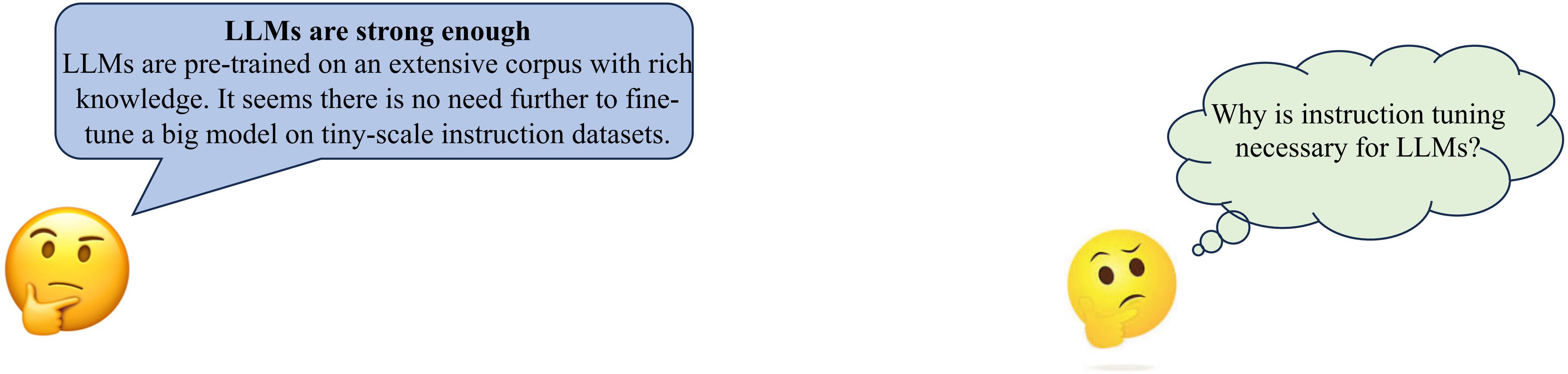
MUFFIN: Curating Multi-Faceted Instructions for Improving Instruction-Following

[♠]Renze Lou, [♢]Kai Zhang, [♣]Jian Xie, [♥]Yuxuan Sun, [♠]Janice Ahn, [†]Hanzi Xu, [♢]Yu Su, [♠]Wenpeng Yin

[♠]The Pennsylvania State University, [♢]The Ohio State University, [♣]Fudan University, [♥]Westlake University, [†]Temple University

{renze.lou, wenpeng}@psu.edu


LLMs are strong enough
LLMs are pre-trained on an extensive corpus with rich knowledge. It seems there is no need further to fine-tune a big model on tiny-scale instruction datasets.
Why is instruction tuning necessary for LLMs?

Purpose of Instruction Tuning

Instruction-following

Teach the models to understand various real-world human queries.

Problem-solving

Tame the models to provide the expected responses to human requests

diverse task instructions

high-quality responses

Instruction Tuning datasets matter a lot!

[3] Wang Y, Ivison H, Dasigi P, et al. How far can camels go? exploring the state of instruction tuning on open resources.

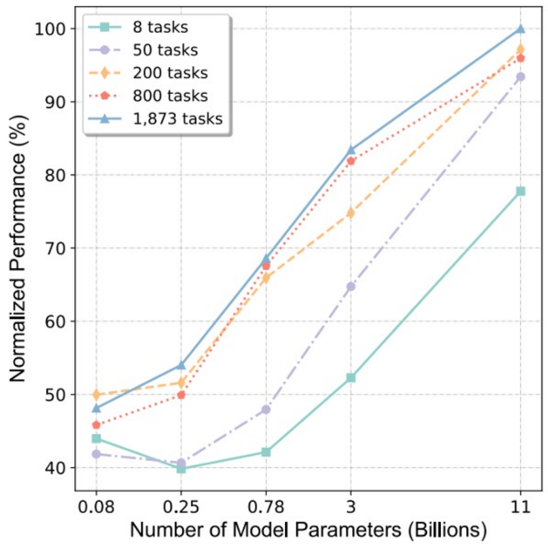
Datasets	Release Time	Scale		Language	Annotator
		# of Tasks	# of Instances (k)		
UnifiedQA (Khashabi et al., 2020)	05/2020	46	750	monolingual	✍ Human
CrossFit (Ye et al., 2021)	04/2021	159	71,000	monolingual	✍ Human
Natural Instructions (Mishra et al., 2022b)	04/2021	61	620	monolingual	✍ Human
Flan 2021 (Wei et al., 2022a)	09/2021	62	4,400	monolingual	✍ Human
P3 (Sanh et al., 2022)	10/2021	62	12,000	monolingual	✍ Human
MetaICL (Min et al., 2022a)	10/2021	142	3,500	monolingual	✍ Human
ExMix (Aribandi et al., 2022)	11/2021	107	500	monolingual	✍ Human
Super-Natural Instructions (Wang et al., 2022d)	04/2022	1,613	5,000	multilingual	✍ Human
GLM (Zeng et al., 2022)	10/2022	77	12,000	bilingual	✍ Human
Flan 2022 (Longpre et al., 2023)	10/2022	1,836	15,000	multilingual	✍ Human
xP3 (Muennighoff et al., 2022)	11/2022	71	81,000	multilingual	✍ Human
Unnatural Instructions (Honovich et al., 2022a)	12/2022	117	64	monolingual	🤖 InstructGPT
Self-Instruct (Wang et al., 2022c)	12/2022	/	82	monolingual	🤖 GPT-3
OPT-IML (Iyer et al., 2022)	12/2022	2,207	18,000	multilingual	✍ Human
Alpaca (Taori et al., 2023)	03/2023	/	52	monolingual	🤖 InstructGPT
Baize (Xu et al., 2023b)	04/2023	/	100	monolingual	🤖 ChatGPT
Koala (Geng et al., 2023)	04/2023	/	/	monolingual	✍ Human 🤖 ChatGPT
GPT4All (Anand et al., 2023)	04/2023	/	808	monolingual	✍ Human 🤖 ChatGPT
Alpaca-gpt4 (Peng et al., 2023)	04/2023	/	113	bilingual	🤖 GPT-4
Vicuna (Chiang et al., 2023)	04/2023	/	76	monolingual	✍ Human 🤖 ChatGPT
Dolly (Conover et al., 2023)	04/2023	/	15	monolingual	✍ Human
Oasst (Köpf et al., 2023)	04/2023	/	84	multilingual	✍ Human
LongForm (Köksal et al., 2023)	04/2023	/	27	monolingual	✍ Human 🤖 InstructGPT
Symbolic-Instruct (Liu et al., 2023b)	04/2023	/	796	monolingual	✍ Human
LaMini (Wu et al., 2023)	04/2023	/	2,580	monolingual	🤖 ChatGPT
WizardLM (Xu et al., 2023a)	04/2023	/	196	monolingual	🤖 ChatGPT
COEDIT (Raheja et al., 2023)	05/2023	/	82	monolingual	✍ Human
UltraChat (Ding et al., 2023)	05/2023	/	1,500	monolingual	🤖 ChatGPT

- *Numerous instruction-tuning datasets exist!*

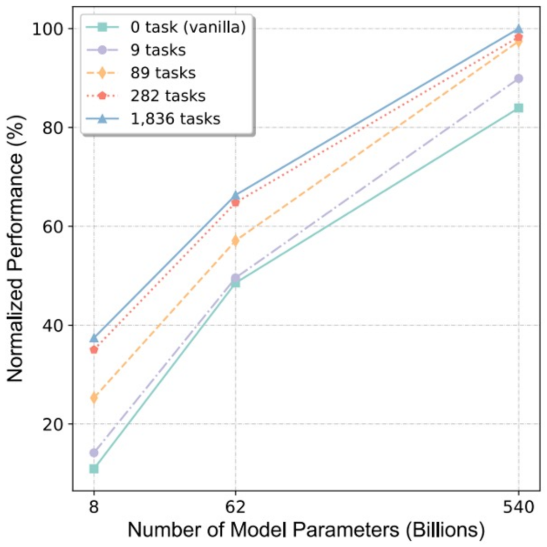
- ***LLM-synthetic data** is the trend!*
  1. Quick.
  2. Low-cost.
  3. More diverse instructions.
  4. Model-friendly.

- ***Scaling up data size becomes much easier.***

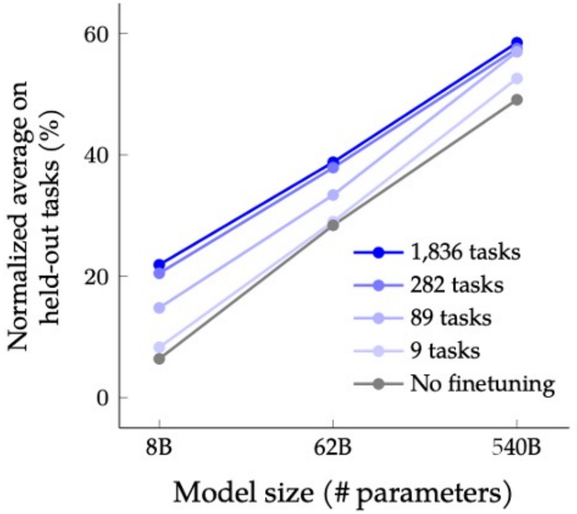
[1] R Lou, et al. Is prompt all you need? no. A comprehensive and broader view of instruction learning.

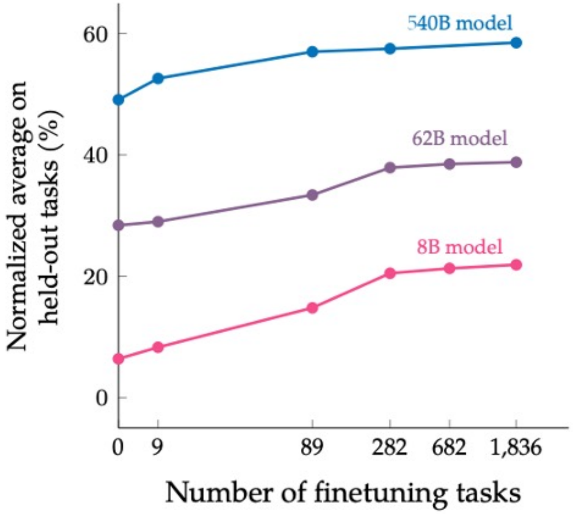


(a) Flan-T5

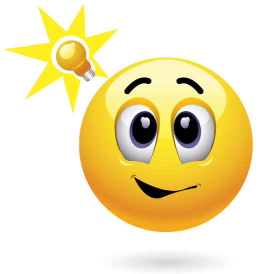


(b) Flan-PaLM



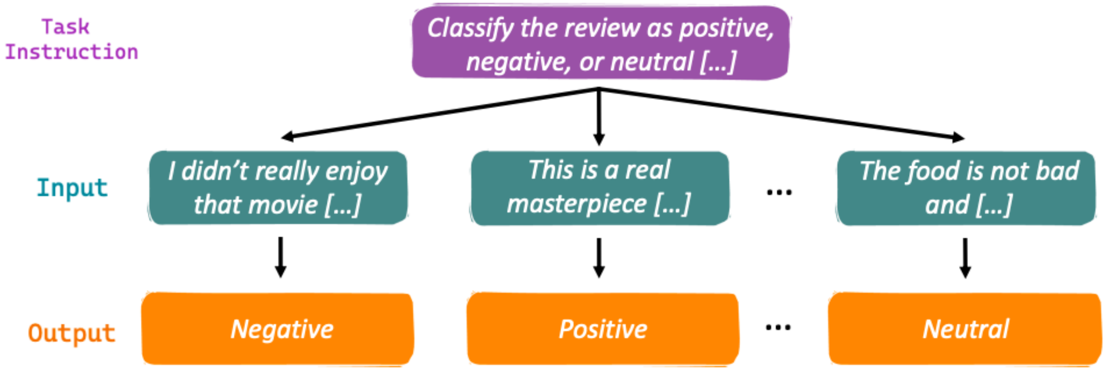


Scaling up dataset size is still the most **straightforward** way to promote the zero-shot problem-solving capacity.



[1] R Lou, et al. Is prompt all you need? no. A comprehensive and broader view of instruction learning.
[2] Chung H W, et al. Scaling instruction-finetuned language models.

Existing instruction tuning dataset paradigms



(a). Scaling-Inputs

For each task instruction, scaling up various input-output pairs (e.g., SuperNI).

➔ A conventional multi-task learning paradigm.

Potential drawbacks:

inputs play a more critical role than instructions for the models. 😩

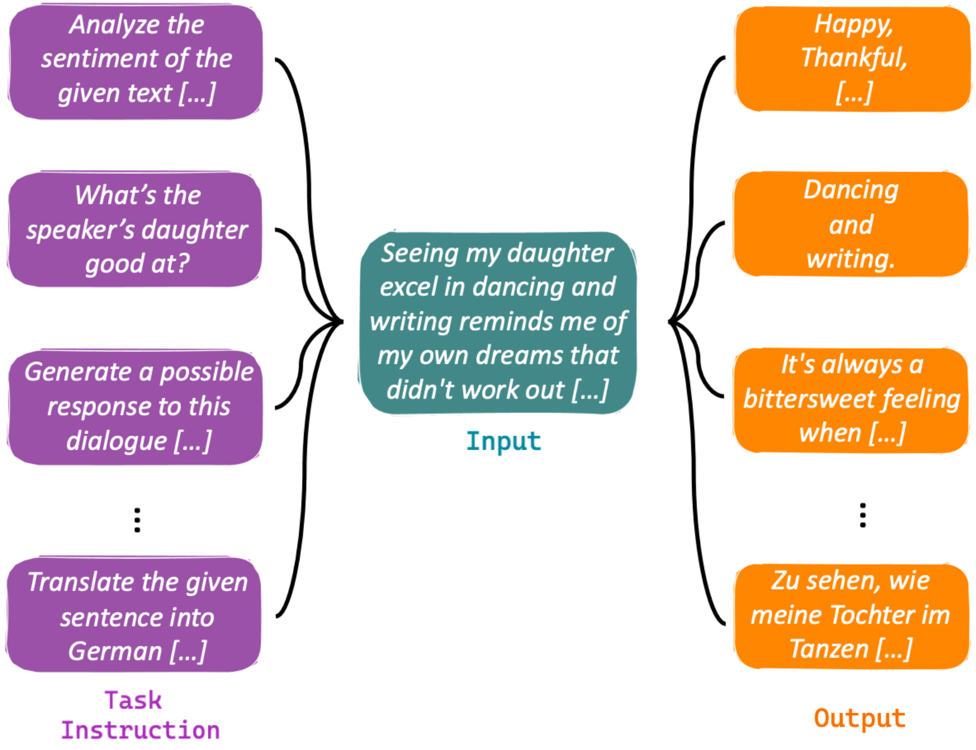
(b). Scaling Input-Free Tasks

Scaling up instruction-output pairs directly.
(e.g., Self-Instruct and Alpaca)

➔ Input contexts are omitted / tightly combined with instructions.

Potential drawbacks

cannot effectively handle downstream tasks with separate / extra context. 🤔



(c). Scaling Tasks per Input (Ours)

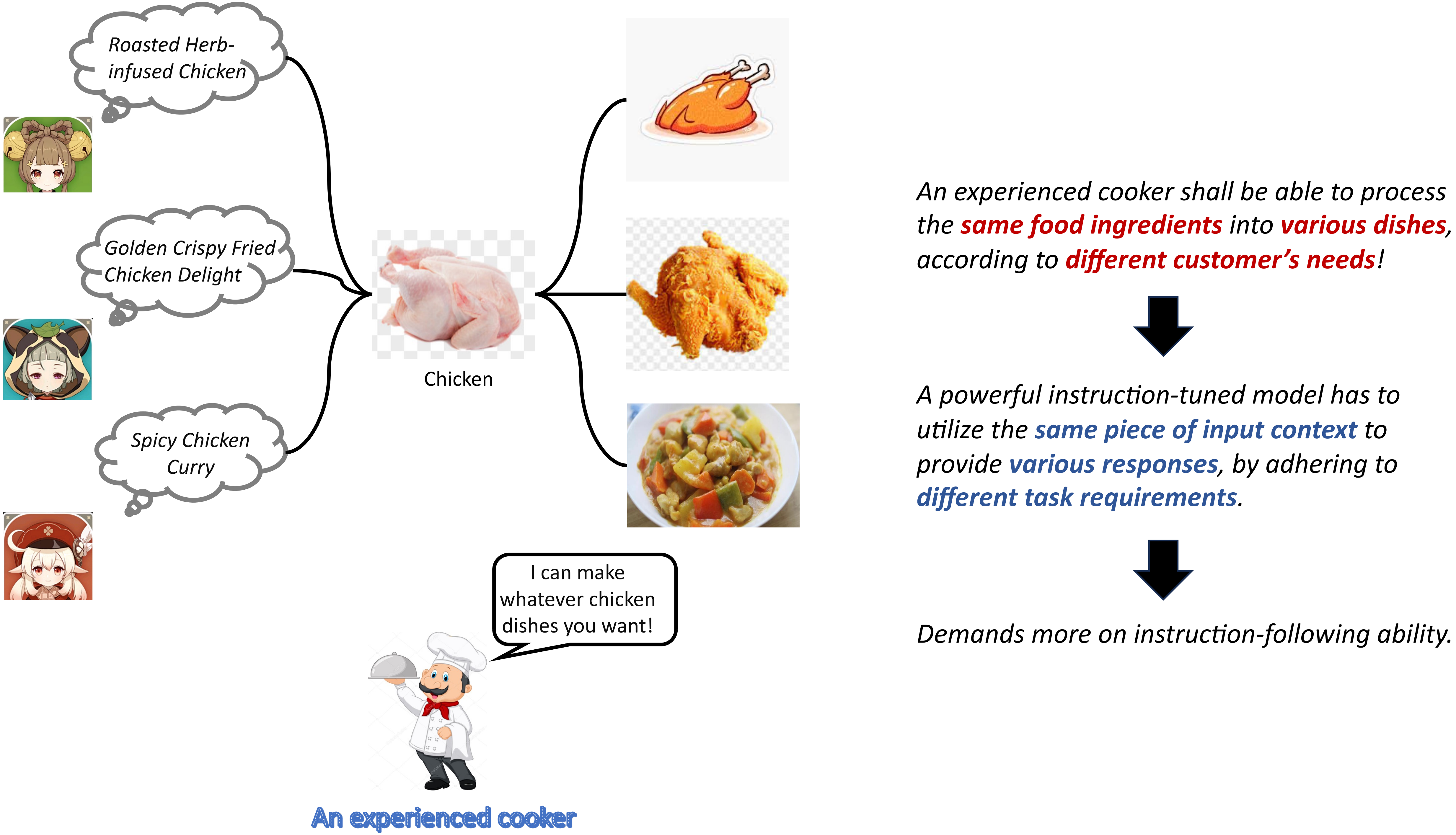
Benefits:

- Input contexts are not omitted.
- Instructions weigh more than input.
- **More challenging instruction-following training.**

Ideally, one input context can be used for diverse task purposes.

➔ e.g., given a paragraph as the context, we can use it for QA, summarization, etc.

➔ the models are trained to generate different outputs by adhering to different instructions, while the input context is fixed.



Roasted Herb-infused Chicken
Golden Crispy Fried Chicken Delight
Spicy Chicken Curry
Chicken
I can make whatever chicken dishes you want!
An experienced cooker
An experienced cooker shall be able to process the same food ingredients into various dishes, according to different customer's needs!
A powerful instruction-tuned model has to utilize the same piece of input context to provide various responses, by adhering to different task requirements.
Demands more on instruction-following ability.

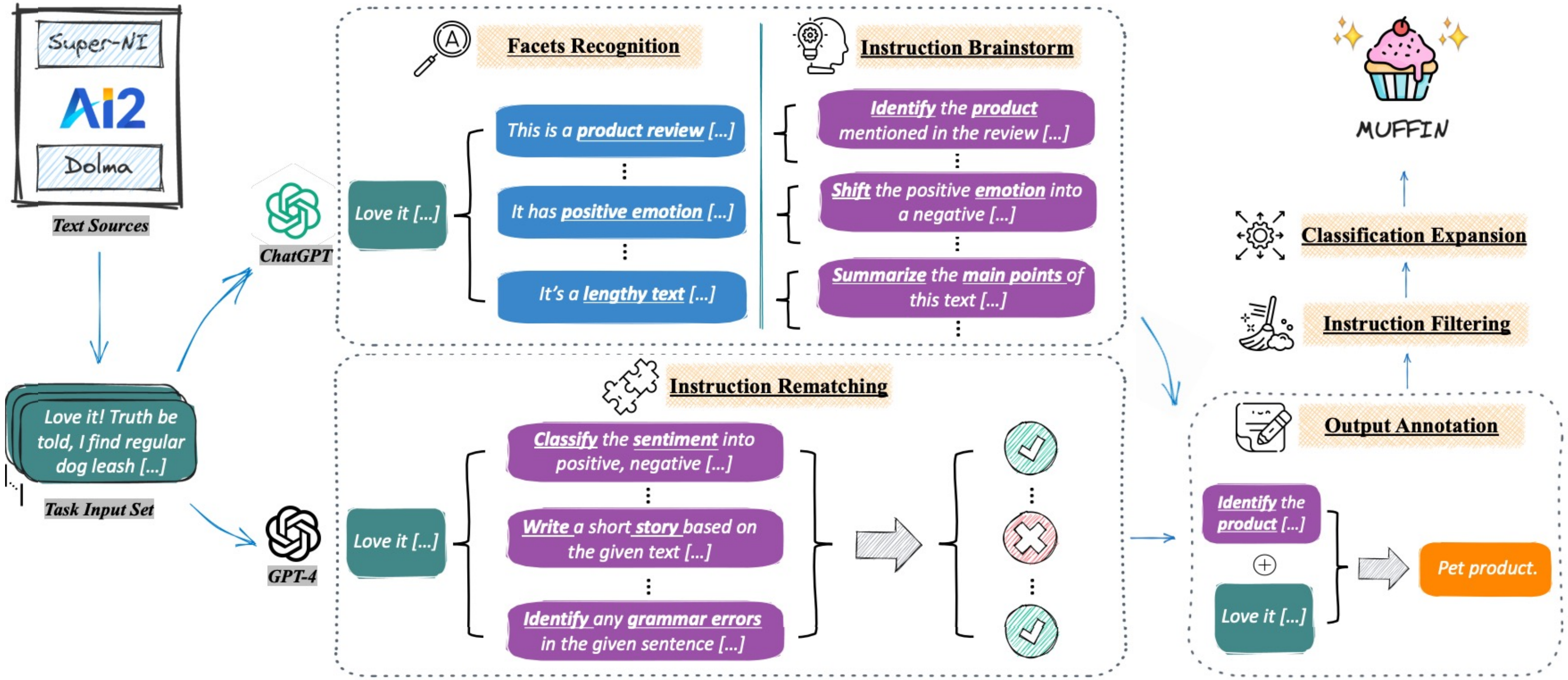


Figure 2: Data construction pipeline of 🧁 MUFFIN.

- <u>**Instruction Brainstorm**</u> 🧠: adopting two-step prompting. For each input context, first let LLMs generate diverse textual **facets** (e.g., length, topic, sentiment, etc.), then ask LLMs to use each facet as a "hint" to brainstorm various task instructions.
- <u>**Instruction Rematching**</u> 🧩: for each input context, gather suitable instructions from existing datasets (i.e., employing LLMs to decide whether an instruction can be compatible with the given input context).

Table 1: Statistics of MUFFIN.

statistic	
# of inputs	1,463
- # of inputs (from SuperNI)	953
- # of inputs (from Dolma)	510
# of instructions	56,953
- # of instructions by "rematching" (from SuperNI)	574
- # of instructions (from brainstorm)	33,720
- # of instructions (from classification expansion)	22,659
# of instructions per input	46.48
# of inputs per instruction[3]	20.27
# of (instruction, input, output) instances	68,014
ave. input length (in words)	119.26
ave. instruction length (in words)	84.74
ave. output length (in words)	71.32

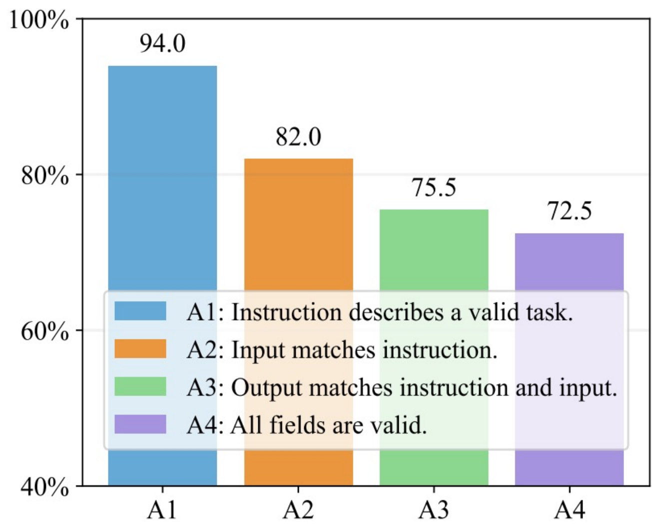


Figure 3: Human evaluation on the data quality. Both valid and invalid instances can be found in Table 17. A4 indicates the joint set of successful cases in A1, A2, and A3.

- 🧁 MUFFIN (Multi-Faceted Instructions) has about 68k instances with diverse textual distribution.
- According to small-scale human evaluation, our data has high data quality and diversity.

	Models	Data Size	SuperNI-Test EM (CLS)	SuperNI-Test ROUGE-L (GEN)	SuperNI-Test ROUGE-L (overall)	MMLU Rank ACC	MMLU EM	T0-Eval Rank ACC	T0-Eval EM	BBH EM	Average
T5-3B	*Human Annotated Data (indirect comparison)*										
	SuperNI-Train	68k	35.46	48.01	43.25	38.42	36.97	49.65	48.73	19.60	40.01
	Generated Data (direct comprison)										
	Dolly	15k	0.49	34.32	14.52	23.05	0.00	39.84	6.78	5.71	15.59
	LongForm	23k	0.00	33.58	11.29	23.07	0.00	39.68	0.62	3.84	14.01
	Alpaca	52k	20.43	46.08	35.25	28.55	8.02	43.26	20.52	11.53	26.71
	Alpaca-GPT4	52k	11.72	41.84	27.49	23.89	0.00	41.51	14.14	8.50	21.14
	WizardLM	68k	5.34	41.09	20.81	25.55	0.00	40.55	5.87	5.16	18.05
	Self-Inst.	82k	29.59	43.70	36.87	27.11	23.55	41.74	38.57	**20.53**	32.71
	Unnatural Inst.	68k	32.56	45.08	41.42	32.65	18.03	43.42	34.49	8.53	32.02
	Dynosaur	66k	26.97	44.27	35.65	26.11	20.38	38.98	38.81	13.68	30.61
	Muffin (Ours)	68k	**33.84**	**49.52**	**42.63**	**36.27**	**29.75**	**46.35**	**44.45**	14.25	**37.13**
T5-11B	*Human Annotated Data (indirect comparison)*										
	SuperNI-Train	68k	41.13	50.05	47.76	54.45	54.37	56.89	54.23	29.80	48.59
	Generated Data (direct comprison)										
	Dolly	15k	2.71	37.12	17.81	22.99	0.06	49.17	23.96	10.18	20.50
	LongForm	23k	1.88	38.05	16.27	23.23	0.00	39.85	2.79	5.53	15.95
	Alpaca	52k	25.36	47.74	39.62	30.17	8.10	54.48	34.90	9.28	30.21
	Alpaca-GPT4	52k	13.65	43.19	31.46	25.58	0.00	49.94	34.79	7.94	25.82
	WizardLM	68k	4.81	40.43	21.26	24.63	0.01	45.10	6.44	4.79	18.43
	Self-Inst.	82k	28.88	44.88	36.53	28.22	32.45	48.61	41.46	**31.39**	36.55
	Unnatural Inst.	68k	41.11	47.46	45.54	34.38	22.39	43.40	41.91	12.84	36.13
	Dynosaur	66k	**42.02**	47.53	46.42	27.60	24.96	42.85	43.39	9.22	35.50
	Muffin (Ours)	68k	40.20	**50.69**	**48.32**	**41.95**	**41.83**	**55.38**	**57.74**	20.53	**44.58**

We adopt distinct evaluation benchmarks with different paradigms:

- *Scaling-Inputs*: SuperNI
- *Scaling Input-Free Tasks*: MMLU
- *Hybrid*: T0-Eval and BBH

Meanwhile, we compare our dataset with previous baseline datasets across different paradigms as well.

- Models tuned on our MUFFIN consistently achieve better performance across 3 out of 4 benchmarks, compared with previous LLM-synthetic datasets.
- MUFFIN can **even surpass human-craft SuperNI in some cases**.

Models	Data Size	SuperNI-Test EM (CLS)	SuperNI-Test ROUGE-L (GEN)	SuperNI-Test ROUGE-L (overall)	MMLU EM	T0-Eval EM	BBH EM
Human Annotated Data (indirect comparison)							
SuperNI	68k	<u>50.73</u>	55.99	<u>52.43</u>	31.38	46.37	12.26
Generated Data (direct comprison)							
Dolly	15k	9.96	43.58	27.25	0.39	22.29	7.76
LongForm	23k	4.30	41.30	19.07	0.12	0.72	5.27
Alpaca	52k	33.34	51.67	43.65	36.01	40.39	21.72
Alpaca-GPT4	52k	18.27	44.27	33.50	1.01	6.29	2.20
WizardLM	68k	10.52	43.36	27.27	0.29	7.20	4.24
Self-Inst.	82k	36.82	46.79	41.04	23.12	31.43	<u>**28.69**</u>
Unnatural Inst.	68k	37.63	50.23	46.03	6.69	8.35	5.05
Dynosaur	66k	**44.35**	49.34	47.08	17.26	34.59	7.11
Muffin (Ours)	68k	40.85	<u>**57.71**</u>	**49.71**	<u>**37.67**</u>	<u>**55.98**</u>	19.01

Table 3: Results based on Llama2-13B.

- We also experiment with Llama2 + LoRA
- The observations and conclusions are similar.

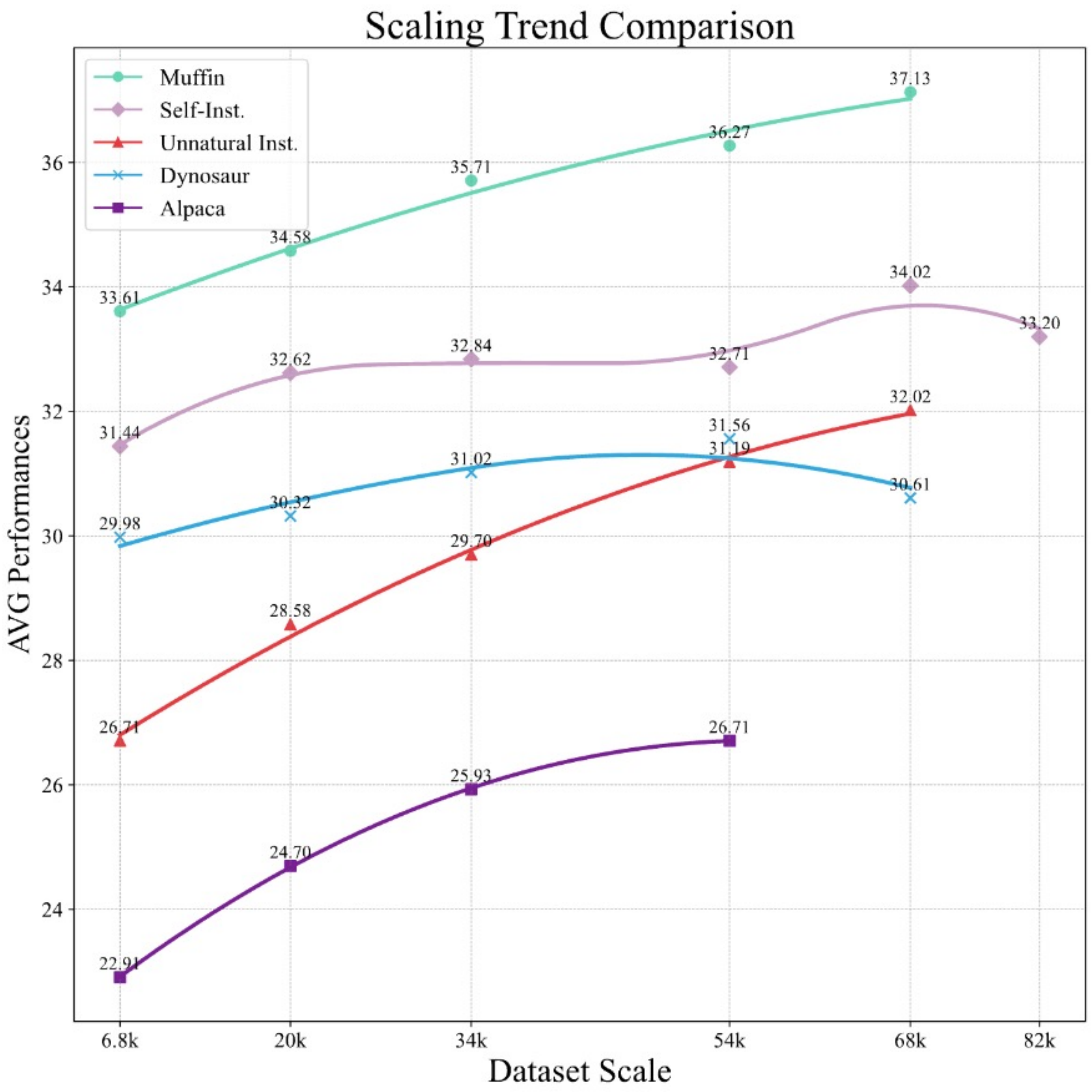
Models	SuperNI-Test	MMLU	T0-Eval	BBH	Average
Dolly	22.5	14.0	36.5	28.0	25.3
LongForm	6.0	15.0	10.0	12.0	10.8
Alpaca	44.5	20.0	42.0	24.0	32.6
Alpaca-GPT4	45.0	11.0	38.0	24.0	29.5
WizardLM	35.0	19.5	36.0	26.0	29.1
Self-Inst.	39.0	23.5	**45.5**	29.5	34.4
Unnatural Inst.	50.5	24.0	34.5	23.0	33.0
Dynosaur	43.0	28.5	30.0	22.0	30.9
Muffin (Ours)	**56.5** (↑ 6.0)	**34.5** (↑ 6.0)	45.0 (↓ 0.5)	**31.0** (↑ 1.5)	**41.8** (↑ 7.4)

Table 4: Human evaluation acceptance ratio. We randomly sample 200 instances from each benchmark and let workers evaluate different systems' outputs.

SuperNI-Test			MMLU			T0-Eval			BBH		
Ours	Self-Inst.	Tie	Ours	Self-Inst.	Tie	Ours	Self-Inst.	Tie	Ours	Self-Inst.	Tie
47.0	41.5	11.5	**39.5**	16.5	44.0	**11.0**	10.0	79.0	**19.5**	15.5	65.0
Ours	Unnatural	Tie	Ours	Unnatural	Tie	Ours	Unnatural	Tie	Ours	Unnatural	Tie
31.5	20.0	48.0	**42.5**	10.0	47.5	**43.5**	16.5	40.0	**21.5**	11.5	67.0
Ours	SuperNI	Tie	Ours	SuperNI	Tie	Ours	SuperNI	Tie	Ours	SuperNI	Tie
31.0	16.0	53.0	**24.0**	21.0	55.0	9.0	**15.0**	76.0	**16.5**	9.5	74.0

Table 5: Pair-wise comparison between MUFFIN (Ours) and three strong baselines, namely Self-Instruction (Self-Inst.), Unnatural Instruction (Unnatural), and SuperNI, across four benchmarks.

- We conduct further human evaluation regarding the model's responses.
- Results reflect MUFFIN's excellent task-solving capacity.
- According to our error case analyses, MUFFIN's responses align more with the task requirements, especially for those complicated evaluation tasks (e.g., in the SuperNI).



- We randomly sample subsets from various datasets and train models on the subsets to show the performance trends (10%, 30%, 50%, 80%, 100%).
- MUFFIN exceeds the baselines by a noteworthy margin (average scores on four evaluation benchmarks).
- Other baselines may only be comparable to our data results when they continue to be scaled to several times the size of our data.

For more experiments and analyses, please refer to our paper. 🤗

References:

[1]. Lou, Renze, Kai Zhang, and Wenpeng Yin. "Is prompt all you need? no. A comprehensive and broader view of instruction learning." arXiv preprint arXiv:2303.10475 (2023).

[2]. Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." arXiv preprint arXiv:2210.11416 (2022).

[3]. *Wang, Yizhong, et al. "How far can camels go? exploring the state of instruction tuning on open resources." Advances in Neural Information Processing Systems 36 (2024).*

[4]. Longpre, Shayne, et al. "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning." (2023).

Thanks.

Q&A

Paper

Website

Model