Text Augmentation using LLMs

Yashwanth Babu Vunnam
34051046
yashwanthbab@umass.edu

Sneha Sree Vavilapalli
34062309
svavilapalli@umass.edu

Abstract

This paper delves into the application of Large Language Models (LLMs), specifically ChatGPT, for text augmentation in the context of addressing class imbalance in text classification problems. The study focuses on biased datasets where certain classes are under-represented, utilizing samples from Amazon, IMDB, and Yahoo review datasets. The experimentation involves downsizing class samples and fine-tuning the DistillBERT model for the primary classification task. Baseline comparisons are made with multiple augmentation models, evaluating performance through traditional methods such as cosine similarity and classification accuracy. The main focus is the classification performance, showcasing ChatGPT's superiority over other augmentation models. This research contributes valuable insights into the effectiveness of LLMs, particularly ChatGPT, for text augmentation in mitigating bias and enhancing classification accuracy.

1. Introduction

In the realm of Natural Language Processing (NLP), text data classification faces formidable challenges, particularly when dealing with imbalanced datasets featuring under-represented classes. The inherent bias in such datasets can result in skewed model predictions, where learning algorithms prioritize majority classes, leading to suboptimal performance for minority or under-represented categories. To overcome these challenges and bolster the overall robustness and accuracy of classification models, this project proposes the integration of Large Language Models (LLMs) for text augmentation.

While various augmentation techniques such as [3] and [5] exist for this purpose, the recent surge in LLMs like ChatGPT has demonstrated remarkable generative capabilities. Recent methods, including back-translation and word vector interpolation, leverage language models for more effective data augmentation. Current augmentation techniques face limitations due to constraints on both the accuracy and diversity of the generated text data.

LLMs are trained through self-supervised methods, scaling up with the available text corpus in open domains. Their expansive parameter space allows for the storage of extensive knowledge, and large-scale pre-training enables them to encode rich factual information for language generation, even in highly specific domains.

This project leverages the power of LLMs to address the challenges of under-represented classes by expanding existing datasets through the generation of synthetic examples. Text augmentation using LLMs facilitates the creation of contextually relevant and coherent text that closely aligns with the semantics and distribution of the original data. This approach aims to tackle data scarcity for minority classes, enhancing the generalizability of the classification model. This approach not only helps to improve the generalizability of the classification model but also promotes fairness and inclusivity in the classification process by ensuring that all classes are equally represented during the training phase.

The primary objective of this project is to maximize the class-wise classification accuracy of the model, surpassing both baseline performance and conventional augmentation strategies. Two types of data, namely sentiment analysis and topic classification, will be utilized for the classification tasks. Furthermore, the project aims to establish LLMs as a powerful tool for data augmentation in NLP, showcasing their potential to address data scarcity challenges and improve the overall performance of classification models.

2. Related Work

The study [8] delving into the application of Large Language Models (LLMs) for data augmentation in commonsense reasoning tasks contributes to the evolving landscape of natural language processing (NLP). While the paper under consideration focuses on a specific aspect, related research in the broader NLP domain provides context and inspiration.

An influential precursor to this work is the research by [7] and [1], where the authors highlight the prowess of pre-trained language models such as GPT-2 and GPT-3 in capturing complex language patterns across diverse domains. [8] build upon this foundation by extending the application

of LLMs to address data scarcity in commonsense reasoning tasks.

[6] contributed to the concept of synthetic data generation to tackle data scarcity in NLP tasks. Their exploration of generating synthetic data aligns with the broader motivation of enhancing model performance when faced with limited annotated datasets. [8]'s work extends this idea into the realm of commonsense reasoning, showcasing the versatility of LLM-powered augmentation.

Beyond synthetic data generation, the study by [4] delves into the effectiveness of multilingual pre-training for cross-lingual NLP tasks. While not directly related to commonsense reasoning, this research underscores the potential of leveraging pre-trained language models in diverse linguistic contexts. [8]'s work, although not explicitly cross-lingual, draws inspiration from the idea of maximizing the utility of pre-trained models.

2.1. Classification using using Chatgpt

The field of text data augmentation in natural language processing (NLP) has witnessed noteworthy developments, and AugGPT, proposed by [2], emerges as a promising addition to this landscape. AugGPT addresses a common challenge in NLP—limited data, which becomes particularly pronounced in scenarios where data scarcity impedes model performance.

AugGPT sets itself apart by harnessing ChatGPT, a prominent large language model recognized for its robust language comprehension. Its rephrasing methodology transforms each sentence into multiple semantically distinct variations, ensuring semantic consistency and diverse augmentation. This innovative approach addresses limitations observed in existing data augmentation methods that struggle to capture the entire spectrum of linguistic variations or create semantically inconsistent samples.

In the domain of text classification tasks, AugGPT demonstrates superiority over state-of-the-art augmentation methods. AugGPT's augmented samples exhibit higher testing accuracy and a well-distributed representation, showcasing its effectiveness in enhancing model performance.

Furthermore, the landscape of large language models in NLP has witnessed significant developments, with GPT-3, a precursor to ChatGPT, showcasing the capabilities of massive language models [1]. AugGPT, by leveraging ChatGPT for data augmentation, aligns with the broader trend of utilizing advanced language models to overcome data limitations in NLP.

3. Datasets

In our experiments, we utilized two distinct domains of datasets, each serving a specific purpose in evaluating the effectiveness of Large Language Models (LLMs) for text augmentation in sentiment analysis and topic classification tasks.

3.1. Amazon dataset

The Amazon dataset comprises customer reviews spanning 24 product categories. The primary objective is to classify reviews into their respective product categories. Given the proverbially large size of the original Amazon product dataset, we judiciously sampled a subset of 300 samples from each category. This sub-sampling approach allows for a manageable yet representative selection of data, enabling effective experimentation while preserving the diversity inherent in the vast Amazon product catalog.

3.2. IMDB dataset

The IMDB dataset encompasses two distinct types of sentiment analysis: Sentiment Polarity and Sentiment Scale. In the Sentiment Polarity dataset, we curated 1000 positive and 1000 negative processed movie reviews. On the other hand, the Sentiment Scale dataset consists of documents labeled using a rating scale, introducing a non-binary aspect to the sentiment analysis. This nuanced approach allows for a more comprehensive evaluation of sentiment understanding within the IMDB dataset.

3.3. Yahoo! dataset

For topic classification, we leveraged the Yahoo! Answers dataset, encompassing a vast array of questions and their corresponding answers. The corpus distributed here encompasses approximately 4 million questions and answers. Alongside the textual content, the dataset includes a modest amount of metadata, such as the designation of the best answer, and category and sub-category assignments for each question. Due to resource constraints, only a subset of this extensive dataset was employed for our experiments. This subset ensures a judicious balance between the computational demands of the experiment and the need for a representative sample, preserving the richness of question-answer dynamics in the Yahoo! Answers dataset.

By employing datasets from Amazon, IMDB, and Yahoo!, our experiment covers a spectrum of domains, enabling a comprehensive assessment of the LLM-based text augmentation approach across diverse contexts of sentiment analysis and topic classification.

4. Experimental Setup

The experimental pipeline begins with the acquisition of the initial dataset, laying the foundation for subsequent analyses. To simulate scenarios characterized by limited data availability, the dataset undergoes underscaling for specific classes. The degree of underscaling is systematically varied, introducing diverse scarcity levels and facilitating a nuanced evaluation. Once the biased datasets are prepared,

Prompts	Generated ChatGPT response
Augmented text for the sentence	Absolutely amazing! I love it; the emotional depth is incredibly touching. It's just unfortunate that the DVD description doesn't mention the absence of subtitles. Despite this drawback, the content itself is so captivating that it's still a must-watch.
Generate augmented texts for the following sentence, we use this data for the text classification task of the product review dataset:	Incredible product! I absolutely love it; the emotional impact is profound. It's a bit disappointing that the DVD description doesn't mention the absence of subtitles, but overall, it's a top-notch experience.

Table 1. The Above table shows different responses ChatGPT gives when the prompt is changed. We can see that the augmented text has more relevance when the context is given into the prompt.

Original Text	Generated ChatGPT response
Great book for travelling Europe: I currently live in Europe, and this is the book I recommend for my visitors. It covers many countries, colour pictures, and is a nice starter for before you go, and once you are there.	A must-read for anyone exploring Europe! Living in this diverse continent, I often suggest this book to friends and family visiting me. Packed with comprehensive information on multiple countries, enriched with colorful visuals, it not only prepares you before your trip but also enhances your experience once you're here. A travel essential!
Some of the best fiddle playing I have heard in a long time: This is an excellent album with some great fiddle playing. Ryan is amazing to watch when he plays in person. You can feel the emotion he puts into each song	Experience the magic of masterful fiddle playing with this outstanding album! Featuring some of the best performances I've heard in a long time, Ryan's skill is truly exceptional. Watching him play live is a mesmerizing spectacle; you can tangibly feel the depth of emotion he pours into every song. A must-have for any fiddle enthusiast!

Table 2. The Above table shows different responses ChatGPT gives when asked to augment the original text. In practise, We generated more than one sentence, which is mentioned in the prompt i.e., Generate 4 augmented texts for the sentence 'Mention the data sample'

they are subjected to the baseline classification model to establish a performance benchmark.

The augmentation phase follows, where the downscaled data is processed using a spectrum of augmentation models. Traditional augmentation methods, alongside the powerful ChatGPT, are employed. The unique strength of ChatGPT lies in its extensive knowledge, enabling semantic-level exploration during the augmentation process. The produced augmented data, encompassing outputs from various models, including ChatGPT, is appended to the downscaled dataset, forming a new entity termed D_{aug}. This augmented dataset serves as the input for the classification model, and its performance is scrutinized and compared against the baseline results.

The experimental setup is designed for iterative execution, allowing for variations in downsampling percentages and augmentation models. This iterative approach ensures a comprehensive exploration of ChatGPT's augmentation capabilities across diverse scarcity levels. By systematically adjusting parameters and evaluating outcomes, the experimental design aims to provide nuanced insights into the model's effectiveness in enhancing classification performance on imbalanced datasets.

4.1. Data Augmentation Baselines

we compare our method with other popular data augmentation methods. For these methods, we use the implementation in open-source libraries.

• **InsertCharAugmentation:** Inserts random characters at random locations in text, which improves the generalization ability of the model by injecting noise into the data.

• **SubstituteCharAugmentation:** Randomly replaces selected characters with other ones.

• **DeleteCharAugmentation:** Randomly deletes characters.

• **WordNet:** This method expands a given text dataset by replacing words with their synonyms or related terms from WordNet, a lexical database of the English language.

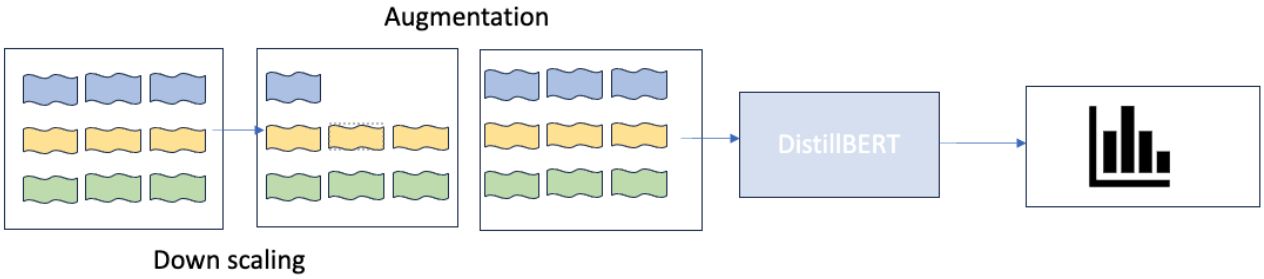


Figure 1. This shows the architecture of the experimental setup.

4.2. Data Augmentation using ChatGpt

For the text augmentation process utilizing ChatGPT, we leverage the capabilities of ChatGPT's open API. To address class imbalance, we focus on the minority class, extracting the required number of samples to align the sample count across all classes. Subsequently, these selected samples serve as the input for augmentation. Using a specified prompt, each data sample is sent to ChatGPT for augmentation. Upon receiving the responses from ChatGPT, these generated data samples are seamlessly integrated into the original downscaled dataset. This augmentation procedure effectively balances the representation of classes, mitigating bias in the dataset and rendering it more diverse. As a result, the augmented dataset reflects a more equitable distribution of samples across classes, contributing to enhanced model training and performance.

4.2.1 Prompts and responses

We utilized the power of Large Language Models (LLMs) to create synthetic data points for all datasets through targeted prompts. Employing various instructions, we randomly selected samples from the training split of the downscaled data and presented them to the LLM for augmentation. Through a systematic evaluation of the classification performance using different instructions as shown in Table.1 and Table.2, we identified the set of instructions that yielded the most favorable results. Notably, we observed that providing contextual instructions to the LLM resulted in the generation of more analogous and relevant data points. This improvement in relevance was further reflected in the enhanced classification performance. The table above provides a glimpse into our experimentation with different instructions, showcasing the impact of contextual guidance on the LLM's ability to generate synthetic data points that align closely with the original data distribution.

4.3. Classification

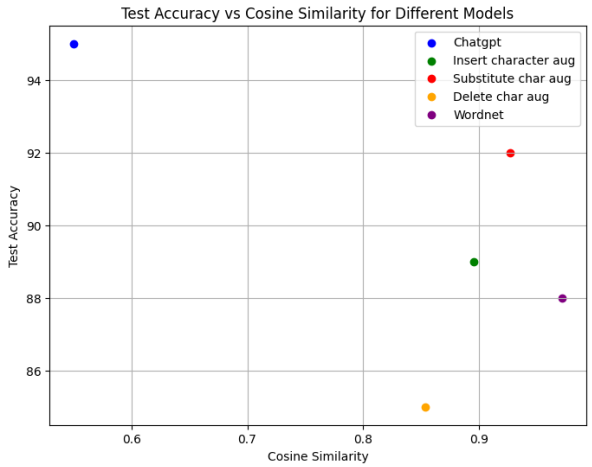
In our classification task, we opted for DistillBERT as the model of choice to assess the synergy between Large Language Models (LLMs) and the augmentation techniques elucidated in the preceding sections. The experimental design encompassed several key steps to comprehensively evaluate the impact of LLM-based augmentation on the downstream classification performance. Integration with ChatGPT's open API facilitated seamless augmentation, enabling the generation of diverse and contextually relevant synthetic data points. The augmented dataset, now comprising both original and synthetic samples, underwent a standard split into training and testing sets to ensure an unbiased evaluation of the classification model's performance.

DistillBERT, chosen for its balance between computational efficiency and performance, served as the classification model. Fine-tuning the DistillBERT model on the augmented training set involved standard configurations, including learning rate, batch size, and training epochs. The model was trained to predict class labels based on the respective datasets.

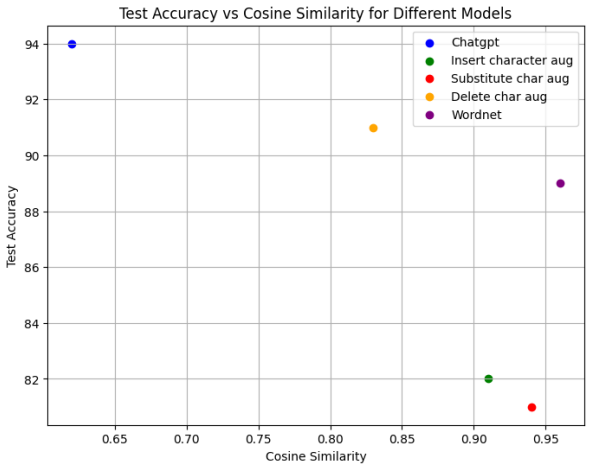
This systematic framework enabled a thorough investigation into the influence of LLM-based text augmentation on DistillBERT's classification performance across diverse datasets. By integrating contextual information into the augmentation process, our approach aimed to enhance the relevance and diversity of synthetic data points, contributing to improved model generalization and classification accuracy.

4.4. Evaluation Metrics

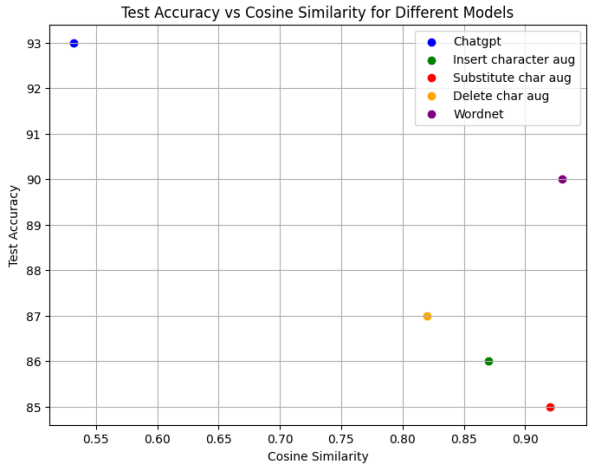
We utilized cosine similarity as a measure to evaluate faithfulness, determining how closely the generated data samples align with the original samples. Additionally, we assessed compactness, gauging whether samples within each class exhibit sufficient closeness for effective discrimination.



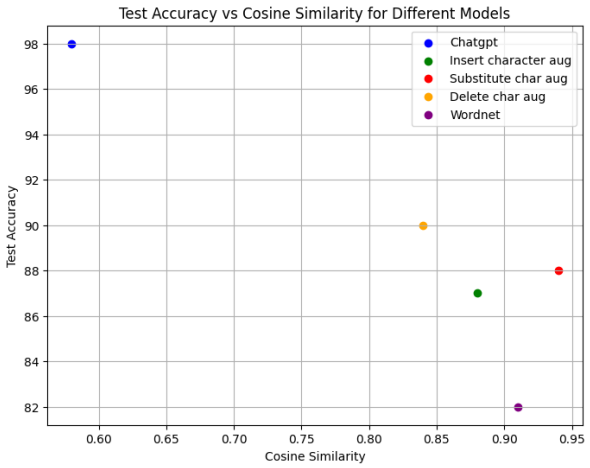
(a) Cosine Similarity measure for IMDB polarity dataset



(b) Cosine Similarity measure for Amazon review dataset



(c) Cosine Similarity measure for IMDB review dataset



(d) Cosine Similarity measure for Yahoo dataset

Figure 2. Accuracy vs Cosine Similarity for various augmentation techniques and averaged across different classes

Percentage of downsize	Imbalanced	Aug data accuracy	ChatGPT aug data accuracy
10	88.64	91.33	96.67
20	83.21	88.55	95.21
30	85.54	89.65	97.94
40	86.66	89.33	96.91
50	90.33	92.75	97.00

Table 3. Average Test metrics for IMDB polarity datasets for different classes downsized

4.4.1 Cosine similarity

Text embedding is a process where words or phrases are represented as numerical vectors in a high-dimensional space. This transformation allows for the capture of semantic relationships and contextual information. In the context of text augmentation, embedding is employed to create a numerical representation of both the original and augmented text samples.

Cosine similarity is a metric used to assess the similarity between two vectors in a multi-dimensional space. In the context of text augmentation, cosine similarity is employed to quantify how closely the augmented text aligns with the original text. It measures the cosine of the angle between two vectors and provides a value between -1 and 1, where a higher value indicates greater similarity.

Model	IMDB polarity	Amazon review dataset	IMDB classification	Yahoo dataset
Imbalance	83	84	87	81
Aug data accuracy(WordNet)	88	89	82	90
Chat GPT accuracy	95	94	98	93
InsertCharAugmentation	89	82	87	86
SubstituteCharAugmentation	92	81	88	85
DeleteCharAugmentation	85	91	90	87

Table 4. Average test metrics across different models for 20 percent downsampling

Algorithm 1 Architecture for Text classification

Input: Dataset D
Downsize the data by various percentages D_{set}
for D' in D_{set} **do**
    D_{Aug} = $Augment(D')$
    $D_0 = D' + D_{Aug}$
    for epoch in epochs **do**
        train(model, D_0)
    end for
end for

When applied to text vectors, cosine similarity becomes a valuable tool to evaluate the faithfulness of augmented data. High cosine similarity suggests that the augmented text maintains a similar semantic context to the original text, signifying a successful augmentation process.

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \tag{1}$$

5. Results and Discussion

The above images show that Cosine similarity of chatGPT is lowest of all than the other three methods: Custom library augmentor (WordNet from nltk), InsertCharAugmentation, SubstituteCharAugmentation, DeleteCharAugmentation, WordNet. We see that ChatGPT augmented text has low cosine similarity than other because of the efficient augmentation technique and also leading to high test accuracies. In InsertCharAugmentation, SubstituteCharAugmentation, DeleteCharAugmentation we do character level augmentation hence the texts would not vary much when we calculate the cosine similarity metric. In WordNet the word gets replaces one of the similar words in their corpus. Even here we are not varying the text length or have the ability to generate extra few sentences which is slightly different from the text given to augment. Value of 0.5 indicates that not there is perfect balance of diversity and similarity in the texts which prove to be useful in our text classification tasks enabling the model to learn new context. Also we observe that the model is performing constantly better when we use ChatGPT for augmentation resulting in consistent accuracies above 90 % .

6. Future Work and Conclusion

In this study, we delved into the augmentation capabilities of ChatGPT for text classification as downstream tasks. The findings revealed that ChatGPT surpasses conventional augmentation methods when applied to classification tasks. Unlike existing augmentation techniques, our architecture explores the semantic nuances of limited data, enhancing data consistency and robustness, ultimately resulting in superior performance compared to many current methods.

Looking ahead, there are promising avenues for future research. First, the ChatGPT augmentation approach can be systematically tested across a diverse array of downstream tasks to gauge its adaptability and effectiveness. This broader exploration will shed light on the versatility of ChatGPT augmentation in various domains.

Additionally, given the scarcity and under-representation of data in specific domains like medical datasets, applying our method to such challenging datasets presents an intriguing avenue. The unique semantic-level exploration by ChatGPT may prove particularly beneficial in domains where traditional augmentation methods face limitations. Automation of the prompting process represents another avenue for future exploration. Establishing a pipeline where a model suggests prompts for augmentation and downstream tasks can streamline the process. By training a model to optimize prompt suggestions, we aim to improve the overall accuracy and efficiency of the augmentation process.

The future directions outlined, including diverse task exploration, domain-specific applications, and automated prompting, underscore the potential for further advancements in leveraging large language models for effective data augmentation in natural language processing.

Access the video link here

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom

Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1, 2

[2] Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, W. Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. Auggpt: Leveraging chatgpt for text data augmentation. 2023. 2

[3] Biyang Guo, Songqiao Han, and Hailiang Huang. Selective text augmentation with word roles for low-resource text classification, 2022. 1

[4] Songbo Hu, Han Zhou, Zhangdie Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. A systematic study of performance disparities in multilingual task-oriented dialogue systems, 2023. 2

[5] Amit Jindal, Arijit Ghosh Chowdhury, Aniket Didolkar, Di Jin, Ramit Sawhney, and Rajiv Ratn Shah. Augmenting NLP models using latent feature interpolations. pages 6931–6936, Barcelona, Spain (Online), 2020. 1

[6] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022. 2

[7] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[8] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. Llm-powered data augmentation for enhanced cross-lingual performance, 2023. 1, 2