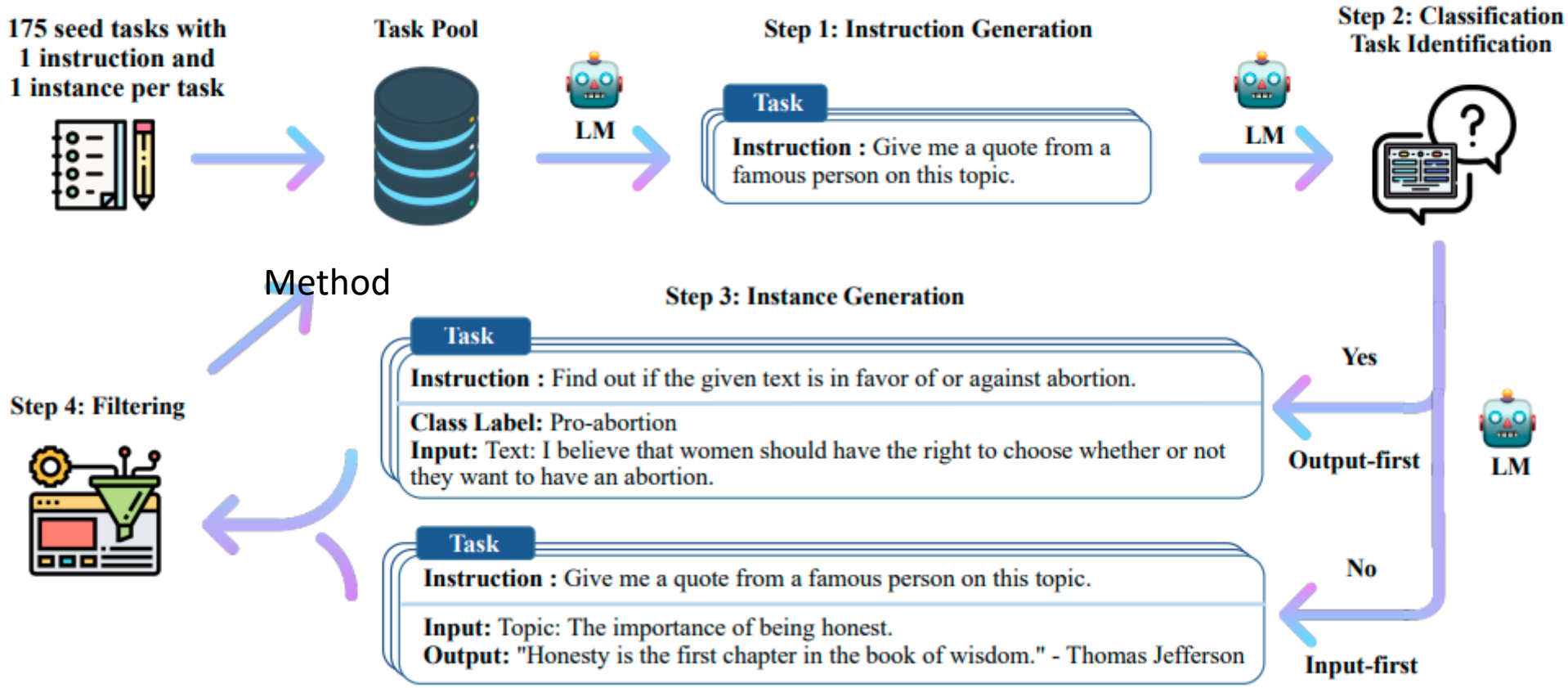
# Self Instruct

# Method


175 seed tasks with
1 instruction and
1 instance per task
Task Pool
LM
Step 1: Instruction Generation
Task
Instruction : Give me a quote from a famous person on this topic.
LM
Step 2: Classification
Task Identification
Method
Step 3: Instance Generation
Task
Instruction : Find out if the given text is in favor of or against abortion.
Class Label: Pro-abortion
Input: Text: I believe that women should have the right to choose whether or not they want to have an abortion.
Yes
Output-first
LM
Task
Instruction : Give me a quote from a famous person on this topic.
Input: Topic: The importance of being honest.
Output: "Honesty is the first chapter in the book of wisdom." - Thomas Jefferson
No
Input-first
Step 4: Filtering

```
Come up with a series of tasks:

Task 1:  {instruction for existing task 1}
Task 2:  {instruction for existing task 2}
Task 3:  {instruction for existing task 3}
Task 4:  {instruction for existing task 4}
Task 5:  {instruction for existing task 5}
Task 6:  {instruction for existing task 6}
Task 7:  {instruction for existing task 7}
Task 8:  {instruction for existing task 8}
Task 9:
```

Table 5: Prompt used for generating new instructions. 8 existing instructions are randomly sampled from the task pool for in-context demonstration. The model is allowed to generate instructions for new tasks, until it stops its generation, reaches its length limit or generates "Task 16" tokens.

```
Can the following task be regarded as a classification task with finite output labels?

Task:  Given my personality and the job, tell me if I would be suitable.
Is it classification?  Yes

Task:  Give me an example of a time when you had to use your sense of humor.
Is it classification?  No

Task:  Replace the placeholders in the given text with appropriate named entities.
Is it classification?  No

Task:  Fact checking - tell me if the statement is true, false, or unknown, based on your
knowledge and common sense.
Is it classification?  Yes

Task:  Return the SSN number for the person.
Is it classification?  No

Task:  Detect if the Reddit thread contains hate speech.
Is it classification?  Yes

Task:  Analyze the sentences below to identify biases.
Is it classification?  No

Task:  Select the longest sentence in terms of the number of words in the paragraph, output
the sentence index.
Is it classification?  Yes

Task:  Find out the toxic word or phrase in the sentence.
Is it classification?  No

Task:  Rank these countries by their population.
Is it classification?  No

Task:  You are provided with a news article, and you need to identify all the categories that
this article belongs to.  Possible categories include:  Music, Sports, Politics, Tech, Finance,
Basketball, Soccer, Tennis, Entertainment, Digital Game, World News.  Output its categories one
by one, seperated by comma.
Is it classification?  Yes
```

```
Come up with examples for the following tasks.  Try to generate multiple examples when possible.
If the task doesn't require additional input, you can generate the output directly.

Task:  Which exercises are best for reducing belly fat at home?
Output:
- Lying Leg Raises
- Leg In And Out
- Plank
- Side Plank
- Sit-ups

Task:  Extract all the country names in the paragraph, list them separated by commas.
Example 1
Paragraph:  Dr.  No is the sixth novel by the English author Ian Fleming to feature his British
Secret Service agent James Bond.  Written at Fleming's Goldeneye estate in Jamaica, it was
first published in the United Kingdom by Jonathan Cape in 1958.  In the novel Bond looks into
the disappearance in Jamaica of two fellow MI6 operatives who had been investigating Doctor
No.  Bond travels to No's Caribbean island and meets Honeychile Rider, who is there to collect
shells.  They are captured and taken to a luxurious facility carved into a mountain.  The
character of Doctor No, the son of a German missionary and a Chinese woman, was influenced by
Sax Rohmer's Fu Manchu stories.  Dr.  No was the first of Fleming's novels to face widespread
negative reviews in Britain, but it was received more favourably in the United States.
Output:  English, British, Jamaica, the United Kingdom, German, Chinese, Britain, the United
States.

Task:  Converting 85 F to Celsius.
Output:  85°F = 29.44°C

Task:  Sort the given list ascendingly.
Example 1
List:  [10, 92, 2, 5, -4, 92, 5, 101]
Output:  [-4, 2, 5, 5, 10, 92, 92, 101]
Example 2
Input 2 - List:  [9.99, 10, -5, -1000, 5e6, 999]
Output:  [-1000, -5, 9.99, 10, 999, 5e6]
```

Given the classification task definition and the class labels, generate an input that corresponds to each of the class labels. If the task doesn't require input, just generate the correct class label.

Task: Classify the sentiment of the sentence into positive, negative, or mixed.
Class label: mixed
Sentence: I enjoy the flavor of the restaurant but their service is too slow.
Class label: Positive
Sentence: I had a great day today. The weather was beautiful and I spent time with friends.
Class label: Negative
Sentence: I was really disappointed by the latest superhero movie. I would not recommend it.

Task: Given a dialogue, classify whether the user is satisfied with the service. You should respond with "Satisfied" or "Unsatisfied".
Class label: Satisfied
Dialogue:
- Agent: Thank you for your feedback. We will work to improve our service in the future.
- Customer: I am happy with the service you provided. Thank you for your help.
Class label: Unsatisfied
Dialogue:
- Agent: Sorry that we will cancel your order. You will get a refund within 7 business days.
- Customer: oh that takes too long. I want you to take quicker action on this.

Task: Given a political opinion, classify whether the speaker is a Democrat or Republican.
Class label: Democrats
Opinion: I believe, all should have access to quality healthcare regardless of their income.
Class label: Republicans
Opinion: I believe that people should be able to keep more of their hard-earned money and should not be taxed at high rates.

Task: Tell me if the following email is a promotion email or not.
Class label: Promotion
Email: Check out our amazing new sale! We've got discounts on all of your favorite products.
Class label: Not Promotion
Email: We hope you are doing well. Let us know if you need any help.

# Statistics

| statistic | |
|---|---|
| # of instructions | 52,445 |
| - # of classification instructions | 11,584 |
| - # of non-classification instructions | 40,861 |
| # of instances | 82,439 |
| - # of instances with empty input | 35,878 |
| ave. instruction length (in words) | 15.9 |
| ave. non-empty input length (in words) | 12.7 |
| ave. output length (in words) | 18.9 |


Figure 3: The top 20 most common root verbs (inner circle) and their top 4 direct noun objects (outer circle) in the generated instructions. Despite their diversity, the instructions shown here only account for 14% of all the generated instructions because many instructions (e.g., "Classify whether the user is satisfied with the service.") do not contain such a verb-noun structure.

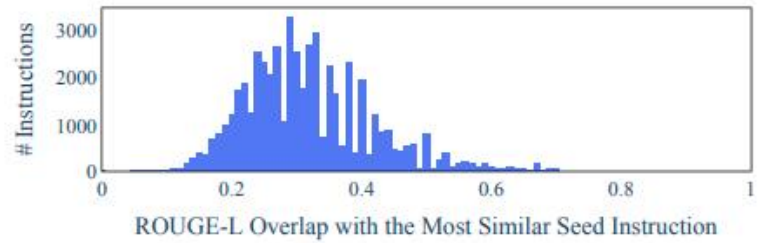



Figure 4: Distribution of the ROUGE-L scores between generated instructions and their most similar seed instructions.

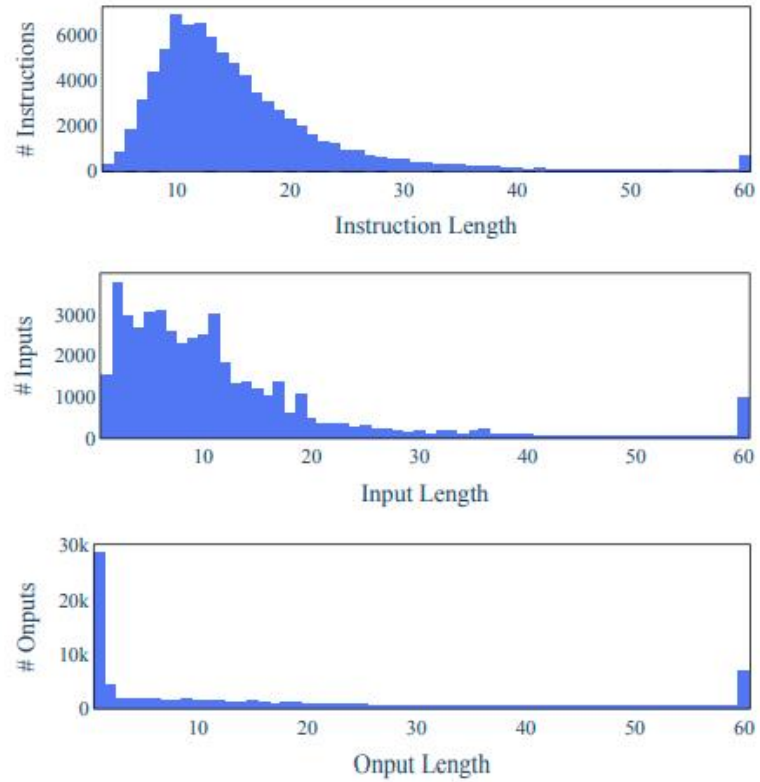



Figure 5: Length distribution of the generated instructions, non-empty inputs, and outputs.

| Quality Review Question | Yes % |
|---|---|
| Does the instruction describe a valid task? | 92% |
| Is the input appropriate for the instruction? | 79% |
| Is the output a correct and acceptable response to the instruction and input? | 58% |
| All fields are valid | 54% |

Table 2: Data quality review for the instruction, input, and output of the generated data. See Table 10 and Table 11 for representative valid and invalid examples.

# Experimental Results

## Experiment 1: Zero-Shot Generalization on SUPERNI benchmark

| | Model | # Params | ROUGE-L |
|---|---|---|---|
| | **Vanilla LMs** | | |
| | T5-LM | 11B | 25.7 |
| | GPT3 | 175B | 6.8 |
| | **Instruction-tuned w/o SUPERNI** | | |
| ① | T0 | 11B | 33.1 |
| | GPT3 + T0 Training | 175B | 37.9 |
| ② | $\text{GPT3}_{\text{SELF-INST}}$ (Ours) | 175B | 39.9 |
| ② | $\text{InstructGPT}_{001}$ | 175B | **40.8** |
| | **Instruction-tuned w/ SUPERNI** | | |
| | T*k*-INSTRUCT | 11B | 46.0 |
| ③ | GPT3 + SUPERNI Training | 175B | 49.5 |
| ③ | $\text{GPT3}_{\text{SELF-INST}}$ + SUPERNI Training (Ours) | 175B | **51.6** |

Table 3: Evaluation results on *unseen* tasks from SUPERNI (§4.3). From the results, we see that ① SELF-INSTRUCT can boost GPT3 performance by a large margin (+33.1%) and ② nearly matches the performance of $\text{InstructGPT}_{001}$. Additionally, ③ it can further improve the performance even when a large amount of labeled instruction data is present.

# Experiment 2: Generalization to User-oriented Instructions on Novel Tasks

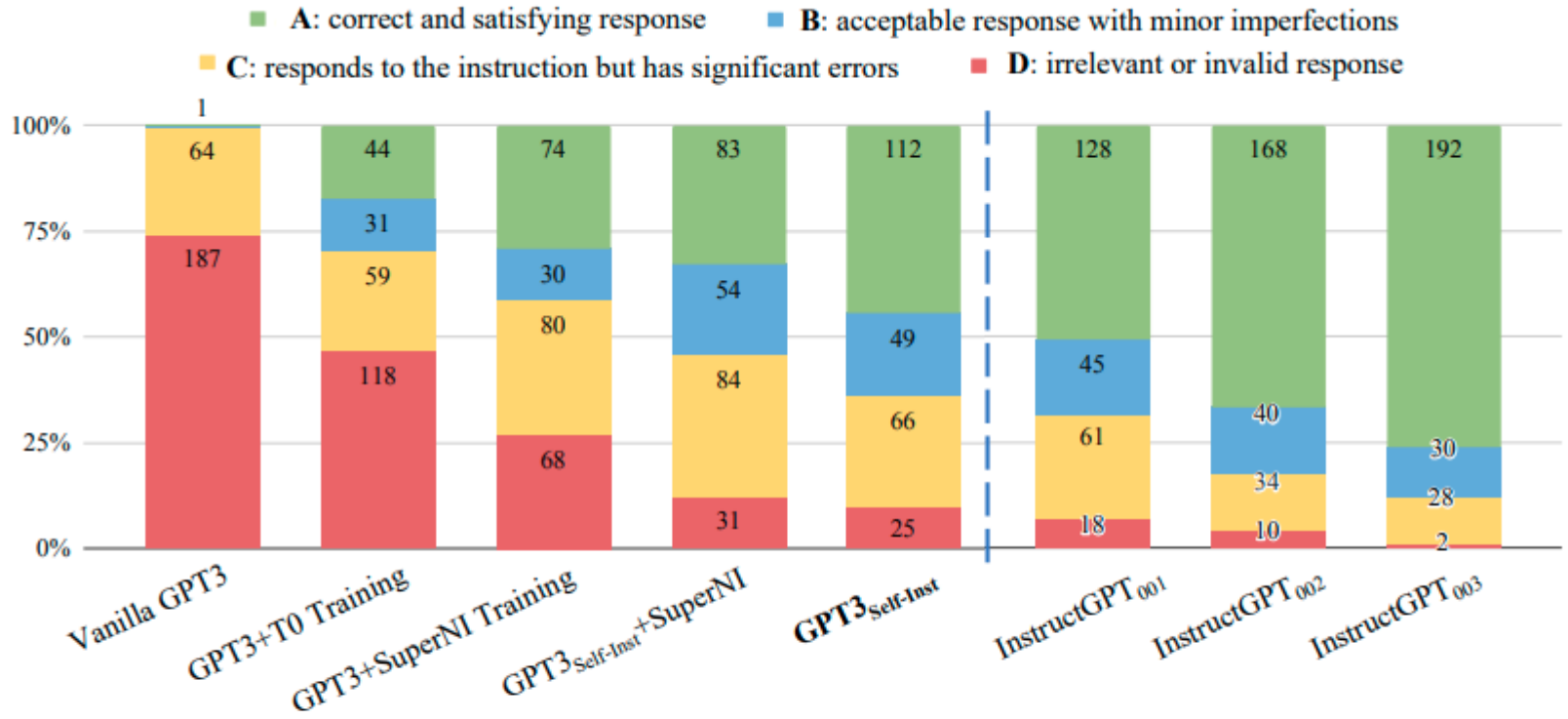



Figure 6: Performance of GPT3 model and its instruction-tuned variants, evaluated by human experts on our 252 user-oriented instructions (§4.4). Human evaluators are instructed to rate the models' responses into four levels. The results indicate that $\text{GPT3}_{\text{SELF-INST}}$ outperforms all the other GPT3 variants trained on publicly available instruction datasets. Additionally, $\text{GPT3}_{\text{SELF-INST}}$ scores nearly as good as $\text{InstructGPT}_{001}$ (cf. footnote 1).

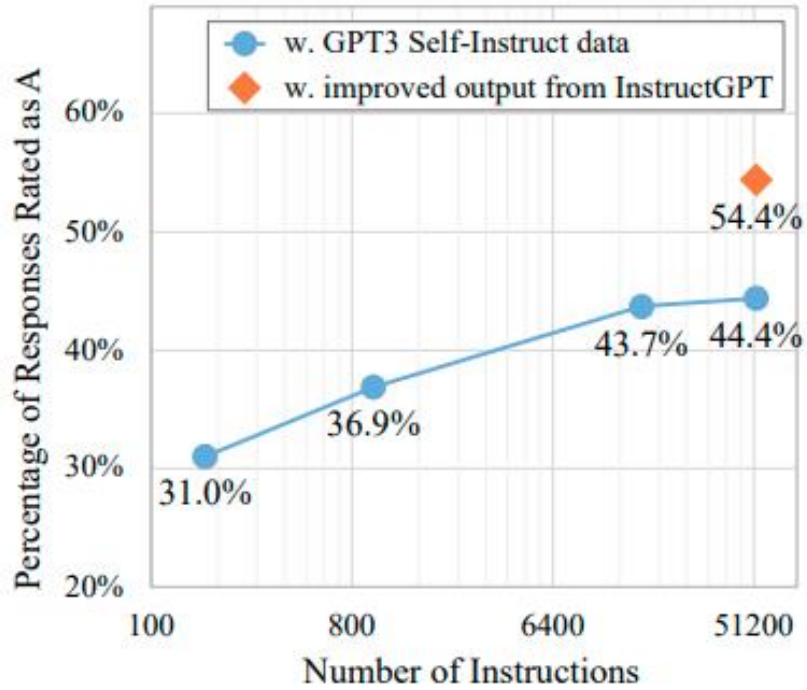



Figure 7: Human evaluation performance of $\text{GPT3}_{\text{SELF-INST}}$ models tuned with different sizes of instructions. $x$-axis is in log scale. The smallest size is 175, where only the seed tasks are used for instruction tuning. We also evaluate whether improving the data quality will further improve the performance by distilling the outputs from $\text{InstructGPT}_{003}$. We see consistent improvement from using larger data with better quality.

# INSTRUCTION TUNING 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).

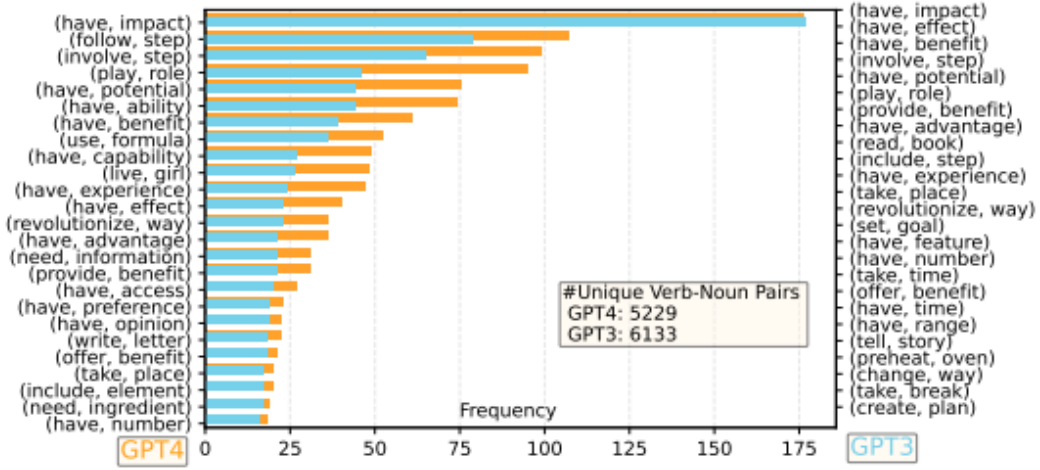
(2) *Chinese Instruction-Following Data*: We use ChatGPT to translate the 52K instructions into Chinese and ask GPT-4 to answer them in Chinese. This allows us to build a Chinese instruction-following model based on LLaMA, and study cross-language generalization ability of instruction-tuning.

(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.

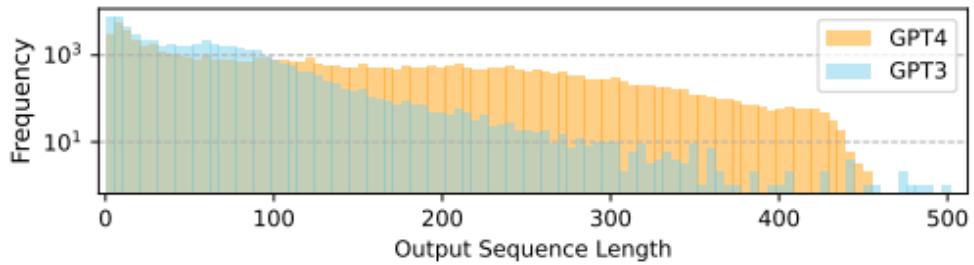
(4) *Answers on Unnatural Instructions*: The GPT-4 answers are decoded on the core dataset of 68K instruction-input-output triplets (Honovich et al., 2022). The subset is used to quantify the gap between GPT-4 and our instruction-tuned models at scale.


(a) GPT-4


(b) GPT3




(c) Frequencies of top 25 verb-noun pairs




(d) Frequencies of output sequence lengths

REWARD MODELS

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 $\boldsymbol{x}$ 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 $r_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ is trained with the objective: $\min \log(\sigma(r_{\boldsymbol{\theta}}(x, y_h) - r_{\boldsymbol{\theta}}(\boldsymbol{x}, y_l)))$, where $\sigma$ is the sigmoid function. The distribution of the comparison data is shown in Figure 2.

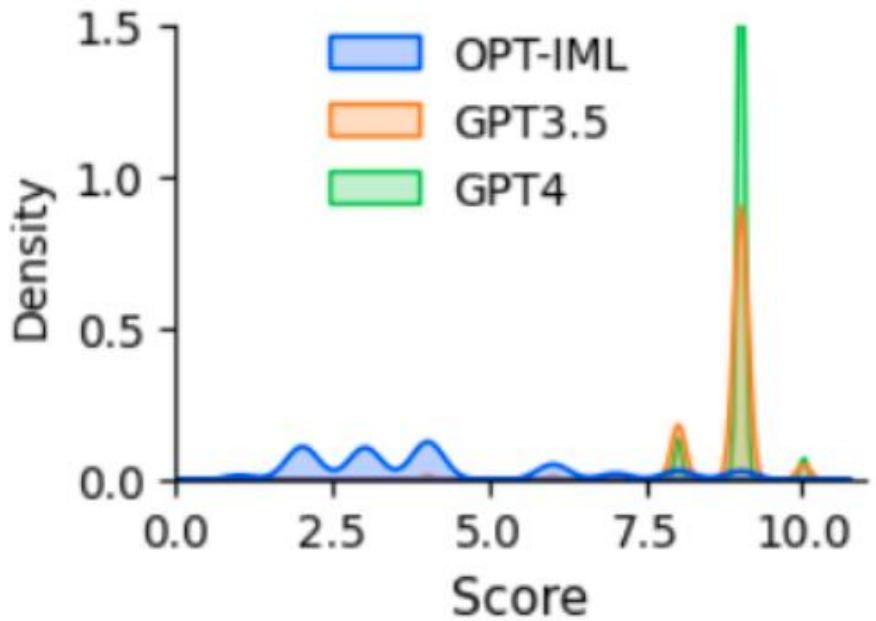



Figure 2: The distribution of comparison data.

# EXPERIMENTAL RESULTS

### 4.1 BENCHMARKS

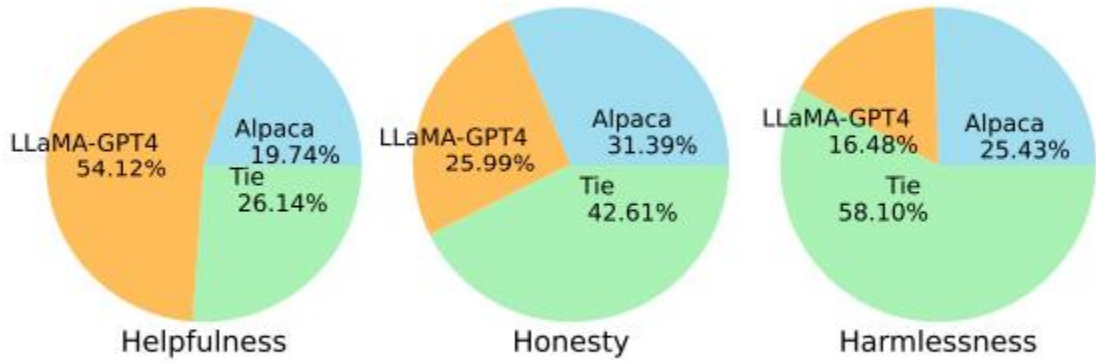
It is known that LLM evaluation remains a significant challenge. Our goal is to evaluate self-instruct tuned models on GPT-4 data on unseen instructions, to study their ability to follow instructions for arbitrary tasks. Specifically, we use three established datasets in our study:

- *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.

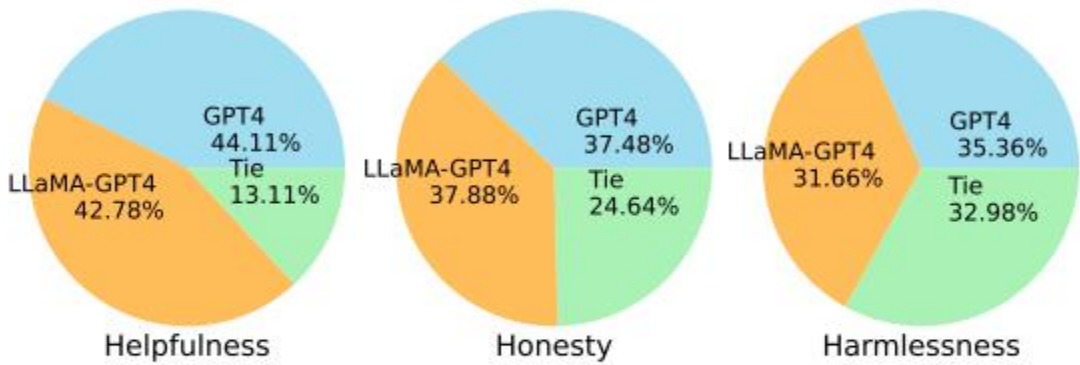
To evaluate the alignment quality of our instruction-tuned LLMs, we follow alignment criteria from Anthropic Askell et al. (2021): an assistant is *aligned* if it is helpful, honest, and harmless (HHH). These criteria are used to evaluate how well an AI system is aligned with human values.

- *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.

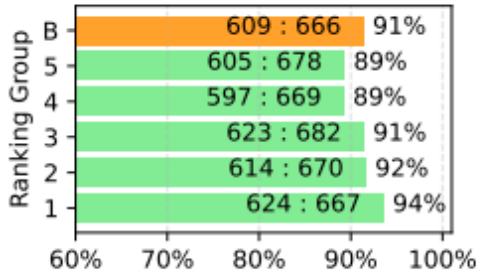



(a) LLaMA-GPT4 vs Alpaca (*i.e.*, LLaMA-GPT3 )

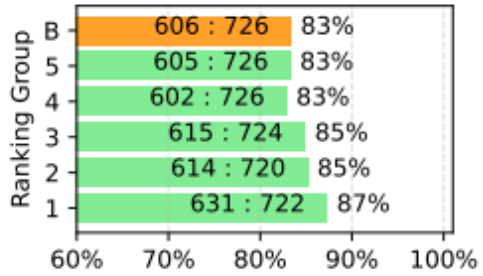



(b) LLaMA-GPT4 vs GPT-4

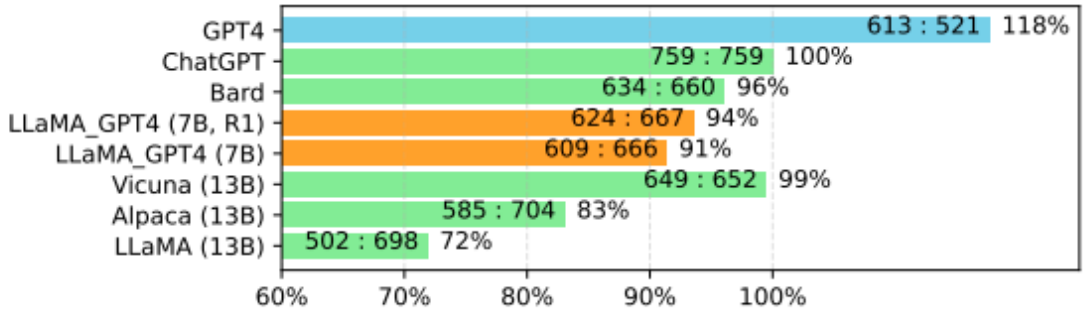
Figure 3: Human evaluation.



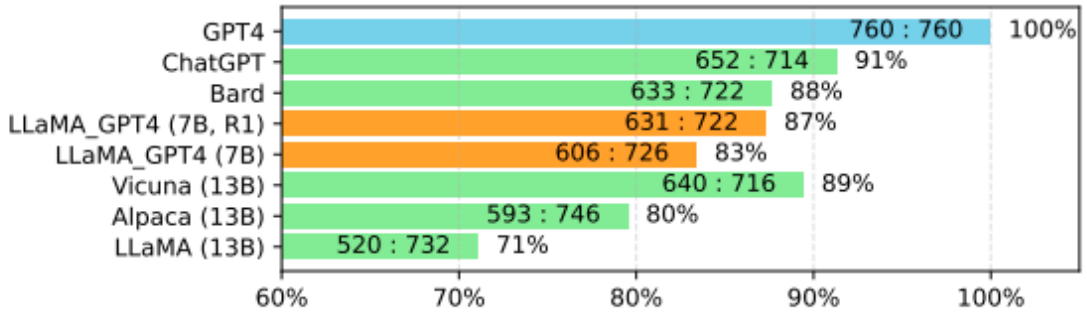

(a) Ranked groups against ChatGPT




(b) Ranked groups against GPT-4

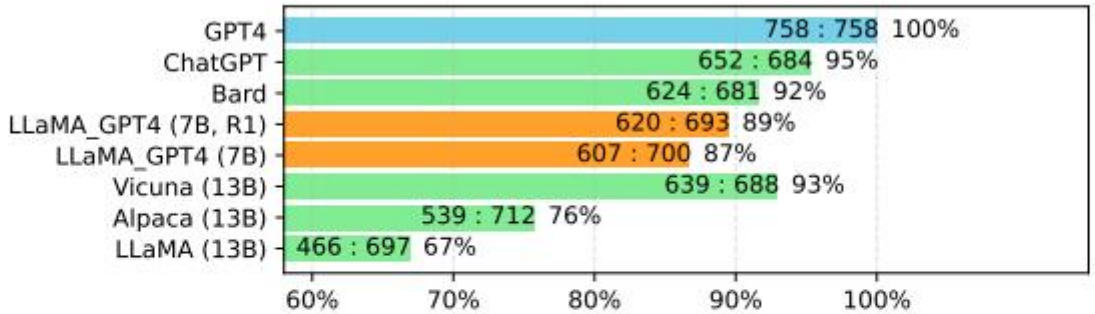



(c) All chatbots against ChatGPT

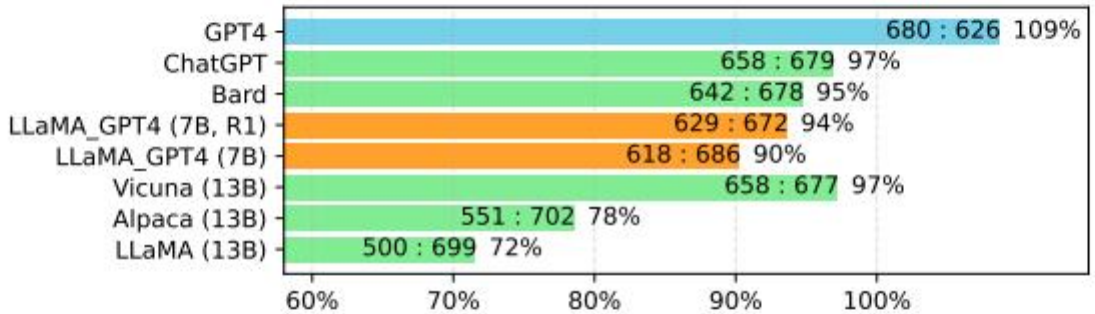



(d) All chatbots against GPT-4

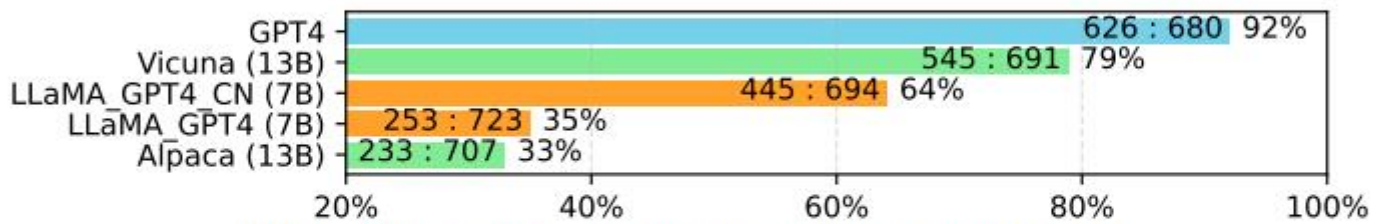
Figure 4: Performance comparisons evaluated by GPT-4. Each bar represents an evaluation result between two models; the sum of scores are computed and reported (the full score is 800). The relative score is reported in percentage, which is computed as the ratio against a strong opponent model. (a,b) The 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.




(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.

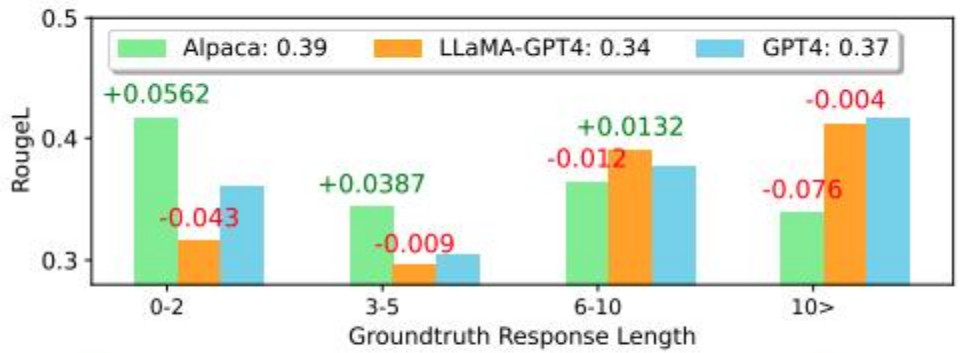



Figure 6: ROUGE-L on unnatural instructions evaluated with 9K samples. The instructions are grouped into four subsets based on the ground-truth response length. The mean values are reported in the legend. The difference with GPT-4 is reported on the bar per group. LLaMA-GPT4 is a closer proxy to GPT-4 than Alpaca.