High Quality Data

두개의 논문

High Quality Training Data

ICML 2024

Long Is More for Alignment: A Simple but Tough-to-Beat Baseline for Instruction Fine-Tuning

Hao Zhao [1] Maksym Andriushchenko [1] Francesco Croce [1] Nicolas Flammarion [1]

https://arxiv.org/pdf/2402.04833

ICML 2024

QuRating: Selecting High-Quality Data for Training Language Models

Alexander Wettig [1] Aatmik Gupta [1] Saumya Malik [1] Danqi Chen [1]

https://arxiv.org/pdf/2402.09739

Preliminary

전제

ALPAGASUS: TRAINING A BETTER ALPACA WITH FEWER DATA

Lichang Chen[*†], **Shiyang Li** [*‡], **Jun Yan**[♯], **Hai Wang** [‡], **Kalpa Gunaratna**[‡], **Vikas Yadav**[‡], **Zheng Tang**[‡], **Vijay Srinivasan**[‡], **Tianyi Zhou**[†], **Heng Huang**[†], **Hongxia Jin**[‡]
[†] University of Maryland, College Park [‡] Samsung Research America [♯] University of Southern California
{bobchen, tianyi, heng}@umd.edu
{shiyang.li, h.wang2, k.gunaratna, vikas.y, zheng.tang, v.srinivasan, hongxia.jin}@samsung.com
yanjun@usc.edu

LIMA: Less Is More for Alignment

Chunting Zhou[μ*] **Pengfei Liu**[π*] **Puxin Xu**[μ] **Srini Iyer**[μ] **Jiao Sun**[λ]
Yuning Mao[μ] **Xuezhe Ma**[λ] **Avia Efrat**[τ] **Ping Yu**[μ] **Lili Yu**[μ] **Susan Zhang**[μ]

LLM을 학습할 때

좋은 품질의 데이터를 선별하는 것은

매우 중요한 작업이다

Preliminary

Alpagasus (ICLR2024)

System Prompt:
We would like to request your feedback on the performance of AI assistant in response to the instruction and the given input displayed following.

Instruction: [Instruction]
Input: [Input]
Response: [Response]

User Prompt:
Please rate according to the [dimension] of the response to the instruction and the input. Each assistant receives a score on a scale of 0 to 5, where a higher score indicates higher level of the [dimension]. Please first output a single line containing the value indicating the scores. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias.

Figure 3: Prompt p_G to ChatGPT for rating and filtering training data in Eq. (1).

ChatGPT에게

데이터 품질에 대해

직접 물어보고

점수를 구해서

Preliminary

Alpagasus (ICLR2024)

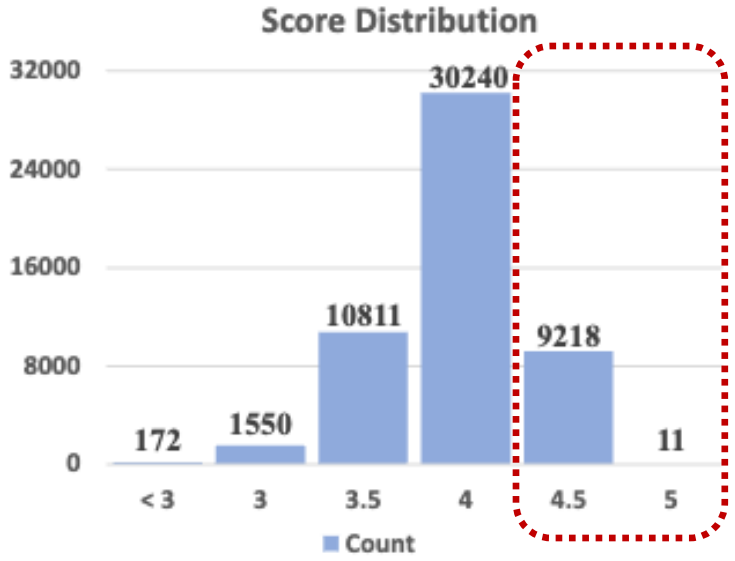
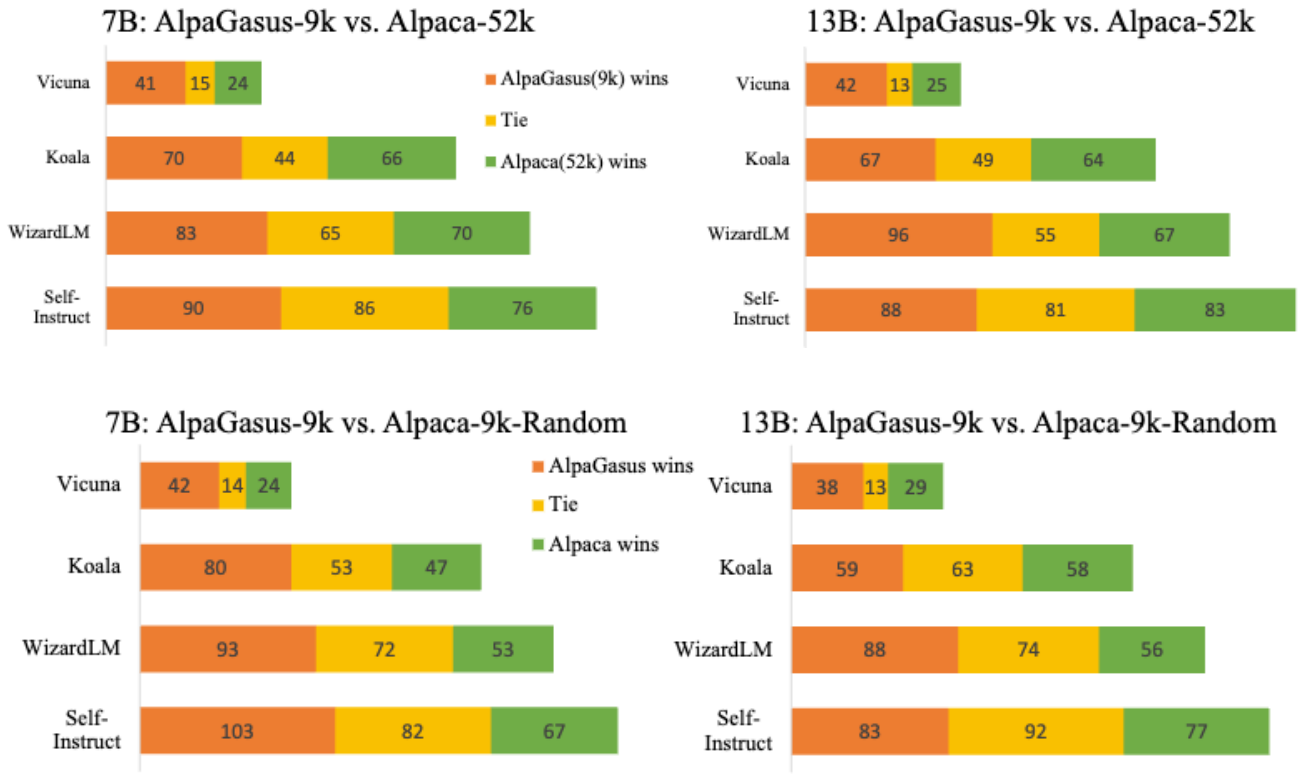


Figure 4: Histogram of Scores (Alpaca Dataset).



4.5점 이상을 받은 데이터만

골라서 학습했더니

모든 데이터로 학습 하는 것 보다

분명히 좋더라

Preliminary

LIMA (Neurips2023)

데이터 품질 지표를

아주 잘 고려해서

데이터를 선정했더니

- Diverse input
- Helpful AI assistant style output
- Input-output Alignment

Stack Exchange

First, we divide the exchanges into 75 STEM exchanges (including programming, math, physics, etc.) and 99 other (English, cooking, travel, and more); we discard 5 niche exchanges. We then sample 200 questions and answers from each set using a temperature of tau = 3 to get a more uniform sample of the different domains. Within each exchange, we take the questions with the highest score that are self-contained in the title (no body). We then select the top answer for each question, assuming it had a strong positive score (at least 10). To conform with the style of a helpful AI assistant, we automatically filter answers that are too short (less than 1200 characters), too long (more than 4096 characters), written in the first person ("I", "my"), or reference other answers ("as mentioned", "stack exchange", etc); we also remove links, images, and other HTML tags from the response, retaining only code blocks and lists. Since Stack Exchange questions contain both a title and a description, we randomly select the title as the prompt for some examples, and the description for others.

wikiHow

We sample 200 articles from wikiHow, sampling a category first (out of 19) and then an article within it to ensure diversity. We use the title as the prompt (e.g. "How to cook an omelette?") and the article's body as the response. We replace the typical "This article..." beginning with "The following answer...", and apply a number of preprocessing heuristics to prune links, images, and certain sections of the text.

Reddit Dataset

Due to its immense popularity, Reddit is geared more towards entertaining fellow users rather than helping; it is quite often the case that witty, sarcastic comments will obtain more votes than serious, informative comments to a post. We thus restrict our sample to two subsets, r/AskReddit and r/WritingPrompts, and manually select examples from within the most upvoted posts in each community. From r/AskReddit we find 70 self-contained prompts (title only, no body), which we use for the test set, since the top answers are not necessarily reliable. The WritingPrompts subreddit contains premises of fictional stories, which other users are then encouraged to creatively complete. We find 150 prompts and high-quality responses, encompassing topics such as love poems and short science fiction stories, which we add to the training set. All data instances were mined from the Pushshift Reddit Dataset [Baumgartner et al., 2020].

Preliminary

LIMA (Neurips2023)

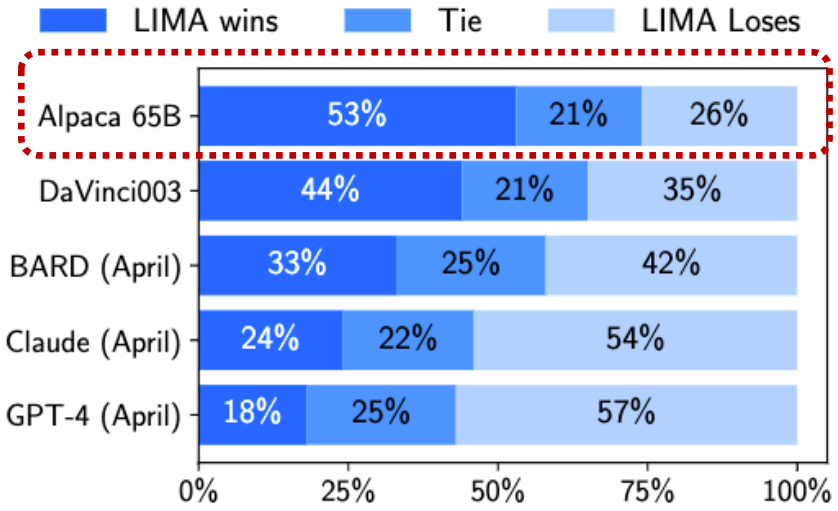


Figure 1: Human preference evaluation, comparing LIMA to 5 different baselines across 300 test prompts.

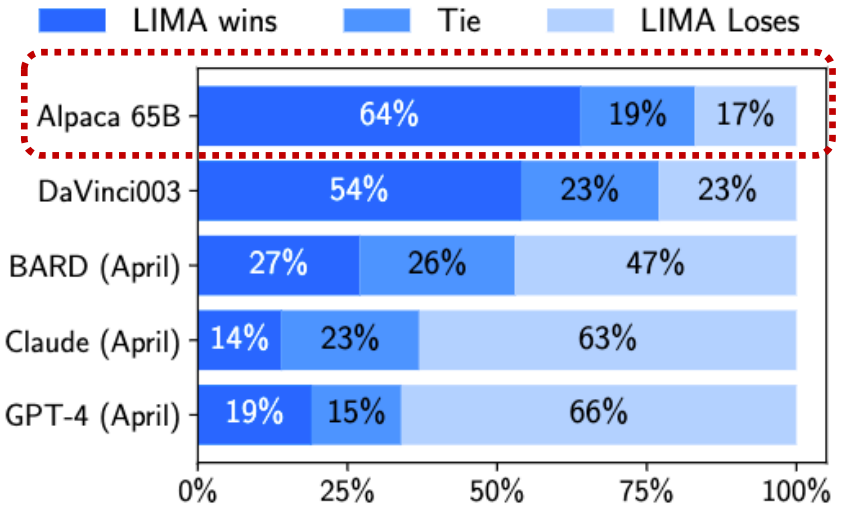


Figure 2: Preference evaluation using GPT-4 as the annotator, given the same instructions provided to humans.

1,000개로만 학습해도 좋더라

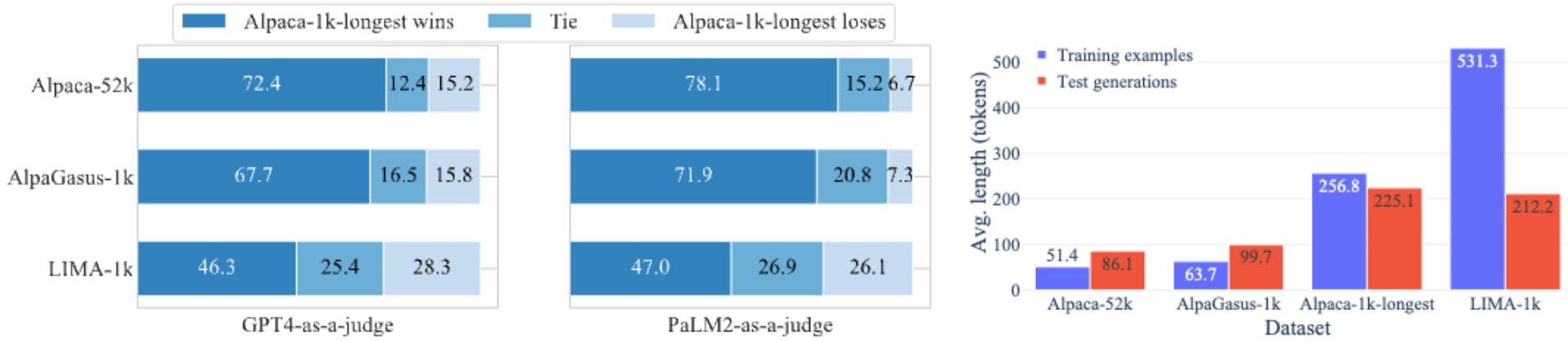
L (Long) IMA

While the quality of the instructions seems to play a major role for IFT,

it remains unclear which are the distinguishing features of high quality demonstrations.

L (Long) IMA

1,000 instructions with longest responses



(a) Head-to-head comparisons (in %) with two different LLM judges

(b) Average number of tokens in responses

Figure 1. **Selecting the longest responses leads to a strong IFT dataset.** We fine-tune LLaMA-2-7B models on Alpaca-52k (Taori et al., 2023), AlpaGasus-1k (Chen et al., 2023), LIMA-1k (Zhou et al., 2023) and our Alpaca-1k-longest datasets. **(a)** Alpaca-1k-longest beats three baselines in instruction-following performance according to both GPT-4 and PaLM-2 as judges. **(b)** Alpaca-1k-longest leads to an average response length at test time higher than Alpaca-52k and AlpaGasus-1k, but similar to LIMA-1k: then its higher win rate cannot be solely attributed to the model having learnt to generate long responses.

Discussion

Fine-tuning on long instructions is a very strong baseline

: Inexpensive yet strong baseline for future works on alignment

Why?

- Usually more informative and thus contain more features relevant to human intentions
- Intuitively harder for LLMs to fit, which forces the model to actually learn the response style rather than just memorize the answer.
- Encourages the model to capture long-distance semantic connections, and stay on-topic when answering complicate instructions.

Model	1k-longest wins	Tie	1k-longest loses
Base dataset: Alpaca-52k			
1k-shortest	**97.0**	2.6	0.4
1k-random	**72.1**	18.0	9.9
Base dataset: Evol-Instruct-70k			
1k-shortest	**93.4**	4.9	1.7
1k-random	**39.7**	29.1	31.2
Base dataset: Open-Hermes-1M			
1k-shortest	**95.9**	3.5	0.6
1k-random	**84.3**	10.4	5.3

Discussion

Analysis

(upper): Alpaca-52k 중 선택

(lower): Evol-Instruct-70k중 선택

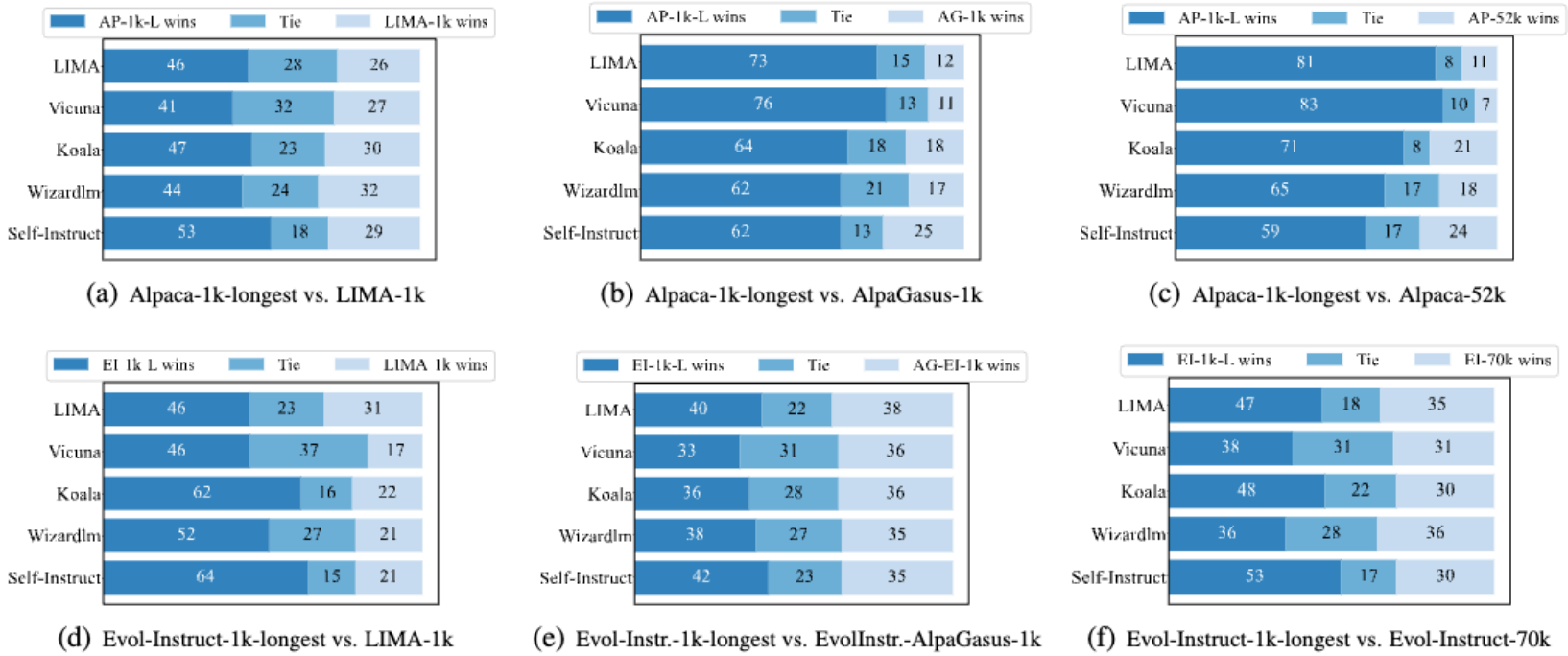


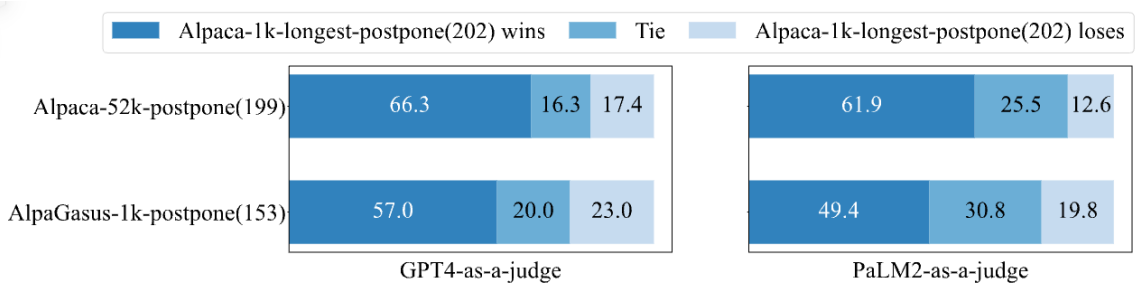
Figure 2. **Detailed preference evaluation (in %).** For each pair of LLMs we report the win rate on 5 datasets (LIMA, Vicuna, Koala, WizardLM, Self-Instruct) according to GPT-4-as-a-judge. **Top:** we compare fine-tuning on Alpaca-1k-longest (AP-1k-L) to Alpaca-52k, AlpaGasus-1k, and LIMA-1k. **Bottom:** we compare fine-tuning on Evol-Instruct-1k-longest (EI-1k-L) to Evol-Instruct-70k, Evol-Instruct-AlpaGasus-1k (i.e. using the method of Chen et al. (2023) to subsample Evol-Instruct-70k), and LIMA-1k. Our datasets of long responses consistently lead to higher preferences (higher win rate) than the existing methods.

Evol-Instruct contains higher-quality data than Alpaca,
thus even selecting examples using GPT-3.5-Turbo scores can find relatively effective training examples

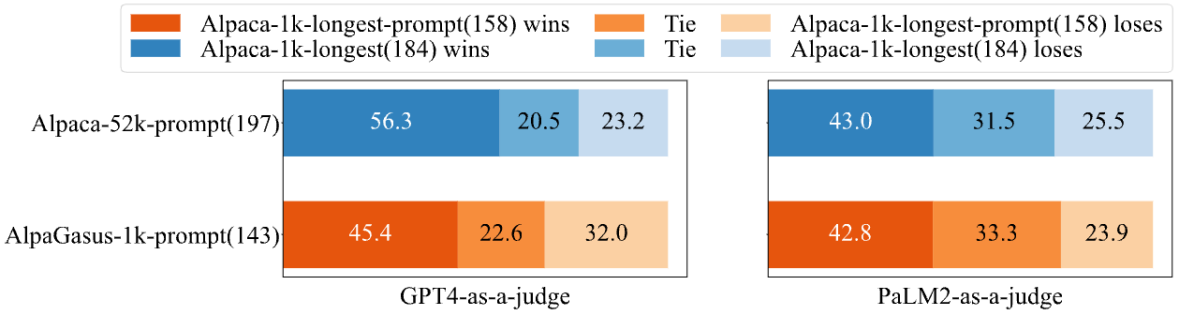
Discussion

Analysis

LLM Evaluator들이, 길이가 긴 답변을 선호해서 더 좋게 나오는 것 아닌가?



(a) Postpone the EOS token (base model: Llama-2-7B)



(b) Prompting strategy (base model: Mistral-7B-v0.1)

Minimum response length를 150 이상으로 정의

Baseline model들의 답변 길이를 늘렸을 때

➔ 여전히 Longest-1k가 더 좋음

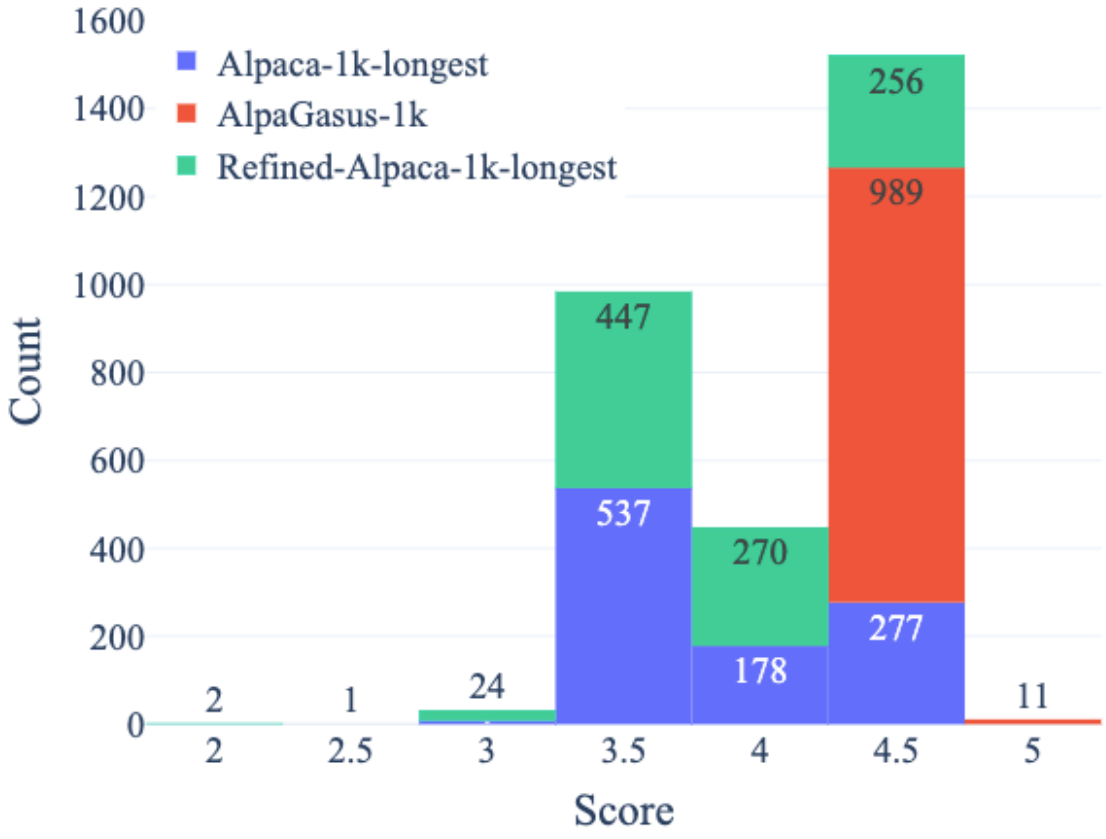
Alpaca-52k / Alpagasus-1k 에게는 “answer in N paragraphs”

Longest-1k에게는 “answer in as few words as possible”

➔ 여전히 Longest-1k가 더 좋음

Discussion

Analysis



AlpaGasus에서 시도했던 데이터 평가 방식을 그대로 적용했을 때,

Longest-1k에는 여전히 낮은 품질로 평가되는 데이터가 존재

But, Longest-1k는 AlpaGasus-1k보다 학습 데이터로서 효과가 좋음

➔ This suggests that other factors come into play when determining the effectiveness of IFT dataset

➔ As a result, it remains uncertain which specific components in the fine-tuning dataset are crucial for achieving the best model performance.

What Quality Measure?

Human intuitions about data quality

- Writing style
- Required expertise
- Facts & trivia
- Educational value

이 기준으로 데이터를 선별한다면
LLM학습에 유효한 데이터를 골라낼 수 있을까?

실험 설계

- **GPT3.5 활용. 250K개 데이터 annotation**
  - 앞선 4개의 quality measure를 기준으로 (docA, docB, preference) 데이터셋을 구축

- **구축한 데이터를 활용해서, Evaluator 모델 학습 (ShearedLLama1.3B)**
  - Pairwise Evaluation 수행 (docA 〉 docB // docA 〈 docB)

- **SlimPajama 데이터셋 260B token 데이터 중, 30B token 데이터 선별**
  - 임의로 두개의 document 추출
  - Evaluator활용, 두개의 document 중 더 선호되는 데이터 선택

Method

Why Pairwise Comparison?

We observe that LLMs are better at comparing texts than they are at judging individual texts

- **Writing style** 관련하여, human annotated 10개 document의 ranking을 준비

- 10개 document에 대하여, ChatGPT에게 **Writing style**을 평가하도록 지시
  - **1-10 scale** ➔ 0.61 ± 0.06 alignment
  - **Pairwise comparison** ➔ 0.79 ± 0.01 alignment

이를 데이터 선별 작업에 사용하기 위해서?

- 2개의 document를 랜덤 선별 (without replacement)
- 각 pair에 대해서, **우수한 document를 하나 선정**

➔ 260B token 중 30B token데이터셋을 구축

Method

QuRator

GPT3.5 Distillation → ShearedLLAMA 학습

- 500K unique document를 랜덤 선별 (without replacement)
- 이를 통해 임의로 250K document pair를 생성
- GPT3.5 모델을 통해 binary evaluation 결과 얻음 (각 Pair당 20회)
- **Bredly-Terry model**을 사용해서, binary judgement를 수치화
  → (doc A, doc B, docA 점수, docB 점수)

이 결과를 기반으로,
두개의 document에 대한 판단을 수행하는
ShearedLLAMA 기반의 Evaluator 학습

Bredly-Terry model

A,B,C,D 의 승/패 결과가 있을 때,

A,B,C,C 의 강함을 수치화 하는 방법

[1,1,1,1]로 초기값 세팅하고,
승/패결과를 기반으로 강함을 최적화

	A	B	C	D
A	-	2	0	1
B	3	-	5	0
C	0	3	-	1
D	4	0	3	-

$$p_1 = \frac{\sum_{j(\neq 1)} w_{1j} p_j/(p_1+p_j)}{\sum_{j(\neq 1)} w_{j1}/(p_1+p_j)} = \frac{2\frac{1}{1+1} + 0\frac{1}{1+1} + 1\frac{1}{1+1}}{3\frac{1}{1+1} + 0\frac{1}{1+1} + 4\frac{1}{1+1}} = 0.429.$$

Now, we apply (5) again to update p_2, making sure to use the new value of p_1 that we just calculated:

$$p_2 = \frac{\sum_{j(\neq 2)} w_{2j} p_j/(p_2+p_j)}{\sum_{j(\neq 2)} w_{j2}/(p_2+p_j)} = \frac{3\frac{0.429}{1+0.429} + 5\frac{1}{1+1} + 0\frac{1}{1+1}}{2\frac{1}{1+0.429} + 3\frac{1}{1+1} + 0\frac{1}{1+1}} = 1.172$$

Similarly for p_3 and p_4 we get

$$p_3 = \frac{\sum_{j(\neq 3)} w_{3j} p_j/(p_3+p_j)}{\sum_{j(\neq 3)} w_{j3}/(p_3+p_j)} = \frac{0\frac{0.429}{1+0.429} + 3\frac{1.172}{1+1.172} + 1\frac{1}{1+1}}{0\frac{1}{1+0.429} + 5\frac{1}{1+1.172} + 3\frac{1}{1+1}} = 0.557$$

$$p_4 = \frac{\sum_{j(\neq 4)} w_{4j} p_j/(p_4+p_j)}{\sum_{j(\neq 4)} w_{j4}/(p_4+p_j)} = \frac{4\frac{0.429}{1+0.429} + 0\frac{1.172}{1+1.172} + 3\frac{0.557}{1+0.557}}{1\frac{1}{1+0.429} + 0\frac{1}{1+1.172} + 1\frac{1}{1+0.557}} = 1.694$$

Method

Choice of Criteria

Abstract Qualities

(1) are applicable to a wide variety of text,

(2) require a deeper understanding of a text's content, which cannot easily be derived from surface features

(3) result in fine-grained rankings with few ties

(4) are complementary to each other.

Writing style

Which text has a more polished and beautiful writing style (favor literary and academic writing)

Facts & Trivia

Which text contains more facts and trivia? (have a high density of long-tail factual knowledge.)

Educational Value

Which text has more educational value? (e.g. it includes clear explanations, step-by-step reasoning, or questions and answers) particularly valuable for inducing reasoning capabilities in LLMs,

Required Expertise

Which text requires greater expertise and prerequisite knowledge to understand it? (difficulty level of the training corpus)

Method

Choice of Criteria

(4) are complementary to each other.

Method

Prompt Validation

Prompt Template

```
Compare two text excerpts and choose the text
which {criterion}

Aspects that should NOT influence your
judgement:
1.  Which language the text is written in
2.  The length of the text
3.  The order in which the texts are presented

Note that the texts are cut off, so you have
to infer their contexts.  The texts might have
similar quality, but you should still make a
relative judgement and choose the label of the
preferred text.

[Option A]
...  {text_a} ...

[Option B]
...  {text_b} ...

Now you have to choose between either A or B.
Respond only with a single word.
```

Writing Style

```
has a more polished and
beautiful writing style.
```

Facts & Trivia

```
contains more facts and trivia.
Prefer specific facts and
obscure trivia over more
common knowledge.
```

Educational Value

```
has more educational value,
e.g., it includes clear
explanations, step-by-step
reasoning, or questions and
answers.
```

Required Expertise

```
requires greater expertise
and prerequisite knowledge to
understand it.
```

GPT3.5 Annotation에 사용된 Prompt

Human annotated 40개 document에 대해서 GPT3.5의 annotation agreement를 확인

➔ fact&trivia에서는 92% agreement

➔ 이외는 97% 이상의 agreement

Experiments

Results

Selection Method		Perplexity	Reading Comprehension (5 tasks)	Commonsense Reasoning (3 tasks)	World Knowledge (2 tasks)	Average (10 tasks)
			Reading Comprehension	Commonsense Reasoning	World Knowledge	
			ARC-E ARC-C SciQ LogiQA BoolQ	HellaSw. PIQA W.G.	NQ MMLU	
Uniform		8.96	50.9	55.0	14.9	44.9
DSIR	*with Wiki*	10.67 ↑1.71	50.1 ↓0.8	49.8 ↓5.2	14.7 ↓0.2	42.9 ↓2.0
	with Book	11.00 ↑2.04	47.9 ↓3.0	**56.6** ↑1.6	14.1 ↓0.8	43.8 ↓1.1
Perplexity	*lowest*	11.92 ↑2.96	48.3 ↓2.6	49.6 ↓5.4	13.7 ↓1.2	41.7 ↓3.2
	highest	9.97 ↑1.01	49.6 ↓1.3	53.5 ↓1.5	13.4 ↓1.5	43.5 ↓1.4
Writing Style	*top-k*	10.53 ↑1.57	49.3 ↓1.6	53.3 ↓1.7	13.5 ↓1.4	43.4 ↓1.5
	$\tau = 2.0$	**8.90** ↓0.06	51.0 ↑0.1	55.8 ↑0.8	14.1 ↓0.8	45.0 ↑0.1
Facts & Trivia	*top-k*	10.56 ↑1.60	54.3 ↑3.4	51.7 ↓3.3	15.5 ↑0.6	45.8 ↑0.9
	$\tau = 2.0$	8.91 ↓0.05	52.7 ↑1.8	55.6 ↑0.6	15.6 ↑0.7	46.2 ↑1.3
Educational Value	*top-k*	10.59 ↑1.63	**54.7** ↑3.8	54.9 ↓0.1	14.4 ↓0.5	**46.7** ↑1.8
	$\tau = 2.0$	8.91 ↓0.05	53.3 ↑2.4	56.3 ↑1.3	**15.7** ↑0.8	**46.7** ↑1.8
Required Expertise	*top-k*	11.54 ↑2.58	52.8 ↑1.9	48.7 ↓6.3	14.3 ↓0.6	43.9 ↓1.0
	$\tau = 2.0$	8.93 ↓0.03	52.7 ↑1.8	55.5 ↑0.5	15.0 ↑0.1	46.0 ↑1.1
Criteria mix	$\tau = 2.0$	**8.90** ↓0.06	52.1 ↑1.2	55.5 ↑0.5	15.2 ↑0.3	45.7 ↑0.8
Uniform +50% data		*8.46* ↓0.50	*52.9* ↑2.0	*57.0* ↑2.0	*15.9* ↑1.0	*46.8* ↑1.9

기존 방법들은, 데이터 임의 추출보다 성능이 떨어지는 경우가 많음 (DSIR, PPL)

Educational value is the strongest criterion.

Experiments

Results

ShareGPT데이터셋 내 1000개 데이터셋으로 Instruction Tuning

AlpacaFarm 데이터를 통해 instruction following 능력 평가 (GPT4 judge)

➔ 유일하게 Educational Value만이 Random Baseline 능가

Conclusion

아무런 기준 없이

GPT4와 같은 Super-LLM에게

데이터 품질에 대해서 물어보는 것은

능사가 아닐 수 있다

감사합니다