TUM School of Computation, Information and Technology
Chair of Theoretical Information Technology
IBM Research

Fixing It in Post: A Comparative Study of LLM Post-Training Data Quality and Model Performance

NeurIPS 2025 - Track on Datasets and Benchmarks
Spotlight Paper

Aladin Djuhera, Swanand Ravindra Kadhe, Syed Zawad,
Farhan Ahmed, Heiko Ludwig, Holger Boche

1. Motivation: Post-Training Datasets are a Mess!

Proprietary Datasets

- mostly inaccessible
- no documentation or details on curation recipes

Open Datasets

- publicly accessible
- high-performance

… but

- **no systematic performance comparisons** (across models and datasets)
- **lack of transparency** (curation recipes)
- **no sample-level tags** (difficult reusability)

How to curate and/or choose good datasets in practice then?

2. Contribution of this Work

Side-by-Side Performance Evaluation for

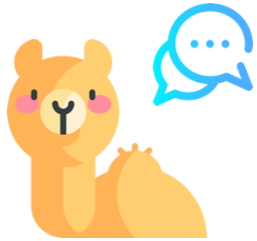
- Tulu-3-SFT-Mix and SmolTalk SFT datasets
- fixed models and hyperparameters
- 14 diverse benchmarks

High-Quality Annotations with Magpie for

- conversational structure (single- vs. multi-turn)
- prompt and response quality
- task categories, language, safety

New Dataset: TuluTalk

- quality- and task-based curation recipe
- 14% - 23% smaller + better performance
- 100% transparent and annotated



3. Dataset Analysis

Task Diversity

- Tulu: STEM-oriented
- SmolTalk: conversational

→ **complementary task distributions!**

Conversation Lengths

- Tulu: single-turn
- SmolTalk: multi-turn

Input (Prompt) Quality

- mostly good and excellent
- non-negligible “bad samples”

→ **potential redundancy**

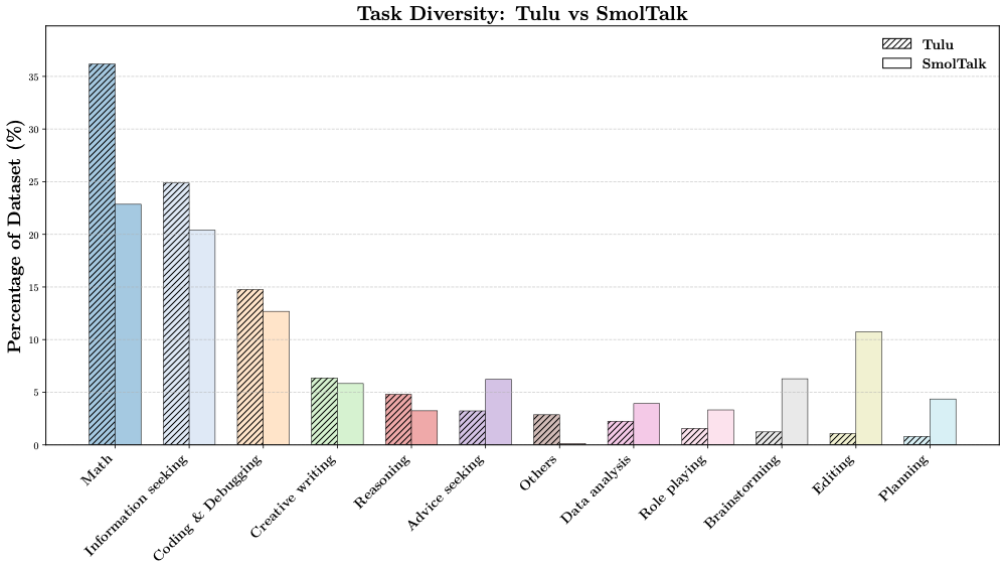


Fig.1: Task Distribution for Tulu and SmolTalk.

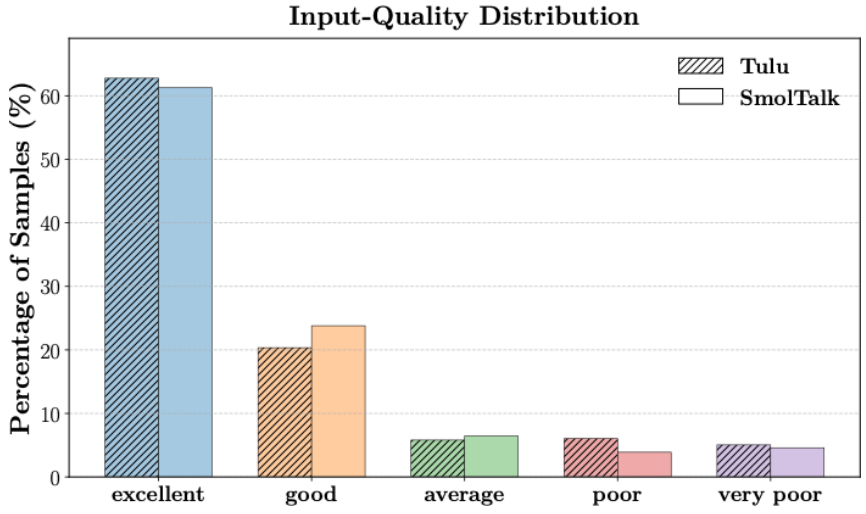
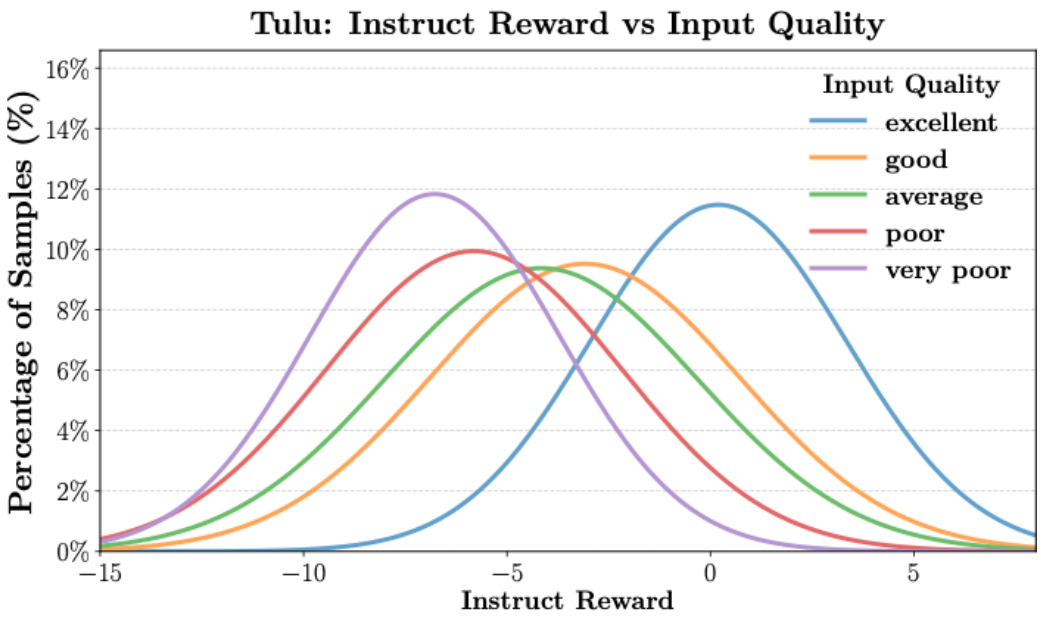


Fig.2: Input Quality Distribution for Tulu and SmolTalk.

3. Dataset Analysis

Input Quality vs. Instruct (Response) Reward

- single-turn: high input quality → high response quality
- multi-turn: no correlation



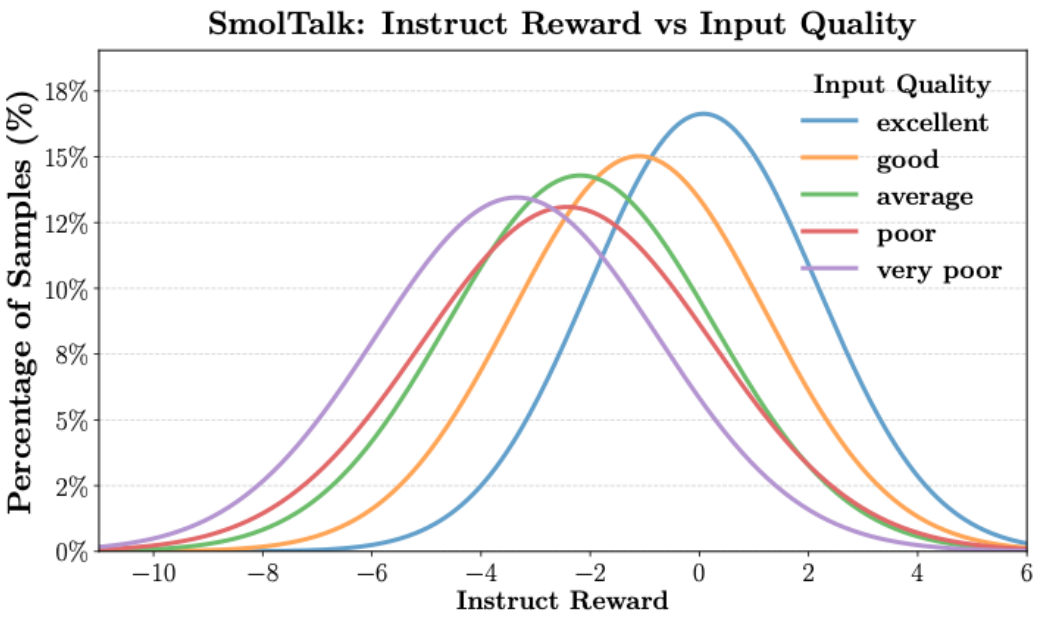


Fig.3: Input Quality vs. Instruct Reward for Tulu (left) and SmolTalk (right).

4. Quality- and Task-Aware Data Curation

Step 1: Quality-Based Filtering

- very good / excellent input quality
- high response reward

Step 2: Task-Aware Adaptation

- add more instruction following samples
- improve task diversity

→ **increases cross-task performance**

Result: Faster Training + Higher Performance

- smaller datasets imply faster training
- removing redundancies improves accuracy

Quality- and Task-Aware Data Curation Recipe

Input: Annotated dataset $\mathcal{D}$ with Magpie tags for input quality (`input_quality`), single-turn/multi-turn response quality (`st_reward`/`mt_reward`), and task category (`task_category`); task diversity threshold τ.
Output: Curated subset $\mathcal{D}_c$ that is both high-quality and task-diverse.

Recipe:

1. Compute quantiles:

$$Q_1^e,\ Q_2^e \leftarrow \text{1st/2nd quantiles of } \{S[\texttt{st\_reward}] \mid S[\texttt{input\_quality}] = \texttt{excellent},\ S[\texttt{turn}] = \texttt{single\_turn}\},$$
$$Q_3^g \leftarrow \text{3rd quantile of } \{S[\texttt{st\_reward}] \mid S[\texttt{input\_quality}] = \texttt{good},\ S[\texttt{turn}] = \texttt{single\_turn}\}.$$

2. For each $S \in \mathcal{D}$, add S to $\mathcal{D}_c$ if

$$S[\texttt{input\_quality}] = \texttt{excellent} \wedge \Big((S[\texttt{turn}] = \texttt{multi\_turn} \wedge S[\texttt{mt\_reward}] = 5) \vee (S[\texttt{turn}] = \texttt{single\_turn} \wedge S[\texttt{st\_reward}] > Q_2^e)\Big).$$

3. Let $\mathcal{C}$ be the set of task categories whose coverage in $\mathcal{D}_c$ drops by more than $\tau\%$ relative to $\mathcal{D}$.
4. For each $S \in \mathcal{D} \setminus \mathcal{D}_c$, add S to $\mathcal{D}_c$ if

$$S[\texttt{task\_category}] \in \mathcal{C} \wedge \Big(\underbrace{S[\texttt{input\_quality}] = \texttt{excellent} \wedge \big((S[\texttt{turn}] = \texttt{multi\_turn} \wedge S[\texttt{mt\_reward}] = 4) \vee (S[\texttt{turn}] = \texttt{single\_turn} \wedge Q_1^e < S[\texttt{st\_reward}] < Q_2^e)\big)}_{\text{high-quality fallback}}$$
$$\vee \Big(\underbrace{S[\texttt{input\_quality}] = \texttt{good} \wedge \big((S[\texttt{turn}] = \texttt{multi\_turn} \wedge S[\texttt{mt\_reward}] = 5) \vee (S[\texttt{turn}] = \texttt{single\_turn} \wedge S[\texttt{st\_reward}] > Q_3^g)\big)}_{\text{diversity boost}}\Big)\Big)$$

5. Numerical Results

23% smaller than SmolTalk and 14% smaller than Tulu!

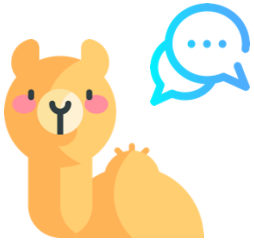
	Llama-3.1-8B					SmolLM2-1.7B				
Benchmark	Base	Tulu	SmolTalk	Orca	TuluTalk	Base	Tulu	SmolTalk	Orca	TuluTalk
Knowledge										
MMLU (5-shot)	**65.03**	62.90	62.88	62.64	63.91	50.09	49.71	47.88	**51.65**	49.34
MMLU-Pro (5-shot)	**32.71**	28.73	31.76	31.89	30.17	21.26	19.61	20.37	**23.40**	20.67
TruthfulQA (0-shot)	45.22	46.41	**55.74**	52.08	53.16	36.61	44.04	**44.74**	42.84	43.65
GPQA (0-shot)	37.96	**42.86**	38.49	40.21	40.62	**34.66**	33.33	33.86	33.20	33.28
Reasoning										
ARC-C (25-shot)	54.69	54.61	**59.04**	53.07	57.42	**51.54**	44.54	48.46	46.25	47.27
BBH (3-shot)	**46.48**	39.06	45.50	45.74	43.50	34.04	36.66	37.81	38.05	**38.33**
MuSR (0-shot)	37.96	**42.86**	38.49	40.21	40.62	**34.66**	33.33	33.86	33.20	33.28
Commonsense										
HellaSwag (10-shot)	61.44	60.87	61.54	60.60	**62.98**	**53.65**	51.01	52.10	51.61	51.36
WinoGrande (5-shot)	76.87	76.64	77.19	71.19	**79.22**	68.19	**65.90**	65.27	64.96	66.06
Instruction Following										
IF-Eval (0-shot)	12.45	74.09	74.51	57.73	**74.84**	23.91	60.25	56.83	35.17	**60.85**
Math										
GSM8K (5-shot)	50.64	74.37	74.75	60.58	**74.84**	29.64	49.43	52.46	29.34	**54.13**
MATH (4-shot)	5.97	**12.31**	10.42	11.86	11.96	2.64	**6.27**	5.89	5.82	6.16
Code										
HumanEval (pass@1)	34.76	**58.54**	54.51	51.37	56.49	0.61	**1.83**	**1.83**	0.61	**1.83**
HumanEval+ (pass@1)	28.66	**45.37**	44.27	40.29	44.33	0.61	**1.83**	**1.83**	0.61	**1.83**
Leaderboards										
Open LLM Leaderboard 1	58.98	62.63	65.19	60.03	**65.26**	48.29	50.77	51.82	47.78	**51.97**
Open LLM Leaderboard 2	27.84	37.47	38.24	36.05	**38.40**	24.14	30.66	30.39	27.67	**31.16**
Overall	41.74	50.32	51.38	47.72	**51.62**	31.13	35.16	35.49	32.42	**35.89**

Tab.1: SFT Results for Llama and SmolLM models for Tulu, SmolTalk, Orca, and TuluTalk.

6. Conclusion

What we did

- used Magpie to annotate open SFT datasets
- developed a principled data curation strategy



→ **TuluTalk: smaller + high-performance**

What we showed

- quality > quantity
- data curation is task dependent

What we gave the Community

- annotated datasets
- more robust Magpie annotation pipeline