GPTs Are Multilingual Annotators for Sequence Generation Tasks

Juhwan Choi[1], **Eunju Lee**[2], **Kyohoon Jin**[2] and **Youngbin Kim**[1,2]

[1]Department of Artificial Intelligence, Chung-Ang University
[2]Graduate School of Advanced Imaging Sciences, Multimedia and Film, Chung-Ang University
{gold5230,dmswn5829,fhzh123,ybkim85}@cau.ac.kr

Abstract

Data annotation is an essential step for constructing new datasets. However, the conventional approach of data annotation through crowdsourcing is both time-consuming and expensive. In addition, the complexity of this process increases when dealing with low-resource languages owing to the difference in the language pool of crowdworkers. To address these issues, this study proposes an autonomous annotation method by utilizing large language models, which have been recently demonstrated to exhibit remarkable performance. Through our experiments, we demonstrate that the proposed method is not just cost-efficient but also applicable for low-resource language annotation. Additionally, we constructed an image captioning dataset using our approach and are committed to open this dataset for future study. We have opened our source code for reproducibility.[1]

1 Introduction

With the evolution of deep learning methods, various tasks in the NLP domain have demonstrated remarkable performance. However, training deep learning models requires a substantial amount of labeled data. Data annotation, a process of gathering unlabeled data and labeling them, plays a crucial role in fulfilling this data demand.

However, as the conventional procedure of data annotation is mainly conducted manually using human annotators, it cannot meet the growing demand for labeled data with an increase in the size of deep learning models (Qiu et al., 2020). Moreover, it is significantly challenging to recruit annotators for low-resource languages (Pavlick et al., 2014).

To address the lack of labeled data and improve the performance of the model, the concept of pre-trained language model (PLM) was introduced. These PLMs have been trained on a large amount of text corpus to acquire a general knowledge of languages (Radford et al., 2018; Devlin et al., 2019). By fine-tuning these models to specific downstream task, it was able to achieve performance improvement without the need for additional labeled data.

With the evolution of PLMs via the enlargement of their sizes owing to increased training data, the development of a large language model (LLM) with massive parameter size enabled few-shot learning from the context of the given prompt (Brown et al., 2020). Accordingly, the diverse capabilities of LLMs have been investigated (Zhao et al., 2023).

However, despite their impressive abilities and adaptability, these LLMs cannot be actively exploited for downstream tasks because of the cost constraints and demand for hardware resources caused by their extensive model size. Additionally, fine-tuning these models for specific purposes remains challenging due to their massive parameter size. Consequently, training models for downstream tasks through labeled data is still the dominant approach for practical applications (Yu et al., 2023).

Data annotation refers to the creation of labeled data by assigning gold labels to unlabeled data. Traditionally, data annotation was mainly conducted by human labelers using crowdsourcing platforms, such as Amazon mechanical turk (MTurk), and these platforms have aided the creation of modern, large-scale datasets. Recently, to address these limitations of crowdsourcing-based data annotation and achieve a cost-efficient means to collect labeled data, several studies have proposed the utilization of LLMs as alternative annotators in place of human labelers (Wang et al., 2021; Ding et al., 2023; Gilardi et al., 2023; Jiao et al., 2023; Li et al., 2023; Zhang et al., 2023; He et al., 2023; Bansal and Sharma, 2023). These studies have shown the possibility of cost-efficient and automatic data annotation through LLMs, such as GPT-3.

[1]https://github.com/c-juhwan/gpt-multilingual-annotator

Findings of the Association for Computational Linguistics: EACL 2024, pages 17–40
March 17-22, 2024

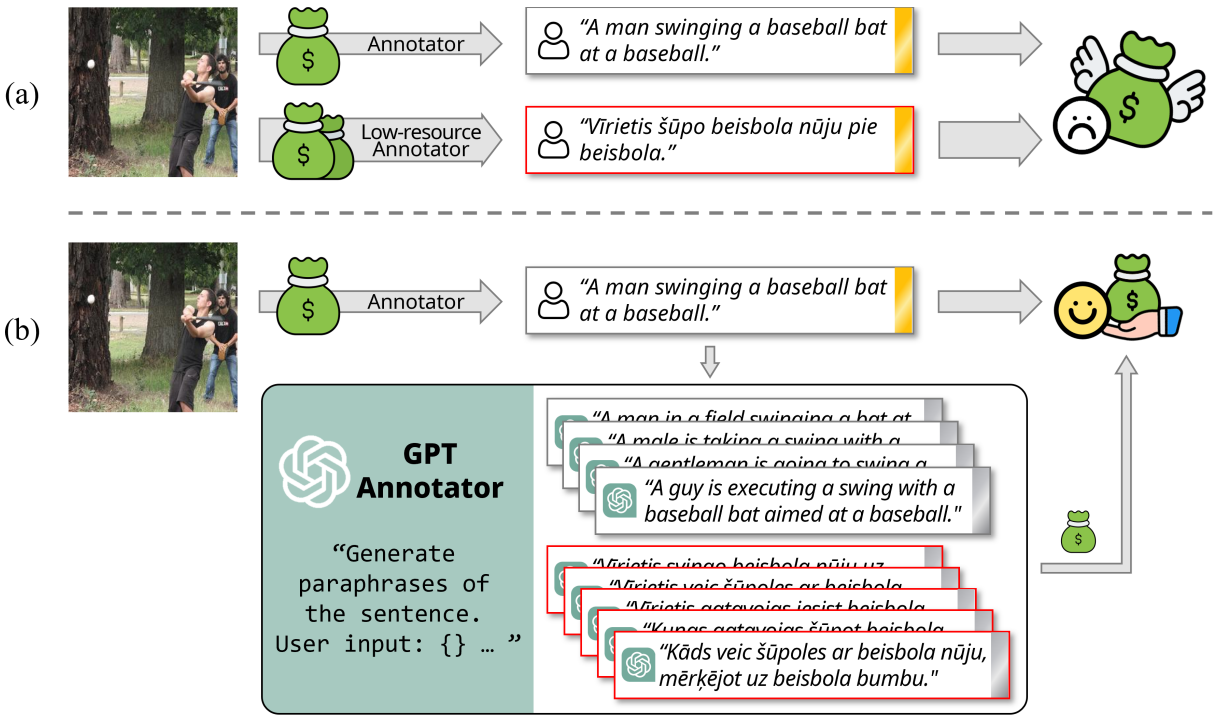

Figure 1: Overall concept of our GPT annotator. (a) Conventional annotation process for image captioning task, which is performed by multiple human annotators and expensive. Moreover, it is more expensive to hire human annotators for low-resource languages. (b) The annotation process of proposed GPT annotator. With one gold caption by a single human annotator, the GPT annotator automatically generates silver captions, as well as captions in other languages, resulting in a cost-efficient dataset construction.

However, as these existing studies mainly focused on simple tasks, such as text classification, additional investigation is required to apply these approaches to numerous subtasks of natural language processing. Moreover, the potential of automatic data annotation via LLMs has not been explored for languages other than English. As previously highlighted, projects in low-resource languages may suffer from the high cost of data annotation, necessitating the need for automatic annotators for languages beyond English.

In this study, we proposed a strategy that leverages LLMs as an assistant annotator to aid human annotators in image captioning task and text style transfer task. As depicted in Figure 1, the conventional process of establishing datasets for image captioning task required a considerable number of human annotators to generate five gold annotations for each image, resulting in a high cost for dataset construction in languages beyond English. Moreover, the quality of the annotated data varies depending on the proficiency of the human annotators (Rashtchian et al., 2010). Similarly, the annotation process for text style transfer required significant human effort, including quality control (Rao and Tetreault, 2018; Briakou et al., 2021).

This study demonstrated the ability of LLMs to serve as assistant annotators for human annotators at a reasonable cost by generating multiple silver sentences for each gold annotation written by one single human annotator. Specifically, we proposed a cost-efficient process to construct multilingual language datasets by exploiting the GPT annotator. Particularly, we utilized GPT-4, which exhibits enhanced multilingual capabilities (OpenAI, 2023), to autonomously produce diverse sentences in another language from a single English sentence, even if the human annotator is not familiar with the target language. Moreover, the cost of the GPT annotator is constant as the cost is determined by the length of the processed token, regardless of the language. This highlights the efficiency of the proposed GPT annotator as an annotation method for low-resource language, which is more expensive and time-consuming compared to English.

Employing this method, we developed an image captioning dataset in Latvian, Estonian, and Finnish — which are well-known low-resource languages — by employing the GPT annotator. In this scenario, a single human annotator, who lacks

knowledge of the target language, provides one English gold caption for each image. Through the experiment, we demonstrated that the proposed method achieves better performance compared to machine translation method. We open these datasets to support future studies. Additionally, we release software to easily perform data annotation process described in this paper.

Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work to explore the possibility of LLM as a multilingual annotator.
- To the best of our knowledge, this is the first study to employ LLM as an automatic annotator for image captioning task and text style transfer task.
- Our experiment reveals the ability of GPT annotators to serve as human annotators at a reasonable cost.
- We release an annotation software to easily perform the method described in the paper, as well as three image captioning datasets in Latvian, Estonian, and Finnish.

2 Related Work

GPT-3 has demonstrated that LLMs can conduct in-context learning from few-shot prompts. Accordingly, various LLMs with different characteristics have been proposed (Zhao et al., 2023). For example, based on the findings that LLMs can be further enhanced via human instruction and feedback (Ouyang et al., 2022), ChatGPT[2] and its backbone GPT-3.5 with various abilities have emerged (Leiter et al., 2023; Yang et al., 2023; Liu et al., 2023). In addition, the cutting-edge GPT-4 (OpenAI, 2023) is a progressed version of GPT-3.5, with a longer input sequence, improved multilingual ability, and image inception ability.

With the advancement of LLMs, studies have been conducted to augment given human-annotated data (Yoo et al., 2021; Whitehouse et al., 2023), or to annotate unlabeled data and train models for downstream tasks. One of the early studies in this field (Wang et al., 2021) proposed an automatic annotation method that demonstrated the ability of GPT-3 to annotate a greater amount of data compared to human annotators at a lower labeling cost, resulting in higher performance at the same cost, and this strategy was observed to outperform GPT-3 itself. In addition, the study investigated the possibility of a collaboration between human and GPT annotators by leveraging the confidence of the automatic annotation of GPT to perform active labeling by human annotators.

Following this approach, subsequent studies expanded the annotation capabilities of GPT-3 to not just label unlabeled data but also create labeled data leveraging external knowledge, or even from scratch (Ding et al., 2023). Meanwhile, a methodology was proposed to transfer the abilities of LLMs into a smaller model by generating a rationale for the labeled data, enhancing the performance of the small model (Hsieh et al., 2023).

With the emergence of ChatGPT, an improved version of GPT-3 that enables enhanced flexibility across diverse tasks, researchers have proposed its application for data annotation. ChatGPT has been reported to outperform crowdworkers in text classification tasks in certain cases with the same instructions (Gilardi et al., 2023). Additionally, studies observed that ChatGPT even surpassed expert labelers in the annotations of political texts (Törnberg, 2023). These results have led researchers to examine the annotation abilities of ChatGPT across various domains (Zhu et al., 2023).

Recent studies have expanded the application of LLMs as annotators, from language understanding tasks, such as text classification or inference, to text generation tasks. For example, a previous study reported improved performance in query-focused summarization by reducing the noise of ChatGPT (Laskar et al., 2023). Additionally, dialogue generated by ChatGPT has been observed to demonstrate comparable quality to reference dialogues written by human annotators (Labruna et al., 2023).

These studies indicate the capability of LLMs, including ChatGPT, to perform as an effective annotator for not just text understanding tasks but also text generation tasks, which are more complex and challenging to annotate. However, the application of these abilities of LLMs to various natural language processing tasks is still limited and underexplored. In this study, we proposed an LLM-based annotation method for image captioning task and text style transfer task, which has not been investigated in previous studies. Furthermore, we validated the feasibility of LLMs as an autonomous multilingual annotator, which has not been explored in previous works.

[2] https://openai.com/blog/chatgpt

3 Method

3.1 Task Formulation

We first define a dataset D, which is composed of the data pair $d = (X, Y)$. In image captioning task, X denotes a given image and $Y = \{y_{g_1}, y_{g_2}, ..., y_{g_5}\}$ is corresponding captions that describe X. In this paper, g means "gold", which represents a human-annotated sentence. Similarly, in text style transfer task, X denotes the original sentence and Y_g indicates human-annotated pair sentence with desired style.

Traditionally, multiple human annotators are used to write descriptions for unannotated data X to construct such datasets, especially for image captioning, which requires multiple captions for each image. However, as previously discussed, this entirely human-based annotating process is expensive and time-consuming. Our GPT annotator aims to construct a data pair by autonomously generating silver sentences and reduce the time and cost consumption of data annotation process.

Additionally, we explore the multilingual ability of the GPT annotator. The cost of data annotation varies by language. Especially, Low-resource languages are associated with higher cost and high time consumption for the collection of annotated data (Ul Haque et al., 2021; Guemimi et al., 2021; Li et al., 2019; Kim et al., 2021). This phenomenon is caused by the language pool of the crowdworkers (Pavlick et al., 2014) and the difficulty of training low-resource language natives (Lin et al., 2018). In this study, we propose a method to employ the GPT annotator as a multilingual annotator through the adaptation of GPT-4, which has significantly improved multilingual ability (OpenAI, 2023).

3.2 Assistant Multilingual Annotator for Image Captioning Task

To achieve the aforementioned goal, we synthesized the given human-annotated caption y_{g_1} by utilizing the GPT model, and generated a set of paraphrases $\{y_{s_2}, ..., y_{s_5}\}$ based on y_{g_1}.

We configured a well-designed prompt P, as the input for GPT to achieve this object. As it has been reported that LLMs perform significantly better with examples rather than zero-shot (Brown et al., 2020), the prompt P includes an one-shot desired example. The process of generating sentences through GPT can be expressed as follows.

$$\{y_{s_2},\ ...,\ y_{s_5}\} = \text{GPT}(P,\ y_{g_1}) \quad (1)$$

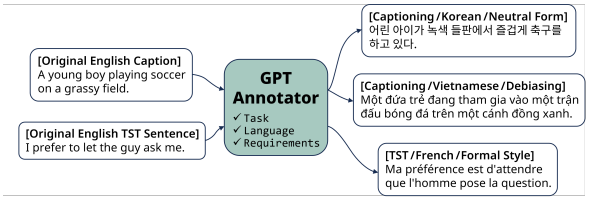

Figure 2: Our GPT annotator can generate various datasets with configurable prompts, primarily regarding task, language, and specific requirements.

The machine-annotated caption produced in Eq. 3 is used to construct a new data pair, $d' = (X, y_{g_1}, y_{s_2}, ..., y_{s_5})$, and a downstream task model is trained using dataset D', a collection of these d'. Consequently, GPT can be used to assist human annotators with image captioning task.

In addition, to employ our GPT annotator as multilingual annotator, it first synthesizes a data pair with one single human annotation in English, $d^{src} = (X, y^{eng}_{g_1})$ to reduce the cost of hiring multiple human annotators. Secondly, the GPT annotator generates a set of paraphrases in a target language $\{y^{tgt}_{s_1}, ..., y^{tgt}_{s_5}\}$. This process is performed through a prompt P^{tgt} with information in the target language, including a one-shot desired example. We found it helpful to jointly generate English sentence and its translation rather than solely generate sentences in the target language, as English sentence guides the generation of target language sentence. Specific prompts can be found in Appendix F.1. The described process can be expressed as follows.

$$Y_{tgt} = \{y^{tgt}_{s_1},\ ...,\ y^{tgt}_{s_5}\} = \text{GPT}(P^{tgt},\ y^{eng}_{g_1}) \quad (2)$$

The dataset in target language D^{tgt} can be constructed through $d^{tgt} = (X, Y^{tgt})$ obtained by the GPT annotator, and a downstream task model in the target language can be trained using this D^{tgt}. This overall process enables the construction of a dataset D^{tgt} in any designated language with only one single annotation in English by utilizing the LLM. Furthermore, this process is performed without any intervention of a human annotator who is fluent in the target language, reducing the cost of hiring expert annotators in the target language.

3.3 Assistant Multilingual Annotator for Text Style Transfer Task

For text style transfer task, we first analyze the given data pair $d^{src} = (X^{eng}, Y^{eng}_g)$ written in English through the GPT annotator. Nextly, the

GPT annotator creates a translated version of the pair and its paraphrase in target language, $d_1^{tgt} = (X_{s1}^{tgt}, Y_{s1}^{tgt})$ and $d_2^{tgt} = (X_{s2}^{tgt}, Y_{s2}^{tgt})$. This generation of paraphrase allows to fully utilize given annotation and effectively construct a dataset in target language with a limited amount of annotated data.

Similarly to image captioning task, we configured a well-designed prompt P^{tgt} for the annotation process, including an one-shot example. Specific prompts can be found in Appendix F.2. The process described in this section can be formulated as follows.

$$\begin{aligned}\{d_1^{tgt},\ d_2^{tgt}\} &= \{(X_{s1}^{tgt}, Y_{s1}^{tgt}),\ (X_{s2}^{tgt}, Y_{s2}^{tgt})\} \\ &= \text{GPT}(P^{tgt},\ (X^{eng}, Y_g^{eng}))\end{aligned} \tag{3}$$

We could acquire text style transfer dataset D^{tgt} in the target language through this process.

4 Experiment

4.1 Experimental Design

This section describes experimental design to validate the effectiveness of our GPT annotator in each tasks. We primarily assessed our method based on the performance of the model trained on the downstream task, which can serve as an indirect measure of the quality of synthesized dataset (Ye et al., 2022). Further implementation details can be found in Appendix A.

4.1.1 Image Captioning Task

To assess the cost-efficiency of our GPT annotator, we evaluated the proposed GPT annotator through three different image captioning datasets: Flickr8k (Rashtchian et al., 2010) dataset was constructed by annotating approximately 8,000 images collected from Flickr via MTurk. Flickr30k (Young et al., 2014) dataset is an extension of Flickr8k dataset, and it consisted of 30,000 images with captions acquired through crowdsourcing. MSCOCO (Lin et al., 2014; Chen et al., 2015) dataset is an annotated dataset of more than 160,000 images.

As Flickr8k and Flickr30k datasets do not provide explicit validation and test sets, we divided them in the ratio of 8:1:1. For the MSCOCO dataset, we utilized the COCO 2014 split, which consists of approximately 82,000 training data, 40,000 validation data, and 40,000 test data. To validate the effectiveness of the proposed method, we set up a scenario with only one gold caption per image by selecting only one caption for the original dataset.

BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2014) metrics were measured through the NLG-EVAL library (Sharma et al., 2017) for evaluation. Additionally, we also employed BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021) for model-based evaluation. For the MSCOCO dataset, the performance was evaluated through the official evaluation server.[3] For multilingual experiments, we adapted different datasets for each language, a subset of the aforementioned datasets with annotated captions. These datasets will be accordingly discussed in each section. We report the average performance of the model trained on three different random seeds, except the result on MSCOCO 2014 dataset.

4.1.2 Text Style Transfer Task

For text style transfer task, we conducted our experiments based on XFormal (Briakou et al., 2021) dataset, which encompasses French, Brazilian Portuguese, and Italian. First, we selected 6,000 data for the GYAFC (Rao and Tetreault, 2018) dataset, an English dataset that performs the same text formality style transfer, and translated them into each language using the NLLB (Costa-jussà et al., 2022) model and Google Translator[4] to build a baseline dataset. Second, we built a dataset with only 3,000 English data using our GPT Annotator as it generates two target language data for each English data. Using each dataset built by the Translation model and GPT Annotator respectively, we fine-tuned mBART (Tang et al., 2021) model to perform text style transfer task, and compared its performance and the formality of the generated text. Simliarly to image captioning task, NLG-EVAL library, as well as BERTScore and BARTScore were deployed for measuring metrics. Throughout the manuscript, we report the average performance of the model trained on three different random seeds.

4.2 Cost-Efficiency of GPT Annotator

Based on the concept of a previous study (Wang et al., 2021), we evaluated the difference in the performance of human annotators and GPT annotator under a fixed budget. The previous study (Rashtchian et al., 2010) suggested that it takes

[3] https://codalab.lisn.upsaclay.fr/competitions/7404

[4] https://translate.google.com

Flickr8k	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Human Annotator w/ Limited Budget	28.96	38.76	17.83	0.7817	-18.379
Synonym Replacement	30.30	38.61	17.61	0.7802	-18.457
Back-Translation	30.02	39.02	17.32	0.7795	-18.413
HRQ-VAE	21.62	29.53	15.83	0.7542	-18.641
GPT Annotator w/ GPT-3.5	33.13	39.98	18.41	0.7892	-18.374
Flickr30k	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Human Annotator w/ Limited Budget	25.72	34.14	15.66	0.7539	-18.350
Synonym Replacement	26.78	35.28	15.54	0.7556	-18.329
Back-Translation	27.32	36.70	15.67	0.7591	-18.321
HRQ-VAE	20.94	27.53	12.97	0.7385	-18.542
GPT Annotator w/ GPT-3.5	30.57	37.68	16.02	0.7669	-18.298
MSCOCO 2014	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Human Annotator w/ Limited Budget	40.40	46.60	18.90		
Synonym Replacement	45.10	50.30	23.90		
Back-Translation	41.35	46.70	21.80		
HRQ-VAE	45.59	50.10	24.20		
GPT Annotator w/ GPT-3.5	46.38	50.40	24.50		

Table 1: Experimental results to validate the cost-efficiency of the proposed GPT annotator. We only report BLEU, ROUGE, and METEOR for MSCOCO 2014 dataset as the official evaluation server does not provide BERTScore and BARTScore result.

0.05$ to create five gold captions per image, which is equivalent to 0.01$ for each gold caption. In the experiment, approximately 1000 tokens were used to generate annotated data pair.

According to this cost analysis, the method proposed in this study required 0.012$ to generate one gold caption and four silver captions for each image using GPT-3.5, as it takes approximately 1,000 tokens to generate silver captions.[5] Based on this configuration, it would cost approximately 76.8$ to exploit GPT annotator to annotate the 6,400 images in the Flickr8k train set. In contrast, only 1,500 images can be annotated by purely human annotators under the same fixed budget. Similarly, for Flickr30k dataset, annotating 24,000 train data using the proposed method would cost approximately 288$, whereas for the same amount, human annotators can only annotate 5,800 images to generate five gold captions. Following the same configuration, in the MSCOCO dataset, only 19,680 images can be annotated by human annotators under the budget that can annotate 82,000 images with GPT annotator.

Under this scenario, we compared the results of training the model by selecting only 1,500 fully human-annotated data from Flickr8k dataset, 5,800 fully human-annotated data from Flickr30k dataset, and 19,680 fully human-annotated data from MSCOCO dataset with the results obtained by training the model using the GPT-annotated data for the entire images of each dataset. Additionally, we also exploited other data augmentation baselines such as synonym replacement (Zhang et al., 2015), Back-Translation (Sennrich et al., 2016) and HRQ-VAE (Hosking et al., 2022) to augment one gold data for extensive comparison.

Table 1 shows the results of the experiment. The experimental results suggest that under the same budget, annotating a larger number of images with one gold caption and multiple silver captions resulted in improved performance compared to annotating a smaller number of images with multiple gold captions using only human annotators. This outcome is consistent with the findings of previous work (Wang et al., 2021), indicating the cost efficiency of GPT annotators, and indicates that these characteristics of GPT annotators are applicable to a wider range of tasks including image captioning. Furthermore, GPT annotator has shown superior performance against other augmentation baselines, suggesting that GPT annotator can generate better and diverse sentences.

4.3 Multilingual Experiment

4.3.1 Korean Experiment

Korean is a language that is attracting increasing attention owing to its approximately 80 million native speakers and rising Korean content. Nevertheless, the resource to fulfill this demand is limited (Gu et al., 2018; Sennrich and Zhang, 2019; Kim et al., 2021; Sahoo et al., 2023). For example, there is no dedicated Korean dataset for the image captioning task. Although there are data that applied machine translation to existing English datasets, they are not fully open and have limited availability.[6]

Considering these characteristics of the Korean language, we first evaluated the multilingual ability of the proposed method based on Korean. In this experiment, we assessed the effectiveness of a Korean image captioning model which was trained on two separate datasets: the AiHub dataset, which applies machine translation to the English dataset, and the Korean dataset constructed by GPT-4 using the approach described in this study. Due to the absence of dedicated evaluation set for a fair

[5] As of the time of this study, GPT-3.5 charged 0.002$ per 1000 tokens. Currently, it charges 0.001$ per 1000 tokens of prompt and 0.002$ per 1000 tokens of generation.

[6] https://aihub.or.kr operated by the Korean government offers a machine-translated version of COCO captioning dataset; however, the public usage of this dataset is limited as it is only available to Korean citizens.

Korean	Precision↑	Recall↑	Fluency↓	THUMB↑
AiHub (Machine-Translated)	4.3	4.09	0.03	4.17
GPT Annotator w/ GPT-4	4.72	4.59	0.02	4.64

Table 2: Human evaluation results of the validation of the effectiveness of the proposed GPT annotator on Korean language. We follow the evaluation process and metric of THUMB (Kasai et al., 2022), and report the average THUMB score of three Korean native speakers. Please refer to Appendix C for quantitative analysis.

Vietnamese	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Original (Human-Annotated)	48.62	53.82	32.16	0.8309	-14.511
NLLB (Machine-Translated)	31.76	40.49	26.61	0.8114	-14.645
HRQ-VAE + NLLB	21.26	28.64	23.48	0.7720	-15.342
Google Translator	37.22	46.24	26.86	0.8196	-14.534
GPT Annotator w/ GPT-4	41.32	47.83	30.57	0.8235	-14.537

Table 3: Experimental results in Vietnamese based on UiT-ViIC dataset.

comparison, human evaluation was conducted on 100 captions generated by each model from the test image set. The human evaluation was performed in accordance with the previously proposed protocol (Kasai et al., 2022), and we report the average THUMB score of three Korean native speakers.

Table 2 presents the results of the human evaluation. The outcomes of the evaluation indicate that the model trained on the dataset using GPT annotator performed better than the machine-translated dataset in terms of ratings by humans. In addition, our GPT annotator demonstrated a lower penalty on fluency, which suggests that our method generates more natural sentences.

These evaluation results confirmed that the model can achieve improved performance when trained with the dataset constructed using the method proposed in this study. Furthermore, as our GPT annotator generates five Korean captions using only one gold English caption by a human annotator, it is more cost-efficient compared to applying machine translation to five gold captions in English. Moreover, our GPT annotator has additional advantages that could ensure consistency in sentence structure compared to machine translation. Specifically, we instructed the annotator to generate sentences in the neutral form ("-하다") rather than the polite form ("-합니다") through the prompt. We can maintain consistency in tone and style of the dataset through this configuration, leading to better for the quality of the annotated data and reduce the need for post-processing and human intervention.

4.3.2 Vietnamese Experiment

Vietnamese also has more than 85 million native speakers, but suffering from lack of annotated data (Ngo et al., 2020; Huynh et al., 2022). To demonstrate the versatility of our approach in another language, we expanded our experiments to Vietnamese. For the experiment, we adapted UiT-ViIC dataset (Lam et al., 2020). This dataset consists of images selected from the MSCOCO dataset relating to sports, each with five Vietnamese captions manually annotated by a human annotator. We applied NLLB model and Google Translator to build a baseline by translating English captions from the original MSCOCO dataset into Vietnamese. Additionally, we adopted the data generated by HRQ-VAE in Section 4.2 and translated them into Vietnamese using NLLB model.

Table 3 presents the results on Vietnamese. The experimental result suggests that our approach is valid in Vietnamese, leading to better performance of the model compared to a machine translation-based approach.

4.3.3 Polish Experiment

Polish is another language that has challenge of low-resource language (Dadas et al., 2020; Augustyniak et al., 2022). To further validate our method's applicability, we also conducted experiments on the AIDe dataset for Polish (Wróblewska, 2018). This dataset is composed of 1,000 images selected from the Flickr8k dataset, each with two human-annotated captions in Polish. For this experiment, we configured our prompt to generate two caption pairs for each image. Similarly to Vietnamese experiment, for the Polish translation baseline, we utilized the NLLB model and Google Translator to translate two English captions from the original Flickr8k dataset into Polish. We also adopted the data generated by HRQ-VAE in Section 4.2 and translated them into Polish using NLLB model.

Table 4 indicates the results on Polish. The experimental result demonstrates the effectiveness of our approach, showcasing not just better performance compared to translation baseline but also comparable performance to human-annotated data.

Polish	BLEU	ROUGE	METEOR	BERTS.	BARTS.
Original (Human-Annotated)	8.68	19.38	9.38	0.7405	-18.162
NLLB (Machine-Translated)	4.14	14.46	6.78	0.6466	-18.279
HRQ-VAE + NLLB	3.21	13.15	5.99	0.6495	-18.331
Google Translator	4.64	14.14	6.91	0.6507	-18.244
GPT Annotator w/ GPT-4	5.17	18.90	8.92	0.6962	-18.197

Table 4: Experimental results in Polish based on AIDe dataset.

French	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	48.59	50.26	31.42	0.8103	-17.596	72.37
Google Translator	51.69	54.02	32.62	0.8076	-17.541	75.38
GPT Annotator w/ GPT-4	54.81	56.83	33.98	0.8175	-17.519	**85.12**
Brazilian Portuguese	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	52.73	55.81	32.44	0.8286	-18.955	68.58
Google Translator	55.98	57.74	34.19	0.8318	-18.938	74.27
GPT Annotator w/ GPT-4	57.94	60.72	35.60	0.8363	-18.864	**79.21**
Italian	BLEU	ROUGE	METEOR	BERTS.	BARTS.	Formality
NLLB (Machine-Translated)	47.97	49.34	30.12	0.7839	-18.843	68.03
Google Translator	49.13	51.73	30.89	0.7873	-18.805	71.86
GPT Annotator w/ GPT-4	52.34	53.71	32.02	0.7994	-18.702	**74.29**

Table 5: Experimental results on text style transfer in French, Brazilian Portuguese, and Italian.

4.4 Text Style Transfer Experiment

Table 5 presents the experimental result of our GPT annotator for text style transfer task in French, Brazilian Portuguese, and Italian. The results not only highlight the performance of our GPT Annotator with fewer original human-annotated samples but also underscore its ability to enhance text formality against translation. This achievement was possible through the consistent generation of sentences with formal and informal styles, owing to the flexibility of LLMs and instructible prompts.

4.5 Employing GPT Annotator for Dataset Construction

Latvian, Estonian, and Finnish have approximately 1.5, 1.1, and 4.8 million native speakers, which make them hard to hire annotators and construct datasets. To address the practical challenges in the field of data annotation, we constructed an image captioning dataset in these languages, which did not have any image captioning dataset, using our GPT annotator. We first selected 3,850 images and their English captions from the MSCOCO dataset and splited them into 2,695 train images, 924 validation images, and 231 test images, following the configuration of the Vietnamese UiT-ViIC dataset.

To build a baseline, we utilized NLLB and Google Translator to translate the English caption

Latvian	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	6.39	17.53	10.13	0.6803	-16.061
HRQ-VAE + NLLB	5.14	16.61	10.21	0.6728	-16.127
Google Translator	8.53	17.09	10.67	0.6848	-16.067
GPT Annotator w/ GPT-4	10.35	18.61	10.79	0.6911	-16.054
Estonian	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	4.97	13.12	7.89	0.6893	-15.409
HRQ-VAE + NLLB	3.37	7.84	5.87	0.6876	-15.409
Google Translator	6.04	12.51	8.75	0.7008	-15.408
GPT Annotator w/ GPT-4	6.62	13.47	9.22	0.7050	-15.407
Finnish	BLEU	ROUGE	METEOR	BERTS.	BARTS.
NLLB (Machine-Translated)	4.19	10.43	7.74	0.7122	-16.392
HRQ-VAE + NLLB	3.74	10.23	7.06	0.6965	-16.401
Google Translator	4.28	10.84	7.88	0.7128	-16.394
GPT Annotator w/ GPT-4	4.96	12.29	8.64	0.7143	-16.389

Table 6: Experimental results of our constructed dataset in Latvian.

of each training image, similar to previous experiments. The validation and test captions were constructed by translating using mBART model, for a fair comparison.

Table 6 clearly showcases the efficiency of our GPT annotator when human-annotated data is scarce, as observed in case of these low-resource languages. The human investigation of annotated data remains for future work. We plan to release the training, validation, and testing datasets for wider access and further study. This experimental result demonstrates the possibility of the GPT annotator to easily construct dataset in any designated language, enhancing the accessibility of various languages.

5 Conclusion

In this study, we have demonstrated the possibility of exploiting LLM as a multilingual assistant annotator by generating multiple silver data from a single gold data in different languages. The experimental results showcased that the proposed method is cost-efficient compared to entirely human annotation, and can be effectively employed to construct datasets in various languages and tasks.

The approach described in this work can be widely adapted to various languages, as it utilizes the multilingual fluency and flexibility of LLMs. We constructed an image captioning in Latvian as a practical application of our GPT annotator. Furthermore, the cost-efficiency of the GPT annotator suggested in this paper will be improved in the future, as the price of LLMs is expected to decline as recent cost reductions of GPT-3.5 and GPT-4 have

shown. Future studies will focus on improving the proposed method by utilizing the image inception ability and expanding this method to other tasks.

Limitations

Extreme low-resource languages may still encounter difficulty producing high-quality sentences even with the use of GPT-4. To examine the responses of GPT-4 in translating into extremely low-resource languages, we conducted an error analysis in two extremely low-resource languages, Basque and Māori. Basque has a small amount of speakers, and it is also a unique language isolate, that does not have a distinct relationship with other languages such as Spanish and French, making it harder to process. Māori has a very small amount of language users, posing a challenge as an extremely low-resource language. Please refer to Appendix E.7 for the analysis result.

Additionally, the approach demonstrated in this work generates silver sentences as paraphrases of the given gold sentences, thus they might not fully capture the information that exists in the image but is not mentioned in the gold sentences. Consequently, the gold captions produced by multiple human annotators can be more diverse than silver captions. To address this issue, human annotators could create gold captions that contain as much detailed and diverse information as possible while constructing a new dataset through this method.

Ethics Statement

As this work proposes the utilization of LLMs as an assistant data annotator and for the automatic generation of sentences, it may suffer from the potential bias of LLMs. To mitigate this concern, we added explicit instructions to prevent the generation of biased sentences in the prompts. However, the human supervisor is still essential to examine and validate the absence of biased expressions in the generated data. Specifically, the human supervisor should ensure that there is not any biased gold sentence produced by the human annotator, as it directly affects the bias of generated sentences using LLMs.

Furthermore, in addition to the error analysis presented in the previous section, we have conducted supplementary error analysis on Basque and Māori languages in Appendix E.8. This additional investigation aims to explore the potential ethical biases exhibited by GPT-4. Our findings suggest that GPT-4 may exhibit unexpected ethical biases, particularly in extremely low-resource languages, where its knowledge about the language may be limited compared to high-resource languages such as English.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

References

Lukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Arkadiusz Janz, Piotr Szymański, Marcin Wątroba, Mikołaj Morzy, et al. 2022. This is the way: designing and compiling lepiszcze, a comprehensive nlp benchmark for polish. *Advances in Neural Information Processing Systems*, 35:21805–21818.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3199–3216.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Evaluation of sentence representations in Polish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1674–1680.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Meryem Guemimi, Daniel Camâra, and Ray Genoe. 2021. Iterative learning for semi-automatic annotation using user feedback. In *International Conference on Intelligent Technologies and Applications*, pages 31–44.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2489–2501.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478.

Bosung Kim, Juae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Commonsense knowledge augmentation for low-resource languages via adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6393–6401.

Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. *arXiv preprint arXiv:2305.14556*.

Quan Hoang Lam, Quang Duy Le, Van Kiet Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Uit-viic: A dataset for the first evaluation on vietnamese image captioning. In *International Conference on Computational Collective Intelligence*, pages 730–742.

Md Tahmid Rahman Laskar, Mizanur Rahman, Israt Jahan, Enamul Hoque, and Jimmy Huang. 2023. Can large language models fix data annotation errors? an empirical study using debatepedia for query-focused text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10245–10255.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505.

Xinjian Li, Zhong Zhou, Siddharth Dalmia, Alan W Black, and Florian Metze. 2019. Santlr: Speech annotation toolkit for low resource languages. In *Proceedings of Interspeech 2019*, pages 3681–3682.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755.

Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

TorchVision maintainers and contributors. 2016. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`.

Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen. 2020. Improving multilingual neural machine translation for low-resource languages: French, English - Vietnamese. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 55–61.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *Preprint*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147.

Nihar Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13316–13330.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Md Afnan Ul Haque, Ashiqur Rahman, and M. M. A Hashem. 2021. Sentiment analysis in low-resource bangla text using active learning. In *5th International Conference on Electrical Information and Communication Technology*, pages 1–6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Alina Wróblewska. 2018. Polish corpus of annotated descriptions of images. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, and Reihaneh Rabbany. 2023. Open, closed, or small language models for text classification? *arXiv preprint arXiv:2308.10092*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Implementation Details

A.1 Model Implementation

PyTorch (Paszke et al., 2019) and Huggingface Transformers library (Wolf et al., 2020) have been employed for the implementation process.

For image captioning task, Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Transformer (Vaswani et al., 2017) were deployed as the encoder and decoder of the model, respectively. Particularly, pretrained *vit_b_16* from torchvision library (maintainers and contributors, 2016) was adapted as an encoder, and the decoder consisted of 12 heads and

12 layers, with a hidden layer size and embedding layer size of 768.

For text style transfer task, we fine-tuned *mbart-50-large* model using each dataset to convert informal text into formal text. Additionally, we separately trained another mBART model as formality classifier using XFormal training data for each language to measure the formality of the generated text. The text formality was measured by the average logit of the classifier.

Every model was trained using AdamW (Loshchilov and Hutter, 2018) with a batch size of 16 and a learning rate of 5e-5 through 10 epochs, while the weight decay of the optimizer was set to 1e-5, and a CosineAnnealingLR (Loshchilov and Hutter, 2017) scheduler was deployed.

A.2 GPT Annotator Implementation

We utilized the official API from OpenAI to implement the proposed GPT annotator. The versions of the models used are *gpt-3.5-turbo-0301* and *gpt-4-0314*, respectively. The prompts used can be found in Appendix F. If an error occurred while generating an annotation using a given prompt, the API was called again with a patience of three times. If this patience was exceeded, the data pair was excluded from the annotation process.

A.3 Further Details

We employed the *facebook/nllb-200-distilled-600M* model, which comprises 600M parameters, to create a training dataset using the NLLB baseline. Similarly, we utilized the *facebook/mbart-large-50-many-to-many-mmt* model, with approximately 611M parameters, to construct validation and test sets for Latvian, Estonian, and Finnish. This choice was made to ensure a fair and equitable comparison between the baseline models and our proposed GPT annotator. For evaluation with BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), we exploited *bert-base-multilingual-cased* and *facebook/mbart-large-50*, respectively. Note that we reported BERTScore-F1 in the manuscript.

Label smoothing (Szegedy et al., 2016) was applied with a smoothing epsilon of 0.05. The training procedure was conducted on a single Nvidia RTX 3090 GPU.

For the tokenizing of text input, we employed tokenizer of pre-trained model available on Huggingface for each language. Specifically, *facebook/bart-base*, *cosmoquester/bart-ko-base*, *vinai/bartpho-syllable*, *sdadas/polish-bart-base*, and *joelito/legal-latvian-roberta-base*, *tartuNLP/EstBERT*, *TurkuNLP/bert-base-finnish-uncased-v1* were adapted as the tokenizer for English, Korean, Vietnamese, Polish, Latvian, Estonian, and Finnish. For text style transfer task, as it is based on *facebook/mbart-large-50* model, each language shares same tokenizer.

For the test procedure of the Flickr8k and Flickr30k datasets, all five available human-annotated captions of the test set were utilized as reference sentences for evaluation. Beam search (Freitag and Al-Onaizan, 2017) was applied as a decoding strategy to generate sentences at inference time, with a beam size of 5.

B GPT Annotator Software

In order to streamline the annotation process outlined in this paper, we have developed specialized software tailored for multilingual data annotation, leveraging OpenAI GPT models. This software currently supports tasks such as image captioning, text style transfer, and machine translation. Although these functionalities are not discussed in detail in this paper due to space constraints, they are available within the software.

The annotator software takes a JSON file as input and generates a new JSON file containing multilingual annotations in the target language. This is achieved by utilizing the specified prompt and the chosen version of the GPT model. Moreover, the software is designed to facilitate faster data annotation through multiprocessing capabilities. For a more comprehensive understanding of the software's functionality, please refer to the attached code.

C Quantitative Experiments on Korean

We have included the human evaluation results in Table 2 within the main manuscript. This was done because there is no dedicated evaluation set available in Korean, which is essential for a fair evaluation. In this section, we present additional quantitative evaluation results to provide a more comprehensive perspective on our model's performance.

To conduct this quantitative evaluation, we utilized the validation set from the AiHub dataset since there is no specific test set available in Korean within the official COCO dataset. In addition to this evaluation, we also translated the model's inferences on the test image set into English. This

Evaluation Method	Validation Set (Korean)			Test Set (Translated to English)		
Metric	BLEU	ROUGE	METEOR	BLEU	ROUGE	METEOR
AiHub (Machine-Translated)	11.20	20.64	19.41	34.85	41.60	19.80
GPT Annotator w/ GPT-4	7.01	15.84	18.56	32.70	39.90	19.20

Table 7: Quantitative experimental results of the machine-translated dataset and proposed GPT annotator on Korean language. The left column ('Validation Set') refers to the inference result of the validation set provided in Korean. The right column ('Test Set') is the evaluation result of the Korean model, but as there is no Korean data for the test set, we translated the Korean inference result into English and uploaded it to the official evaluation server.

Metric	Precision	Recall	Fluency	THUMB
Human #1	4.61	4.26	0.01	4.43
Human #2	4.3	4.21	0.05	4.21
Human #3	4.62	4.56	0.01	4.58

Table 8: For transparency of human evaluation, we report the average value of each metric as rated by three raters.

allowed us to assess the model's performance on the test set using the official evaluation server. The quantitative analysis results are presented in Table 7.

However, it is important to note that while quantitative analysis is relatively straightforward to perform, it may not provide an accurate measure of the Korean model's performance. The AiHub dataset's validation set relies on machine translation, which may be too coarse to gauge the model's capabilities precisely. Similarly, assessing the quality of a generated Korean sentence by translating it into English is not a direct evaluation method. This is the primary rationale for conducting a human evaluation, which offers a more robust assessment of the model's performance.

D Detailed Information on Human Evaluation

Human raters were recruited from volunteered students who are native in Korean. Three raters are native Korean speakers in their 20s who majored in engineering. The detailed information about THUMB score (Kasai et al., 2022), the metric used in this study for the assessment of the generated caption, was provided to raters. After the explanation of the metric, process, and purpose of the study, raters were asked to evaluate the precision, recall, and fluency penalty that composes THUMB score. Figure 3 is a screenshot as an example of the human evaluation form. To prevent rater fatigue, We instructed them to pause the evaluation process if they felt exhausted and not to finish it all at once. 100 images for evaluation were randomly selected from the generated output by each model from the COCO2014 test image set. Table 8 shows the average evaluation result of each rater.

E Case Analysis

To evaluate the excellence and contextual precision of the produced captions, we conducted a direct comparison between captions originating from each dataset for identical images. This assessment unveiled significant enhancements in both caption quality and contextual alignment within our recently generated dataset compared to the baselines.

E.1 Korean Analysis

- **Quality of Generated Sentence**
  - MSCOCO Image ID: `237944`
    - English Reference:
      A person under a dryer wearing a towel
    - AiHub (Machine-Translated):
      드레이더 (*Drader* - This word does not exist in Korean.)
    - GPT Annotator w/ GPT-4:
      수건을 두른 사람이 드라이어 아래에 있다. (*A person with a towel is under the dryer.*)
  - MSCOCO Image ID: `215878`
    - English Reference:
      A white microwave oven a pot holder and some books
    - AiHub (Machine-Translated):
      하얀 전자 레인지에 냄비 뚜껑과 책 몇권을 넣어 (*Put a pot lid and some books in a white microwave*)
    - GPT Annotator w/ GPT-4:
      하얀 전자레인지 오븐, 냄비 받침이랑 몇 권의 책들이 있다. (*There is a white microwave oven, pot holders, and some books.*)

- **Context of Generated Sentence**
  - MSCOCO Image ID: `190556`
    - English Reference:
      Close up images of bikes parked next to the highway.
    - AiHub (Machine-Translated):
      고속 도로 옆에 주차된 자전거의 이미지를 닫아라. (*Close the image of a bicycle parked on the side of the high way.*)
    - GPT Annotator w/ GPT-4:
      고속도로 옆에 주차된 자전거의 근접한 이미지들이다. (*Close-up images of a bicycle parked on the side of the highway.*)
  - MSCOCO Image ID: `273929`
    - English Reference:
      A far away shot of Big Ben and the nearby complex.
    - AiHub (Machine-Translated):
      멀리서 빅 벤과 인근 콤플렉스를 총으로 쐈어요 (*I shot Big Ben and the nearby complex from a distance with a gun*)
    - GPT Annotator w/ GPT-4:
      빅 벤과 인근 건물들을 멀리서 찍은 사진이다. (*This is a photo of Big Ben and nearby buildings from a distance.*)

E.2 Vietnamese Analysis

- **Quality of Generated Sentence**
  - MSCOCO Image ID: `213669`

* English Reference:
  A young man holding a tennis racquet on a tennis court.
* Vietnamese Reference:
  Người đàn ông đang cầm vợt tennis chạy tới đánh bóng. (*A man holding a tennis racket runs to hit the ball.*)
* NLLB (Machine-Translated):
  một người đàn ông đứng trên một thức ăn với một tên lửa (*a man standing on a food with a rocket*)
* GPT Annotator w/ GPT-4:
  Một người trẻ tuổi đang ở trên sân tennis với cây vợt trong tay. (*A young person is on the tennis court with a racket in his hand.*)

E.3 Polish Analysis

- **Context of Generated Sentence**
  - Flickr File Name:
    `1153704539_542f7aa3a5`
    * English Reference:
      A girl playing trumpet in a marching band.
    * Polish Reference:
      Dziewczyna w sportowym stroju i czapce z daszkiem stoi na trawniku i gra na trąbce w towarzystwie innych muzyków. (*A girl in sports clothes and a baseball cap stands on the lawn and plays the trumpet in the company of other musicians.*)
    * NLLB (Machine-Translated):
      Dziewczyna grająca na trąbce w zespole. (*A girl playing the trumpet in a band.*)
    * GPT Annotator w/ GPT-4:
      Dziewczyna grająca na trąbce w orkiestrze marszowej. (*A girl playing the trumpet in the march orchestra.*)

- **Quality of Generated Sentence**
  - Flickr File Name:
    `1386251841_5f384a0fea`
    * English Reference:
      A woman is looking at dressed, headless mannequins in a store display.
    * Polish Reference:
      Kobieta ogląda wystawę z ubranymi w damskie stroje manekinami. (*A woman looks at an exhibition with mannequins dressed in women's clothes.*)
    * NLLB (Machine-Translated):
      Kobieta patrzy na ubrane, bezgłowe manieki w sklepach. (*A woman looks at clothed, headless maniacs in stores.*)
    * GPT Annotator w/ GPT-4:
      Kobieta patrzy na ubrane, bezgłowe manekiny w wystawie sklepowej. (*A woman looks at clothed, headless mannequins in a store window.*)
  - Flickr File Name:
    `1387785218_cee67735f5`
    * English Reference:
      A child pushes a doll in a baby carriage.
    * Polish Reference:
      Dziecko idzie drogą, prowadząc przed sobą mały wózek z lalką. (*A child walks along the road, pushing a small stroller with a doll in front of him.*)

    - NLLB (Machine-Translated):
      Mała dziewczynka wpychająca koło lalki (*A little girl pushing a doll wheel*)
    - GPT Annotator w/ GPT-4:
      Dziecko popycha lalkę w wózku dla dzieci. (*A child pushes a doll in a baby stroller.*)
  - Flickr File Name:
    `1499495021_d295ce577c`
    - English Reference:
      A dark haired woman wearing a brown jacket and fatigue bottoms and a balding man wearing a green sweater and blue jeans with a fishing pole, stand at the foot of the surf.
    - Polish Reference:
      Mężczyzna stoi z wędką nad brzegiem wody, a obok niego stoi kobieta. (*A man stands with a fishing rod at the water's edge and a woman stands next to him.*)
    - NLLB (Machine-Translated):
      Czarnowłosa kobieta w brązowej kurtce i zmęczonym dnie i łysy mężczyzna w zielonej swetrze i niebieskich dżinsów z palcem rybaczem, stoją u stóp pływu. (*A black-haired woman in a brown jacket and a tired bottom and a bald man in a green sweater and blue jeans with a fisherman's finger, stand at the foot of the tide.*)
    - GPT Annotator w/ GPT-4:
      Kobieta o ciemnych włosach, ubrana w brązową kurtkę i spodnie w kamuflaż, oraz łysiejący mężczyzna w zielonym swetrze i niebieskich dżinsach z wędką, stoją u podnóża fali. (*A woman with dark hair, wearing a brown jacket and camouflage pants, and a balding man in a green sweater and blue jeans with a fishing rod, stand at the foot of a wave.*)
  - Flickr File Name:
    `146098876_0d99d7fb98`
    - English Reference:
      A boy and three girls in blue school uniforms walk down a dirt-covered road.
    - Polish Reference:
      Chłopiec i trzy dziewczynki w mundurkach idą, niosąc zeszyty. (*A boy and three girls in uniforms are walking, carrying notebooks.*)
    - NLLB (Machine-Translated):
      Chłopak i trzy dziewczyny w niebieskich mundurach szli po błędnej drodze. (*A boy and three girls in blue uniforms were walking on the wrong path.*)
    - GPT Annotator w/ GPT-4:
      Chłopiec i trzy dziewczyny w niebieskich mundurkach szkolnych idą po drodze pokrytej brudem. (*A boy and three girls in blue school uniforms are walking on a road covered with dirt.*)

E.4 Latvian Analysis

- **Quality of Generated Sentence**
  - MSCOCO Image ID: `46544`
    - English Reference:
      A woman playing tennis on a tennis court.
    - NLLB (Machine-Translated):
      Sieva tenisā tenisā. (*Tennis wife in tennis.*)
    - GPT Annotator w/ GPT-4:
      Sieviete spēlē tenisu tenisa kortā. (*A woman plays tennis on a tennis court.*)
  - MSCOCO Image ID: `43960`
    - English Reference:
      A boy catching a ball while another boy holds a bat.

    - NLLB (Machine-Translated):
      Puikas, kas ieņem lopu, kamēr cits puikas, kas drīkst pieņemt lopu. (*Boys who take livestock, while other boys who are allowed to accept livestock.*)
    - GPT Annotator w/ GPT-4:
      Zēns noķer balls, kamēr cits zēns tur nūju. (*A boy catches the ball while another boy holds the stick.*)

  - MSCOCO Image ID: `47813`
    - English Reference:
      There are four people playing tennis in doubles.
    - NLLB (Machine-Translated):
      Divās grupās spēlē četri cilvēki. (*Four people play in two groups.*)
    - GPT Annotator w/ GPT-4:
      Četri cilvēki spēlē tenisu dubultspēlēs. (*Four people play tennis in doubles.*)

E.5 Estonian Analysis

- **Quality of Generated Sentence**
  - MSCOCO Image ID: `1596`
    - English Reference:
      A person swing a tennis racket at a tennis ball.
    - NLLB (Machine-Translated):
      Üks inimene käigub tennisepalli peal tennise racket. (*One person moves a tennis racket on top of a tennis ball.*)
    - GPT Annotator w/ GPT-4:
      Inimene lööb tennise reketiga tennisepalli. (*A person hits a tennis ball with a tennis racket.*)
  - MSCOCO Image ID: `35818`
    - English Reference:
      A group of boys play soccer in a grassy field.
    - NLLB (Machine-Translated):
      Grupp poisid mängib jalgpalli mägedes. (*A group of boys plays football in the mountains.*)
    - GPT Annotator w/ GPT-4:
      Poiste grupp mängib jalgpalli rohusel väljakul. (*A group of boys plays football on a green field.*)
  - MSCOCO Image ID: `65500`
    - English Reference:
      Two sets of people are at a tennis net.
    - NLLB (Machine-Translated):
      Kaks inimest on tennistöö juures. (*Two people are at tennis work.*)
    - GPT Annotator w/ GPT-4:
      Kaks inimeste rühma on tennisevõrgu juures. (*Two groups of people are at the tennis net.*)

E.6 Finnish Analysis

- **Quality of Generated Sentence**
  - MSCOCO Image ID: `217929`
    - English Reference:
      people in uniforms playing baseball in the field
    - NLLB (Machine-Translated):
      joukkueessa pelaavat joukkueessa (*in the team play in the team*)

    - GPT Annotator w/ GPT-4:
      Ihmiset uniformuissa pelaavat baseballia kentällä. (*People in uniforms are playing baseball on the field.*)

  - MSCOCO Image ID: `226747`
    - English Reference:
      a persong swinging a tennis racket hitting a tennis ball
    - NLLB (Machine-Translated):
      laulaja, joka heiluttaa tenniskäytä, joka lyö tenniskappiin (*the singer who swings the tennis racket, who hits the tennis locker*)
    - GPT Annotator w/ GPT-4:
      Henkilö heiluttaa tennis-mailaa osuen tennispalloon. (*A person swings a tennis racket hitting a tennis ball.*)
    - Comment:
      In this example, we found that the original reference for input to NLLB and GPT annotator has a typo ("persong" instead of "person"), which leads to quality issues in the sentence generated through NLLB model. Whereas, our GPT annotator automatically corrected this typo through the generation process. This behavior underscores the additional advantage of our GPT annotator compared to the baseline.

E.7 Error Analysis

- **Error Analysis on Basque**

  - Case 1
    - Source Text:
      A black dog and a spotted dog are fighting.
    - Google Translator:
      Txakur beltz bat eta txakur orban bat borrokan ari dira
    - GPT Annotator w/ GPT-4:
      Kolore beltzeko txakur bat eta beste bat orbainekin borrokan ari dira. (*A black dog and another with scars are fighting.*)
    - Comment:
      The word "spot" has changed to "scar."

  - Case 2
    - Source Text:
      Several women wait outside in a city.
    - Google Translator:
      Hainbat emakume kanpoan itxaroten hiri batean.
    - GPT Annotator w/ GPT-4:
      Emakume talde bat hiri bateko inguru batean zutik daude. (*A group of women are standing in an area of a city.*)
    - Comment:
      The word "wait" has changed to "standing", which does not fully contain the original meaning.

  - Case 3
    - Source Text:
      A man clad in a safety vest and helmet is standing in the street with a flag.
    - Google Translator:
      Segurtasun txalekoa eta kaskoa jantzita dagoen gizon bat kalean dago bandera batekin.

* GPT Annotator w/ GPT-4:
  Segurtasun gerrikoa eta kaskoa jantzita, gizon bat bandera batekin dago kalean. (*A man wearing a seat belt and a helmet stands on the street with a flag.*)
* Comment:
  The word "safety best" has changed to "seat belt."

- **Error Analysis on Māori**
  - Case 1
    * Source Text:
      A man in green holds a guitar while the other man observes his shirt.
    * Google Translator:
      Ko tetahi tangata he kakariki e mau ana ki te kita, ko tetahi atu tangata e titiro ana ki tana koti.
    * GPT Annotator w/ GPT-4:
      E pupuri ana te tangata i te kita, ko te tangata ke atu e matakitaki ana i tana hāte. (*One is holding a guitar while the other is looking at his shirt.*)
    * Comment:
      The sentence has lost the expression about the color of the clothes the man wearing.
  - Case 2
    * Source Text:
      Boys perform dances on poles during the nighttime.
    * Google Translator:
      Ka kanikani nga tama ki runga pou i te po.
    * GPT Annotator w/ GPT-4:
      Tamariki tāne e mahi ake ana i ngā pou i te po tuturu. (*Boys who work up posts in the real night.*)
    * Comment:
      The word "perform dance" has changed to "work up."
  - Case 3
    * Source Text:
      A ballet class of five girls jumping in sequence.
    * Google Translator:
      He karaehe paoro o nga kotiro tokorima e peke ana i te raupapa.
    * GPT Annotator w/ GPT-4:
      He akomanga parekareka o rima kōtiro e peke tahi ana i roto i te raupapa. (*It's a fun class of five girls jumping together in a series.*)
    * Comment:
      The generated sentence has lost the word "ballet" and changed it to "fun", implying the model has less understanding of proper nouns.

E.8 Ethical Bias Analysis

- **Error Analysis on Ethical Bias**
  - Basque
    * Source Text:
      Two young, White males are outside near many bushes.
    * Google Translator:
      Bi ar zuri kanpoan daude zuhaixka askoren ondoan.
    * GPT Annotator w/ GPT-4:
      Gazte kaukasiko gizon pare bat kanpoan daude, zuhaixka ugariz inguratuta. (*A couple of young Caucasian men are outside, surrounded by many bushes.*)

    - Comment:
      The output from GPT annotator contains the word “Caucasian”, which is a word based on obsolete racial classification.

- Māori
    - Source Text:
      Two young, White males are outside near many bushes.
    - Google Translator:
      Tokorua nga tama tane ma, kei waho tata ki nga ngahere maha.
    - GPT Annotator w/ GPT-4:
      E rua ngā tāne rangatahi, Pākehā, kei waho e tū tata ana ki ngā pūrei. (*Two young European men are standing outside near the playground.*)
    - Comment:
      The output from GPT annotator contains the word “European” instead of “White”, which can be interpreted as a geographical bias.

image_id: 69
Sentence 1: 한 무리의 사람들이 테이블에 둘러앉아 있다.
Sentence 2: 몇몇 사람들이 테이블에 앉아 있다.

	1	2	3	4	5
Sentence 1: Pr...	○	○	○	○	○
Sentence 1: Re...	○	○	○	○	○
Sentence 1: Flu...	○	○	○	○	○
Sentence 2: Pr...	○	○	○	○	○
Sentence 2: Re...	○	○	○	○	○
Sentence 2: Flu...	○	○	○	○	○

Figure 3: The screenshot of human evaluation form. Sentence 1 is the output from the model trained by AiHub dataset, and Sentence 2 is the output from the model trained by the dataset constructed by our GPT annotator.

F Prompt

This section describes the prompt used for the experiment.

F.1 Prompt for Image Captioning Task

`System`

You are a helpful assistant.
User will ask you to generate paraphrases of a sentence.
You will generate paraphrases of the sentence and its translation in Korean language.
VERY IMPORTANT: You must speak '-하다' form in Korean. You must not use '-합니다' or other forms. 한국어 문장을 번역하여 생성할 때, 반드시 '-하다' 체를 사용하여야 한다. '-합니다', '-입니다' 등의 표현을 절대 사용하지 않는다.
You will generate a translation of input sentence in Korean, and also generate 4 paraphrases and its translation in Korean.
Output sentence should be neutral expression. You should not generate phrases like 'You will see' or 'You will find'.
Output sentence will be complete, natural and fluent.
Each output sentence should have different expressions as much as possible.
You will not generate the same sentence as the input sentence.
You must not generate any biased, offensive, or inappropriate paraphrases.
User input example: The men at bat readies to swing at the pitch while the umpire looks on.
Your output example:
Translation: 타석에 있는 남자들이 심판이 지켜보는 동안 스윙할 준비를 한다.
Paraphrase 1: The male players at the bat ready to hit the ball as the umpire watches attentively. / 심판이 주의 깊게 지켜보는 가운데 배트를 든 남자 선수들이 공을 칠 준비를 하고 있다.
Paraphrase 2: The male batters at the bat prepare to hit the pitch as the umpire stands watch. / 타석에 선 남성 타자들이 심판이 지켜보는 가운데 타구를 칠 준비를 하고 있다.
Paraphrase 3: The batters at the plate are poised to swing as the umpire keeps an eye on them. / 타석에 있는 타자가 심판이 지켜보는 가운데 스윙할 자세를 취한다.
Paraphrase 4: The hitters at the plate wait for themselves to take their swings at the ball while the umpire looks on. / 타석에 선 타자들은 심판이 지켜보는 동안 공을 향해 스윙할 준비를 한다.
You will not say 'Sure! here's the output' or any similar phrases.
You will not say 'I don't know' or any similar phrases.
You will just generate the output paraphrases following the output example.

`User`

Input: Living room with furniture with garage door at one end.

F.2 Prompt for Text Style Transfer Task

System
You are a helpful assistant. You are fluent in French and English.
You will generate paraphrases of formal and informal sentences and their translations into French.
Output sentence should be neutral expression.
Output sentence will be complete, natural and fluent.
Each output sentence should have different expressions as much as possible.
You will not generate the same sentence as the input sentence.
You must not generate any biased, offensive, or inappropriate paraphrases.
You will not say 'Sure! here's the output' or any similar phrases.
You will not say 'I don't know' or any similar phrases.
You will just generate the output paraphrases following the output example.
[Input Sentence]
Formal 1: Then kiss her, brother; that works every time.
Informal 1: Then kiss her;) works every time bro!!!!
[Paraphrase]
Formal 2: Subsequently, kiss her, sibling; that method proves effective on each occasion.
Informal 2: So, just give her a smooch, bro! It seriously works every single time ;)
[Translation in French]
Formal 1: Alors embrasse-la, mon frère. Cela fonctionne à chaque fois.
Informal 1: Alors embrasse-la ;) ça marche à chaque fois frérot!!!!
Formal 2: Ensuite, embrasse-la, frère ; cette méthode fonctionne à chaque fois.
Informal 2: Alors, donne-lui un bisou, mec ! Ça marche à tous les coups ;)

User
[Input Sentence]
Formal 1: After that I never bought her another gift.
Informal 1: and enver since then i never bought her another gift