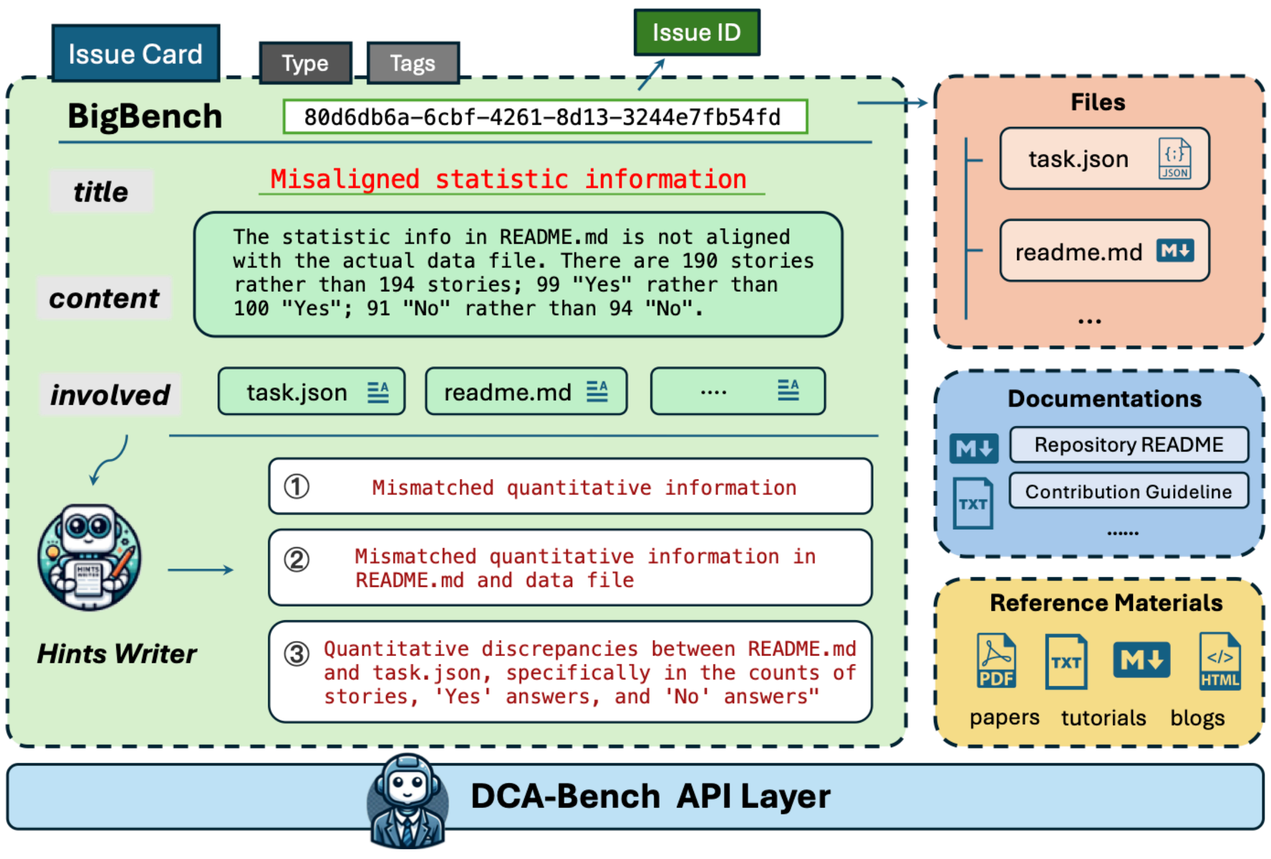
DCA-Bench: A Benchmark for Dataset Curation Agents

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

[1]Carnegie Mellon University [2]University of Illinois Urbana-Champaign

[3]University of Michigan

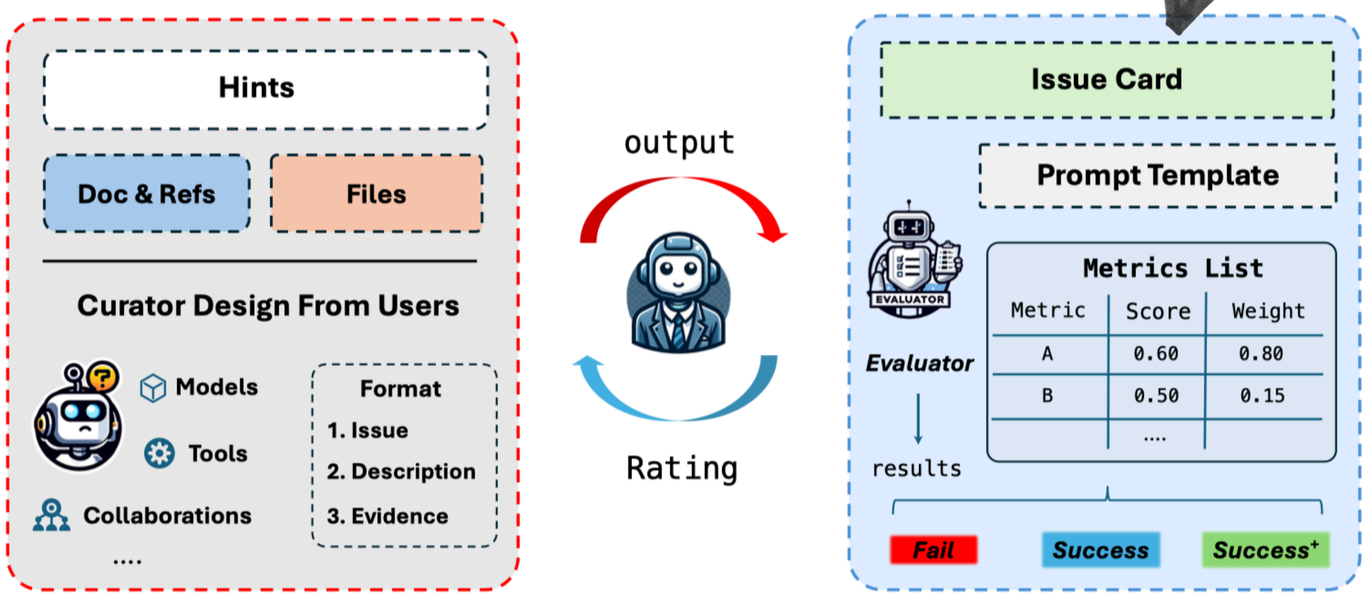


Motivation

- Community-contributed datasets are often affected by **hidden issues like inaccurate annotations, poor documentation and ethical risks.**
- Rule-based curation tools are **too rigid** to detect such subtle problems.
- LLM agents excel in fixing known issues but have not been tested on **discovering hidden ones.**
- **DCA-Bench is the first benchmark to evaluate LLMs' ability to uncover dataset quality issues at scale.**

Approach

- DCA-Bench collects **221 real-world test cases** from 8 popular dataset platforms (Hugging Face, Kaggle, BIG-Bench, etc.). Issues are categorized into **4 types and 18 fine-grained tags** covering data errors, documentation problems, infrastructure issues, and ethical/legal risks and more.
- To make it more manageable and test the Curators' capability in finer granularity, **four levels of hints** (from no hint to detailed guidance) are provided.
- We develop an automatic **evaluation framework** using GPT-4o, achieving high alignment with human annotations.



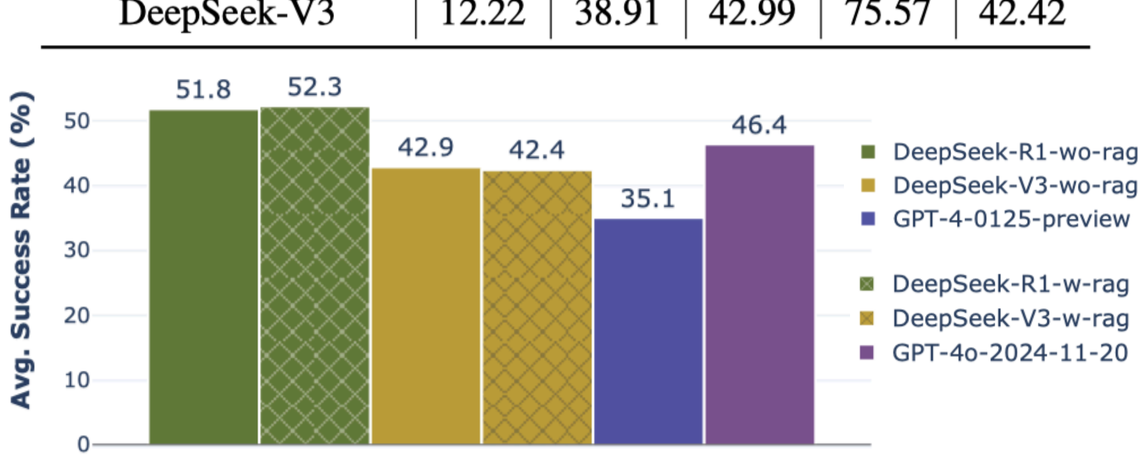
Statistic		Number
Sample-Level	#Samples	221
	Avg. #Files/Sample	2.13
	Avg. #Tokens/Sample	3.58×10^6
Type-Level	Single-Issue Single-File	61
	Single-Issue Multi-File	100
	Multi-Issue Single-File	14
	Multi-Issue Multi-File	46
Tag-Level	data-problem	197
	document-problem	83
	infrastructure-problem	19
	ethical/legal-risk	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	71
		missing-value	15
		data-leakage	2
		apparent-corruption	40
		hidden-corruption	59
document-problem	83	wrong-info	27
		insufficient-info	52
infrastructure-problem	19	data-access	4
		script-code	15

Experiment

Curator's Performance

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



Benchmark Evaluator's Performance

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

1. **The ability to perform evaluation tasks in alignment with human judgments is of crucial value.**
2. **Advanced models struggle to uncover hidden issues without prior indication that problems exist.**

Code Paper Dataset