

LLM-powered Data Augmentation for Enhanced Crosslingual Performance







Chenxi Whitehouse^{1,3}, Monojit Choudhury², Alham Fikri Aji³

¹ City, University of London ² Microsoft ³ MBZUAI

Introduction

- We explore the potential of leveraging LLMs for data augmentation in multilingual commonsense reasoning datasets where the available training data is extremely limited.
- We fine-tune smaller models, mBERT and XLMR, using the synthesised data and showcasing the practical value of the LLM-generated data.
- Human evaluation of the naturalness and logical coherence of the data generated in various languages reveals the challenge in multilingual commonsense reasoning dataset generation.

Data Augmentation

- Establish the desired total number of examples to generate. Then generate following the steps below until sufficient examples.
- Randomly sample a set of n examples from the training datasets (diversity).
- Append these sampled examples to the instructions and prompt the model to generate an additional set of m new examples.
- Post-process and add valid and unique examples to the generated set.
- 4 LLMs: Dolly-v2, StableVicuna-13B, ChatGPT, GPT-4
- 3 Datasets: XCOPA, XWinograd, XStoryCloze
- They show different success rates of generating examples (actual_valid_examples/total_requested_examples):

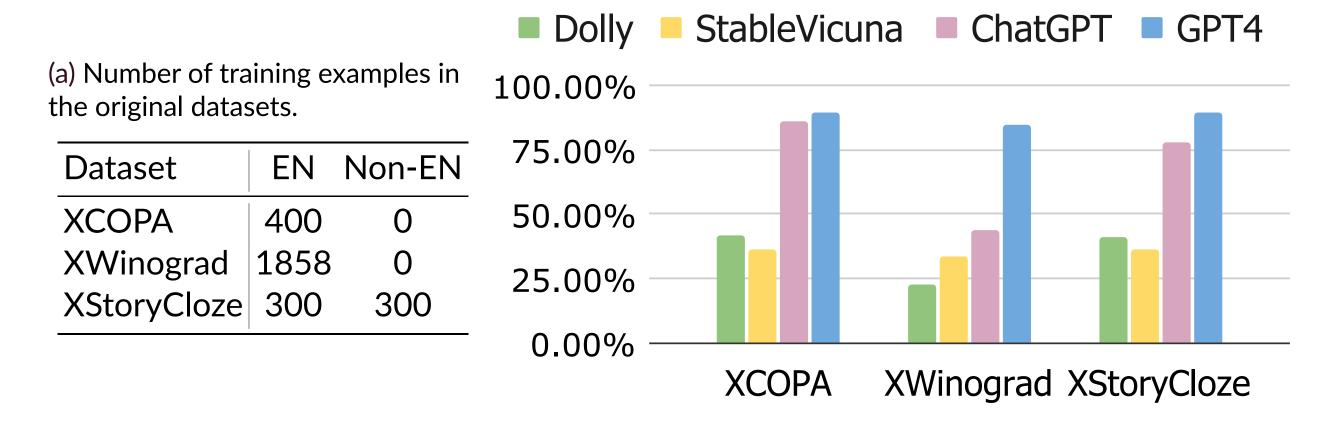


Figure 1. Training sizes in the 3 datasets and generate success% of the 4 LLMs.

Instruction and Generated Examples:

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。

Figure 2. Examples of instructions and ChatGPT-responses on XCOPA.

Fine-tuning Smaller Multilingual Models

- We fine-tune mBERT, XLMR-Base, and XLMR-Large, using the original and different LLM-generated English data.
- We can see that 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 *generally* better than generating examples directly in target languages.

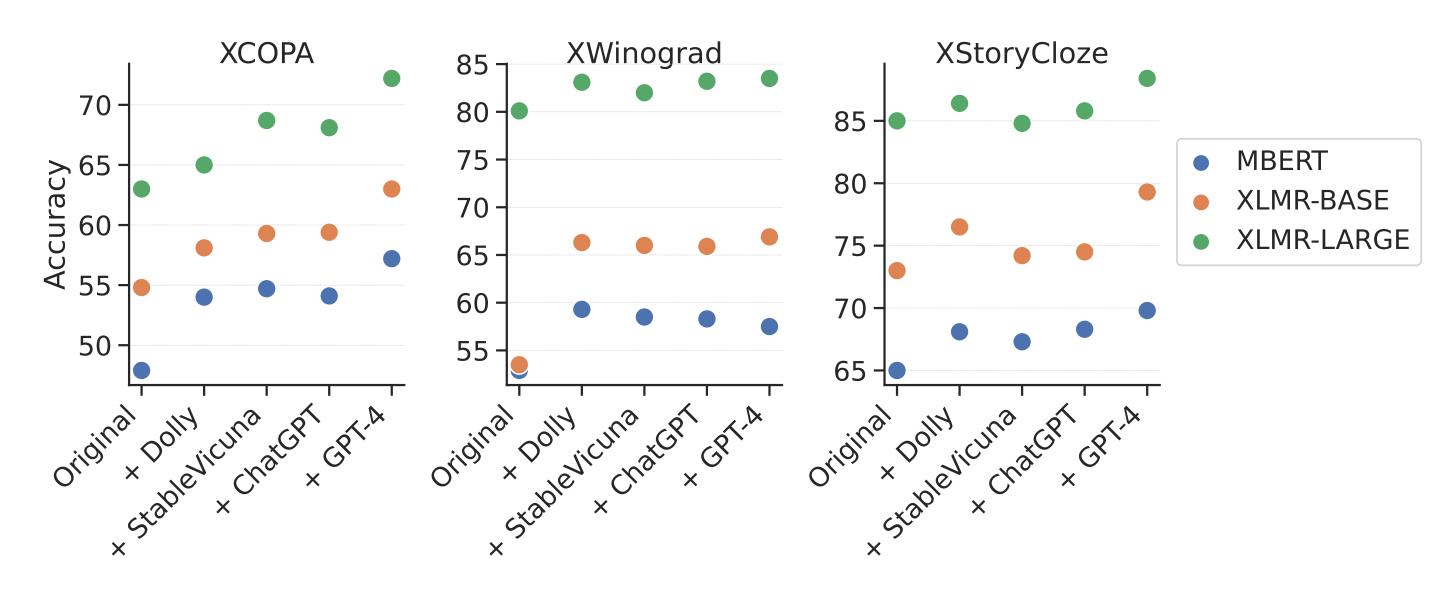


Figure 3. Average Accuracy across *all* languages when training on Original English data and Original+LLM-generated English data.

Human Evaluation

- We ask two native speakers to access the text naturalness and logic soundness of ChatGPT and GPT-4-generated Examples.
- Both models can mostly generate fluent text, GPT-4 stands out in logic soundness.
- Some languages are surprisingly bad, such as Tamil!

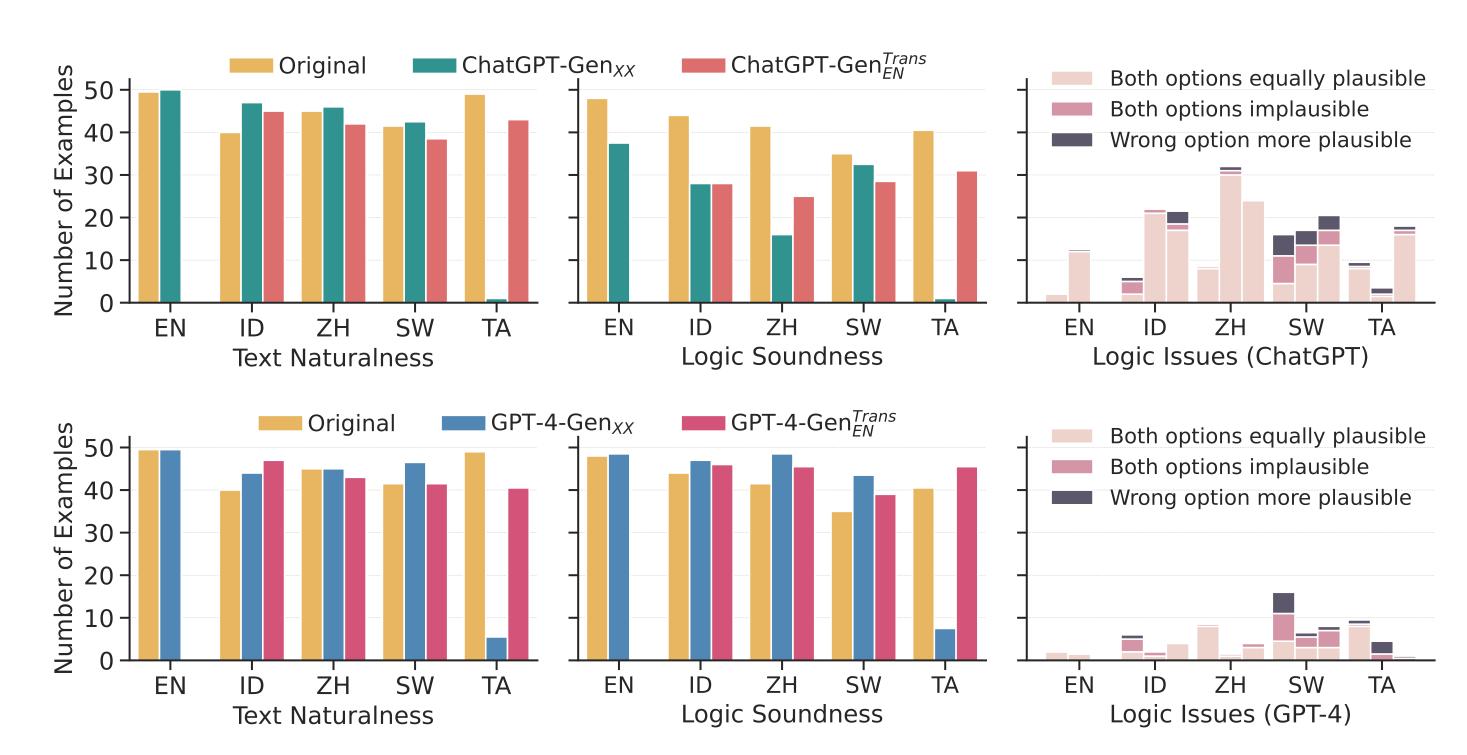


Figure 4. Human evaluation of 50 random examples from the original XCOPA, ChatGPT (top) and GPT-4 (bottom) generated data in target languages, and translation of English generated data. The three bars for each language in the most right subplots represent the logic issues of *Original*, Gen_{XX} , and Gen_{EN}^{Trans} .

Conclusions

- 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.
 - GPT-4 demonstrate the most robustness in data generation.
- Future work could explore the effectiveness of more recent instruction-tuned or aligned open-source LLMs, e.g. LLaMA 2.