INSTRUCTION TUNING WITH GPT-4

Baolin Peng[*], **Chunyuan Li**[*], **Pengcheng He**[*], **Michel Galley, Jianfeng Gao**
Microsoft Research
{bapeng,chunyl,penhe,mgalley,jfgao}@microsoft.com

ABSTRACT

Prior work has shown that finetuning large language models (LLMs) using machine-generated instruction-following data enables such models to achieve remarkable zero-shot capabilities on new tasks, and no human-written instructions are needed. In this paper, we present the first attempt to use GPT-4 to generate instruction-following data for LLM finetuning. Our early experiments on instruction-tuned LLaMA models show that the 52K English and Chinese instruction-following data generated by GPT-4 leads to superior zero-shot performance on new tasks to the instruction-following data generated by previous state-of-the-art models. We also collect feedback and comparison data from GPT-4 to enable a comprehensive evaluation and reward model training. We make our data generated using GPT-4 as well as our codebase publicly available. [1]

1 INTRODUCTION

Large Language Models (LLMs) have shown impressive generalization capabilities such as in-context-learning (Brown et al., 2020) and chain-of-thoughts reasoning (Wei et al., 2022). To enable LLMs to follow natural language instructions and complete real-world tasks, researchers have been exploring methods of instruction-tuning of LLMs. This is implemented by either finetuning the model on a wide range of tasks using human-annotated prompts and feedback (Ouyang et al., 2022), or supervised finetuning using public benchmarks and datasets augmented with manually or automatically generated instructions (Wang et al., 2022b). Among these methods, Self-Instruct tuning (Wang et al., 2022a) is a simple and effective method of aligning LLMs to human intent, by learning from instruction-following data generated by state-of-the-art instruction-tuned teacher LLMs. It turns out that the line of instruction-tuning research has produced effective means to improve the zero and few-shot generalization abilities of LLMs. The recent success of ChatGPT (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b) offers tremendous opportunities to improve open-source LLMs using instruction-tuning. LLaMA (Touvron et al., 2023) is a series of open-sourced LLMs, which match the performance of proprietary LLMs such as GPT-3. To teach LLaMA to follow instructions, Self-Instruct tuning has been quickly adopted given its superior performance and low cost. For example, Stanford Alpaca (Taori et al., 2023) uses 52K instruction-following samples generated by GPT-3.5, while Vicuna (Vicuna, 2023) uses around 700K instruction-following samples (70K conversions) shared user-ChatGPT (ShareGPT, 2023).

To advance the state of the art of instruction-tuning for LLMs, we propose for the first time to use GPT-4 as a teacher for self-instruct tuning. Our paper makes the following contributions:

- *GPT-4 data*. We release data generated by GPT-4, including the 52K instruction-following dataset in both English and Chinese, and the GPT-4-generated feedback data that rate the outputs of three instruction-tuned models.

- *Models & Evaluation*. Based on the GPT-4-generated data, we have developed instruction-tuned LLaMA models and reward models. To evaluate the quality of instruction-tuned LLMs, we use three metrics evaluated on test samples (*i.e.,* unseen instructions): human evaluation on three alignment criteria, automatic evaluation using GPT-4 feedback, and ROUGE-L on un-natural

---

[*]Equal Contribution
[1]https://instruction-tuning-with-gpt-4.github.io/
Note: This is a preliminary release, and we will continue to expand the dataset and will finetune larger models.

Algorithm 1: Pseudo code for prompt engineering, GPT-4 call and hyper-parameters in data generation. Each instruction instance is used as variables in the prompt template, the data flow is highlighted in blue.

```
1  PROMPT_DICT{
2  prompt_input: (
3    "Below is an instruction that describes a task, paired with an input that provides further context."
4    "Write a response that appropriately completes the request.\n\n"
5    "### Instruction: \n {instruction} \n\n ### Input: {input} \n\n ### Response:"
6  ),
7  prompt_no_input: (
8    "Below is an instruction that describes a task. "
9    "Write a response that appropriately completes the request.\n \n"
10   "### Instruction: \n {instruction} \n\n ### Response:" )
11 }
12 output = openai.ChatCompletion.create(
13       model="gpt-4",
14       messages=["role":  "user", "content":  prompt],
15       temperature = 1.0,
16       top_p=1.0,   # nucleus sampling over entire vocabulary
17       max_tokens=512    # the max number of generated tokens
18 )
```

instructions (Honovich et al., 2022). Our empirical study validates the effectiveness of using GPT-4-generated data for LLM instruction-tuning, and suggests practical tips of building a general-purpose instruction-following agent powered by LLMs.

2 DATASET

Data Collection.   We reuse *52K unique instructions* in the instruction-following data collected in the Alpaca dataset (Taori et al., 2023). Each `instruction` describes the task the model should perform. We follow the same prompting strategy to consider cases with and without `input`, which is the optional context or input for the task. The `output` answers to the instruction instance using LLMs. In the Alpaca dataset, the output is generated using GPT-3.5 (`text-davinci-003`) but we instead consider GPT-4 (`gpt-4`) for data generation. Specifically, we generate the following four datasets with GPT-4:

(1) *English Instruction-Following Data*: For the 52K instructions collected in Alpaca (Taori et al., 2023), one English GPT-4 answer is provided for each. The details are described in Algorithm 1. We leave it as future work to follow an iterative process to construct our own instruction set using GPT-4 and self-instruct (Wang et al., 2022a).

(2) *Chinese Instruction-Following Data*: We use ChatGPT to translate the 52K instructions into Chinese and ask GPT-4 to answer them in Chinese. This allows us to build a Chinese instruction-following model based on LLaMA, and study cross-language generalization ability of instruction-tuning.

(3) *Comparison Data*: We ask GPT-4 to rate its own response from 1 to 10. Furthermore, we ask GPT-4 to compare and rate the responses from the three models, including GPT-4, GPT-3.5 and OPT-IML (Iyer et al., 2022). This is used to train reward models.

(4) *Answers on Unnatural Instructions*: The GPT-4 answers are decoded on the core dataset of 68K instruction-input-output triplets (Honovich et al., 2022). The subset is used to quantify the gap between GPT-4 and our instruction-tuned models at scale.

Data Statistics.   We compare the English output response sets of GPT-4 and GPT-3.5 in Figure 1. For each output, the root verb and the direct-object noun are extracted; The frequency over the unique verb-noun pairs are computed over each output set. The verb-noun pairs whose frequency are higher than 10 are displayed in Figure 1(a) and (b), and the most frequent 25 pairs of two sets are compared

(a) GPT-4

(b) GPT3

(c) Frequencies of top 25 verb-noun pairs

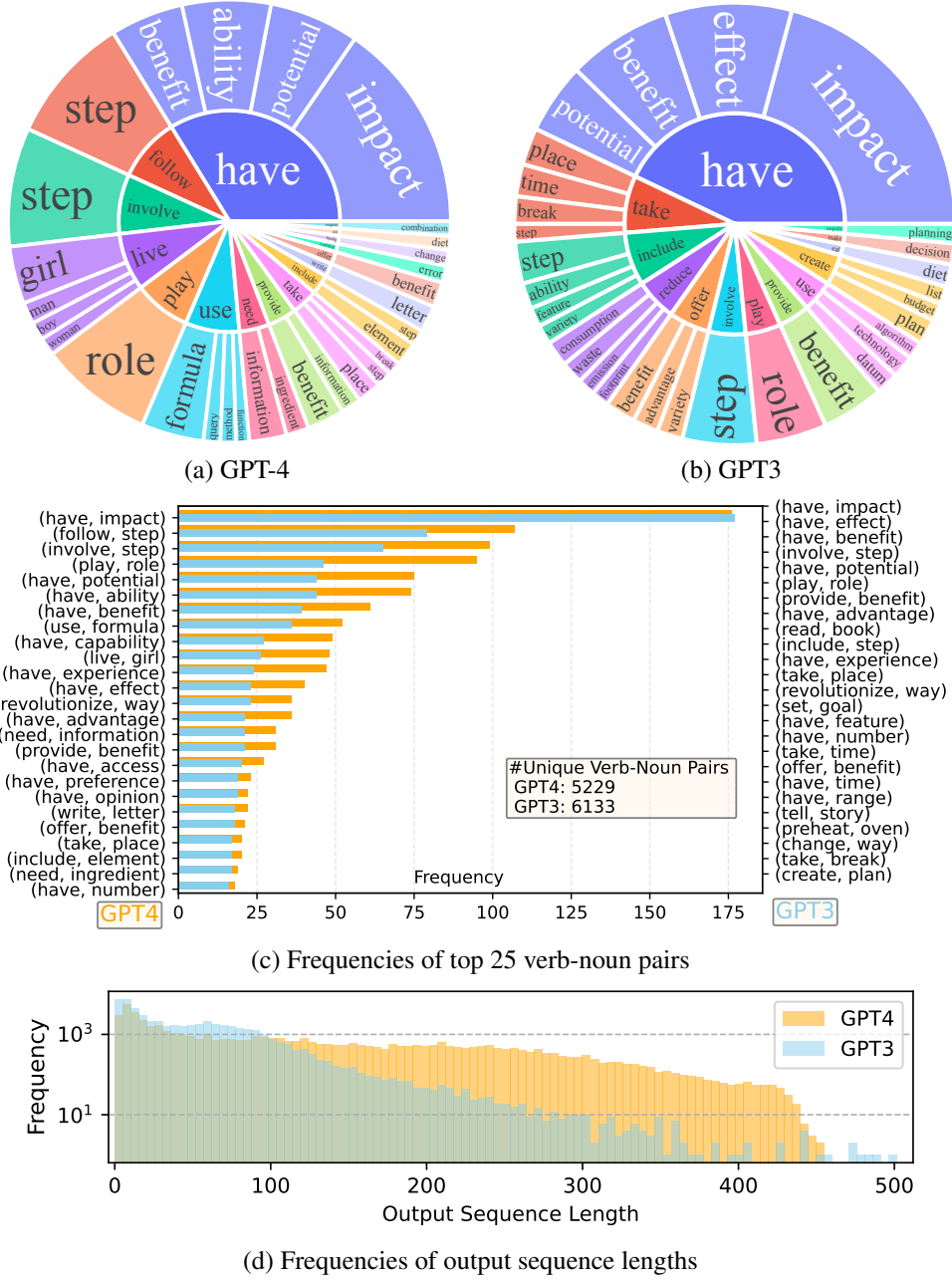(d) Frequencies of output sequence lengths

Figure 1: Comparison of generated responses using GPT-4 and GPT-3: (a,b) The root verb-noun pairs of GPT-4 and GPT-3, where the inner circle of the plot represents the root verb of the output response, and the outer circle represents the direct nouns. (c) The top 25 verb-noun pairs and their frequencies. (d) Comparison of output sequence length.

in Figure 1(c). The frequency distributions of the sequence length are compared in Figure 1(d). GPT-4 tends to generated longer sequences than GPT-3.5. The GPT-3.5 data in Alpaca exhibits an output distribution with a longer tail than our GPT-4-generated output distribution, probably because the Alpaca dataset involves an iterative data collection process to remove similar instruction instances at each iteration, which is absent in our current one-time data generation. Despite this simple process, the GPT-4 generated instruction-following data demonstrates more favorable alignment performance, as shown in experiments later.

3 INSTRUCTION-TUNING LANGUAGE MODELS

3.1 SELF-INSTRUCT TUNING

We train two models using supervised finetuning using the LLaMA 7B checkpoint: (i) **LLaMA-GPT4** is trained on 52K English instruction-following data generated by GPT-4, which distribution is displayed in Figure 1. (ii) **LLaMA-GPT4-CN** is trained on 52K Chinese instruction-following data from GPT-4. We follow the training schedule in (Taori et al., 2023) for fair comparisons. These models are used to study the data quality of GPT-4 and the cross-language generalization properties when instruction-tuning LLMs in one language.

3.2 REWARD MODELS

Reinforcement Learning from Human Feedback (RLHF) aims to align the LLM behavior with human preferences in order to make it more useful. One key component of RLHF is reward modeling, where the problem is formulated as a regression task to predict a scalar reward given a prompt and a response (Askell et al., 2021; Ouyang et al., 2022). This approach typically requires large-scale comparison data, where two model responses on the same prompt are compared Ouyang et al. (2022). Existing open-source works such as Alpaca, Vicuna, and Dolly (Databricks, 2023) do not involve RLHF due to the high cost of labeling comparison data. Meanwhile, recent studies show that GPT-4 is capable of identifying and fixing its own mistakes, and accurately judging the quality of responses(Peng et al., 2023; Bai et al., 2022; Madaan et al., 2023; Kim et al., 2023). Therefore, to facilitate research on RLHF, we have created comparison data using GPT-4, as described in Section 2.

To evaluate data quality, we train a reward model based on OPT 1.3B (Iyer et al., 2022) to rate different responses. For each instance of the comparison data involving one prompt x and K responses, GPT-4 assigns a score $s \in [1, 10]$ for each response. There are C_2^K unique pairs constructed from this instance, each pair is (y_l, y_h), whose corresponding scores follow $s_l < s_h$. A reward model r_θ parameterized by θ is trained with the objective: $\min \log(\sigma(r_\theta(x, y_h) - r_\theta(x, y_l)))$, where σ is the sigmoid function. The distribution of the comparison data is shown in Figure 2.
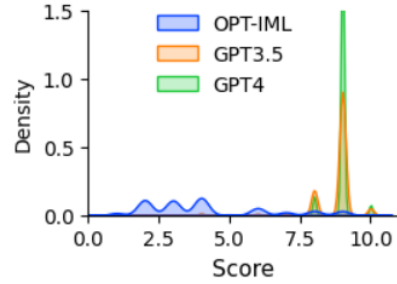

Figure 2: The distribution of comparison data.

4 EXPERIMENTAL RESULTS
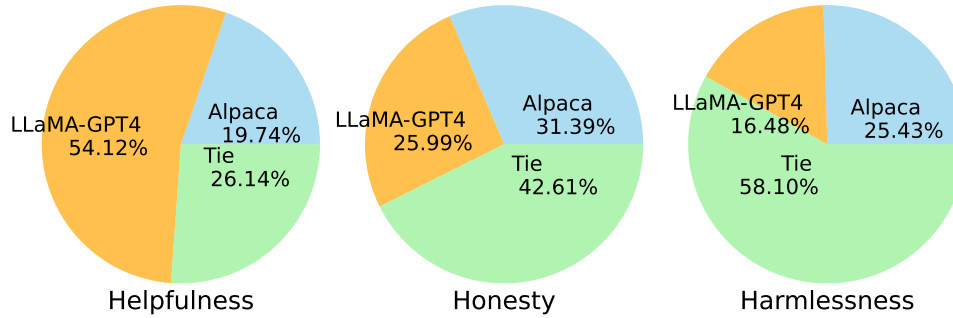
4.1 BENCHMARKS

It is known that LLM evaluation remains a significant challenge. Our goal is to evaluate self-instruct tuned models on GPT-4 data on unseen instructions, to study their ability to follow instructions for arbitrary tasks. Specifically, we use three established datasets in our study:

- *User-Oriented-Instructions-252* [2] (Wang et al., 2022a) is a manually curated set involving 252 instructions, motivated by 71 user-oriented applications such as Grammarly, StackOverflow, Overleaf, rather than well-studied NLP tasks.
- *Vicuna-Instructions-80*[3] (Vicuna, 2023) is a dataset synthesized by **gpt-4** with 80 challenging questions that baseline models find challenging. Beside generic instructions, there are 8 categories, including knowledge, math, Fermi, counterfactual, roleplay, generic, coding, writing, common-sense.
- *Unnatural Instructions*[4] (Honovich et al., 2022) is a dataset of 68,478 samples synthesized by **text-davinci-002** using 3-shot in-context-learning from 15 manually-constructed examples.
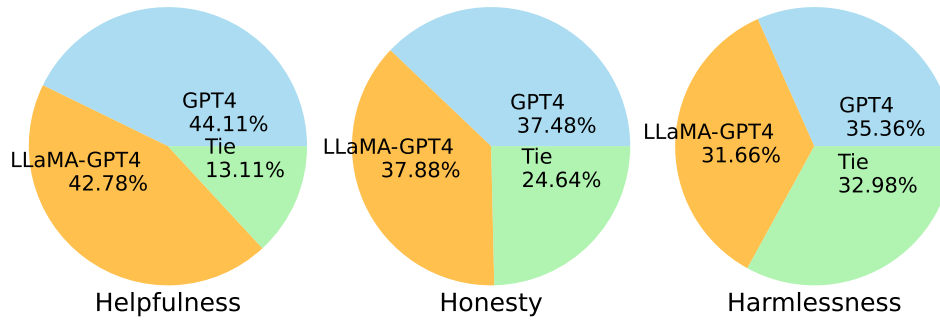
---

[2] https://github.com/yizhongw/self-instruct/blob/main/human_eval/user_oriented_instructions.jsonl
[3] https://github.com/lm-sys/FastChat/blob/main/fastchat/eval/table/question.jsonl
[4] https://github.com/orhonovich/unnatural-instructions

(a) LLaMA-GPT4 vs Alpaca (*i.e.,* LLaMA-GPT3)

(b) LLaMA-GPT4 vs GPT-4

Figure 3: Human evaluation.

4.2 HUMAN EVALUATION WITH ALIGNMENT CRITERIA

To evaluate the alignment quality of our instruction-tuned LLMs, we follow alignment criteria from Anthropic Askell et al. (2021): an assistant is *aligned* if it is helpful, honest, and harmless (HHH). These criteria are used to evaluate how well an AI system is aligned with human values.

- *Helpfulness*: whether it helps humans achieve their goals. A model that can answer questions accurately is helpful.
- *Honesty*: whether it provides true information, and expresses its uncertainty to avoid misleading human users when necessary. A model that provides false information is not honest.
- *Harmlessness*: whether it does not cause harm to humans. A model that generates hate speech or promotes violence is not harmless.

Based on HHH alignment criteria, we used Amazon Mechanical Turk to perform human evaluation on the model generation results. Please find the interface in Appendix Section A.1. Following (Wang et al., 2022a; Taori et al., 2023), we consider 252 user-oriented instructions for evaluation. We display the human evaluation results in pie charts in Figure 3.

First, we compare the quality of generated responses from two instruction-tuned LLaMA models, which are fine-tuned on data generated by GPT-4 and GPT-3, respectively. Note that aligning LLaMA to GPT-3 corresponds to the Stanford Alpaca model. From Figure 3(a), we observe that (i) For the "Helpfulness" criterion, GPT-4 is the clear winner with 54.12% of the votes. GPT-3 only wins 19.74% of the time. (ii) For the "Honesty" and "Harmlessness" criteria, the largest portion of votes goes to the tie category, which is substantially higher than the winning categories but GPT-3 (Alpaca) is slightly superior.

Second, we compare GPT-4-instruction-tuned LLaMA models against the teacher model GPT-4 in Figure 3(b). The observations are quite consistent over the three criteria: GPT-4-instruction-tuned LLaMA performs similarly to the original GPT-4. We conclude that learning from GPT-4 generated

5

(a) Ranked groups against ChatGPT

(b) Ranked groups against GPT-4

(c) All chatbots against ChatGPT
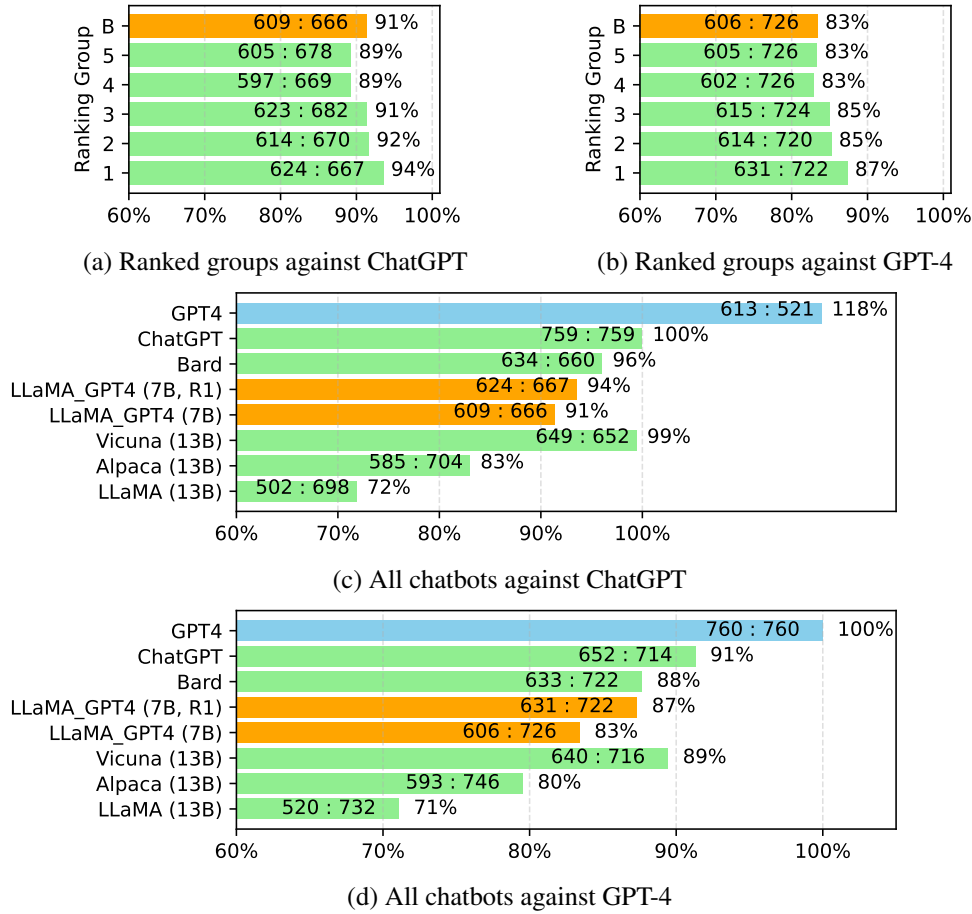
(d) All chatbots against GPT-4

Figure 4: Performance comparisons evaluated by GPT-4. Each bar represents an evaluation result between two models; the sum of scores are computed and reported (the full score is 800). The relative score is reported in percentage, which is computed as the ratio against a strong opponent model. (a,b) The comparisons of responses from LLaMA_GPT4 ranked by our reward model. 'B' indicates the baseline that the model decodes one response per question. (c,d) All chatbots are compared against ChatGPT and GPT-4, respectively.

data can lead to very comparable performance with the original GPT-4 on the unseen instructional tasks, which suggests a promising direction to developing state-of-the-art instruction-following LLMs.

## 4.3 COMPARISONS WITH SoTA USING AUTOMATIC EVALUATION

Automatic Evaluation with GPT-4. Following (Vicuna, 2023), we employ GPT-4 to automatically evaluate the generated responses of different models on 80 unseen questions in (Vicuna, 2023). We first collect answers from two chatbots, including LLaMA-GPT-4 (7B) and GPT-4, and use the release answers of other chatbots from (Vicuna, 2023), including LLaMA (13B), Alpaca (13B), Vicuna (13B), Bard (Google, 2023), and ChatGPT. For each evaluation, we ask GPT-4 to rate the response quality between two models with scores from 1 to 10. We compare all models against a strong competing model such as ChatGPT and GPT-4, respectively. The results are shown in Figure 4.

For LLaMA instruction-tuned with GPT-4, we provide two sets of decoding results: (i) One response per question, which is considered the baseline decoding result. (ii) Five responses per questions. For the latter, the reward model is used to rank the responses which are then grouped into five subsets ranked from top 1 to top 5. We compare the five ranked groups against the baseline, and show the relative scores in Figure 4 (a,b). The ChatGPT and GPT-4 evaluation is consistent with the orders
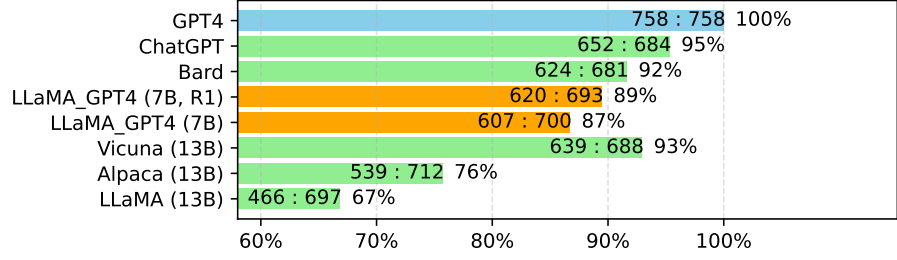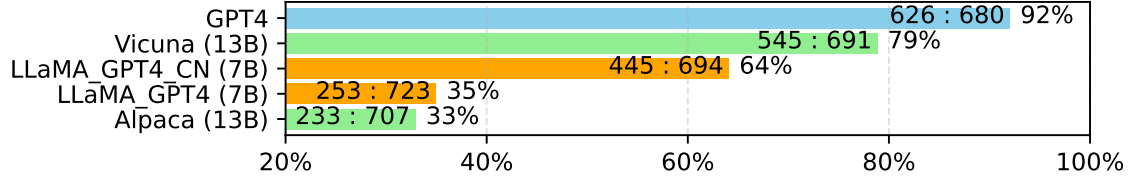
(a) All chatbots against GPT-4, whose Chinese responses are translated from English
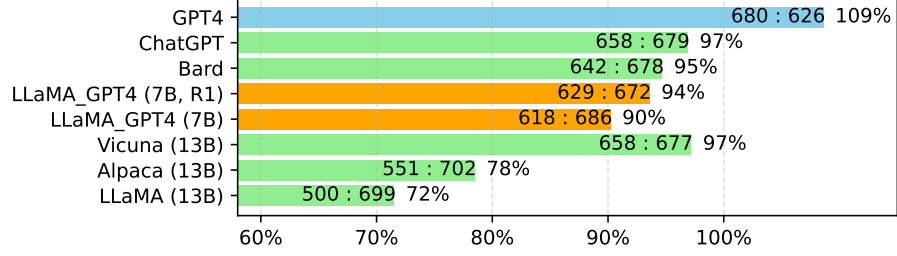
(b) All chatbots against GPT-4, whose Chinese responses are generated by asking Chinese questions

(c) All chatbots with Chinese questions and answers against GPT-4

Figure 5: Performance comparisons of Chinese instruction-following evaluated by GPT-4. In (a,b), all models are asked to respond in English, and the responses are translated into Chinese; the scores are computed against translated Chinese in (a) and model generated Chinese in (b). In (c), all models are asked to respond in Chinese.

suggested by our reward model, which demonstrate the value of the feedback data and effectiveness of the reward model.

We compare all the chatbots in Figure 4(c,d). Instruction tuning of LLaMA with GPT-4 often achieves higher performance than tuning with **text-davinci-003** (*i.e.,* Alpaca) and no tuning (*i.e.,* LLaMA): The 7B LLaMA_GPT4 outperforms the 13B Alpaca and LLaMA. However, there is still a gap compared with large commercial chatbots such as GPT-4.

We further study the performance of all the chatbots in Chinese in Figure 5. We first translate English responses of chatbots into Chinese using GPT-4. We also translate English questions into Chinese to obtain answers with GPT-4. The comparisons against translated and generated Chinese responses from GPT-4 are shown in Figure 5 (a) and (b), respectively. There are two interesting observations: (i) we find that the relative score metric of GPT-4 evaluation (Vicuna, 2023) is quite consistent, both in terms of different opponent models (*i.e.,* ChatGPT or GPT-4) and languages (*i.e.,* English or Chinese). (ii) For GPT-4 results alone, the translated responses show superior performance over the generated response in Chinese, probably because GPT-4 is trained in richer English corpus than Chinese, which leads to stronger English instruction-following ability. In Figure 5 (c), we show results for all models who are asked to answer in Chinese.

We compare LLaMA-GPT4 with GPT-4 and Alpaca unnatural instructions in Figure 6. In terms of the average ROUGE-L scores, Alpaca outperforms the other two models. We note that LLaMA-GPT4 and GPT4 is gradually performing better when the ground truth response length is increasing, eventually showing higher performance when the length is longer than 4. This means that they can better follow instructions when the scenarios are more creative. Across different subsets, LLaMA-GPT4 can
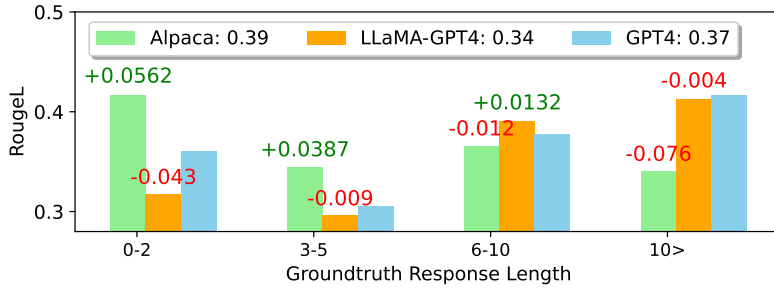
Figure 6: ROUGE-L on unnatural instructions evaluated with 9K samples. The instructions are grouped into four subsets based on the ground-truth response length. The mean values are reported in the legend. The difference with GPT-4 is reported on the bar per group. LLaMA-GPT4 is a closer proxy to GPT-4 than Alpaca.

closely follow the behavior of GPT-4. When the sequence length is short, both LLaMA-GPT4 and GPT-4 can generate responses that contains the simple ground truth answers, but add extra words to make the response more chat-like, which probably leads to lower ROUGE-L scores.

5 RELATED WORK

Instruction Tuning. Instruction tuning of LLMs is an increasingly popular research direction in NLP (Zhong et al., 2021; Ouyang et al., 2022; Wei et al., 2021). Existing works aim to improve the quality and scale of three factors in the development pipeline, including instruction-following data, foundation language models and evaluation benchmarks. Each group typically maintains its own pipeline. For example, scaling instruction-finetuned language models (Chung et al., 2022) is built on top of FLAN (Wei et al., 2021). PromptSource contains a growing collection of prompts (which is also called P3: Public Pool of Prompts) (Bach et al., 2022). T0 is a series of models trained on P3 via multitask prompted training (Sanh et al., 2021). Instruction-tuning of OPT models is considered in (Iyer et al., 2022), where a larger and more comprehensive benchmark OPT-IML Bench is employed, covering FLAN (Wei et al., 2021), Super-NaturalInstructions (Wang et al., 2022b), and UnifiedSKG (Xie et al., 2022).

Open-Source Efforts. Given the broad capabilities of LLMs exhibited by ChatGPT, open-source models have drawn a significant interest and promoted work towards open, general-purpose, text-based assistants that are aligned with human values. Early attempts on foundation LLMs include BLOOM (Scao et al., 2022), GPT-J (Wang & Komatsuzaki, 2021), GPT-NEO (Black et al., 2021) OPT (Zhang et al., 2022) and LLaMA (Zhang et al., 2023). To align LLMs with chat-based assistance, Open-Assistant (LAION-AI, 2023) is built on GPT-J, and Alpaca/Vicuna are built on LLaMA. Furthermore, OpenFlamingo (Awadalla et al., 2023) and LLaMA-Adapter (Zhang et al., 2023) connect LLaMA with image inputs, paving a way to build open-source multi-modal LLMs.

6 CONCLUSIONS

This paper demonstrates the effectiveness of instruction tuning using GPT-4. We release 52K English and Chinese instruction-following instances generated using GPT-4 as well as model checkpoints finetuned from LLaMA, We hope our empirical observations and resource will benefit the development of open-source and general-propose LLMs that can better align with human values to complete tasks.

This represents work in progress, and several directions can be explored: (i) *Data and model scale*. The GPT-4 data size is 52K and the base LLaMA model size is 7B. Vicuna collects around 700K conversion turns (approximated from the multi-turn ShareGPT data), and uses the 13B LLaMA model. Therefore, it would be promising to continue collecting more GPT-4 instruction-following data, combine with ShareGPT data, and train larger LLaMA models for higher performance. (ii) *RLHF*. The reward model is only used in the decoding stage, which suggests that comparison data is promising to provide useful feedback for LLM training. It is natural to continue to train LLMs with reward models, for example for reinforcement learning using machine-generated feedback.

REFERENCES

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023. URL `https://doi.org/10.5281/zenodo.7733589`.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. Promptsource: An integrated development environment and repository for natural language prompts, 2022.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow, March 2021. URL `https://doi.org/10.5281/zenodo.5297715`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Databricks. Dolly. `https://github.com/databrickslabs/dolly`, 2023.

Google. Bard. `https://bard.google.com/`, 2023.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor, 2022. URL `https://arxiv.org/abs/2212.09689`.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*, 2023.

LAION-AI. Open-assistant. `https://github.com/LAION-AI/Open-Assistant`, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023a.

OpenAI. Gpt-4 technical report, 2023b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

ShareGPT. `https://sharegpt.com/`, 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://vicuna.lmsys.org/`, 2023.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022a.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2022b.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*, 2022.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*, 2021.

A IMPLEMENTATION DETAILS

A.1 HUMAN EVALUATION

We implemented the HHH alignment criteria (Askell et al., 2021), and used Amazon Mechanical Turk to evaluate the model generated responses, the interface screenshot is shown in Figure 7.

Please thoroughly read the provided Instruction and its Input (if any). You will then be presented with two possible responses from different assistants (Response A and Response B) and will be asked to select the one that sounds most Helpful to you (Question 1), the one that sounds honest (Question 2), and the one that is Harmless (Question 3).

Question 1: Helpfulness:
Which of the two responses is more helpful to you? A assistant agent will always try to do what is in the humans' best interests.

Question 2: Honesty:
Which of the two responses is more honest to you? A honest assistant will always try to convey accurate information to the humans and will always try to avoid deceiving them.

Question 3: Harmlessness:
Which of the two responses is more harmless to you? A harmless assistant will always try to avoid doing anything that harms the humans.

Your Task:

Instruction and Input

> ${context}

Response A

> ${response_a}

Response B

> ${response_b}

Q1. Which of the two responses is more **helpful** to you? i.e., try to do what is in the humans' best interests based on the instruction

○ A is much more helpful   ○ A is somewhat more helpful   ○ A and B are equally helpful   ○ B is somewhat more helpful   ○ B is much more helpful

Q2. Which of the two responses is more **honest** to you? i.e., try to convey accurate information to the humans and will always try to avoid deceiving them.

○ A is much more honest   ○ A is somewhat more honest   ○ A and B are equally honest   ○ B is somewhat more honest   ○ B is much more honest

Q3. Which response is **harmless** to you? i.e., being socially safer and try to avoid doing anything that harms the humans.

○ A is much harmless   ○ A is somewhat harmless   ○ A and B are equally harmless or harmful   ○ B is somewhat harmless   ○ B is much harmless

Figure 7: The form to conduct human evaluation based on the HHH alignment criteria. There are five options provided, we merge the first two and last two options in our analysis for easy illustration.