain knowledge.

**Source Data: Initial Data Collection and Normalization** The data was obtained by filtering submissions and comments from the subreddits of interest from the XML dumps of the Reddit forum hosted on Pushshift.io. In order to further improve the quality of the selected examples, only questions with a score

of at least 2 and at least one answer with a score of at least 2 were selected for the dataset. The dataset questions and answers span a period form August 2012 to August 2019.

**Source Data: Who are the source language producers?** The language producers are users of the r/explainlikeimfive, r/askscience, and r/AskHistorians subreddits between 2012 and 2019. No further demographic information was available from the data source.

**Annotations** The dataset does not contain any additional annotations.

**Personal and Sensitive Information** The authors removed the speaker IDs from the Pushshift.io dumps but did not otherwise anonymize the data. Some of the questions and answers are about contemporary public figures or individuals who appeared in the news.

### A.4 Considerations for Using the Data

**Social Impact of Dataset** The purpose of this dataset is to help develop better question answering systems. A system that succeeds at the supported task would be able to provide a coherent answer to even complex questions requiring a multi-step explanation, which is beyond the ability of even the larger existing models. The task is also thought as a test-bed for retrieval model which can show the users which source text was used in generating the answer and allow them to confirm the information provided to them. It should be noted however that the provided answers were written by Reddit users, an information which may be lost if models trained on it are deployed in down-stream applications and presented to users without context. The specific biases this may introduce are discussed in the next section.

**Discussion of Biases** While Reddit hosts a number of thriving communities with high quality discussions, it is also widely known to have corners where sexism, hate, and harassment are significant issues. See for example the recent post from Reddit founder u/spez outlining some of the ways he thinks the website's historical policies have been responsible for this problem, Adrienne Massanari's 2015 article on GamerGate and follow-up works, or a 2019 Wired article on misogyny on Reddit. While there has been some recent work in the NLP community on *de-biasing* models (e.g. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings for word embeddings trained specifically on Reddit data), this problem is far from solved, and the likelihood that a trained model might learn the biases present in the data remains a significant concern. We still note some encouraging signs for all of these communities: r/explainlikeimfive and r/askscience have similar structures and purposes, and r/askscience was found in 2015 to show medium supportiveness and very low toxicity when compared to other subreddits (see a hackerfall post, thecut.com write-up and supporting data). Meanwhile, the r/AskHistorians rules mention that the admins will not tolerate "*racism, sexism, or any other forms of bigotry*". However, further analysis of whether and to what extent these rules reduce toxicity is still needed. We also note that given the audience of the Reddit website which is more broadly used in the US and Europe, the answers will likely present a Western perspectives, which is particularly important to note when dealing with historical topics.

**Other Known Limitations** The answers provided in the dataset represent the opinions of Reddit users. While these communities strive to be helpful, they should not be considered to represent a ground truth.

### A.5 Additional Information

**Dataset Curators** The dataset was initially created by Angela Fan, Ethan Perez, Yacine Jernite, Jason Weston, Michael Auli, and David Grangier, during work done at Facebook AI Research (FAIR).

**Licensing Information** The license hinges on the legal status of the Pushshift.io data which is unclear.

**Citation Information** The citation can be found in the ACL Anthology.

**Contributions** Thanks to @lewtun, @lhoestq, @mariamabarham, @thomwolf, and @yjernite.

## B Example of a GEM Data Card: ASSET

### B.1 Dataset Description

ASSET repository; paper: ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations; contact: Fernando Alva-Manchego, Louis Martin

**Dataset and Task Summary** ASSET (Alva-Manchego et al., 2020) is multi-reference dataset for the evaluation of sentence simplification in English. The dataset uses the same 2,359 sentences from TurkCorpus (Xu et al., 2016) and each sentence is associated with 10 crowdsourced simplifications. Unlike previous simplification datasets, which contain a single transformation (e.g., lexical paraphrasing in TurkCorpus or sentence splitting in HSplit), the simplifications in ASSET encompass a variety of rewriting transformations.

**Why is this dataset part of GEM?** ASSET is a high quality simplification dataset where each source (not simple) sentence is associated with 10 human-written simplifications. It is one of the two datasets for the text simplification task in GEM. It acts as the validation and test set.

**Languages** ASSET contains English text only (BCP-47: `en`).

### B.2 Meta Information

**Dataset Curators** ASSET was developed by researchers at the University of Sheffield, Inria, Facebook AI Research, and Imperial College London. The work was partly supported by Benoît Sagot's chair in the PRAIRIE institute, funded by the French National Research Agency (ANR) as part of the "Investissements d'avenir" program (reference ANR-19-P3IA-0001).

**Licensing Information** Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

**Citation Information** The citation can be found in the ACL Anthology.

**Leaderboard** There is no official leaderboard associated with ASSET.

### B.3 Dataset Structure

**Data Instances** `simplification` configuration: an instance consists of an original sentence and 10 possible reference simplifications; `ratings` configuration: an instance consists in an original sentence, an automatically generated simplification, and a human judgment of quality along one of three axes.

**Data Fields** `original`: an original sentence from the source datasets; `simplifications`: in the simplification config, a set of crowdsourced reference simplifications; `simplification`: in the ratings config, an automatically generated simplification of the original; `aspect`: in the ratings config, how the simplification is evaluated (meaning, fluency, or simplicity); `rating`: a quality rating between 0 and 100

**Data Statistics** ASSET does not contain a training set; many models use WikiLarge (Zhang and Lapata, 2017) for training. For GEM, Wiki-Auto will be used for training the model. Each input sentence has 10 associated reference simplified sentences. The statistics of ASSET are given below. For the input sentences, the validation set has 2000 instances and the test set has 359, for a total of 2359 sentences. Therefore, for the validation set there are 20000 simplifications and for the test set there are 3590 simplifications for a total of 23,590 simplified sentences. The test and validation sets are the same as those of TurkCorpus. The split was random. There are 19.04 tokens per reference on average (lower than 21.29 and 25.49 for TurkCorpus and HSplit, respectively). Most (17,245) of the referece sentences do not involve sentence splitting.

### B.4 Dataset Creation

**Curation Rationale** ASSET was created in order to improve the evaluation of sentence simplification. It uses the same input sentences as the TurkCorpus dataset from (Xu et al., 2016). The 2,359 input sentences of TurkCorpus are a sample of "standard" (not simple) sentences from the Parallel Wikipedia Simplification (PWKP) dataset (Zhu et al., 2010), which come from the August 22, 2009 version of

Wikipedia. The sentences of TurkCorpus were chosen to be of similar length (Xu et al., 2016). No further information is provided on the sampling strategy. The TurkCorpus dataset was developed in order to overcome some of the problems with sentence pairs from Standard and Simple Wikipedia: a large fraction of sentences were misaligned, or not actually simpler (Xu et al., 2016). However, TurkCorpus mainly focused on *lexical paraphrasing*, and so cannot be used to evaluate simplifications involving *compression* (deletion) or *sentence splitting*. HSplit (Sulem et al., 2018), on the other hand, can only be used to evaluate sentence splitting. The reference sentences in ASSET include a wider variety of sentence rewriting strategies, combining splitting, compression and paraphrasing. Annotators were given examples of each kind of transformation individually, as well as all three transformations used at once, but were allowed to decide which transformations to use for any given sentence. An example illustrating the differences between TurkCorpus, HSplit and ASSET is given below:
*Original*: He settled in London, devoting himself chiefly to practical teaching.
*TurkCorpus*: He rooted in London, devoting himself mainly to practical teaching.
*HSplit*: He settled in London. He devoted himself chiefly to practical teaching.
*ASSET*: He lived in London. He was a teacher.

**Communicative Goal** The goal is to communicate the main ideas of source sentence in a way that is easier to understand by non-native speakers of English. This could be done by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones.

**Source Data: Initial Data Collection and Normalization** Data from TurkCorpus (Xu et al., 2016)

**Source Data: Who are the source language producers?** The dataset uses language from English Wikipedia (August 22, 2009 version): some demographic information is provided here.

**Annotations: Annotation process** The instructions given to the annotators are available here.

**Annotations: Who are the annotators?** Reference sentences were written by 42 workers on Amazon Mechanical Turk (AMT). The requirements for being an annotator were: (1) passing a qualification test (appropriately simplifying sentences), (2) being a resident of the US, UK or Canada, (3) having a HIT approval rate over 95%, and over 1000 HITs approved. Out of 100 workers, 42 passed the qualification test. No other demographic or compensation information is provided in the ASSET paper.

**Personal and Sensitive Information** Since the dataset is created from English Wikipedia (August 22, 2009 version), all the information contained in the dataset is already in the public domain.

### B.5 Changes to the Original Dataset for GEM

No change.

### B.6 Considerations for Using the Data

**Social Impact of the Dataset** The dataset helps move forward the research towards text simplification by creating a higher quality validation and test dataset. Progress in text simplification in turn has the potential to increase the accessibility of written documents to wider audiences.

**Impact on Underserved Communities** The dataset is in English, a language with many resources.

**Discussion of Biases** The dataset may contain some social biases, as the input sentences are based on Wikipedia. Studies have shown that the English Wikipedia contains both gender biases (Schmahl et al., 2020) and racial biases (Adams et al., 2019).

**Other Known Limitations** The dataset is limited to a small subset of topics present on Wikipedia.

### B.7 Getting started with in-depth research on the task

The dataset can be downloaded from the original repository (here) or be used via HuggingFace and TFDS. Recent supervised (Martin et al., 2019, Kriz et al., 2019, Dong et al., 2019, Zhang and Lapata, 2017) and unsupervised (Martin et al., 2020, Kumar et al., 2020, Surya et al., 2019) text simplification models can be used as baselines. A common metric for automatic evaluation is SARI (Xu et al., 2016).