Appendix

Table of Contents

A FREQUENTLY ASKED QUESTIONS

A.1 IS THERE ANY BIAS CONTAINED IN THE EVALUATION PROMPTS?

We also explore alternate evaluation prompts such as the prompts provided by Zheng et al. (2023), which are shown in Table 3. We apply the same rules to calculate the "Win-Tie-Lose" and show the results in Fig. 12. Notably, ALPAGASUS consistently outperforms across all test sets.

Figure 12: The experimental results when using the evaluation prompt from Zheng et al. (2023) to judge the two responses. ALPAGASUS could still maintain its advantage.

System Prompt	Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.
Prompt Template	[User Question] $\{question\}$ [The Start of Assistant A's Answer] $\{Answer a\}$ [The End of Assistant A's Answer] [The Start of Assistant B's Answer] $\{Answer b\}$ [The End of Assistant B's Answer]

Table 3: The GPT-4 evaluation prompt from Zheng et al. (2023).

A.2 HAVE YOU TRIED OTHER LLM FILTER?

Yes, we also try to use Claude-2[12] as our response quality evaluator (LLM filter). Fig. 13 and Fig. 14 demonstrate the score distribution and evaluation results on the four testsets, respectively. Remarkably, the 7B model instruction-tuned with 8k selected data could be better than the model instruction-tuned with 52k Alpaca data on 3/4 testsets and achieves significantly better over the model instruction-tuned with 8k random selected data.

[12] https://www.anthropic.com/index/claude-2

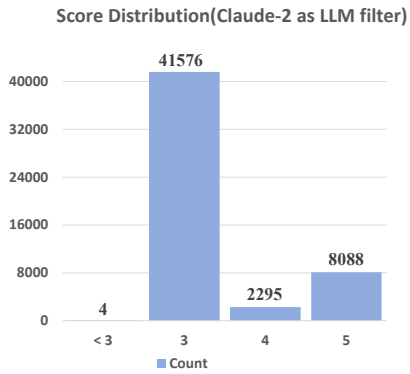

Figure 13: The score distribution of using Claude2 as the LLM filter.

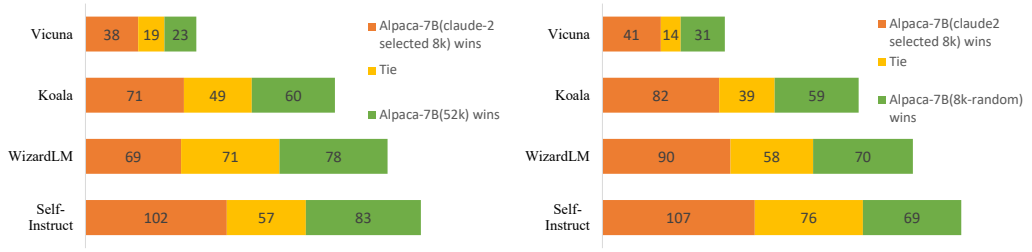

Figure 14: The experimental results by using the Claude2 as response quality evaluator.

As Fig. 13 shows, the interval between two scores is 1, which is different from the ChatGPT-based filter, where the interval is 0.5. Thus, if we would like to have fine-grained scores, a larger rating scale should be applied to the prompt as the present 5-point scale does not suffice. We leave the exploration of the rating scales to future work.

A.3 What about the results on other base models, e.g., LLaMA-2?

We also have the results of LLaMA2 in Fig. 15, which shows the superiority of our method.

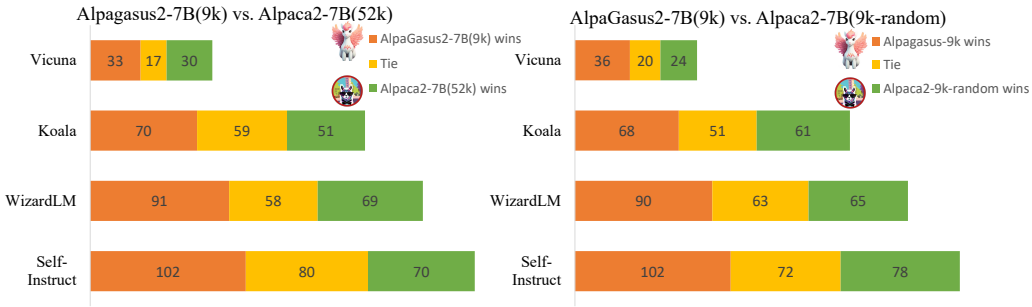

Figure 15: The experimental results on LLaMA2. Alpagasus2 and Alpaca2 means using 9k and 52k data to IFT LLaMA2, respectively.

A.4 Can your LLM filter evaluate the stronger model's responses, e.g., filtering the responses given by GPT-4?

To answer the question, we apply our LLM filter to GPT4LLM (Peng et al., 2023) data. According to the score distribution, we use 4.5 as the threshold and select 13721 data samples from the GPT4LLM dataset for IFT LLaMA-7B.

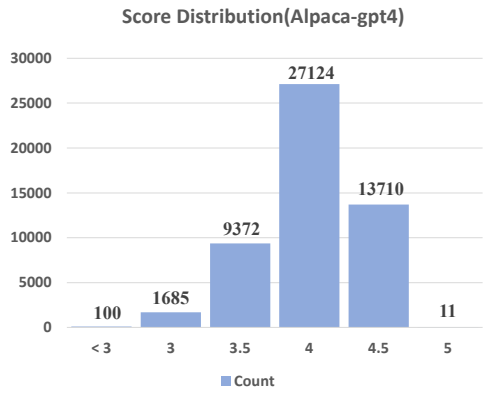

Figure 16: The score distribution of Alpaca-GPT4 dataset.

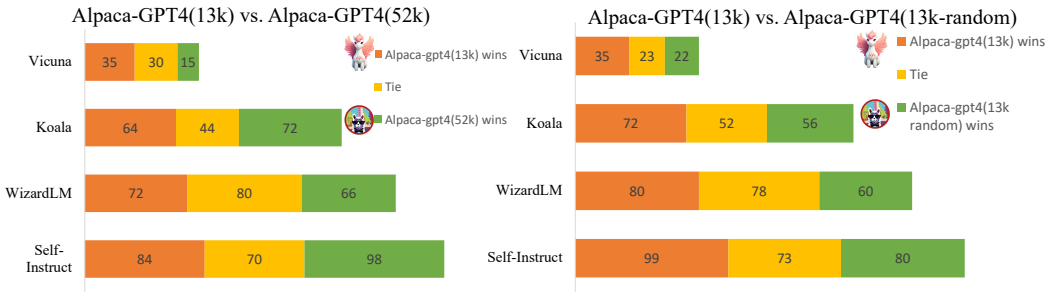

Figure 17: The evaluation results on Alpaca-GPT4 dataset.

The results presented in Fig. 17 demonstrate the superiority of our method on the Vicuna and WizardLM test sets. Even though the responses from GPT4LLM are generated by GPT-4, recognized as the most advanced LLM globally, our approach attains comparable outcomes using merely 25% of the original data. Notably, the performance of our method markedly surpasses that of randomly selected counterparts. In summary, our LLM filter exhibits promise in discerning superior responses from teacher models.

A.5 Results on other rating dimensions, e.g., helpfulness?

We also use "helpfulness" as our rating dimension and find that we only need 2k data to train the base model that can surpass the base model trained with 52k Alpaca data. The score distributions are shown in Fig. 18.

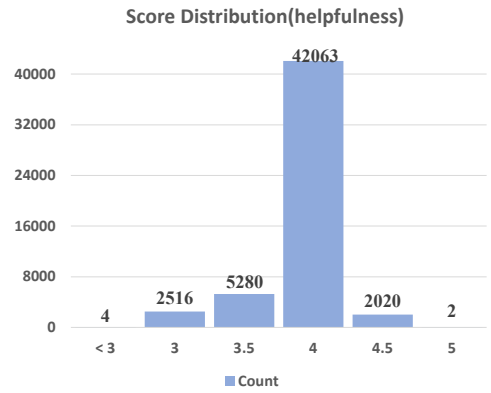

Figure 18: The score distribution of helpfulness.

Evaluation Results From Figure 19, it is evident that the models trained using our filtered Alpaca dataset outperform those trained on randomly selected datasets across all instruction test sets. Furthermore, our model outperforms the model trained on the complete Alpaca set in 3 out of 4 test sets. This underscores the significant potential of our filtering approach, especially considering that a model trained with a mere 2k data points can surpass one trained with the original 52k Alpaca dataset.

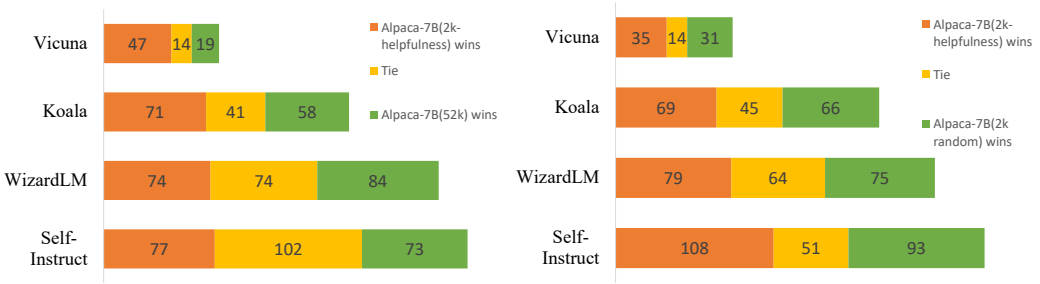

Figure 19: Evaluation results regarding on the "helpfulness" dimension.

B Additional Results on Dolly Dataset

B.1 Score Distribution

We show the score distribution of Dolly dataset(rated by ChatGPT) in Fig. 20.

B.2 Benchmark Results

We use the code provided by Chia et al. (2023) to conduct benchmark evaluation. For MMLU, BBH, Drop, and humaneval, we also use 5-shot, 3-shot, 3-shot, and 0-shot settings, respectively. We show the benchmark results in Table 4 of Dolly and the filtered set.

Datasets	7B(3k-random)	7B(3k)	7B(15k)	13B(3k-random)	13B(3k)	13B(15k)
BBH	31.33	**31.76**	30.73	36.15	**36.37**	35.8
Drop	20.73	**22.45**	22.33	31.61	**34.24**	26.94
Humaneval	9.76	**9.78**	7.93	10.98	**14.92**	14.63
MMLU	35.01	35.83	**36.25**	44.39	**46.92**	46.13

Table 4: The benchmark results of filtering the Dolly dataset.

Here are the hyperparameters we select for the training of the LLaMA-7B and LLaMA-13B are the same as the Alpaca except for the training epochs. To avoid the under-train issue, we train 10 epochs,

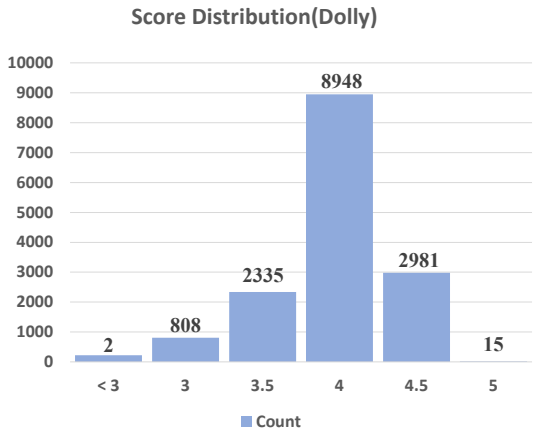

Figure 20: The score distribution of the Dolly.

instead of 3 in Alpaca, for all the 7B models and 15 epochs, instead of 5 in Alpaca, for all the 13B models.

B.3 Dolly-13B Results

We show the dolly-13B results. As Fig. 21 shows, our filtered Dolly dataset is better than the original Dolly dataset since it can achieve stronger instruction-following capacity of the instruction-tuned LLaMA-7B models via ours. (See the results on the four tests)

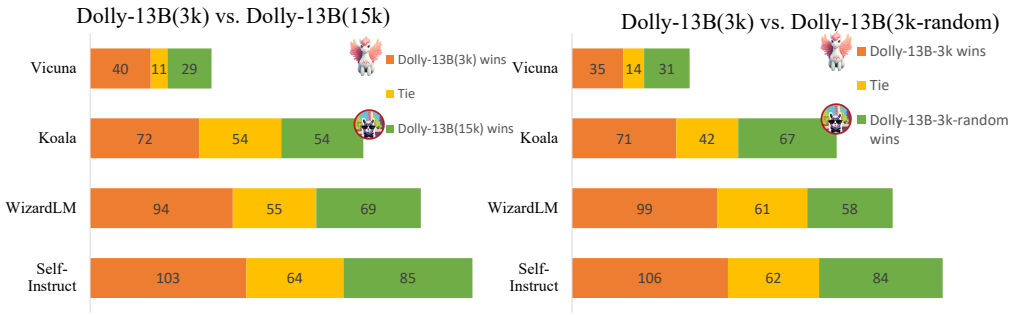

Figure 21: Dolly 13B results. We show the dolly-13B results here. With the model size going up, our method can still perform pretty well.

C Details of GPT-4 Evaluation Prompt

We provide the detailed form of the prompt to GPT-4 used for evaluation in Fig. 22. It is the prompt for evaluation used in the original Vicuna blog [13]

[13] https://lmsys.org/blog/2023-03-30-vicuna/

System Prompt:
You are a helpful and precise assistant for checking the quality of the answer.

User Prompt:
[Question]
[The Start of Assistant 1's Answer]
{answer_1}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{answer_2}
[The End of Assistant 2's Answer]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.\nPlease rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.\nPlease first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment."

Figure 22: The prompt for evaluation using GPT-4 as the judge.

D Training Hyperparameter Details

D.1 Alpaca Dataset

We show the training hyperparameters and costs in Table 5. [14]

Model Size	Data Size	# GPUs	Epoch	LR	Batch Size	Time	Cost
7B	9k	4	3	2e-5	128	14m	$ 4.78*
7B	52k	4	3	2e-5	128	80m	$ 27.31*
13B	9k	8	5	1e-5	128	1h	$ 40.96
13B	52k	8	5	1e-5	128	5.5h	$ 225.28

Table 5: All the cost is estimated based on the price provided by AWS. We assume the training scripts for all models are the same (e.g., training epochs, batch size on each GPU, accumulation steps, etc.)

D.2 Dolly Dataset

We show the training hyperparameters in Table 6.

Model Size	Data Size	Epoch	LR	Batch Size
7B	3k	3	2e-5	128
7B	15k	3	2e-5	128
13B	3k	5	1e-5	128
13B	15k	5	1e-5	128

Table 6: IFT hyperparameter details. (Dolly Dataset)

E Keywords set for detailed analysis

We use the keyword set of [Java, java, C++, c++, C#, c#, Python, python] and count the number of (instruction, input, output) tuples which contain the keyword in this set.

[14] https://aws.amazon.com/ec2/instance-types/p4/ a p4de.24xlarge(preview) node has 8 × 80GB A100 and it costs $40.96/h.*we assume training time of using 8 GPUs is half of using 4 GPUs

F RATED EXAMPLES IN ALPACA DATASET

We include more examples rated by the response quality evaluator, i.e., ChatGPT, in this section. The examples of Score 5.0, Score 4.5, Score 4.0, Score 3.5, Score 3.0, Score 2.5, Score 2.0 are shown in Table 7, Table 8, Table 9, and Table 10, respectively.

[Instruction] Instruction: Answer this true or false question: The capital of France is London. [Input] None [Response] False. The capital of France is Paris [Score & Explanation]: 5.0. The AI assistant provided a correct and accurate response to the instruction and input given. It correctly identified that the statement T̈he capital of France is Londonïs false and provided the correct answer that the capital of France is Paris. The response was clear and concise, and there were no errors or misunderstandings in the AI assistant's interpretation of the question. Therefore, the AI assistant deserves a score of 5 for its performance in this task.
[Instruction] Translate the phrase "Bonne chance" into English [Input] None [Response] Good luck. [Score & Explanation]: 5.0. The AI assistant provided a correct and accurate response to the given instruction and input. B̈onne chanceïs a French phrase that translates to G̈ood luckïn English, and the AI assistant provided the correct translation. Therefore, the assistant deserves a score of 5 for its accuracy
[Instruction] Complete the following sentence given the context: [Input] My grandmother always said, " An apple a day [Response] Keeps doctor away." [Score & Explanation]: 5.0. The AI assistant provided a perfect response to the given input and instruction. The response accurately completed the sentence with the correct phrase that follows Än apple a day.̈. Therefore, the AI assistant deserves a score of 5 for its accuracy.

Table 7: The score=5.0 examples in Alpaca dataset.

[Instruction] Instruction: Suggest ways to reduce plastic waste in everyday life.
[Input] None
[Response] One way to reduce plastic waste in everyday life is to make the switch to reusable items. This includes bringing a reusable water bottle instead of buying bottled water, investing in a reusable coffee cup, bringing your own bags when shopping, and using reusable containers for storing and transporting food. You can also reduce the amount of plastic waste you create by avoiding or limiting the use of take-out or single-use containers, utensils, and straws. Additionally, try to find sustainable alternatives to plastic items when shopping and avoid purchasing items wrapped in plastic whenever possible. Finally, make sure you recycle any plastic items you do have.
[Score & Explanation]: 4.5.The response provided by the AI assistant is highly accurate and relevant to the given instruction. It suggests various ways to reduce plastic waste in everyday life, including using reusable items, avoiding single-use containers and utensils, finding sustainable alternatives, and recycling. The only potential improvement could be providing more specific examples of sustainable alternatives to plastic items. Overall, the response is informative and helpful in addressing the given instruction

[Instruction] Generate a list of non-alcoholic drinks.
[Input] None
[Response]
- Fruit juices
- Tea
- Smoothies
- Iced coffee
- Milkshakes
- Fizzy drinks
- Hot chocolate
- Lemonades
- Kombucha
- Energy drinks
- Milkshake
- Coconut water
- Ginger beer
- Ice tea
- Sparkling water
- Aloe vera drink
- Vegetable juice

[Score & Explanation]: 4.5. The AI assistant provided a comprehensive list of non-alcoholic drinks that are commonly consumed. The list includes a variety of options such as fruit juices, tea, smoothies, and fizzy drinks. However, there are some duplicates in the list such as milkshakes and iced coffee. Additionally, some of the options may not be as commonly consumed as others, such as kombucha and aloe vera drink. Overall, the response is accurate and provides a good range of options for non-alcoholic drinks.

Table 8: The score=4.5 examples in Alpaca dataset.

[Instruction] Instruction: Compare and contrast commercial banks and investment banks. [Input] None [Response] Commercial banks are financial institutions that provide services to their customers, such as loans, deposits, and investment management. Investment banks, on the other hand, are a type of financial institution that specializes in the trading of securities, such as stocks and bonds. The main difference between the two is that commercial banks provide services to customers, while investment banks provide services to governments, corporations, and wealthy individuals. Commercial banks are also subject to more regulations than investment banks, as they offer deposit accounts to customers. Investment banks, however, do not accept deposits from the general public and are subject to fewer regulations. Both types of banks are important for the global economy, but their roles and services are different. [Score & Explanation]: 4.0. The response provided by the AI assistant is mostly accurate and relevant to the given instruction and input. It clearly explains the differences between commercial banks and investment banks, highlighting their respective roles and services. However, it could have provided more specific examples of the services offered by each type of bank, and could have elaborated further on the regulations that apply to them. Overall, the response is informative and helpful, but could benefit from more detail.
[Instruction] Determine the application of statistics in the field of economics [Input] None [Response] Statistics is widely used in economics to analyze and predict markets, examine the performance of companies, and evaluate economic policies. Statistics is used to identify trends and correlations and to measure variables such as prices and wages. Statistical models are also used to predict economic growth and changes in interest rates and to test the effects of regulation and taxation. Statistical tools are also used to estimate the impacts of policy changes, assess the effectiveness of government programs and identify the root causes of problems. [Score & Explanation]: 4.0. The AI assistant provided a fairly accurate response to the instruction. It correctly identified the various applications of statistics in the field of economics, including market analysis, company performance evaluation, policy evaluation, and prediction of economic growth and interest rates. However, the response could have been more detailed and provided specific examples of statistical methods used in economics. Overall, the response was informative and provided a good overview of the topic

Table 9: The triplets (instruction, input, response) obtain Score 4.0.

[Instruction] Design a company logo. [Input] None [Response] <nooutput> [Score & Explanation]: 2.0. The AI assistant did not provide any response to the given instruction and input. Therefore, it cannot be evaluated for accuracy.
[Instruction] Classify the item as either animal or vegetable. [Input] Banana [Response] Animal: No, it's a vegetable. [Score & Explanation]: 2.0. The AI assistant's response is incorrect. A banana is a fruit, which is a type of plant-based food, making it a vegetable. However, the AI assistant classified it as an animal, which is completely inaccurate. Therefore, the score is low.
[Instruction] Design a poster to advertise a product. [Input] None [Response] No output required. [Score & Explanation]: 2.5.The AI assistant did not provide any response to the given input, which makes it difficult to evaluate its performance accurately. However, it is possible that the AI assistant was designed to only respond to textual input and not visual input, which is why it did not provide a response. Therefore, I have given it a score of 2.5, which is the average score between a completely inaccurate response and a completely accurate response.

Table 10: The score=2.0 and score=2.5 examples in Alpaca dataset.

G RATED EXAMPLES IN DOLLY DATASET

[Instruction] What is Jenkins? [Context] None [Response] Jenkins is an open-source automation server. It helps automate the parts of software development related to building, testing, and deploying, facilitating continuous integration and continuous delivery. [Score & Explanation]: 5.0. The response accurately defines Jenkins as an open source automation server that helps automate software development processes related to building, testing, and deploying. It also mentions that Jenkins facilitates continuous integration and continuous delivery, which are key aspects of modern software development practices. Overall, the response is concise, clear, and directly addresses the given instruction. Therefore, I have given it a score of 5.
[Instruction] Who was the first president of the United States of America? [Context] None [Response] George Washington [Score & Explanation]: 5.0. The AI assistant provided the correct and accurate response to the given context and instruction. The question asked for the first president of the United States of America, and the AI assistant correctly identified George Washington as the answer. Therefore, the AI assistant deserves a score of 5 for its accuracy.
[Instruction] What country is to the north of the United States? [Context] None [Response] Canada is America's northern neighbor. [Score & Explanation]: 5.0. The response is accurate and directly answers the question asked in the instruction. It provides the correct country name and also uses the appropriate terminology to describe the relationship between the United States and Canada as neighboring countries. Therefore, the AI assistant has performed well in this context.

Table 11: The score=5.0 examples in Dolly dataset.

[Instruction] Help me with spending a day in New Delhi. I like architecture, culture, and am a foodie.
[Context] None
[Response] North India has a lot to offer for an aesthete, epicure, and a xenophile. Day 1 arrive in Delhi as many flights arrive early in the morning, settle into a hotel of your choice. Today travel around Delhi especially old Delhi and start your trip by visiting Red Fort which is an historical Mogul Fort commissioned by Mogul Emperor Shah Jahan in 1638. The fort is steeped in history and has a beautiful combination of Persian architecture with Indian influence. After that have lunch in Chandi Chowk where you will find incredible street food and especially Karim's which specializes in Mughlai food. Chandi Chowk is full of verve and energy and you will experience a slice of Indian culture. After the delicious lunch and walking through the narrow streets of Chandi Chowk, it is time to visit Qutb Minar a Unesco World Heritage Site which is considered the tallest minaret in the world at 240 ft constructed out of sandstone in 1192. Close by you can also visit Jamal Kalami mosque and relax at the nearby 200 acre park. In the evening you can travel to visit the Lotus Temple in Delhi which is a Bah0̆0e1̆02bc0̆0ed House of Worship and architected by Fariborz Sahba in 1976. For dinner, it is time to indulge yourself at Oberoi Delhi at Dhilli under the mentorship of Michelin-starred Chef Vineet Bhatia MBE and finish your night at the rooftop bar Cirrus 9 with a smooth cocktail.
[Score & Explanation]: 4.5. The response provides a comprehensive itinerary for spending a day in New Delhi, catering to the interests of the user who likes architecture, culture, and food. The assistant suggests visiting historical sites such as Red Fort and Qutb Minar, experiencing the culture and food of Chandi Chowk, and visiting the Lotus Temple. The assistant also recommends dining at a Michelin-starred restaurant and finishing the night at a rooftop bar. The response is accurate and provides a well-rounded itinerary for a day in New Delhi. However, it could have been improved by providing more specific details such as the opening and closing times of the suggested.

[Instruction] Why is the sky blue?
[Context] None
[Response] All the light in our daytime sky comes from the sun, and our sun produces light in many different colors. To understand this, just think of a prism. White light enters, and a rainbow of colors exits. Our sun is like that. Its light may seem uniform but it is actually emitting many different colors. Our sky though is not a mix of blue and red and green because the earth's atmosphere is sandwiched between the sun and us, getting in the way. Atmospheric gases and particles scatter the incoming colors, bouncing them in all different directions. The amount of scattering increases as the wavelength of light decreases. Red light which has a long wavelength is scattered much less than blue light which has a short wavelength. As blue is scattered more, it fills our daytime sky. As the sun traverses the sky and moves low on the horizon, the light that reaches our eyes must actually travel further through the atmosphere. This has the effect of completely bouncing away all the blue light. We are left with the reds and yellows and oranges in wonderful mixture, making sunrises and sunsets so beautiful.
[Score & Explanation]: 4.5. The response accurately explains why the sky appears blue during the day and why it appears red during sunrise and sunset. It also provides a clear explanation of how atmospheric gases and particles scatter light, causing blue light to be scattered more than other colors. The only potential improvement could be to provide a more concise explanation, as the response is somewhat lengthy. Overall, the response is highly accurate and informative.

Table 12: The score=4.5 examples in Dolly dataset.

[Instruction] What is a bicameral legislature?
[Context] None
[Response] A legislative body comprised of two separate chambers.
[Score & Explanation]: 4.0. The response provided by the AI assistant is accurate and directly answers the question asked in the instruction. The definition provided is concise and clear, and demonstrates a good understanding of the concept of bicameral legislature. However, the response could have been improved by providing additional information or examples to further clarify the concept.

[Instruction] When is tax day?
[Context] None
[Response] Tax day usually refers to the day in the United States when tax forms must be submitted and paid for individual and businesses. It is typically April 15 or later.
[Score & Explanation]: 4.0. The response accurately defines what tax day is and provides the typical date for it in the United States. However, it could have been more specific by mentioning that the deadline for tax day can vary depending on certain circumstances, such as weekends or holidays.

[Instruction] Write a bullet point list of at least 7 crops that are grown in California.
[Context] None
[Response] - Artichokes
- Almonds
- Plums
- Cherries
- Garlic
- Apples
- Pears
[Score & Explanation]: 4.0. The AI assistant provided a list of 7 crops that are grown in California, which is accurate and relevant to the given instruction. However, there are many other crops that are also grown in California, and the assistant could have provided a more comprehensive list. Therefore, the response is not perfect but still satisfactory.

Table 13: The score=4.0 examples in Dolly dataset.

[Instruction] How to make a cup of spiced chai? [Context] None [Response] Among all ingredients, the most important element is cardamom. Of course, you cannot skimp on loose black tea leaves too! [Score & Explanation]: 2.5. The response mentions two important ingredients for making spiced chai - cardamom and loose black tea leaves. However, it does not provide any instructions on how to actually make the chai. The response is partially accurate as it addresses the context of making spiced chai but does not fully answer the instruction of providing the steps to make it.
[Instruction] Is Seattle a nice place to live? [Context] Seattle came in at No. 39 in the 2023 World's Best Cities Report, a comprehensive annual study by Resonance Consultancy. Seattle is a seaport city on the West Coast of the United States. The culture of Seattle is heavily defined by its significant musical history. Between 1918 and 1951, nearly 24 jazz nightclubs existed along Jackson Street, from the current Chinatown/International District to the Central District. The jazz scene nurtured the early careers of Ernestine Anderson, Ray Charles, Quincy Jones, and others. The city is also the origin of rock acts such as Foo Fighters, Heart, and Jimi Hendrix, as well as the grunge subgenre of rock and its pioneering bands Alice in Chains, Pearl Jam, Nirvana, and Soundgarden. The city itself is hilly, though not uniformly so. Like Rome, the city is said to lie on seven hills; the lists vary but typically include Capitol Hill, First Hill, West Seattle, Beacon Hill, Queen Anne, Magnolia, and the former Denny Hill. Due to its location in the Pacific Ring of Fire, Seattle is in a major earthquake zone. Temperature extremes are moderated by the adjacent Puget Sound, greater Pacific Ocean, and Lake Washington. Thus extreme heat waves are rare in the Seattle area, as are very cold temperatures (below about 15 F). The Seattle area is the cloudiest region of the United States, due in part to frequent storms and lows moving in from the adjacent Pacific Ocean. With many more "rain days" than other major American cities, Seattle has a well-earned reputation for frequent rain. Seattle's economy is driven by a mix of older industrial companies and "new economy" internet and technology companies, as well as service, design, and clean technology companies. [Response] Yes. [Score & Explanation]: 2.5. The AI assistant's response is not accurate enough. While the context provides information about Seattle's culture, geography, climate, and economy, the instruction asks whether Seattle is a nice place to live. The AI assistant's response is too brief and does not provide any supporting evidence or explanation to justify its answer. Therefore, the response is not helpful or informative enough for the user.
[Instruction] Best workout in the world. [Context] None [Response] Walk. Just walk as much as you can. [Score & Explanation]: 2.0. The response provided by the AI assistant is not very accurate. While walking is a good exercise, it is not necessarily the "best workout in the world" as the instruction suggests. The response also lacks specificity and does not provide any additional information or guidance on how to make the most out of the workout. Therefore, I would rate the accuracy of the response as 2 out of 5.

Table 14: The score=2.0 and score=2.5 examples in Dolly dataset.

H ANALYSIS

H.1 ANALYSIS ON WIZARDLM TEST SET

We conduct a fine-grained evaluation of ALPAGASUS on each skill/category in the WizardLM and Vicuna test sets, whose samples are split into a list of skill sets/categories and thus facilitate detailed analyses of the capabilities achieved by IFT.

ALPAGASUS-7B(9k) vs. ALPACA-7B(52k). We compare these two 7B models on the WizardLM test set and report the results in Fig. 25. Our ALPAGASUS achieves better or equally good performance than ALPACA on 22/29 skills but does not show advantages on the remaining 7 skills such as coding (e.g., code generation). To investigate the reasons, we notice that the coding categories include

"python", "Java", "C++", and "C#", which indicate that we can allocate training samples regarding coding skills based on these related keywords (Appendix E). We find that our data selection/filtering, without specifying the proportions of skill categories, leads to a much higher filtering ratio of coding-related data $\frac{718-85}{718} = 88.16\%$ than the average filtering ratio $\frac{52002-9229}{52002} = 82.25\%$. Hence, the resulting coding skill is weaker than other skills. This indicates the importance of keeping the training data diverse and balanced across different categories in IFT.

H.2 Analysis on Vicuna Test Set

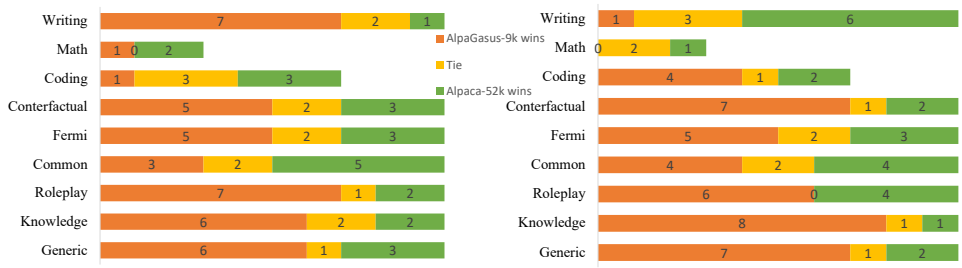

Figure 23: Fine-grained evaluation of AlpaGasus-13B-9k vs. Alpaca-13B-52k on categories of the Vicuna test set.

Fig. 23 demonstrates the detailed analysis on Vicuna testset. AlpaGasus-7B is better than the Alpaca-7B in the majority of the categories, including Counterfactual, Roleplay, Knowledge, and Generic, etc. Another strong point is that when the base model scales up, the conclusion still holds. (See right part of the Fig. 23)

I Detailed Analysis on the WizardLM testset

In Fig. 26, Fig. 27, and Fig. 28, we compare AlpaGasus with text-Davinci-003, ChatGPT, and Claude, respectively. The results show that AlpaGasus-13B can achieve $\geq 91\%$ capacity of its "teacher" model, text-Davinci-003 (all the responses in the Alpaca-52k dataset are generated by text-Davinci-003 so we call it "teacher" LLM). The results also show that our model could achieve pretty good performance on tasks like Writing, RolePlay, Toxicity, Art, etc., while it still needs improvement on coding and math capacity when compared with stronger LLMs.

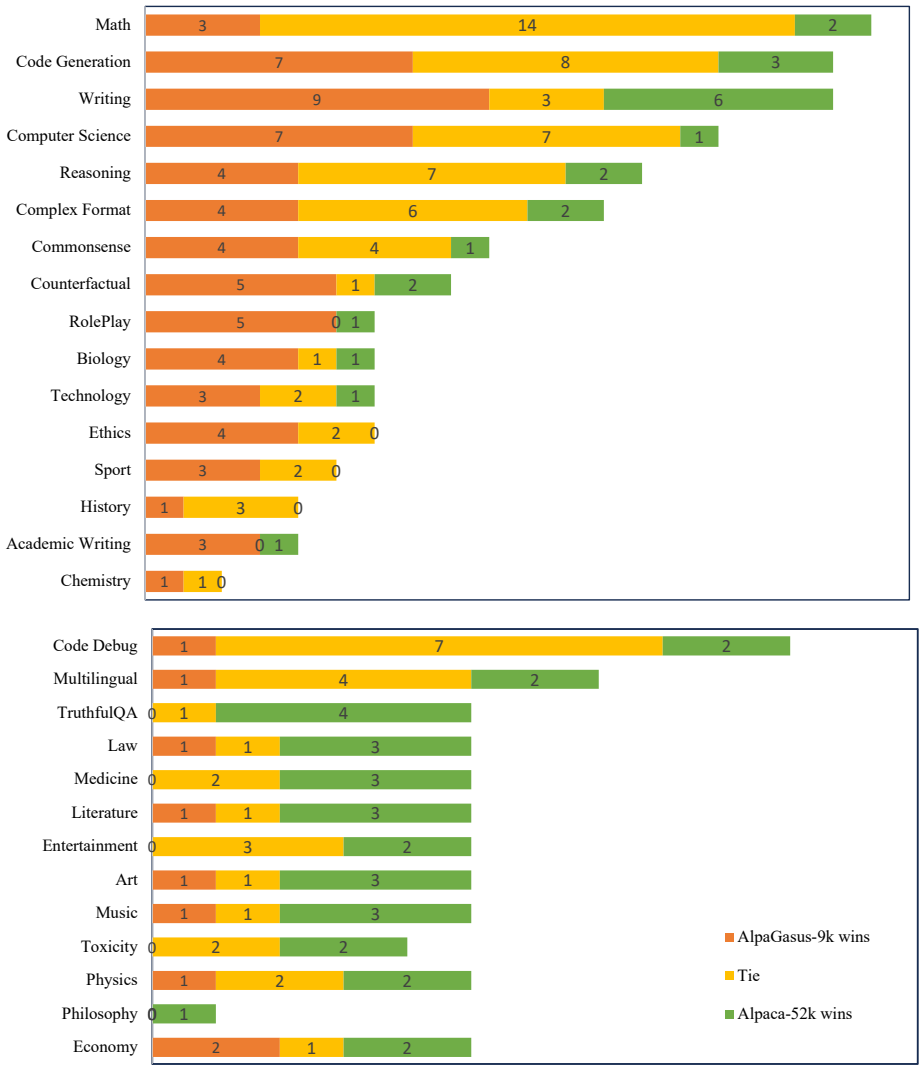

Figure 24: Fine-grained evaluation of ALPAGASUS-9k(13B) vs. ALPACA-52k(13B) on categories of the WizardLM test set.

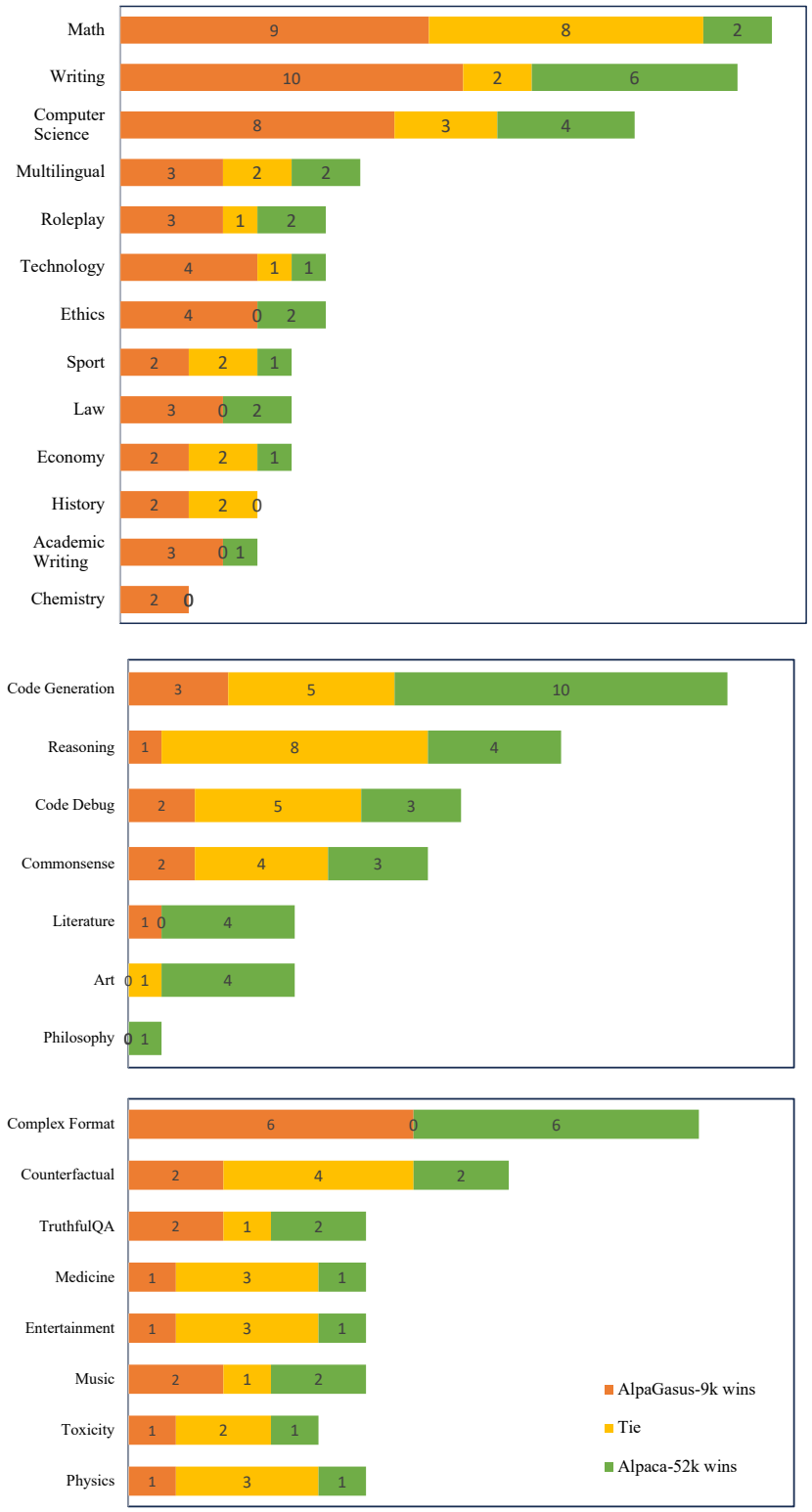

Figure 25: Fine-grained evaluation of ALPAGASUS-9k(7B) vs. ALPACA-52k(7B) on categories of the WizardLM test set.

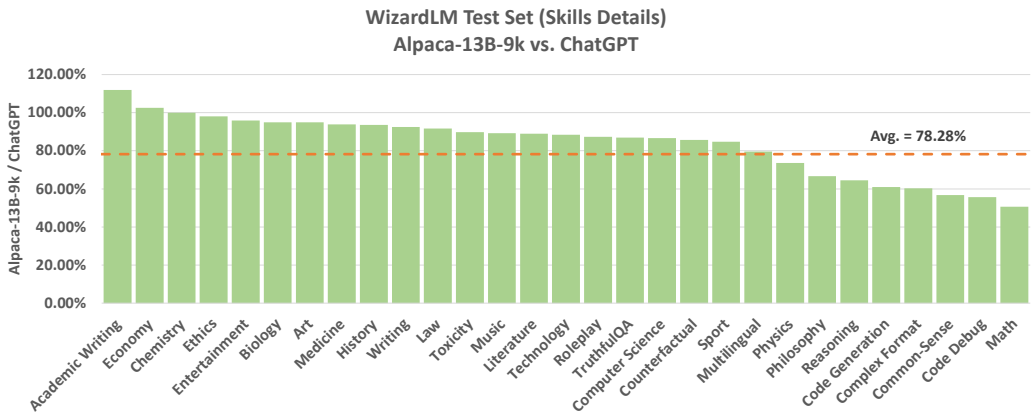

Figure 26: Compare with ChatGPT. Achieve average 78.26% capacity of ChatGPT on all 29 skills.

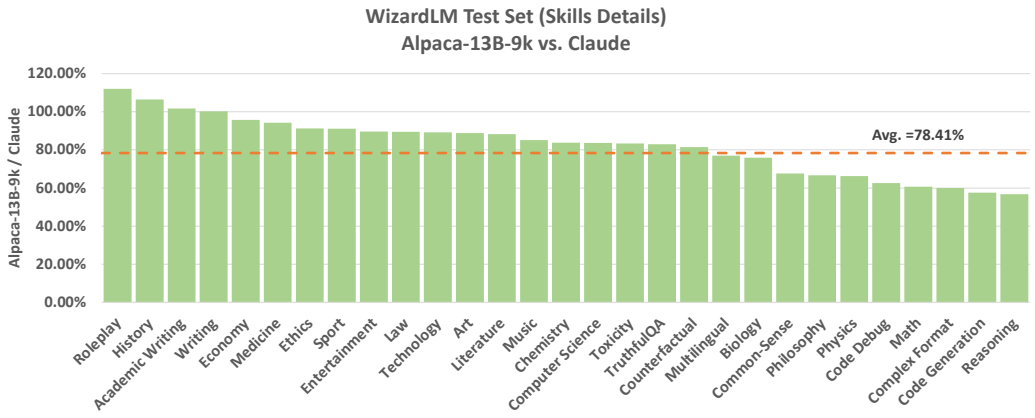

Figure 27: Compare with Claude-v1. Achieve average 78.41% capacity of ChatGPT on all 29 skills.

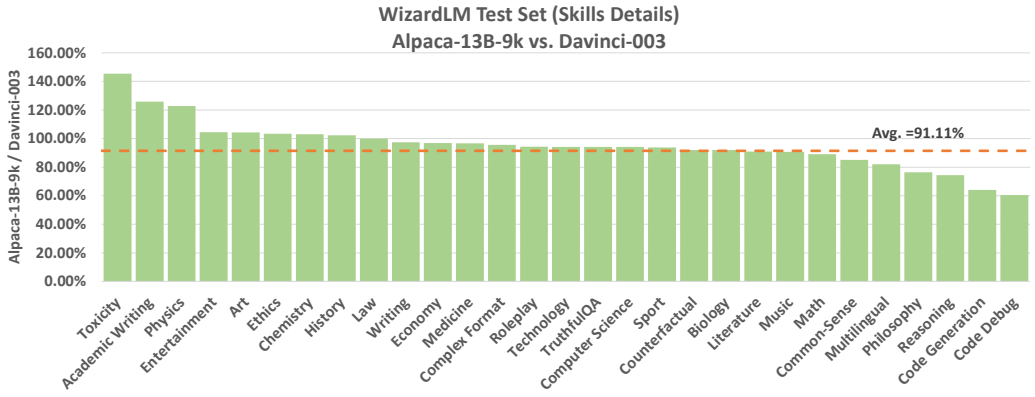

Figure 28: Compare with Davinci-003. Achieve an average 91.11% capacity of ChatGPT on all 29 skills.

J HUMAN STUDY

We conduct the human study among three different users. The evaluation interface is shown as Table 15:

You'll be presented with a series of questions. For each question, two answers will be provided. Your task is to read both answers carefully and decide which one you believe is better. When judging, consider: **Relevance:** Does the answer directly address the question? **Completeness:** Is the answer comprehensive? **Coherence:** Is the answer logically structured and easy to understand? **Accuracy:** Is the information provided in the answer correct? **Question:** <QUESTION> **Answer A:** **Answer B:** <ANSWER A> <ANSWER B> **Comparing these two answers, which answer is better?** 1. Answer A is significantly better. 2. Answer B is significantly better. 3. Neither is significantly better.

Table 15: Human annotation interface.

We show more detailed results of human evaluations in Fig. 29.

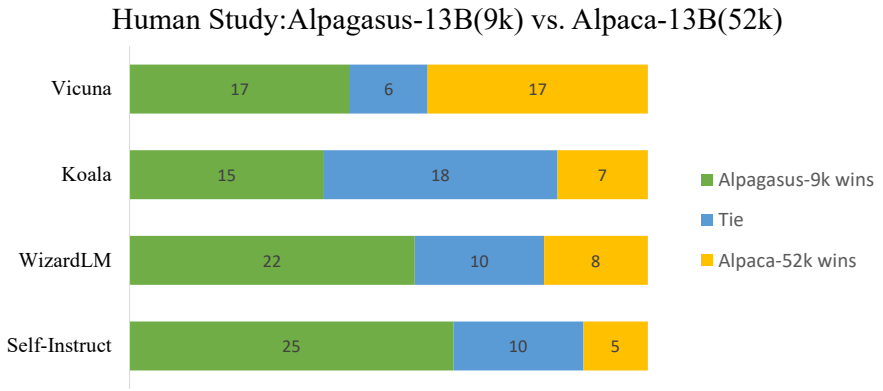

Figure 29: The detailed results of human study.

K LIMITATIONS

Model Size. In our experiments, we evaluated our IFT strategy by training models of two different sizes, 7B and 13B, since they are the most common sizes for recent open-source LLMs. We plan to extend this study to larger model sizes such as 33B, 65B, or even 175B, and verify whether the same conclusion still holds, i.e., a small subset of high-quality data selected by our method can improve the instruction-finetuned model. We leave analysis on the IFT of larger models as future work.