LLM-powered Data Augmentation for Enhanced Cross-lingual Performance

Chenxi Whitehouse, Monojit Choudhury, Alham Fikri Aji

City, University of London, Microsoft, MBZUAI

EMNLP 2023

Introduction

Background and Motivation

- The success of NLP models greatly depends on the **availability and quality of training data**
- It can be challenging to have sufficient labelled data, especially for **multilingual** scenarios
- Recent **powerful LLMs** excel at handling general instructions and have shown promise in data generation tasks
- Can we use LLMs to generate Data for complex multilingual commonsense reasoning tasks?

Data Augmentation

Datasets

- 3 Datasets: XCOPA, XWinograd, XStoryCloze
  - Available training data is extremely limited
  - Complex: commonsense reasoning
  - Multilinguality
  - Baseline models low performance (50% mBERT)

Dataset	EN	Non-EN
XCOPA	400	0
XWinograd	1858	0
XStoryCloze	300	300

Training Examples of the original datasets.

Data Augmentation

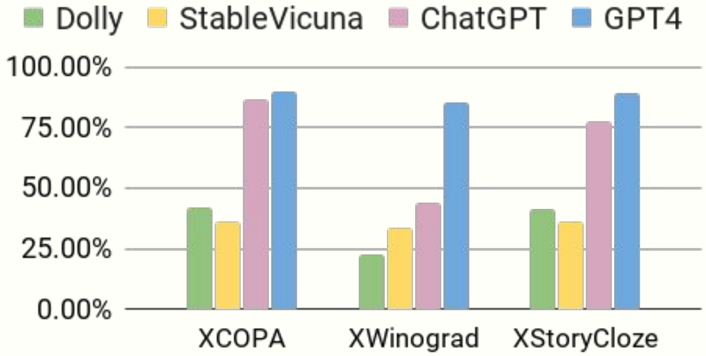
How do we generate more Data with LLMs?

- Start with **instructions** from the original dataset paper and iteratively improve -> improve instructions
- Set the desired total number of **examples** to generate (about 3K in our experiments)
  - *Randomly* sample 5-10 examples from the training datasets (*ensure diversity*)
  - Append these examples to the instructions and prompt the model to generate additional 5-10 new examples
  - Post-process and add valid and unique examples to the generation set

Data Augmentation

Which LLMs do we use?

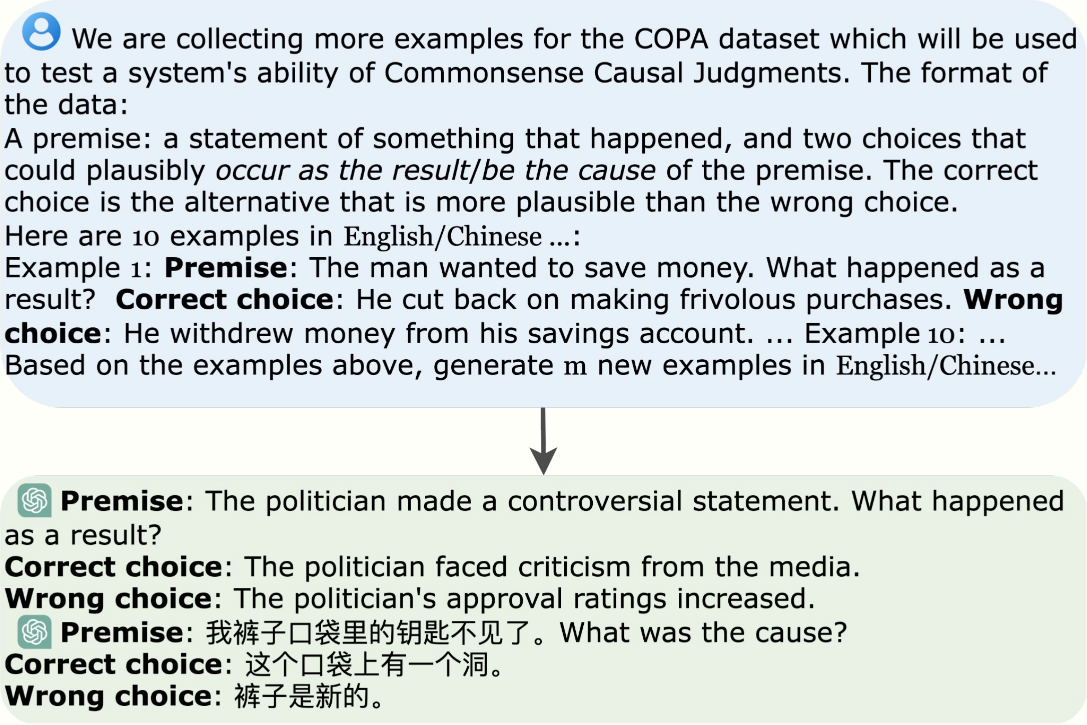
- 4 LLMs: Dolly-v2, StableVicuna-13B, ChatGPT, GPT-4
- They show different data-generation success rates
  actual_valid_examples / total_requested_examples



Data-generation Success Rate

Instruction & Generation Examples

ChatGPT-generated Examples in XCOPA



We are collecting more examples for the COPA dataset which will be used to test a system's ability of Commonsense Causal Judgments. The format of the data:
A premise: a statement of something that happened, and two choices that could plausibly *occur as the result/be the cause* of the premise. The correct choice is the alternative that is more plausible than the wrong choice.
Here are 10 examples in English/Chinese ...:
Example 1: **Premise**: The man wanted to save money. What happened as a result? **Correct choice**: He cut back on making frivolous purchases. **Wrong choice**: He withdrew money from his savings account. ... Example 10: ...
Based on the examples above, generate m new examples in English/Chinese...

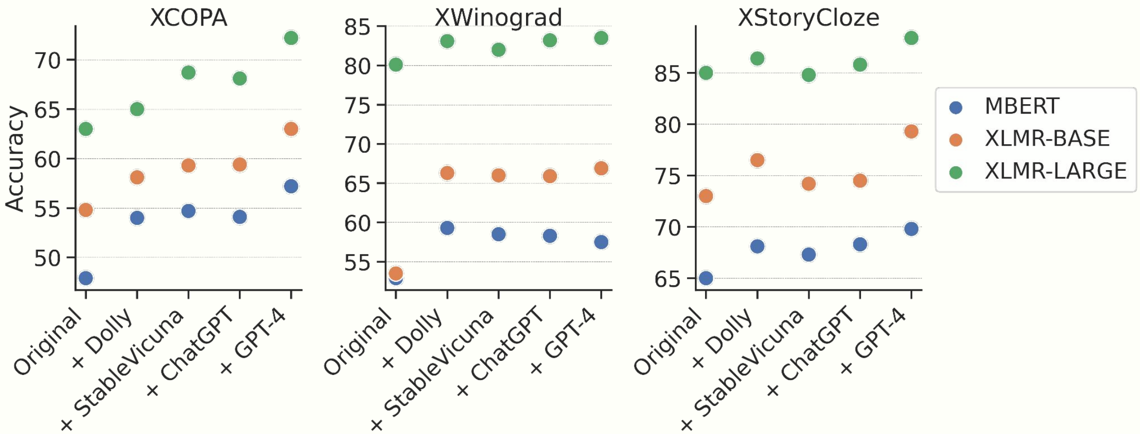
Premise: The politician made a controversial statement. What happened as a result?
Correct choice: The politician faced criticism from the media.
Wrong choice: The politician's approval ratings increased.

Premise: 我裤子口袋里的钥匙不见了。What was the cause?
Correct choice: 这个口袋上有一个洞。
Wrong choice: 裤子是新的。

Fine-tune Smaller Multilingual Models

Fine-tune mBERT, XLMR-Base, XLMR-Large

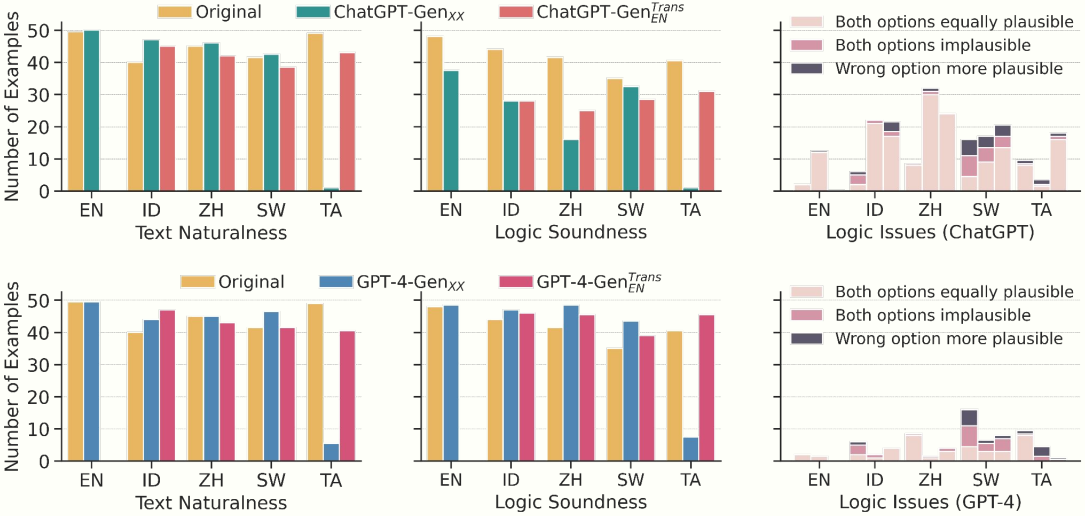
- Compare original & original + different LLM-generated EN data
- Training the models with *relatively large* synthetically generated data yields better performance than training with *limited* manually-created data
- Translating English-generated data with Google API is better than generating examples directly in target languages (paper)



Evaluation by Native Speakers

Text Naturalness & Logic Soundness

- Compare original, ChatGPT and GPT-4 generated XCOPA in target language, and translations of generated English data (50 examples)
- Both models can mostly generate fluent text, GPT-4 stands out in logic soundness
- Some languages are surprisingly bad, such as Tamil!



Conclusion

LLM-powered Data Augmentation is promising!

- LLMs demonstrate promises in Data Augmentation even for challenging multilingual commonsense reasoning tasks
  - Choice of LLM influences the performance of the fine-tuned models
  - LLMs such as ChatGPT and GPT-4 can generate high-quality data in many languages, but surprisingly struggle with certain languages such as Tamil
- Future work could explore the effectiveness of more recent instruction-tuned or aligned open-source LLMs, e.g. LLaMA 2