A ILLUSTRATIONS FOR DATASET CARDS SUGGESTED BY HUGGING FACE COMMUNITY

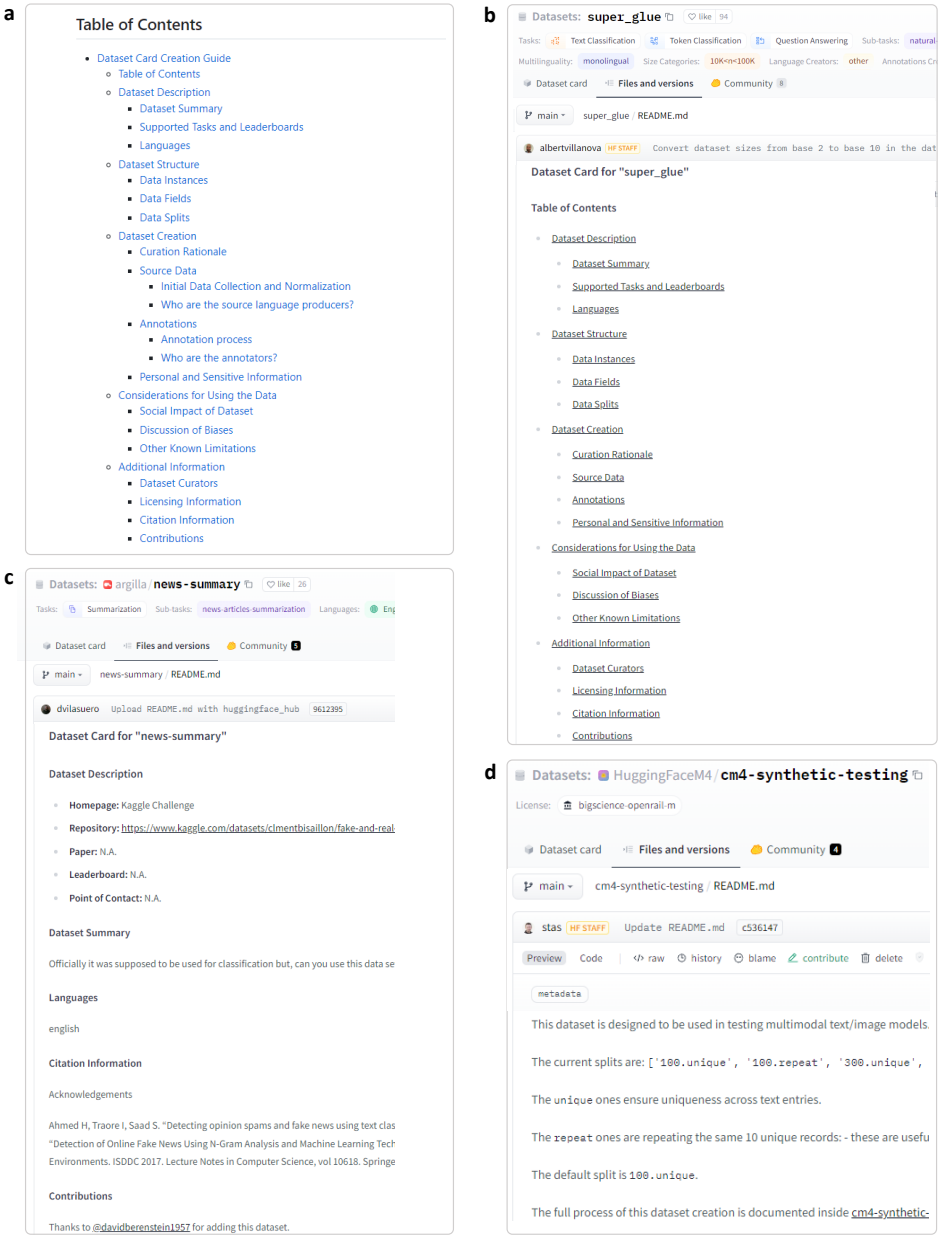

Figure S1: **Illustration of Adherence to Community-Endorsed Dataset Card.** (**a**) *Community-Endorsed Dataset Card Struture:* Hugging Face community provides a suggested dataset card structure, which contains five main sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Considerations for Using the Data*, and *Additional Information*. (**b**) *Example of a Dataset Card Conforming to the Community Guidelines:* A dataset card is considered to conform to the community guidelines when it includes the five main sections outlined in the community guidelines, with the corresponding content provided for each section. (**c**) *Example of Dataset Cards Not Following Community Guidelines (1):* A dataset card is considered non-conforming if it omits any of the five main sections provided in the suggested dataset card structure. (**d**) *Example of Dataset Cards Not Following Community Guidelines (2):* This dataset card contains only a few words and does not follow the structure at all.

B METHOD

B.1 ACCESSING AND PARSING DATASET CARDS

In this work, we analyze datasets hosted on Hugging Face, a popular platform that provides a wealth of tools and resources for AI developers. One of its key features is the Hugging Face Hub API, which grants access to a large library of pre-trained models and datasets for various tasks. With this API, we obtained all 24,065 datasets hosted on the Hub as of March 16th, 2023.

Dataset cards are Markdown files that serve as the README for a dataset repository. They provide information about the dataset and are displayed on the dataset's homepage. We downloaded all dataset repositories hosted on Hugging Face and extracted its README file to get the dataset cards. For further analysis of the documentation content, we utilized the Python package mistune (`https://mistune.readthedocs.io/en/latest/`) to parse the README file and extract the intended content. The structure of dataset cards typically consists of five sections: *Dataset Description*, *Dataset Structure*, *Dataset Creation*, *Additional Information*, and *Considerations for Using the Data*, as recommended by Hugging Face community. Examples of dataset cards, as shown in **Fig. S1**, illustrate the essential components and information provided by dataset cards. We identified and extracted different types of sections through parsing and word matching of the section heading. A significant 84% of the section titles in the 7,433 dataset cards matched one of the 27 titles suggested by the HuggingFace community using the exact keyword matching. This strong alignment underscores the effectiveness of exact keyword matching as an analytical tool.

B.2 HUMAN-ANNOTATED DATASET CARD EVALUATION METHODOLOGY AND CRITERIA

We conducted an evaluation on a sample of 150 dataset cards from a total of 7,433. The assessment involved five human annotators to evaluate the dataset cards, who are PhD students with a solid background in AI fields such as NLP, Computer Vision, Human-AI, ML, and Data Science. Their extensive experience with datasets ensured a deep understanding of dataset documentation. To confirm the reliability of our evaluation, we randomly sampled 30 dataset cards for the annotators to assess and achieved an Intraclass Correlation Coefficient (ICC) of 0.76, which is considered a good agreement (Koo & Li, 2000). This high level of agreement, combined with the annotators' deep expertise in AI research, substantially reinforces the trustworthiness of the annotation results. We focused on seven key aspects of the dataset cards drawing from prior research in dataset documentation and the Hugging Face community-endorsed dataset card:

Aspect	**Description**
Structural Organization	How well is the documentation structured with headings, sections, or subsections?
Content Comprehensiveness	How comprehensive is the information provided in the documentation?
Dataset Description	How effectively does the documentation describe the dataset?
Dataset Preprocessing	How well does the documentation describe any preprocessing steps applied to the data?
Usage Guidance	How well does the documentation offer guidance on using the dataset?
Additional Information	How well does the documentation provide extra details such as citations and references?

Table S1: **Descriptions of Evaluation Aspects**

Each aspect received a score on a scale from 0 to 5, with the following score metrics:

Score	Description	Comment
5	Exceptionally comprehensive and effective	Covers all subsections in detail
4	Very good and thorough	Includes many subsections comprehensively
3	Moderately satisfactory	Covers some subsections adequately
2	Insufficient	Provides a basic, general overview
1	Poor and inadequate	Offers minimal, vague content
0	Absent	Lacks relevant content

Table S2: **Metrics of the Scores**

C Additional Analysis of *Usage* Section

Among 7,433 dataset cards, there are 567 dataset cards uploaded by 52 distinct practitioners that contain a *Usage* section, instructing how to use the dataset through text and codes. A specific example of *Usage* section is from ai4bharat/naamapadam, which has 469 downloads and has a *Usage* section to instruct how to use the dataset (**Fig. S2**).

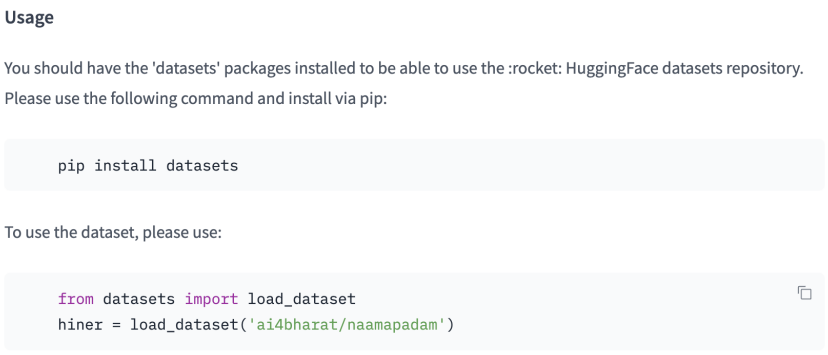

Figure S2: **Example of a *Usage* Section**

Intuitively, a *Usage* section could give users quick instructions on how to use the dataset, which could make the dataset more accessible, transparent, and reproducible. To verify this intuition, we conduct an experiment to quantify how the *Usage* section will affect the dataset's popularity.

In our experiment, we trained a BERT (Devlin et al., 2018) Model using the content of dataset cards and their corresponding download counts. To ensure comparability, the download counts were normalized to a range of [0,1] and stratified monthly based on the dataset's creation time. This ranking system assigned a rank of 1 to the dataset with the highest downloads within a given month, and a rank of 0 to the dataset with the lowest downloads.

Using the dataset card content, the trained BERT Model predicted the download counts. Subsequently, we conducted a test using 567 dataset cards that included a *Usage* section. For this test, we deliberately removed the *Usage* section from the dataset cards and employed the BERT Model to predict the download counts for these modified cards. The resulting predictions are summarized in **Table. S3**. The average predicted score of downloads after removing the *Usage* section is 0.0185 lower compared to the original dataset card. This indicates a decrease in the number of downloads, highlighting the negative impact of not including a *Usage* section.

In future research, it would be valuable to further investigate the effect of adding a *Usage* section to the dataset cards that do not have one originally. A randomized controlled trial (RCT) experiment could be conducted to assess whether the inclusion of a *Usage* section leads to an increase in downloads.

Condition	Predicted Score of Downloads
With *Usage* Section	0.3917
Without *Usage* Section	0.3732
Change in Score	-0.0185

Table S3: **Predicted Impact of *Usage* Section on Dataset Downloads.** This table presents a comparative analysis of predicted download scores for dataset cards, distinguishing between those that include a *Usage* Section and those from which it has been removed. It indicates a potential decrease in download rates following the removal of the *Usage* Section.

D Optional Metrics for Datasets

In our analysis, we employ downloads as a metric to gauge the popularity of the dataset. Numerous factors can influence the download count, including the dataset's publication date and its associated research field. Moreover, aside from dataset downloads, we can incorporate other indicators of dataset popularity, such as the count of models utilizing the datasets and the corresponding download counts.

To address the concerns of factors that might affect downloads, we expanded our dataset analysis by extracting more metadata from the Hugging Face dataset information. We collected data such as the models utilizing the corresponding dataset, the total number of downloads for these models, and the dataset's task domain. The primary dataset tasks recognized by Hugging Face encompass Multimodal, Computer Vision, Natural Language Processing, Audio, Tabular and Reinforcement Learning. Among the total of 7,433 dataset cards, 1,988 are categorized as NLP dataset cards, 198 are related to computer vision, and 102 pertain to multimodal datasets. We proceeded with additional analysis by employing the following metrics:

1. We integrated dataset downloads (*"direct usage"*) with the downloads of models employing the dataset (*"secondary usage"*).
2. A time range (measured in months) was selected, encompassing dataset cards created within the designated time frame and specified task domain.
3. Selected dataset cards were ranked within each domain for each time range and then normalized to a range of $[0, 1]$.

By adopting this approach, we were able to compare dataset cards created in the same month and task domain, assessing them based on the metrics of direct and secondary usage metrics. We conducted a word count analysis using this new metric and attained results consistent with our prior analysis that datasets with higher rankings tend to have longer dataset cards, as shown in **Fig. S3**.

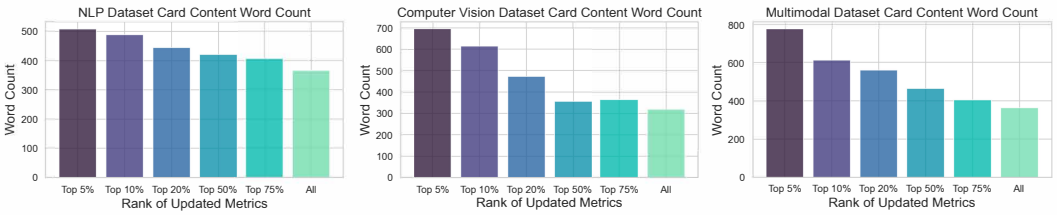

Figure S3: **Word Count Variation Based on Direct and Secondary Usage Rankings.** This figure demonstrates the relationship between the length of dataset cards and their rankings in terms of direct and secondary usage. It reveals a distinct pattern: dataset cards with higher rankings tend to have a greater word count, suggesting a correlation with more thorough and detailed content.

The finding enables us to contemplate an alternative metric option, factoring in publication time, research area, and secondary dataset usage. However, the results remain aligned with our previous analysis, which solely considered download counts, highlighting the reasonableness of using download counts as metrics.

E APPLICABILITY ACROSS PLATFORMS: ADAPTING TO GITHUB

Our study demonstrates strong potential for application across various platforms. The foundational format of Hugging Face's dataset cards, essentially README files, is a prevalent documentation standard shared by many platforms, notably GitHub. This commonality implies that our approach to parsing and analyzing dataset cards can be readily adapted for broader studies. To illustrate, we present an example of how our analysis methodology can be effectively applied to GitHub, a widely recognized open-source platform for data and code sharing.

Our expanded analysis involved sourcing datasets from a GitHub repository of Papers With Code[1]. We chose repositories linked to dataset-relevant papers and processed their README files using the pipeline proposed in our paper on Hugging Face dataset card analysis. This exploration revealed a more varied structure in GitHub's dataset cards. For example, 57% of the section titles on GitHub are unique, compared to just 3% on Hugging Face. Due to their specificity, we excluded these unique sections and created a categorization list based on Hugging Face's community-endorsed dataset card structure, mapping GitHub's titles through keyword matching. This method successfully categorized 74% of GitHub's section titles.

As shown in **Table. S4**, our analysis reveals that both platforms excel in *Dataset Description* and *Additional Information* sections but underperform in *Dataset Creation* and *Considerations for Using the Data*, underscoring points raised in our paper. A notable difference is GitHub's lower emphasis on *Dataset Structure*, highlighting the potentially positive impact of Hugging Face's community-endorsed dataset structure. Furthermore, the prevalence of *Usage* and *Experiment* sections on GitHub, absent in Hugging Face, highlights the practical value of these sections in promoting the usability of datasets. Adopting these sections, as suggested in our paper, could enrich the structure of Hugging Face's dataset cards, making them more comprehensive and practically useful.

These results indicate our method's adaptability to other platforms and provide a benchmark for evaluating dataset documentation elsewhere. The insights from our Hugging Face study can guide the categorization and enhancement of dataset documentation across various platforms, especially in the current situation that most other platforms don't have a standardized dataset card structure.

Section Type	**GitHub**	**Hugging Face**	**Description**
Dataset Description	0.62	0.46	Summary, leaderboard, languages, etc.
Dataset Structure	0.09	0.34	Format, fields, splits, etc.
Dataset Creation	0.08	0.15	Motivation, collection procedures, etc.
Considerations for Using the Data	0.02	0.08	Limitations, biases, disclaimers, etc.
Additional Information	0.62	0.58	Citations, acknowledgements, licensing, etc.
Experiment	0.57	-	Model experiments, training, evaluation on the dataset, etc.
Usage	0.38	-	Instructions for setup, installation, requirements, etc.

Table S4: **Comparison of Fill-out Rate of Dataset Documentation on GitHub and Hugging Face.** Dataset cards from both GitHub and Hugging Face perform well in the *Dataset Description* and *Additional Information* sections, but fall short in the *Dataset Creation* and *Considerations for Using the Data* sections. While GitHub places less emphasis on *Dataset Structure*, it shows a higher occurrence of *Usage* and *Experiment* sections.

[1] https://github.com/paperswithcode/paperswithcode-data

F ADDITIONAL FIGURES AND TABLES

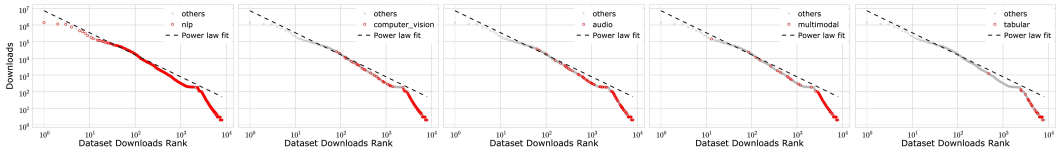

Figure S4: **Power Law Distribution Patterns in Dataset Usage across Task Domains.** This figure illustrates the dataset usage distribution within each task domain, demonstrating a consistent power law distribution, despite the variations in the number of datasets across different domains.

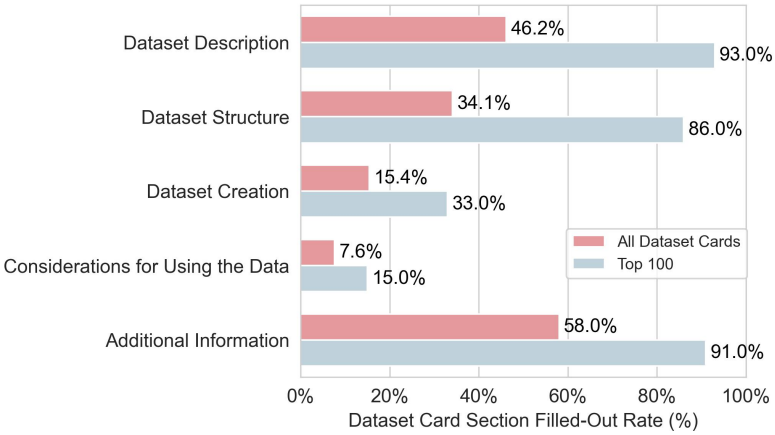

Figure S5: **Highly Downloaded Dataset Cards Exhibit Greater Completion across All Sections.** This figure indicates that the top 100 downloaded dataset cards exhibit a higher completion rate compared to all dataset cards in the sections recommended by the Hugging Face community. However, there is a consistently low completion rate in the *Dataset Creation* and *Considerations for Using the Data sections*, regardless of the dataset cards' popularity.

Category	**Description**	**Dataset Card Number**	**Adherence to Guidelines**	**Avg. Word Count**
Industry organization	Companies (e.g. Hugging Face, Facebook)	2,527	**0.34**	219
Academic organization	Universities, Research Labs (e.g. Stanford CRFM, jhu-clsp)	985	0.31	**427**
Community	Non-profit Communities (e.g. allenai, bio-datasets)	1,387	0.27	190
Industry professional	Engineers, Industry Scientists	985	0.25	256
Academic professional	Students, Postdocs, Faculty	672	0.16	180
All dataset cards	7,433 dataset cards analyzed	7,433	0.29	234

Table S5: **Differences in the Practices of Dataset Documentation across Creators from Different Backgrounds.** This table highlights the diverse documentation practices across creators from different backgrounds. Industry organizations, with the most creators, adhere to the guidelines best. Academics, though fewer, offer the most comprehensive documentation, while academic professionals exhibit lower guideline adherence and shorter word counts. The information about these creators is gathered from their linked GitHub, Twitter, and personal websites on their Hugging Face profiles.