Rapidly Developing High-quality Instruction Data and Evaluation Benchmark for Large Language Models with Minimal Human Effort: A Case Study on Japanese

Yikun Sun*, Zhen Wan*, Nobuhiro Ueda, Sakiko Yahata

Fei Cheng, Chenhui Chu, Sadao Kurohashi

Kyoto University

LREC-COLING 2024

Background

- Motivation
  - Generating instruction data for LLMs(Large Language Models) involves enormous human effort, especially in non-English languages
  - For LLMs evaluation, human reference is also costly

- Related work to these gaps
  - Translate English Alpaca(Taori et al., 2023) instruction data into Japanese, e.g., MT (Machine Translation) Alpaca (Kunishou, 2023)
  - For English LLMs, Vicuna chatbot(Chiangetal.,2023) automatically evaluates different LLMs by GPT-4

Three Parts of Contributions

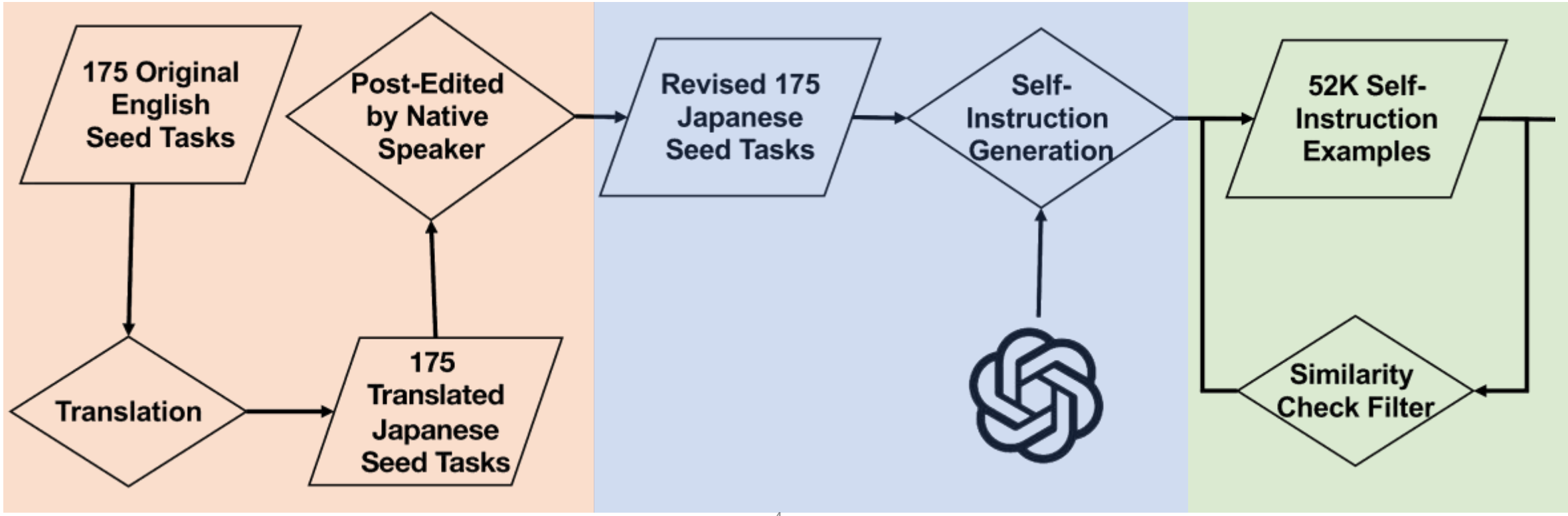
- Propose an **efficient self-instruct** method to generate **native Japanese instruction data** with GPT-4
- Develop an **automatic evaluation benchmark** for Japanese LLMs
- **Empirical results** show:
  - Our self-instruction dataset has a **higher quality** than MT Alpaca
  - Our automatic evaluation is **consistent with human preferences**

Part I: Japanese Self-instruct Generation

Step1:
Translate the original seed tasks

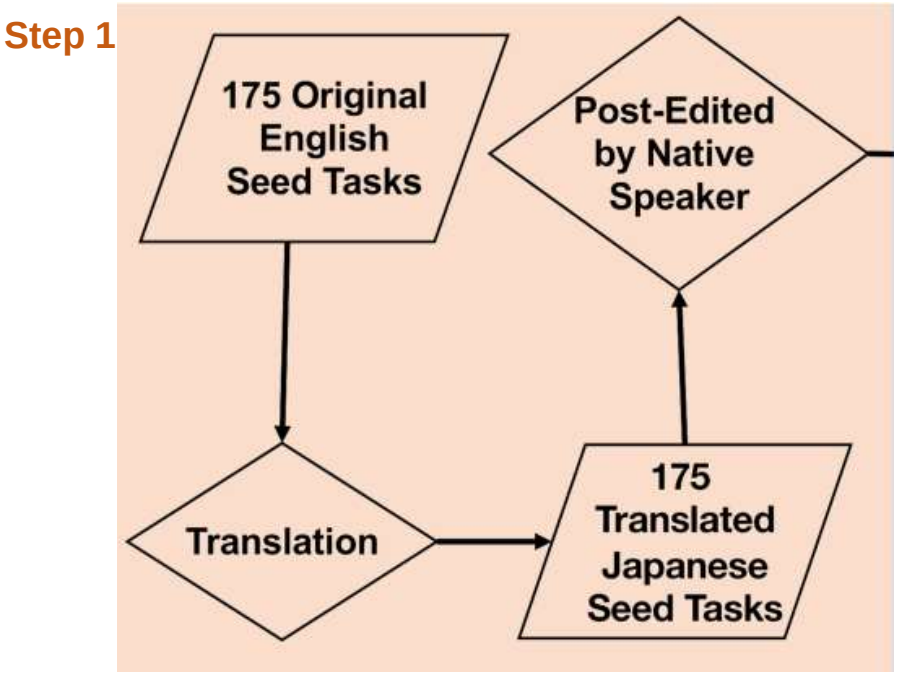
Step2:
Generate instruction data by GPT-4 self-instruction

Step3:
Filter out similar examples



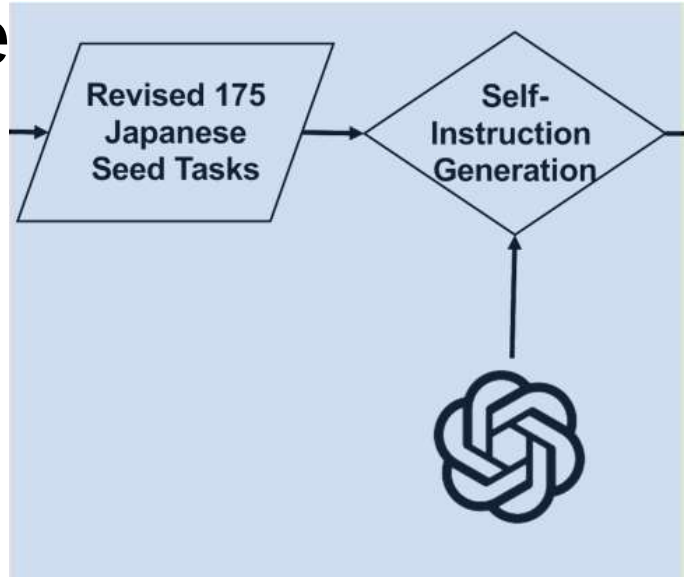
Step 1: Translate seed tasks

- We used GPT-4 to translate **175** human-written instruction seeds (Taori et al., 2023) from English to Japanese
- Two native Japanese speakers helped to post-edit the translations
- Finally, we obtain native-quality Japanese seed tasks



Step 2: GPT-4 Self-Instruction

- Construct a **comprehensive prompt** to guide GPT-4 to generate a total of **20** instruction samples
- Randomly sample **3 demo tasks** from **high-quality Japanese seed tasks** and append them to the prompt
- GPT-4 follows the prompt and imitates these **3** sampled tasks and generates **17** samples



An Example of Prompt

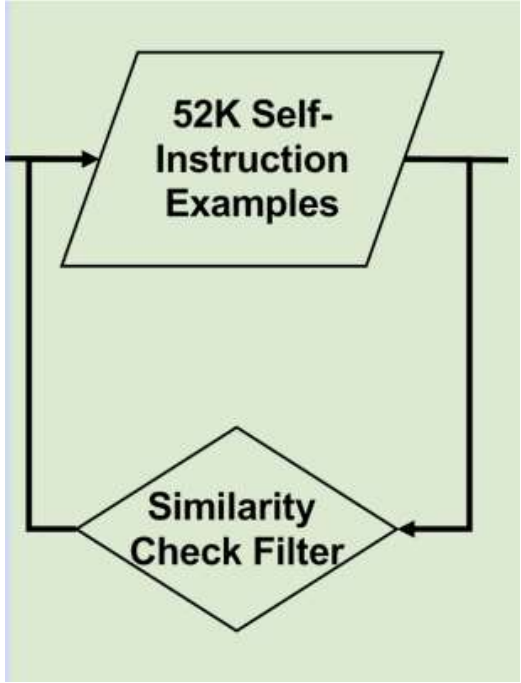
Prompt	3 seed tasks
多様性を重視して20個のタスクを考え、そのタスクを実行するための指示を考えてください。これらのタスク指示はGPTモデルに与えられ、その結果をもとにGPTモデルの性能を評価します。 以下にその要件を示します： 1. 多様性を高めるため、各指示では動詞の繰り返しを避けてください。 2. 使用する言語表現も多様であるべきです。例えば、質問形式と命令形式の指示を組み合わせてみてください。 3. 指示の種類も多様であるべきです。自由記述形式、分類、入力テキストの編集など、多種多様なタスクをリストに含めてください。 4. GPT言語モデルが達成可能な指示でなければなりません。例えば、映像や音声の出力を要求したり、午後5時に誰かを起こすようなリマインダーを設定するような指示は避けてください。 5. 指示は日本語で記述してください。 6. 指示は1〜2文の長さにしてください。命令形でも質問形でも構いません。 7. 指示に対する適切な入力も同時に考えてください。入力フィールドには、指示に対する具体的な例を含めてください。現実的な例を使用し、単純なプレースホルダーの使用は避けてください。指示の難易度を上げるため、しっかりと解きごたえのある入力が望ましいですが、理想的には100語を超えないようにしてください。 8. すべての指示に入力が必要なわけではありません。例えば、「世界で一番高い山は何か」という一般的な情報を問う指示の場合、特定の文脈を提供する必要はありません。このような場合、入力フィールドには「<noinput>」と記述してください。 9. 出力は、指示と入力に対する適切な応答でなければなりません。出力は200単語以下にしてください。 20個のタスク：	"instruction": "以下の単語を使って俳句を生成してください。" "input": "夏” "output": "冷たさ、浸透する\n衝撃、喜び、内側で爆発\n夏の舌が目を覚ます "instruction": "与えられたペアの関係は何ですか？ ”input": "夜：昼間 :: 右：左", "output": "与えられたペアの関係は反意語です。” "instruction": "人間の行動を説明してください。 "input": "行動：泣く” "output": "人が泣く理由はさまざまです。悲しい、怖い、怒っている、イライラしているといった感情があるかもしれません。時には幸せや安堵のために泣くこともあります。人々が自分たちの行動をする理由は一つではありません。 **-> Let GPT-4 complete the rest 17 samples**

Step 3: Filter out similar examples

- Utilize ROUGE-L score to **filter similar generated examples**
  - Recursively check the similarity between newly generated data and the accumulated instruction data pool
  - Remove all the data with ROUGE-L exceeding 0.7

- Add another blacklist for filtering instruction data
  - Unsuitable for SFT(Supervised Fine-tuning)
  - Cannot be processed by LLMs
    - For example: audio, video, images

Step 3

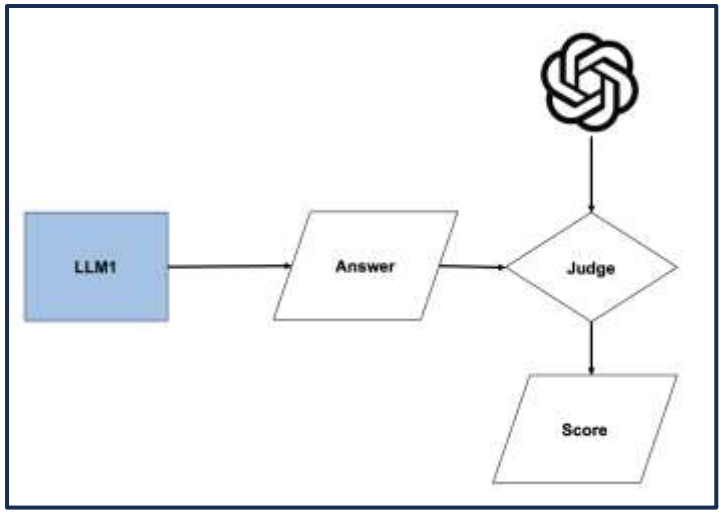


Part II: Japanese LLMs Evaluation Benchmark

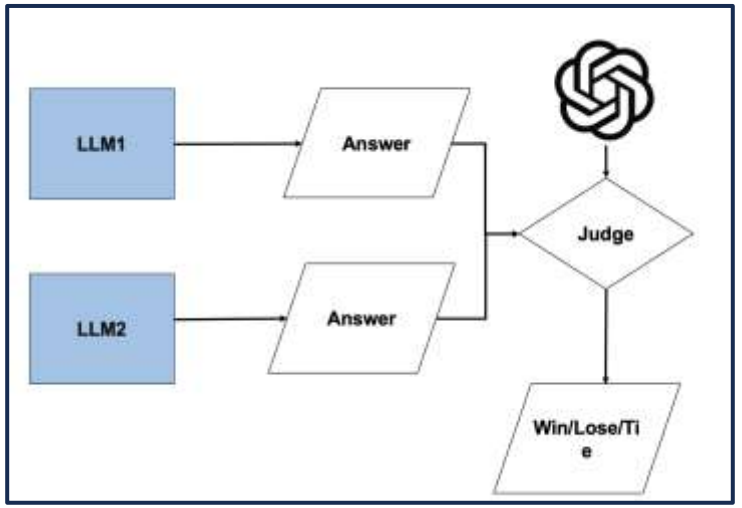
- The target is evaluating Japanese LLMs' instruction-following ability
- We translated the Vicuna benchmark into Japanese, which includes **80** questions within **8** categories:
  - *Writing, generic, knowledge, roleplay, common-sense, fermi, counterfactual, coding and math*
- Our benchmark contains two modes: **Single mode** and **Pairwise mode**

Part II: Japanese LLMs Evaluation Benchmark

- **Single Mode**:
  - Based on the proper prompt, GPT-4 directly **assigns a score** to an answer generated by LLMs

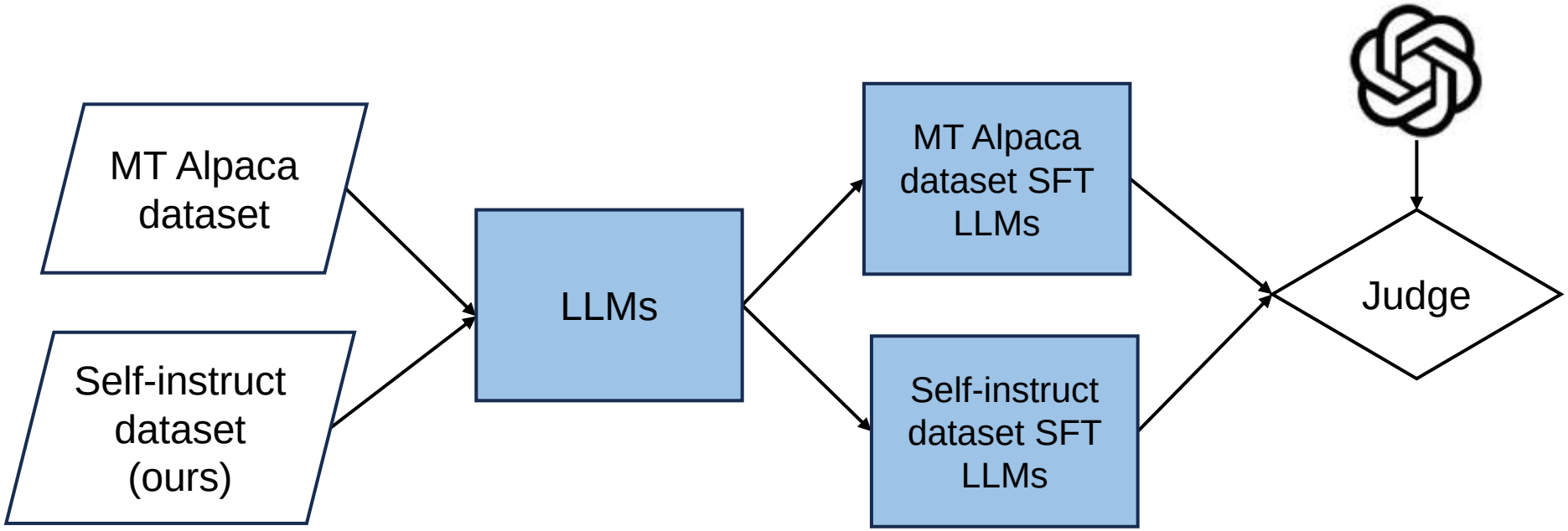


- **Pairwise Mode**:
  - Based on the proper prompt, GPT-4 **pair-wisely compare** answers from two LLMs and assign “Win/Loss/Tie”
  - Final win-rate: $$\text{win-rate} = \frac{\text{win} + \frac{\text{tie}}{2}}{\text{win} + \text{loss} + \text{tie}}$$



Part III: Experimental Settings

- The experiment is designed to compare LLMs fine-tuned with:
  - **Self-instruct** dataset (Proposed)
  - **MT Alpaca** dataset (Directly translated from Alpaca)
- **4 LLMs:** LLaMA-7b, OpenCALM-7b,LLaMA2-7b, LLaMA2-13b



Main Results on Japanese LLM Benchmark

In both **single mode** and **pairwise mode**, compared with the baseline (GPT-3.5 Davinci-003), our self-instruct data consistently achieves better performance

Base model	Instruction	Score
Davinci-003	-	5.86
LLaMA 7B	MT Alpaca	2.05
LLaMA 7B	Self-instruct	2.36
LLaMA2 7B	MT Alpaca	4.45
LLaMA2 7B	Self-instruct	5.71
Open-calm 7B	MT Alpaca	3.36
Open-calm 7B	Self-instruct	4.75
LLaMA2 13B	MT Alpaca	5.30
LLaMA2 13B	Self-instruct	**6.06**

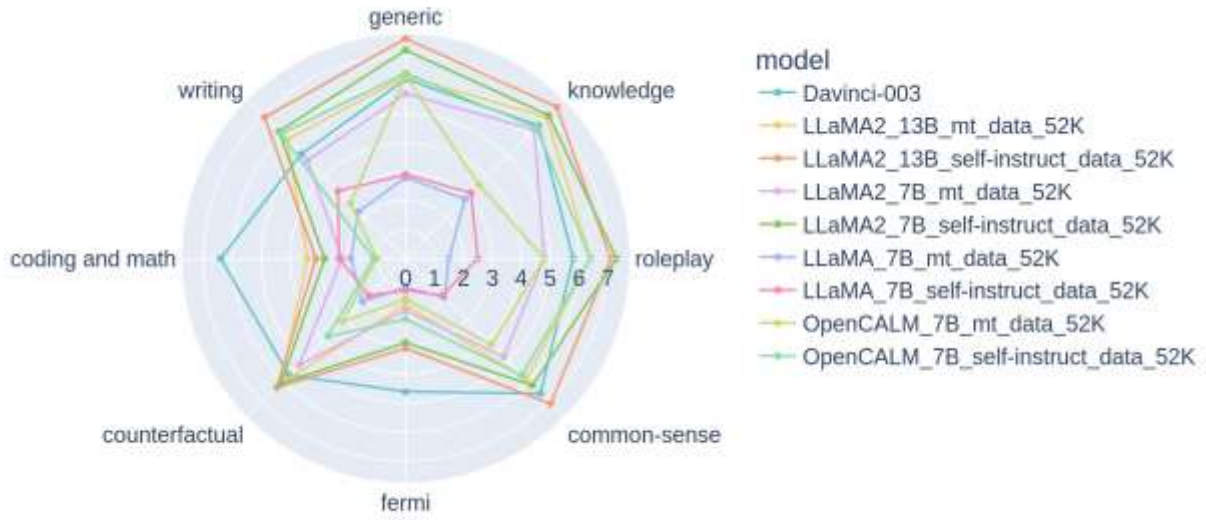
Single Mode Score

Base model	Instruction	Win-rate
LLaMA 7B	MT Alpaca	5.99
LLaMA 7B	Self-instruct	13.12
LLaMA2 7B	MT Alpaca	33.12
LLaMA2 7B	Self-instruct	46.25
Open-calm 7B	MT Alpaca	15.62
Open-calm 7B	Self-instruct	34.37
LLaMA2 13B	MT Alpaca	32.50
LLaMA2 13B	Self-instruct	**54.37**

Pairwise Mode win-rate

Analysis on Each Category

- GPT-4 self-instruct data outperforms MT Alpaca in each category
- LLaMA2_13B_self-instruct_52K is even better than Davinci-003 in **generic, knowledge, roleplay, writing, common-sense ,** and **counterfactual** domains
- All LLMs except for GPT-3.5 still perform poorly in **coding and math** and **fermi** questions

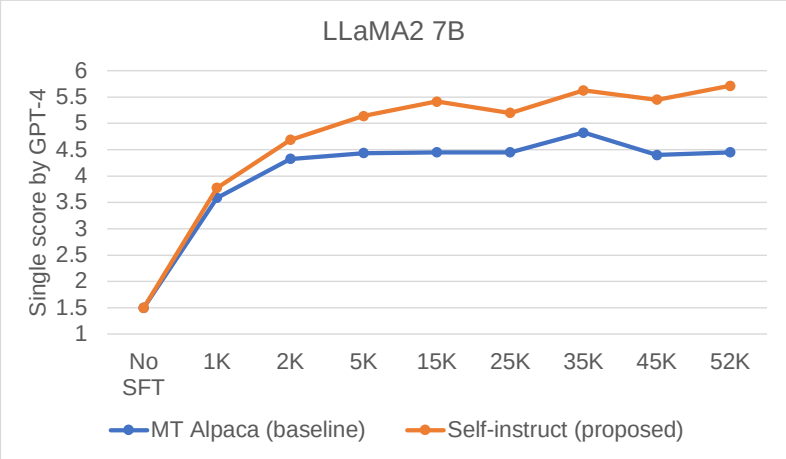


Ablation Study I: Different Sizes of Training Data

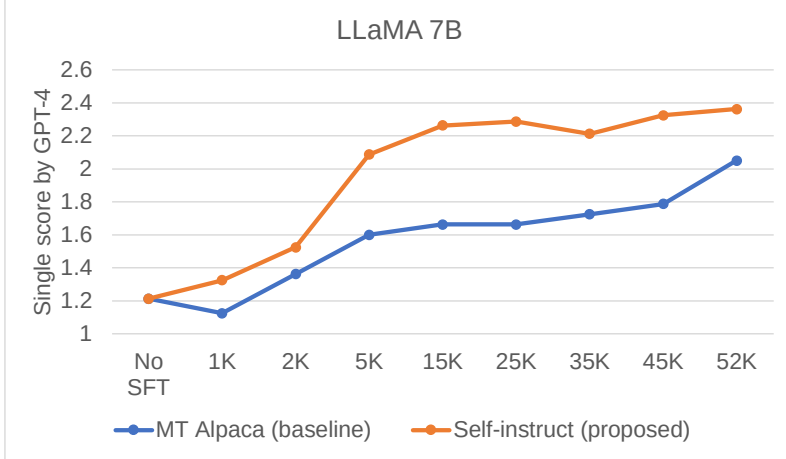
We used the **single mode** to evaluate each model with changing the size of the training data

➢ LLMs SFT with self-instruct data is **always superior** to the MT Alpaca

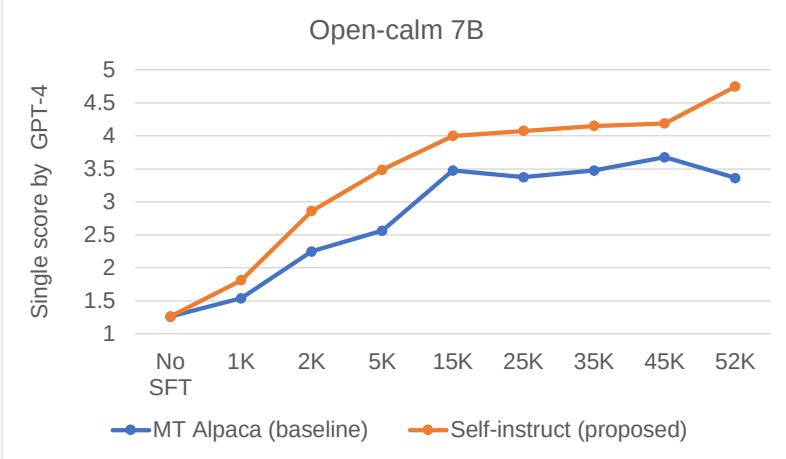
SFT LLaMA2 7B



SFT LLaMA 7B



SFT Open-CALM 7B



Ablation Study II

Concern: **MT Alpaca** is translated from English Alpaca, which is generated by **GPT-3.5**, while our instruction data uses **GPT-4**

- We fine-tuned LLaMA2 7B with both data generated by GPT-4:
  - **Self-instruct (5K):** *5K instruction data generated by our approach*
  - **MT Alpaca (5K):** *5K GPT-4 Alpaca instruction translated by DeepL*
- Result:
  - The difference made by the proposed approach leads to the principal improvements

Proposal	GPT-4 baseline	Win-rate
Self-instruct ($5K$)	MT Alpaca ($5K$)	55.6

Manual Evaluation

- We sampled 100 generated instructions and categorize them into 3 classes
  - **High quality (HQ):** Instruction data are both fluent and natural
  - **Non-fluent text (NT):** Instruction data are incorrect or unnatural
  - **Format violation (FV):** Data is non-instructive or is not coherent to an instruction
- Human pairwise comparison on **LLaMA2-13B** fine-tuned with 2 datasets
  - Self-instruction data **against** MT Alpaca data
- Self-instruction **always performs better**

Dataset	HQ	NT	FV
MT Alpaca	42	28	30
Self-instruct (ours)	67	27	6

Manual assessment of data quality

category	win	loss	tie
generic	3	2	5
knowledge	4	1	5
roleplay	3	3	4
common-sense	4	1	5
fermi	2	1	7
counterfactual	5	0	5
coding	3	0	4
math	0	0	3
writing	7	0	3
Total	**31**	**8**	**41**

Comparison of LLaMA2-13B outputs

Conclusion

We proposed a package of efficient approaches for rapidly developing resources in non-English languages

- Generate high-quality instruction data with GPT-4
  - It shows **better quality** and **minimal human effort**
- We developed a GPT-4-based LLMs evaluation benchmark **without human references** automatically
  - It shows the **consistency** between GPT-4-based assessments and human preferences