KDD2025
August 3–7, 2025

DCA-Bench: A Benchmark for Dataset Curation Agents

Benhao Huang[1], Yingzhuo Yu[2], Jin Huang[3], Xingjian Zhang[3], Jiaqi W. Ma[2]

[1] Carnegie Mellon University, USA

[3] University of Michigan, USA

[2] University of Illinois Urbana-Champaign, USA

Motivations and Background

kaggle

BIG-bench

TensorFlow

Wrong Labels

Ethical/Legal Risk

Data Problems

Insufficient Documentations

Document Problems

Wrong Format

Infrastructure problem

Inter/Cross File Discrepancy

Inappropriate files

There are so many problems !

Real-World Examples of Issues in Dataset Repositories

Example 2 An issue example reported on BIG-Bench that involves a discrepancy between dataset files.

Title `Miss aligned static information`

Meta-Info

- ID: `80d6db6a-6cbf-4261-8d13-3244e7fb54fd`
- Platform: `BIG-Bench`
- Issue Type: single-issue & multi-file
- Issue Tags: cross-file-discrepancy document-problem/wrong-info
- Source: https://github.com/google/BIG-bench/pull/498

Content

`The stastic info in README.md is not aligned with the actual data file. There are 190 stories rather than 194 stories; 99 "Yes" rather than 100 "Yes"; 91 "No" rather than 94 "No".`

Involved Files

1. name: `task.json`
   - context: `the number of datapoints in data files.`
2. name: `README.md`
   - context: `We collected 194 stories from 30 papers published in the span of 1989 to 2021. Each story has a causal judgment question associated with it with a "Yes" or "No" answer. We carefully balanced the dataset -- there are 100 "Yes" answers (52%) and 94 "No" answers (48%). Each paper that we collected from has conducted rigorous human experiments. We follow a simple binarization strategy to reflect the majority of human agreement and use it as the ground truth to evaluate the AI model.`

Example: Cross-file discrepancy — when documentation and data go out of sync

Example 3 An issue example which has a wrong target label that needs precise factual knowledge to discern

Title `Error in 118th Congress data`

Meta-Info

- ID: `51e12546-8bf3-473c-9ed6-f85d63c357ce`
- Platform: `FiveThirtyEight`
- Issue Type: single-issue & multi-file
- Issue Tags: data-problem/hidden-corruption, data-problem/wrong-value
- Source: https://github.com/fivethirtyeight/data/issues/336

Content

`The "congress-demographics" data includes Benjamin Eric Sasse as being a member of the 118th Congress but he resigned after the 117th.`

Involved Files

- name: `data_aging_congress.csv`
- context: `The "congress-demographics" data includes Benjamin Eric Sasse as being a member of the 118th Congress but he resigned after the 117th.`

- Confusing and risky when using the dataset
- Non-trivial effort needed to detect

Example: Factual data corruption — requiring real-world knowledge to catch

Motivations and Background

- Today's AI agents have shown impressive capabilities across a wide range of **complex tasks**—such as coding, web navigation, and deep reasoning.

🧠 *Can we leverage AI agents to <u>detect hidden issues within existing dataset repositories</u>?*

"Dataset Curator"

"Dataset Curation"

Related Works and Challenges

KELVIN WATERS · POSTED 2 YEARS AGO

Boston House Prices B feature is RACIST

B: 1000(Bk−0.63)2 where Bk is the proportion of blacks by town

No other race is featured in this dataset. Red-lining anyone?

Example 1 An issue example reported on Kaggle which involves racial bias

Title `Boston House Prices B feature is RACIST`

Meta-Info

- ID: `7e8f31cb-8c2a-4676-b3d4-941a64184a26`
- Platform: `Kaggle`
- Issue Type: single-issue & multi-file
- Issue Tags: ethical-legal-risk document-problem
- Source: https://www.kaggle.com/datasets/vikrishnan/boston-house-prices/discussion/429030

Content
`B: 1000(Bk-0.63)2 where Bk is the proportion of blacks by town No other race is featured in this dataset. Red-lining anyone?`

Involved Files

- name: `datacard.md`
- context: `PTRATIO:pupil-teacher ratio by town 12. B: 1000(Bk-0.63)2 where Bk is the proportion of blacks by town 13. LSTAT:% lower status of the population`

Example: Ethical Concerns. Rule-based scripts cannot discover this, while AI has the potential to detect such risks

Relevant work falls into a few categories

- Rule-based scripts for specific issues
- Model-based pipelines for data scoring or filtering
- Agent-based systems for software tasks

Analysis of the Task

- Rule-based scripts can only detect **predefined and known patterns**
- Our task focuses on **detecting issues (Unknown),** rather than fixing known issues.
- Detection is a prerequisite for any meaningful fix
- **No clear ground truth**, making supervision and evaluation difficult

Framing Our Research Question and Key Challenges

How well can Curators detect hidden issues in existing open dataset repositories?

🔧 Challenge 1: Test Case Design

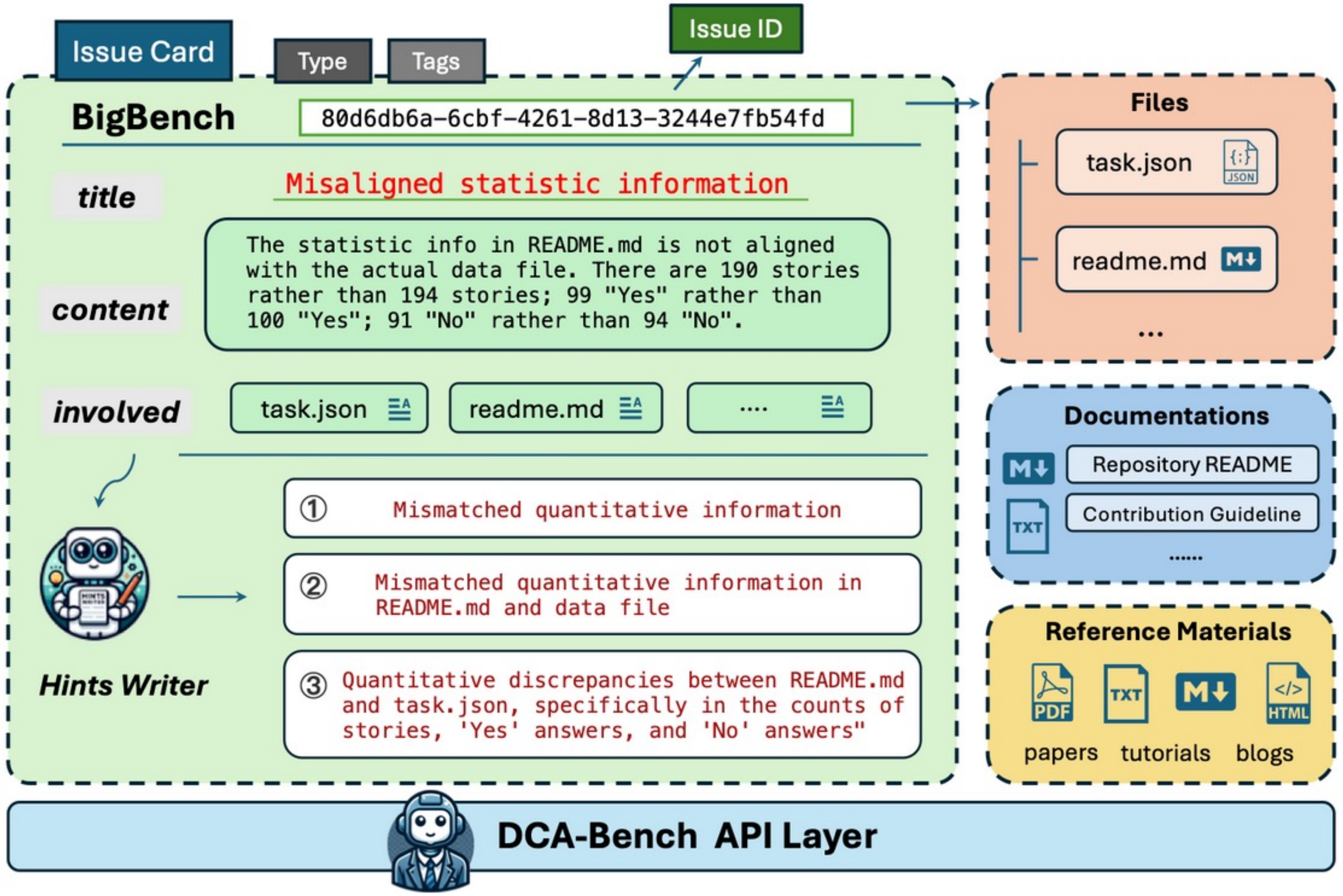
- Dataset issues are subtle, undocumented, and highly varied
- Need realistic, diverse, and manually verified test cases
- Requires human curation, verification, and domain knowledge

🧪 Challenge 2: Evaluation Protocol

- No standard ground truth for what counts as a correct detection
- Evaluation must be scalable and reliable:
  - Alignment with human experts
  - Minimal bias

These challenges motivate the design **of DCA-Bench**, a benchmark and evaluation framework for studying dataset-curation agents.

Introducing DCA-Bench



- 221 real-world curation cases
- 8 dataset platforms
- 4 issue categories
- 18 fine-grained tags

Statistic		Number
Sample-Level	#Samples	221
	Avg. #Files/Sample	2.13
	Avg. #Tokens/Sample	3.58×10^6
Type-Level	Single-Issue Single-File	61
	Single-Issue Multi-File	100
	Multi-Issue Single-File	14
	Multi-Issue Multi-File	46
Tag-Level	data-problem	197
	document-problem	83
	infrastructure-problem	19
	ethical/legal-risk	10

Category	Number	Sub-category	Number
typo	18	—	—
wrong-format	14	—	—
inappropriate-file	4	—	—
ethical/legal-risk	10	—	—
internal-discrepancy	21	—	—
cross-file-discrepancy	44	—	—
data-problem	197	wrong-value missing-value data-leakage apparent-corruption hidden-corruption	71 15 2 40 59
document-problem	83	wrong-info insufficient-info	27 52
infrastructure-problem	19	data-access script-code	4 15

Multi-level Hints

h0: No hint provided. In this case, the Curator is required to detect the issue fully on its own.

h1: General description of the issue, without any specific details or hints on the location.

h2: Information about which files are involved in the issue, in addition to information from h1

h3: Partial contextual information about the issue, in addition to information from h2

Example 3 An issue example which has a wrong target label that needs precise factual knowledge to discern

Title `Error in 118th Congress data`

Meta-Info

- ID: `51e12546-8bf3-473c-9ed6-f85d63c357ce`
- Platform: `FiveThirtyEight`
- Issue Type: single-issue & multi-file
- Issue Tags: data-problem/hidden-corruption, data-problem/wrong-value
- Source: `https://github.com/fivethirtyeight/data/issues/336`

Content

`The "congress-demographics" data includes Benjamin Eric Sasse as being a member of the 118th Congress but he resigned after the 117th.`

Involved Files

- name: `data_aging_congress.csv`
- context: `The "congress-demographics" data includes Benjamin Eric Sasse as being a member of the 118th Congress but he resigned after the 117th.`

Hints

h_1 inaccurate data entry

h_2 an inaccurate data entry in a CSV file

h_3 an entry in 'data_aging_congress.csv' inaccurately includes a member as part of the 118th Congress

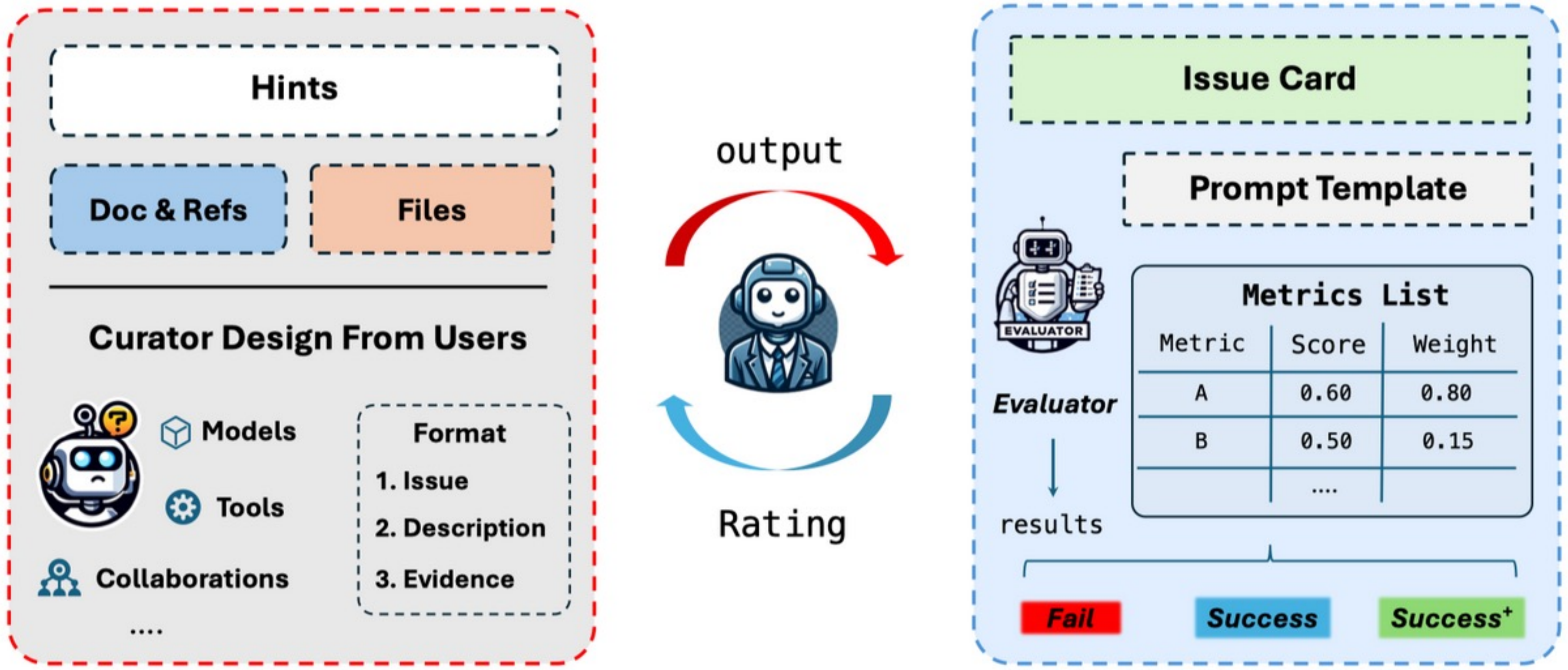
From a higher level hint, the Curator gains more information about the content and location of the issue.

Example: Multi-level hints of an issue which has a wrong target label that needs precise factual knowledge to discern

Evaluation Framework: Scalable & Trustworthy Judging via LLMs

We replace costly human grading with an **LLM evaluator** equipped with carefully designed prompts and majority voting strategies. Results on test cases demonstrate 95% alignment with human experts, confirming its reliability. Additionally, we conduct experiments to showcase its minimal bias (self-preference, length-bias) characteristics in our paper.

Evaluation Protocols of DCA-Bench

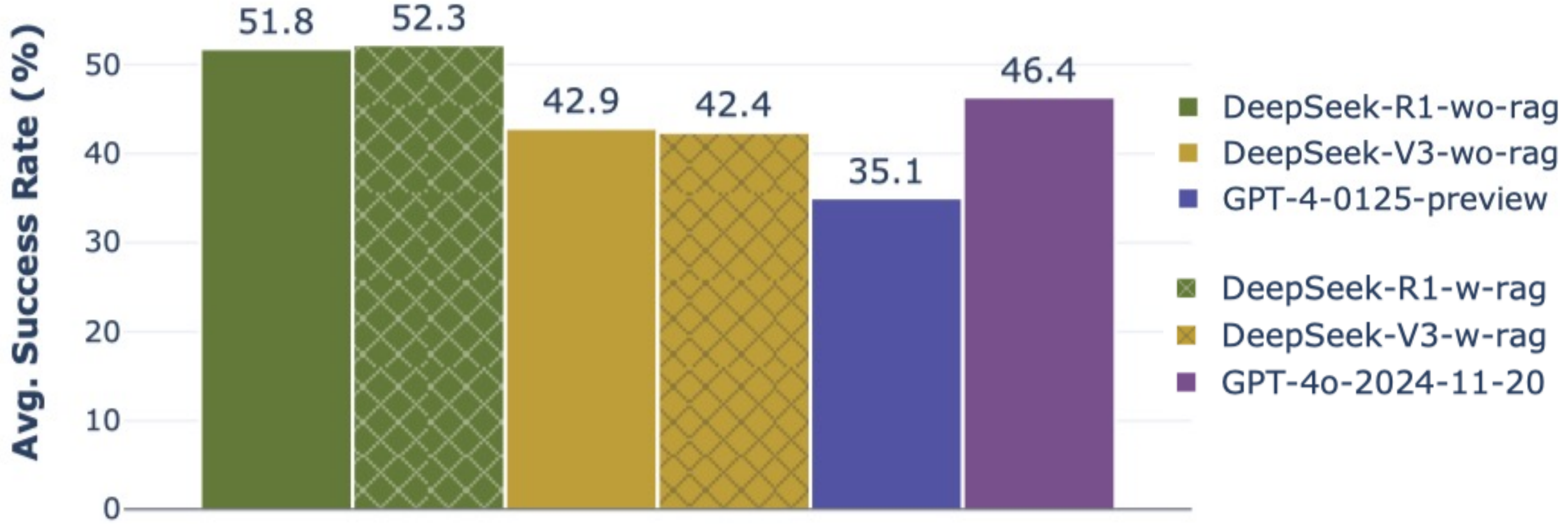


LLM Evaluator Agreement with Human Labels (%)

Model Name	Success Rate / %				
	Accuracy	Precision	Recall	F1 Score	κ Value
gpt-4o-2024-11-20	97.83	100.00	92.59	96.15	94.64
gpt-4-0125-preview	96.74	92.86	96.30	94.55	92.22
gpt-4o-2024-05-13	92.39	81.25	96.30	88.14	82.59
DeepSeek-R1	92.39	81.25	96.30	88.14	82.59
DeepSeek-V3	93.48	88.89	88.89	88.89	84.27
gpt-3.5-turbo	68.48	48.21	100.00	65.06	42.15
Meta-Llama-3-70B-Instruct	69.57	49.09	100.00	65.85	43.68
Meta-Llama-3.3-70B-Instruct	88.04	72.22	96.30	82.54	73.73
o3-mini-2025-01-31	91.30	100.00	70.37	82.61	77.04

Benchmark Results

Model Name	Success Rate / % h_0	h_1	h_2	h_3	*Avg.*
w/o Knowledge RAG					
DeepSeek-R1	29.86	52.04	52.04	73.30	51.81
DeepSeek-V3	15.84	39.82	39.82	76.02	42.87
GPT-4-0125-preview	10.86	27.15	34.84	67.42	35.07
GPT-4o-2024-11-20	20.36	41.63	47.51	76.02	46.38
w/ Knowledge RAG					
DeepSeek-R1	29.41	45.25	56.11	78.28	52.26
DeepSeek-V3	12.22	38.91	42.99	75.57	42.42



- Even some of the most advanced models **uncover barely 30% of issues without hints.** With highest level of hints, none exceed 80%.
- Interestingly, we found the **usage of RAG doesn't guarantee a boost in the performance** in this task.

Limitations

- Limited Coverage
  - Our test cases represent only a portion of real-world dataset issues
- Unlabeled Issues
  - Some problems in test cases may remain undetected
- Text-Only Benchmark
  - Currently excludes multimodal datasets (e.g., image/audio)

Future Works

- Develop stronger and more autonomous curator agents
- Extend DCA-Bench to multimodal datasets
- Create realistic simulation environments for agent training & evaluation

Conclusion

- **We introduce** DCA-Bench**:** a benchmark for testing dataset-curation agents
- Built from **221** real-world data quality issues with **4** hint levels across **8** open platforms
- Tasks focus on **issue detection**, not fixes to known issues with clear target
- LLM-based Evaluator enables scalable and reliable performance assessment
- **Benchmark results show:** current models have potential, but great improvement

remains to be made.

Thanks for Listening

Code Paper Dataset