# Is GPT-3 a Good Data Annotator?

**Bosheng Ding**[*1,2] **Chengwei Qin**[*1] **Linlin Liu**[†1,2]
**Yew Ken Chia**[2] **Boyang Li**[1] **Shafiq Joty**[1] **Lidong Bing**[‡2]
[1]Nanyang Technological University, Singapore [2]DAMO Academy, Alibaba Group
{bosheng001, chengwei003, linlin001, boyang.li, srjoty}@ntu.edu.sg
{bosheng.ding, yewken.chia, l.bing}@alibaba-inc.com

## Abstract

Data annotation is the process of labeling data that could be used to train machine learning models. Having high-quality annotation is crucial, as it allows the model to learn the relationship between the input data and the desired output. GPT-3, a large-scale language model developed by OpenAI, has demonstrated impressive zero- and few-shot performance on a wide range of NLP tasks. It is therefore natural to wonder whether it can be used to effectively annotate data for NLP tasks. In this paper, we evaluate the performance of GPT-3 as a data annotator by comparing it with traditional data annotation methods and analyzing its output on a range of tasks. Through this analysis, we aim to provide insight into the potential of GPT-3 as a general-purpose data annotator in NLP [1].

## 1 Introduction

The democratization of artificial intelligence (AI) (Garvey, 2018; Rubeis et al., 2022) aims to provide access to AI technologies to all members of society, including individuals, small- and medium-sized enterprises (SMEs), academic research labs, and nonprofit organizations. Achieving this goal is crucial for the promotion of innovation, economic growth, and fairness and equality. As typical AI models are usually data-hungry, one significant obstacle of AI democratization is the preparation of well-annotated data for training AI models.

Specifically, supervised learning critically depends on sufficient training data with accurate annotation, but data annotation can be a costly endeavor, particularly for small-scale companies and organizations (Bunte et al., 2021). The cost of data annotation typically includes the labor costs associated with the labeling process, as well as the time and resources required to hire, train and manage annotators. Additionally, there may be costs associated with the annotation tools and infrastructure needed to support the annotation process. Individuals or small-scale organizations may not have resources to annotate sufficient training data, thereby are unable to reap the benefits of contemporary AI technologies. Although the development of pre-trained language models such as BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) eases the data-hungry issue to some extent, data annotation remains an unavoidable challenge for supervised model training.

GPT-3 (Brown et al., 2020; Ouyang et al., 2022)[2] is a powerful large language model developed by OpenAI. Evaluations show that GPT-3 has gained through pretraining a surprisingly wide range of knowledge, which can be transferred to downstream tasks through knowledge distillation (Kim et al., 2022). We present some examples in Appendix A.12. Due to the model architecture and pretraining tasks designed for auto-regressive generation, GPT-3 is capable of generating human-like text and performing a broad array of NLP tasks, such as machine translation, summarization, and question-answering. However, the direct use of GPT-3 for inference in a production setting remains challenging due to its size and computational requirements. Moreover, such large language models often lack the flexibility of local deployment, since their parameters are usually not publicly available. In contrast, it is often more feasible to use smaller language model models, such as $\mathrm{BERT}_{\mathrm{BASE}}$ (Devlin et al., 2019), in production environments.

In this paper, we investigate the ability of GPT-3 to annotate training data for training machine learn-

[*] Equal contribution, order decided by coin flip. Bosheng Ding are under the Joint PhD Program between Alibaba and Nanyang Technological University.

[†] Work done while under the Joint PhD Program between Alibaba and Nanyang Technological University.

[‡] Correspondent Author.

[1]Our code is available at https://github.com/DAMO-NLP-SG/LLM-Data-Annotator.

[2]For brevity, we refer to both the original GPT-3 and InstructGPT as GPT-3.



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 11173–11195
July 9-14, 2023 

ing models, which can substantially lower the annotation cost and level the playing field for individuals or small organizations, so that they can harness the power of AI in their own missions. The process can be considered as distilling the knowledge of GPT-3 to small networks that can be straightforwardly deployed in production environments.

We conduct extensive experiments to evaluate the performance, time, and cost-effectiveness of 3 different GPT-3 based data annotation approaches for both sequence- and token-level NLP tasks. Our main contributions can be summarized as follows:

- We conduct comprehensive analysis of the feasibility of leveraging GPT-3 for data annotation for complex NLP tasks.
- We study 3 different GPT-3 based data annotation approaches, and then conduct extensive experiments on both sequence- and token-level NLP tasks to evaluate their performance.
- We find that directly annotating unlabeled data is suitable for tasks with small label space while generation-based methods are more suitable for tasks with large label space.
- We find that generation-based approaches tend to be more cost-effective compared with directly annotating unlabeled data.

## 2 Related Work

**Large Language Models** Large language models (LLMs) have made significant progress on natural language processing tasks in recent years. These models are trained with self-supervision on large, general corpora and demonstrate excellent performance on numerous tasks (Brown et al., 2020; Rae et al., 2021; Taylor et al., 2022; Hoffmann et al., 2022; Black et al., 2022; Zhang et al., 2022; Chowdhery et al., 2022; Thoppilan et al., 2022; Touvron et al., 2023). LLMs possess the ability to learn in context through few-shot learning (Brown et al., 2020; Ouyang et al., 2022). Their capabilities expand with scale, and recent research has highlighted their ability to reason at larger scales with an appropriate prompting strategy (Lester et al., 2021; Wei et al., 2022; Chowdhery et al., 2022; Liu et al., 2021c; Kojima et al., 2022; Lewkowycz et al., 2022; Qin et al., 2023b; Zhao et al., 2023; Li et al., 2023; Jiao et al., 2023).

Wang et al. (2021) investigate methods to utilize GPT-3 to annotate unlabeled data. However, they mainly focus on the generation and sequence classification tasks. In this work, we conduct more comprehensive experiments and analysis on a wider range of settings, covering both sequence- and token-level tasks. In a recent work, Liu et al. (2022) demonstrate a worker-and-AI collaborative approach for dataset creation with a few seed examples, while we also analyze approaches that support zero-shot training data generation, which do not require any seed examples.

**Prompt-Learning** Prompt-Learning, also known as Prompting, offers insight into what the future of NLP may look like (Lester et al., 2021; Liu et al., 2021c; Ding et al., 2021b). By mimicking the process of pre-training, prompt-learning intuitively connects pre-training and model tuning (Liu et al., 2021d). In practice, this paradigm has proven remarkably effective in low-data regimes (Scao and Rush, 2021; Gao et al., 2021; Qin and Joty, 2022b). For instance, with an appropriate template, zero-shot prompt-learning can even outperform 32-shot fine-tuning (Ding et al., 2021a). Another promising characteristic of prompt-learning is its potential to stimulate large-scale pre-trained language models (PLMs). When applied to a 10B model, optimizing prompts alone (while keeping the parameters of the model fixed) can yield comparable performance to full parameter fine-tuning (Lester et al., 2021; Qin et al., 2023a). These practical studies suggest that prompts can be used to more effectively and efficiently extract knowledge from PLMs, leading to a deeper understanding of the underlying principles of their mechanisms (Li et al., 2022).

**Data Augmentation** There has been a significant amount of research in NLP on learning with limited labeled data for various tasks, including unsupervised pre-training (Devlin et al., 2019; Peters et al., 2018; Yang et al., 2019; Raffel et al., 2020; Liu et al., 2021b), multi-task learning (Glorot et al., 2011; Liu et al., 2017), semi-supervised learning (Miyato et al., 2016), and few-shot learning (Deng et al., 2019; He et al., 2021; Qin and Joty, 2022a). One approach to address the need for labeled data is through data augmentation (Feng et al., 2021; Meng et al., 2022; Chen et al., 2023), which involves generating new data by modifying existing data points using transformations based on prior knowledge about the problem's structure (Yang et al., 2020). The augmented data can be generated from labeled data (Ding et al., 2020; Liu et al., 2021a; Ding et al., 2022) and used directly in supervised learning (Wei and Zou, 2019) or em-

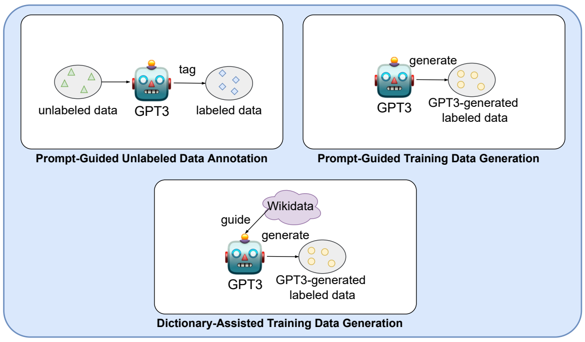


Figure 1: Illustrations of our proposed methods.

**Choose the sentiment of the given text from Positive and Negative.**
**Text:** a feast for the eyes
**Sentiment:** Positive
**...**
**Text:** boring and obvious
**Sentiment:** Negative
**Text:** [Unlabeled Data]
**Sentiment:** [Label]

Figure 2: An example of Prompt-Guided Unlabeled Data Annotation (PGDA) for SST2.

ployed in semi-supervised learning for unlabeled data through consistency regularization (Xie et al., 2020).

## 3 Methodology

We study 3 different approaches to utilize GPT-3 for data annotation: 1) prompt-guided unlabeled data annotation (PGDA); 2) prompt-guided training data generation (PGDG); and 3) dictionary-assisted training data generation (DADG). Illustrations are shown in Figure 1. Overall, these 3 approaches can be regarded as in-context learning (Wei et al., 2022), a new paradigm that is getting popular in NLP. Under this paradigm, a language model "learns" to do a task simply by conditioning on $l_{\text{IOP}}$, a list of input-output pairs (IOP). [3] More formally,

$$y_i = \text{GPT-3}(l_{\text{IOP}}, x_i) \tag{1}$$

where $x_i$ is the query input sequence and $y_i$ is the text generated by GPT-3. For comparison, the performance, cost, and time spent on the three methods are monitored. We also report the results of **Prompted Direct Inference (PGI)**, which is to instruct GPT-3 to directly annotate the test data.

### 3.1 Prompt-Guided Unlabeled Data Annotation (PGDA)

The first approach involves the creation of prompts to guide GPT-3 in annotating unlabeled data. To this end, task-specific prompts are designed to elicit labels from GPT-3 for a given set of unlabeled data. In our experiments, the unlabeled data is derived from human-labeled datasets by removing the existing labels. The resulting GPT-3-labeled data is then used to train a local model to predict human-labeled test data, with the performance of this model being evaluated. As shown in Figure 2, an instruction with few-shot examples is given to GPT-3, followed by unlabeled data. GPT-3 is then prompted to predict labels for the unlabeled data.

### 3.2 Prompt-Guided Training Data Generation (PGDG)

The second approach is to utilize GPT-3 to autonomously generate labeled data for the specified task. This method involves the creation of prompts that guide GPT-3 to self-generate labeled data, which is subsequently used to train a local model to predict on human-labeled test data for the purpose of evaluation. For example, to generate training data with the relation "head of government", we can first "teach" GPT-3 to generate head-tail entity pairs that have the specified relation as illustrated in Figure 3. After we obtain the generated triplets (head-tail entity pairs with specified relation), as shown in Figure 4, we can then instruct GPT-3 to generate a sentence with the given entities and relation. Compared with tagging approach, a significant benefit of the generation-based approach is that it does not require a long list of label definitions specified in the prompt. For example, to generate NER data, it can first generate entities of each entity type (e.g. organization, person, etc.) and then generate a sentence with mixed entities.

### 3.3 Dictionary-Assisted Training Data Generation (DADG)

The third method is designed to utilize a dictionary as an external source of knowledge to assist GPT-3 to generate labeled data for a specific domain. In our experiments, we choose Wikidata[4] as the dictionary. The data generated through this Wikidata-guided process is subsequently used to

[3] Under the zero-shot settings, where $l_{\text{IOP}}$ is not provided, our methods become instruction-tuning (Wei et al., 2021).

[4] https://www.wikidata.org

**Generate 20 different Head Entity and Tail Entity with the given Relation.**
**Relation:** head of government
**Relation Definition:** head of the executive power of this town, city, municipality, state, country, or other governmental body
**Relation:** head of government
**Head Entity:** United States; **Tail Entity:** Chester Alan Arthur
**...**
**Head Entity:** Entity1; **Tail Entity:** Entity2

Figure 3: An example of prompting GPT-3 to generate entities for the relation "head of government" for FewRel.

**Generate a sentence with the given entities and relation.**
**Relation:** head of government
**Head Entity:** United States; **Tail Entity:** Chester Alan Arthur
**Text:** Chester Alan Arthur , 21st President of the United States , died of this disease , November 18 , 1886
**...**
**Relation:** head of government
**Head Entity:** Entity1; **Tail Entity:** Entity2
**Text:** [Generated Sentence]

Figure 4: An example of prompting GPT-3 to generate a sentence with the given entities and the relation "head of government" for FewRel.

train a local model to predict human-labeled test data for the purpose of evaluating performance. For instance, to generate training data with the relation "head of government", we first query the head-tail entity pairs under the relation *P6*, relation ID of "head of government", from Wikidata. Upon obtaining the entity pairs from Wikidata, GPT-3 can then be instructed to generate a sentence with the specified entity pairs and relation. An advantage of this approach is that it can leverage knowledge base in specific domains, particularly when the domains are not present in the pre-trained corpus, thus allowing for the incorporation of external knowledge into GPT-3 without the need for fine-tuning.

# 4 Experiments

## 4.1 Experiment Settings

In this study, we conduct extensive experiments on both sequence- and token-level NLP tasks[5]. The sequence-level tasks include sentiment analysis (SA) and relation extraction (RE). The token-level tasks include named entity recognition (NER) and aspect sentiment triplet extraction (ASTE).

More specifically, we use the SST2 dataset (Socher et al., 2013) for sentiment analysis, a well-known dataset comprising movie reviews. For relation extraction, we use FewRel (Han et al., 2018), a large-scale relation extraction dataset. For NER, we use the AI domain split from the CrossNER dataset (Liu et al., 2020), which is the most difficult domain within the dataset and more closely mirrors real-world scenarios with its 14 entity types. For aspect sentiment triplet extraction, we use the laptop domain split released by (Xu et al., 2020).

To simulate the production scenario, we assume that the user has access to the off-shelf GPT-3 API. In all our experiments, we use *text-davinci-003*[6], the latest GPT-3 model. In addition, we assume that the user uses $\text{BERT}_{\text{BASE}}$ for production and has access to a few data points and Wikidata for each task. For each task, the resulting data of each approach is post-processed and reformatted into the same format of human-labeled data before being used to fine-tune a $\text{BERT}_{\text{BASE}}$ model. In order to accurately determine the cost and time required for human labeling, we conduct interviews and consultations with linguists and professional data annotators to obtain a precise estimation.

## 4.2 Sequence-Level Task

### 4.2.1 SST2

SST2 dataset is used for sequence-level sentiment analysis experiments. We fine-tune $\text{BERT}_{\text{BASE}}$ on the data created by the three approaches for 32 epochs with early stopping. After model fine-tuning, we evaluate the model on human-labeled test data to assess the quality of data created by each approach. We conduct experiments on zero-shot, 2-shot, and 10-shot settings. Here we discuss the results for 10-shot settings. Please refer to Appendix A.13 for the results of the other two settings.

**Annotation Approaches** In PGDA, we randomly sample 10-shot data of the train set of the SST2 dataset to construct a prompt template, as illustrated in Figure 2. The prompt is used to guide GPT-3 in generating sentiment labels for the unlabeled data. In PGDG, the same 10-shot data used in the PGDA is used to guide GPT-3 to generate sentences with specified sentiments. Please refer to

[5]Please refer to Appendix A.11 for the discussion on more complex tasks like semantic parsing

[6]Released on 28 Nov 2022. Please refer to https://beta.openai.com/docs/models for more details.

Appendix A.2 for the prompt example. In DADG, the ability of GPT-3 to perform Wikidata-guided few-shot generation is tested. We query entities in Wikidata from the movie domain. We then use the entities together with the same 10-shot data to prompt GPT-3 to generate sentences with a specified sentiment. Please refer to Appendix A.3 for the prompt example.

**Results** Table 1 presents the results of three different approaches. Overall, PGDA demonstrates the best performance among the three approaches. By labeling the same 3,000 data points, PGDA achieves an accuracy of 87.75, which is only 0.72 lower than that of human-labeled data. However, the cost and time consumed for PGDA are significantly lower than those for human labeling. By labeling 6,000 data, PGDA achieves a better performance than the human-labeled 3,000 data, while the cost is approximately 10% of the cost of human labeling. PGDG performs much worse than PGDA and human-labeled data. However, it also demonstrates a distinct advantage in terms of cost and time efficiency when generating the same amount of data compared with alternative approaches. DADG approach, which involves generating data with in-domain entities, does not result in better performance. This is because entities are not typically key factors in the sentiment classification task, as most entities are neutral and do not provide additional information relevant to sentiment. Furthermore, since a large portion of the data in SST2 does not contain any entities, the sentences generated using DADG do not follow the same distribution as the test data in SST2, leading to poorer performance. For comparison purposes, the result of PGI is also presented. It is suggested that, for small-scale applications, it is practical to use GPT-3 to directly label unlabeled data.

#### 4.2.2 FewRel

The FewRel dataset is used for RE experiments. The original FewRel dataset, proposed for meta-learning, is re-formulated to a supervised learning setting. The train data of FewRel, which comprises 64 distinct relations and 700 labeled instances for each relation, is divided into a new train/dev/test split (560/70/70). It is to simulate the real-world application of GPT-3 to annotate data for tasks with large label spaces. For FewRel experiments, we follow (Devlin et al., 2019) to fine-tune $\text{BERT}_{\text{BASE}}$ on the data created by the three approaches for 3 epochs. Subsequently, the fine-tuned model is evaluated on the human-labeled test data to assess the quality of data produced by the proposed approaches. The number of samples annotated or generated by each approach is determined by assuring the costs of each approach are comparable.

| Approach | Num. of Samples | Cost (USD) | Time (Mins) | Results |
|---|---|---|---|---|
| PGDA | 3000 | 11.31 | 14† | 87.75 |
| | 6000 | 22.63 | 27† | **89.29** |
| PGDG | 3000 | 0.91 | 4† | 73.81 |
| | 6000 | 1.83 | 8† | 76.55 |
| DADG | 3000 | 7.18 | 23† | 68.04 |
| | 6000 | 14.37 | 46† | 71.51 |
| Human Labeled | 3000 | 221 - 300 | 1000 | 88.47 |
| | 67349 | 4800 - 6700 | 22740 | 93.52 |
| PGI | 1821 | 7.33 | 12 | 95.77 |

Table 1: Costs, time spendings and results of SST2. †means multiprocessing (5 processes) is enabled. Time for manual labeling excludes the time spent on instruction preparation and training.

**Annotation Approaches** The FewRel dataset poses significant challenges for the PGDA approach, primarily due to the complexity of instructing GPT-3 to comprehend the 64 relations. Due to the cost and maximum token length constraints of the GPT-3 API, we can only include 1-shot data for each relation within the prompt, which can make it difficult for GPT-3 to "understand" each relation. To address these challenges, we try 5 different prompts for PGDA, with the goal of exploring whether different prompts could be effective for tasks with large label space. Please refer to Appendix A.10 for the prompt examples. As mentioned in Section 3.2, in PGDG, we conduct the annotation for RE in two steps. The first step is to instruct GTP-3 to generate head-tail entity pairs for a specified relation and the second step is to generate sentences with the generated triplets. We generate 200 labeled data for each relation. As mentioned in Section 3.3, DADG for RE is also conducted in two steps. The first step is to query WikiData to obtain head-tail entity pairs for a specified relation and the second step is to generate sentences with the generated triplets. We generate 200 labeled data for each relation.

**Results** Table 2 presents the results of three different approaches. All five proposed prompts for PGDA perform badly on the FewRel task due to the task difficulty and large label space. In contrast, the generation-based approaches, namely PGDG

| Approach | Num. of Samples | Cost (USD) | Time (Mins) | P | R | F1 |
|---|---|---|---|---|---|---|
| PGDA1 (1-shot) | 384 | 28.55 | 13† | 0.03 | 1.56 | 0.05 |
| PGDA2 (1-shot) | 384 | 25.40 | 10† | 0.14 | 1.7 | 0.18 |
| PGDA3 (1-shot) | 384 | 25.19 | 11† | 0.09 | 1.65 | 0.13 |
| PGDA4 (1-shot) | 384 | 25.57 | 10† | 0.02 | 1.56 | 0.05 |
| PGDA5 (1-shot) | 384 | 25.56 | 11† | 0.02 | 1.56 | 0.05 |
| PGDG (1-shot) | 12800 | 30.58 | 285† | 47.82 | 45.58 | **44.11** |
| DADG (1-shot) | 12800 | 17.16 | 220† | 45.41 | 42.41 | 40.02 |
| PGDG (5-shot) | 12800 | 99.35 | 340† | 70.59 | 67.99 | **67.71** |
| DADG (5-shot) | 12800 | 88.91 | 265† | 59.76 | 60.85 | 57.98 |
| Human Labeled | 704 | 101 - 200 | 640 | 41.92 | 41.45 | 34.22 |
| | 12800 | 1828 - 3584 | 11636 | 85.19 | 85.07 | 84.95 |
| | 35840 | 6400 - 10,000 | 32582 | 87.55 | 87.43 | 87.34 |
| PGI | 4480 | 33.30 | 160† | 29.86 | 29.82 | 25.85 |

Table 2: Costs, time spendings, and results of FewRel. Time for manual labeling excludes the time spent on instruction preparation and training. The number of samples annotated or generated by each approach is determined by assuring **comparable costs**. We use ChatGPT instead of GPT-3 to perform PGI on FewRel data as a proxy as the cost of using GPT-3 for PGI is obviously much higher. †means multiprocessing (5 processes) is enabled.

and DADG, achieve much better performance with comparable costs. Even with access to only 1-shot data, PGDG and DADG yield F1 scores of around 44 and 40 points respectively in comparison to PGDA. With access to 5-shot data, the performances of PGDG and DADG are further improved with the increased diversity of the generated data. Under comparable costs, PGDG and DADG outperform the human-labeled data (704 data points) with 33-point and 23-point F1 scores respectively. It is worth noting that the PGDG approach consistently outperforms the DADG approach. Through analysis, it is determined that the head-tail entity pairs generated by PGDG possess greater diversity than those generated by DADG for specific relations such as religion and the language of the work. We do not perform PGI on FewRel data as the cost is obviously much higher.

### 4.3 Token-Level Task

#### 4.3.1 CrossNER

The AI domain split in CrossNER has 14 entity classes, namely product, field, task, researcher, university, programming language, algorithm, misc, metrics, organisation, conference, country, location, person. We fine-tune $\text{BERT}_{\text{BASE}}$ on the CrossNER task with corresponding data for 100 epochs with early stopping.

**Annotation Approaches** In PGDA, as shown in Appendix A.4, for each entity type, we initiate GPT-3 to generate its definition and provide a selection of data (no more than 10-shot) with entities belonging to the specified entity type in the prompt to assist GPT-3 in recognizing entities belonging to the same class within the unlabeled data. It is observed that the same entity may be labeled as different entity types with different prompts. Therefore, we also include an additional prompt, as illustrated in Figure 12 in Appendix A.4, to determine the final entity type for each identified entity. Both PGDG and DADG for CrossNER are conducted in two steps. The first step for PGDG is to prompt GPT-3 to generate entities for each entity type as shown in Appendix A.5. On the other hand, the first step for DADG is to query Wikidata to get the entities of each entity type. Notice that we use no more than 200 generated entities for each entity type in our experiments for both PGDG and DADG. The second step of both approaches is to use the generated entities to generate sentences within a specific domain using GPT-3 as shown in Figure 14 in Appendix A.4. In the process of generating sentences for both PGDG and DADG, we randomly select a few entities from all the entities to generate each sentence.

**Results** Table 3 presents the results of the three approaches. We find the train data labeling method using PGDA has the worst performance yet the highest costs among the three proposed approaches. It should be noted that there are only 100 gold train data points in the AI domain split in the CrossNER dataset, and these same 100 data points are labeled using PGDA. However, the cost of labeling these 100 data points is higher than the cost of using the generation approaches to generate 3000 data points. It is observed that GPT-3 is effective at identifying entities in the text, but it may also identify entities that are not of the specified entity type, resulting in incorrect labeling. Additionally, GPT-3 may not accurately identify the boundaries of the entities. These two disadvantages make it impractical to use PGDA for labeling data for named entity recognition (NER) in a production setting, especially when the label space becomes bigger. The PGDG approach is able to achieve a result comparable to the 100 human-labeled gold train data at a lower cost. When utilizing Wikidata, the DADG approach is able to achieve a higher result than PGDG, likely due to its ability to leverage more unique entities and in-domain entities extracted from Wikidata. This shows that the ability to access in-domain entities is crucial for creating high-quality training

| Approach | Num. of Samples | Cost (USD) | Time (Mins) | Results |
|---|---|---|---|---|
| PGDA (10-shot) | 100 | 15.39 | 21 | 23.08 |
| PGDG (Zero-shot) | 1500 | 7.78 | 17† | 42.63 |
| | 3000 | 13.56 | 33† | 41.35 |
| DADG (Zero-shot) | 1500 | 6.77 | 20† | 46.90 |
| | 3000 | 13.61 | 40† | **47.22** |
| Human Labeled | 100 | 17 - 42.85 | 65 | 42.00 |
| PGI | 431 | 63.23 | 20† | 46.65 |

Table 3: Cost, time spending and results of CrossNER (AI Domain Split). Time for manual labeling excludes the time spent on instruction preparation and training. †means multiprocessing (5 processes) is enabled.

| Approach | Num. of Samples | Cost (USD) | Time (Mins) | P | R | F1 |
|---|---|---|---|---|---|---|
| PGDA1 | 906 | 11.34 | 18 | 57.93 | 44.38 | **50.26** |
| PGDA2 | 906 | 9.02 | 17 | 50.78 | 24.13 | 32.71 |
| PGDA3 | 906 | 12.84 | 19 | 50.73 | 38.31 | 43.65 |
| PGDG1 | 1000 | 9.41 | 15† | 44.36 | 22.47 | 29.83 |
| PGDG2 | 1000 | 7.68 | 14† | 54.93 | 14.36 | 22.77 |
| PGDG3 | 1000 | 13.77 | 18† | 45.10 | 12.71 | 19.83 |
| DADG | 1000 | 13.74 | 18† | 48.61 | 6.45 | 11.38 |
| Human Labeled | 91 | 13 - 20 | 180 | 45.14 | 38.49 | 41.55 |
| | 906 | 130 - 200 | 1800 | 63.07 | 55.99 | 59.32 |
| PGI | 328 | 3.92 | 9 | 50.10 | 48.43 | 49.25 |

Table 4: Costs, time spendings and results of ASTE (laptop domain split). Time for manual labeling excludes the time spent on instruction preparation and training. †means multiprocessing (5 processes) is enabled.

data for NER.

#### 4.3.2 ASTE

We follow (Xu et al., 2021) to fine-tune $\text{BERT}_{\text{BASE}}$ on the ASTE task using data created by each approach for 10 epochs and evaluate the fine-tuned models on human-labeled test data. We conduct our experiment under 10-shot settings.

**Annotation Approaches** In PGDA, we randomly sample 10-shot data from gold train data and use them to guide GPT-3 to tag the unlabeled data. Given the complexity of ASTE, which requires the identification of aspect, opinion, and sentiment triplets, we try 3 different prompts to assess the impact of different prompts on the overall performance of the tagging process. Please refer to Appendix A.8 for more details. In PDGD, for comparison purposes, the same 10-shot data used for PGDA is used in the experiments for PGDG. We first instruct GPT-3 to generate aspect-opinion-sentiment triplets and then instruct GPT-3 to generate sentences with the generated triplets. We also try on 3 prompts under PGDG as specified in Appendix A.9. In DADG, we query entities in laptop and computer hardware domains from WikiData and used them as aspects. We use the prompt that achieved the best performance for PGDG as the prompt to generate opinions and sentiments for the aspects. Then we use the obtained triplets for sentence generation.

**Results** Table 4 presents the results of three different approaches. PGDA achieves the best performance compared with the other approaches. We also notice that performance varies with different prompts, which aligns with the previous research (Luo et al., 2022). PGDG tends to generate data with explicit sentiment, as shown in Appendix A.6. Similar to SST2, as entities are not the key factors for ASTE and provide little help to this task, DADG is also outperformed by PGDA.

## 5 Further Analysis

### 5.1 Impact of Label Space

The results of our experiments indicate that the tagging-based approach (PGDA) is more appropriate for tasks with smaller label spaces and clearly defined labels. Examples of such tasks include sentence-level sentiment analysis and ASTE, which both have small label space (2-3 labels) that can be easily distinguished, e.g. positive, negative, neutral. In contrast, the generation-based approaches (PGDG and DADG) are better suited for tasks with larger label spaces or labels that possess a certain degree of ambiguity. Examples of such tasks include CrossNER and FewRel, which have 14 and 64[7] labels respectively, and some of which may be difficult to identify or differentiate (e.g. Misc, etc.). Both the tagging-based and generation-based approaches have their own advantages and disadvantages. The tagging-based approach allows for direct access to in-domain unlabeled data, while the generation-based approaches may generate data that contains information that was "learned" during pre-training and may not align with the distribution of in-domain data. However, as the label space becomes larger, the tagging-based approach requires a lengthy prompt with examples to guide GPT-3, which can lead to catastrophic forgetting and increase annotation costs. On the other hand, the generation-based approaches can reformulate the task by first generating spans with labels (e.g.

[7]We refer to the train split of the FewRel used in our experiments. The original FewRel data has 100 labels in total.

**Generated Entities:** Chiang Mai International Airport; Chiang Mai, Thailand;
**Generated Sentence:** Chiang Mai International Airport is the main gateway for air travels to and from Chiang Mai, Thailand.

Figure 5: An example to demonstrate the generation ability of GPT-3.

| Model | Numb. of Sampes | Cost | Results |
|---|---|---|---|
| GPT-3 | 3000 | 11.31 | **87.75** |
| ChatGPT | 3000 | **1.50** | 87.31 |

Table 5: Preliminary Comparison between GPT-3 and ChatGPT on SST2.

entities and triplets), and then generating a sentence with the labeled spans. These approaches reduce label errors and avoid the challenges of span boundary detection. In addition, generation-based approaches tend to be more cost-effective. as the prompts used can be significantly shorter when compared to those used in the tagging-based approach and multiple data can be generated with a single prompt at a time.

### 5.2 Comparision with Human Annotators

Through extensive experiments, we find that GPT-3 demonstrates promising ability to generate domain-specific data (e.g., entities in AI), structured data (e.g., triplets), as well as unstructured sequences at a fast speed. As discussed above, GPT-3 can even be used to generate data from scratch or to convert structured knowledge into natural sentences (Figure 5), eliminating the requirement of unlabeled data. While for human annotators, it usually takes longer time to train them for domain-specific data annotation, and their annotation speed is not comparable with machines in most cases. Moreover, it is often more challenging for humans to construct training data without unlabeled data, or when the size of label space is very large. Therefore, in terms of speed and domain-specific data annotation, and in the setting of labeled data generation, large language models (LLMs) exhibit encouraging potential. Machines are good at quickly labeling or generating a large amount of training data. However, if we limit the number of data samples for model training, the per-instance quality of the data annotated by humans is still higher in most cases.

### 5.3 Impact of Number of Shots

We conduct experiments on the following two datasets, SST2 and FewRel to explore the impact of the number of shots. We find that increasing the number of shots does not necessarily lead to better annotation results for all approaches. As shown in Figure 6, for SST2, tagging approach (PGDA) can benefit from more examples in the context, which enhances GPT-3's ability to tag unlabeled data. However, for the PGDG and DADG approaches, GPT-3 tends to generate data similar to the given examples. As shown in Figure 7, for SST2, the data is usually not a complete sentence and tend to be short and carry less information. Thus, with more data examples, GPT-3 will "learn" to generate similar data with less information and lead to poorer data quality. However, for FewRel, the data is a complete sentence and carry lots of information and the relations between the head entity and tail entity tend to be more implicit. Thus, with 5-shot data in the context, GPT-3 can generate data that also contain more implicit relations than only with 1-shot or zero-shot in the context[8].

### 5.4 Preliminary Comparison between GPT-3 and ChatGPT

Based on the findings presented in Table 5, our analysis reveals that ChatGPT exhibits a performance level that is on par with GPT-3 when it comes to the SST2 task. Notably, the results obtained from our observations demonstrate comparable outcomes between ChatGPT and GPT-3 in terms of task performance. Moreover, from a cost-efficiency standpoint, ChatGPT emerges as a more economically viable alternative when compared to GPT-3, which may make it a preferable choice. A study conducted by Gilardi et al. (2023) further illustrates the superior performance of ChatGPT compared to crowd-workers for various annotation tasks. By employing a dataset consisting of 2,382 tweets, the research demonstrates that ChatGPT surpasses the capabilities of crowd-workers across multiple annotation tasks, including relevance assessment, stance analysis, topic identification, and frame detection. These findings suggest that large language models may outperform human annotators when it comes to these specific tasks, highlighting their potential as a highly effective and reliable tool for annotation purposes.

[8]Please refer to Appendix A.7 for the examples of generated data with different number of shots for SST2 and FewRel.

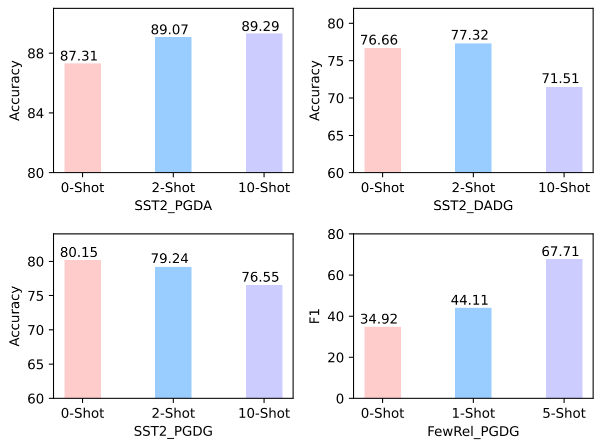


Figure 6: Experiments on the impact of number of shots. We reported the results of 6,000 data on SST2 and 12,800 data (200 data per class) on FewRel.

**SST Example:** a smile on your face (positive)
**FewRel Example:** Winscombe is a lightly populated locality in the southern part of the Canterbury region of New Zealand 's South Island . (Relation: located on terrain feature)

Figure 7: Examples to show the differences between the data distributions of SST2 and FewRel data.

### 5.5 Case Study on Multilingual Data Annotation

As shown in Appendix A.14, we meticulously examined the annotation capabilities of state-of-the-art language models, namely GPT-3, ChatGPT, and GPT-4, within the context of multilingual training data. Our observations revealed that these models possess the remarkable ability to annotate such data effectively, even when presented with minimal or no prior exposure to the target languages. By employing a zero shot or few shot setting, where the models were not explicitly fine-tuned on the specific languages in question, we witnessed their capacity to accurately annotate and comprehend diverse linguistic inputs from a multitude of languages. This notable achievement underscores the potential of these language models to transcend language barriers and facilitate efficient multilingual data processing, making them invaluable tools for a wide range of language-related tasks and applications.

## 6 Conclusions

In this work, we investigate the effectiveness of GPT-3 as a data annotator for various natural language processing (NLP) tasks using three main approaches. Our experimental results show that GPT-3 has the potential to annotate data for different tasks at a relatively lower cost, especially for individuals or organizations with limited budgets. With the limited budget, performance of model trained on the GPT-3 annotated data is often comparable to or even better than that trained on human-annotated data. However, it should be noted that the quality of data annotated by GPT-3 still has room for improvement when compared to human-annotated data. We hope the findings in this work can shed the light on automatic data annotation using large language models and provide some insights so that more methods can be proposed to enhance the quality of data created by these models. With everyone being able to create data for their model training, we can pave the way for the democratization of AI.

## Acknowledgements


This research is supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. We would like to thank Aljunied Mahani for her contributions on her valuable advice in this project from a linguist perspective.

I (Bosheng) would like to express my heartfelt gratitude and deepest appreciation to the memory of my beloved father, whose untimely departure last year has left an indelible void in my life. His unwavering support, encouragement, and wisdom were instrumental in shaping my character and nurturing my academic pursuits. I am forever indebted to his enduring love and belief in my abilities. Though he is no longer physically present, his spirit continues to guide and inspire me as I embark on this research journey. This work is dedicated to his cherished memory, a constant reminder of his profound impact on my life.


## 7 Limitations

Our work is subject to certain limitations, one of which pertains to financial constraints that hindered the ability to conduct large-scale experimentation with the data annotation methods proposed. As a result, the findings of this study may not be fully representative of larger datasets or populations. Additionally, the utilization of GPT-3 as a model presents challenges in terms of interpretability, as it operates as a "black box" system. To further investigate this subject, it would be beneficial to conduct larger-scale experiments and to compare the performances of GPT-3, ChatGPT[9], and GPT-4 (OpenAI, 2023) and the open-sourced LLMs like LLaMA (Touvron et al., 2023).

## Ethics Consideration

One of the significant issues associated with GPT-3 is the potential for it to reinforce existing biases present in the data sets it annotated. This is due to GPT-3 being pre-trained on a vast amount of unlabelled data, which may include bias and stereotypes (Li et al., 2022). To address this concern, it is crucial to guarantee that the data used to train GPT-3 is diverse and representative of various viewpoints and experiences. Furthermore, consistent monitoring and evaluation of the output generated by GPT-3 should be implemented to identify and rectify any possible biases.

## References


Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

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.

Andreas Bunte, Frank Richter, and Rosanna Diovisalvi. 2021. Why it is hard to find ai in smes: A survey from the practice and how to promote it. In *ICAART*.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Shumin Deng, Ningyu Zhang, Zhanlin Sun, Jiaoyan Chen, and Huajun Chen. 2019. When low resource nlp meets unsupervised language model: Meta-pretraining then meta-learning for few-shot text classification (student abstract). In *AAAI Conference on Artificial Intelligence*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2020. Daga: Data augmentation with a generation approach for low-resource tagging tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Haitao Zheng, Zhiyuan Liu, Juan-Zi Li, and Hong-Gee Kim. 2021a. Prompt-learning for fine-grained entity typing. *ArXiv*, abs/2108.10604.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2021b. Openprompt: An open-source framework for prompt-learning. *ArXiv*, abs/2111.01998.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2023. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*.


[9]https://chat.openai.com/chat

Steven Feng, Varun Prashant Gangal, Jason Wei, Soroush Vosoughi, Sarath Chandar, Teruko Mitamura, and Eduard Hovy. 2021. A survey on data augmentation approaches for nlp.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723.

Colin Shunryu Garvey. 2018. A framework for evaluating barriers to the democratization of artificial intelligence. In *AAAI Conference on Artificial Intelligence*.

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

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *International Conference on Machine Learning*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Fangkai Jiao, Zhiyang Teng, Shafiq R. Joty, Bosheng Ding, Aixin Sun, Zhengyuan Liu, and Nancy F. Chen. 2023. Logicllm: Exploring self-supervised logic-enhanced training for large language models. *ArXiv*, abs/2305.13718.

Su Young Kim, Hyeon ju Park, Kyuyong Shin, and KyungHyun Kim. 2022. Ask me what you need: Product retrieval using knowledge from gpt-3. *ArXiv*, abs/2207.02516.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *ArXiv*, abs/2104.08691.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *ArXiv*, abs/2206.14858.

Xingxuan Li, Yutong Li, Shafiq R. Joty, Linlin Liu, Fei Huang, Linlin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq R. Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *ArXiv*, abs/2305.13269.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021a. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Annual Meeting of the Association for Computational Linguistics*.

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq R. Joty, and Luo Si. 2021b. Enhancing multilingual language model with massive multilingual knowledge triples.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021c. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys (CSUR)*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021d. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv*, abs/2110.07602.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. Crossner: Evaluating cross-domain named entity recognition. In *AAAI Conference on Artificial Intelligence*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv: Machine Learning*.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics*.

Chengwei Qin and Shafiq Joty. 2022a. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Chengwei Qin and Shafiq Joty. 2022b. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. In *International Conference on Learning Representations*.

Chengwei Qin, Shafiq Joty, Qian Li, and Ruochen Zhao. 2023a. Learning to initialize: Can meta learning improve cross-task generalization in prompt tuning? *arXiv preprint arXiv:2302.08143*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023b. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Giovanni Rubeis, Keerthi Dubbala, and Ingrid Metzler. 2022. "democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Frontiers in Genetics*, 13.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? In *North American Chapter of the Association for Computational Linguistics*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *ArXiv*, abs/2211.09085.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766, Online. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

# A Appendix

## A.1 PGDA for SST2

**Choose the sentiment of the given text from Positive and Negative.**
**Text:** a feast for the eyes
**Sentiment:** Positive
**...**
**Text:** boring and obvious
**Sentiment:** Negative
**Text:** [Unlabeled Data]
**Sentiment:** [Label]

Figure 8: An example of prompt-guided unlabeled data annotation for SST2.

## A.2 PGDG for SST2

**Write 20 different movie reviews with positive sentiments with no more than 20 words.**
**Sentiment:** Positive
**Text:** a feast for the eyes
**...**
**Sentiment:** Positive
**Text:**

Figure 9: An example of prompt-guided data generation for SST2.

## A.3 DADG for SST2

**Sentiment:** Positive
**Text:** a feast for the eyes
**...**
**Write a movie review with the given entity with positive sentiment.**
**Entity:** [Entity1]
**Sentiment:** Positive
**Text:**

Figure 10: An example of dictionary-assisted training data generation for SST2.

## A.4 PGDA for CrossNER

**Researcher:** A researcher in AI domain is an individual who conducts research and experiments related to Artificial Intelligence and its related fields, such as ...

**Text:** Advocates of procedural representations were mainly centered at MIT , under the leadership of Marvin Minsky and Seymour Papert .
**Researcher entity:** Marvin Minsky; Seymour Papert;
**...**
**Text:** [Unlabeled Data]
**Researcher entity:**

Figure 11: An example of prompt-guided unlabeled data annotation for CrossNER.

**Choose the right entity type from the candidate list for the given entity in the text context.**
**Text:** Advocates of procedural representations were mainly centered at MIT, under the leadership of Marvin Minsky and Seymour Papert .
**Entity:** Marvin Minsky
**Candidate List:** product, task, researcher, university, organisation, person
**Entity Type:** researcher
**...**
**Text:** [Unlabeled Data]
**Entity:** [Entity]
**Candidate List:** [Entity_Type1, Entity_Type2, Entity_Type3, ...]
**Entity Type:**

Figure 12: An example of prompt to determine the entity type of an entity in CrossNER.

### A.5 PGDG and DADG for CrossNER

**Researcher:** A researcher in AI domain is an individual who conducts research and experiments related to Artificial Intelligence and its related fields, such as Machine Learning ...
**Researcher:** David Silver, Fei-Fei Li, Claude Shannon, Marvin Minsky, Ruslan Salakhutdinov
Generate 15 different researchers in the AI domain.
**Researcher:**
1. David Silver
2. ...

Figure 13: An example of prompting GPT-3 to generate entities for the type 'Researcher' for PGDG.

**Generate text with all the given entities in the AI domain.**
**Entities:** Entity1_Type: Entity1; Entity2_Type: Entity2; ...
**Text:**

Figure 14: An example of prompting GPT-3 to generate a sentence with given entities for both PGDG and DADG.

### A.6 Generated Samples for ASTE by GPT-3

**Gold train data:** The biggest problem is that the box had no instructions in it .
**Data generated by PGDG:** The port layout is good and the processor is good for the price .
**Data generated by DADG:** The Edge device is quite lightweight , the PC speaker is mediocre, but great for a Toshiba T3100 and good for other peripherals.

Figure 15: Examples to compare the gold train data and the sentences generated by GPT-3. GPT-3 tends to generate data with more explicit sentiment expressions compared with gold train data.

### A.7 Generated Samples for SST2 and FewRel for Different Number of Shots

**Zero-shot:** Fantastic! Great performances, an incredible soundtrack, and a captivating plot.
**1-shot:** A heartfelt and sincere film that will leave you feeling uplifted
**5-shot:** a real crowd-pleaser

Figure 16: Examples to show the sentences generated by GPT-3 under Zero-shot, 1-shot, and 5-shot settings for SST2 with PDPG.

**Zero-shot:** The Dallas Airport is a transport hub that serves the city of Dallas.
**1-shot:** Narita Airport ( NRT ) serves as the main transport hub for flights to and from Narita.
**5-shot:** It serves as Manila's main international gateway , being located at the heart of Manila International Airport Complex at Ninoy Aquino International Airport in Manila , Philippines.

Figure 17: Examples to show the sentences generated by GPT-3 under Zero-shot, 1-shot, and 5-shot settings for FewRel with PDPG.

## A.8 PGDA for ASTE

**Identify the target, opinion, and sentiment triplets in the given text.**
**Text:** The biggest problem is that the box had no instructions in it .
**Target0:** instructions; Opinion0: problem; Sentiment0: negative
**Target1:** instructions; Opinion1: no; Sentiment1: negative
**...**
**Text:** [Unlabeled Data]
**Target0:** [Label] ...

Figure 18: Prompt for PGDA1 for ASTE.

**Identify the target, opinion, and sentiment triplets in the given text.**
**Text:** The biggest problem is that the box had no instructions in it .
**Target:**instructions; instructions;
**Opinion:** problem; no;
**Sentiment:** negative; negative;
**...**
**Text:** [Unlabeled Data]
**Target:** [Label], ...

Figure 19: Prompt for PGDA2 for ASTE.

**Identify the target, opinion, and sentiment triplets in the given text.**
**Text:** The biggest problem is that the box had no instructions in it .
**Target0:** is instructions. Its opinion span is problem. Its sentiment is negative.
**Target1:** is instructions. Its opinion span is no. Its sentiment is negative.
**...**
**Text:** [Unlabeled Data]
**Target0:** is [Label] ...

Figure 20: Prompt for PGDA3 for ASTE.

## A.9 PGDG and DADG for ASTE

**Generate 20 different sentiment, target and opinion triplets.**
**1. Target0:** instructions; Opinion0: problem; Sentiment0: negative; Target1: instructions; Opinion1: no; Sentiment1: negative;
**...**
**Target0:** [Target0] ...

Figure 21: Prompt for PGDG1 for ASTE.

**Generate 20 different sentiment, target and opinion triplets.**
**1. Target:** instructions; instructions; Opinion: problem; no; Sentiment: negative; negative;
**...**
**11. Target:** [Target0]; ...

Figure 22: Prompt for PGDG2 for ASTE.

**Generate 20 different targets and opinions in positive sentiment. Sentiment:** positive; Target: features; Opinion: nice;
**Sentiment:** positive; Target: priced; Opinion: reasonable;
**...**
**Sentiment:** positive; Target:[Target0] ...

Figure 23: Prompt for PGDG3 for ASTE.

**Generate a sentence with the given target, opinion and sentiment triplets in the laptop domain.**
**Target0:** instructions; Opinion0: problem; Sentiment0: negative; Target1: instructions; Opinion1: no; Sentiment1: negative;
**Text:** The biggest problem is that the box had no instructions in it .
**...**
**Target0:** [Target0]; Opinion0: [Opinion0]; Sentiment0: [Sentiment0];
**...**
**Text:** [Generated Sentence]

Figure 24: An example of Prompting GPT-3 to generate a sentence with given triplets for ASTE using PGDG and DADG.

### A.10 PGDA for FewRel

**Identify the relation between the head entity and the tail entity in the given sentence.**
**Relation:** place served by transport hub; mountain range; religion; participating team; contains administrative territorial entity; head of government; country of citizenship; original network; heritage designation; performer; participant of; position held; has part; location of formation; located on terrain feature; architect; country of origin; publisher; director; father; developer; military branch; mouth of the watercourse; nominated for; movement; successful candidate; followed by; manufacturer; instance of; after a work by; member of political party; licensed to broadcast to; headquarters location; sibling; instrument; country; occupation; residence; work location; subsidiary; participant; operator; characters; occupant; genre; operating system; owned by; platform; tributary; winner; said to be the same as; composer; league; record label; distributor; screenwriter; sports season of league or competition; taxon rank; location; field of work; language of work or name; applies to jurisdiction; notable work; located in the administrative territorial entity;

**Sentence:** Merpati flight 106 departed Jakarta ( CGK ) on a domestic flight to Tanjung Pandan ( TJQ ) . **Head Entity:** TJQ; Tail Entity: Tanjung Pandan
**Relation:** place served by transport hub
**Sentence:** It is approximately 8 km away from Mount Korbu , the tallest mountain of the Titiwangsa Mountains .
**Head Entity:** Mount Korbu; Tail Entity: Titiwangsa Mountains
**...**
**Sentence1:** [unlabeled data]
**Head Entity1:** [head entity]; Tail Entity1:[tail entity]
**Relation:** [label]

Figure 25: Prompt for PGDA1 used for FewRel Experiemtns.

**Identify the relation between the head entity and the tail entity in the given sentence.**
**Relation:** place served by transport hub; mountain range; religion; participating team; contains administrative territorial entity; head of government; country of citizenship; original network; heritage designation; performer; participant of; position held; has part; location of formation; located on terrain feature; architect; country of origin; publisher; director; father; developer; military branch; mouth of the watercourse; nominated for; movement; successful candidate; followed by; manufacturer; instance of; after a work by; member of political party; licensed to broadcast to; headquarters location; sibling; instrument; country; occupation; residence; work location; subsidiary; participant; operator; characters; occupant; genre; operating system; owned by; platform; tributary; winner; said to be the same as; composer; league; record label; distributor; screenwriter; sports season of league or competition; taxon rank; location; field of work; language of work or name; applies to jurisdiction; notable work; located in the administrative territorial entity;

**Sentence:** Merpati flight 106 departed Jakarta ( CGK ) on a domestic flight to Tanjung Pandan ( TJQ ) . **the relation between** TJQ and Tanjung Pandan is place served by transport hub
**Sentence:** It is approximately 8 km away from Mount Korbu , the tallest mountain of the Titiwangsa Mountains .
**the relation between** Mount Korbu and Titiwangsa Mountains is mountain range
**Sentence:** In 1689 , Konstanty was one of the judges who sentenced Kazimierz Łyszczyński to death for atheism .
**the relation between** Kazimierz Łyszczyński and atheism is religion
...
**Sentence1:** [unlabeled data]
**the relation between** [head entity] and [tail entity] is [label]

Figure 26: Prompt for PGDA2 used for FewRel Experiemtns.

**Identify the relation between the head entity and the tail entity in the given sentence.**
**Relation:** place served by transport hub; mountain range; religion; participating team; contains administrative territorial entity; head of government; country of citizenship; original network; heritage designation; performer; participant of; position held; has part; location of formation; located on terrain feature; architect; country of origin; publisher; director; father; developer; military branch; mouth of the watercourse; nominated for; movement; successful candidate; followed by; manufacturer; instance of; after a work by; member of political party; licensed to broadcast to; headquarters location; sibling; instrument; country; occupation; residence; work location; subsidiary; participant; operator; characters; occupant; genre; operating system; owned by; platform; tributary; winner; said to be the same as; composer; league; record label; distributor; screenwriter; sports season of league or competition; taxon rank; location; field of work; language of work or name; applies to jurisdiction; notable work; located in the administrative territorial entity;

**Merpati** flight 106 departed Jakarta ( CGK ) on a domestic flight to [Tanjung Pandan TAIL ENTITY] ( [TJQ HEAD ENTITY] ) . Relation: place served by transport hub
**It is** approximately 8 km away from [Mount Korbu HEAD ENTITY] , the tallest mountain of the [Titiwangsa Mountains TAIL ENTITY] . Relation: mountain range
...
[unlabeled data [[head entity] HEAD ENTITY] [[tail entity] TAIL ENTITY]] Relation: [label]

Figure 27: Prompt for PGDA3 used for FewRel Experiemtns.

**Identify the relation between the head entity and the tail entity in the given sentence.**
**Relation:** place served by transport hub; mountain range; religion; participating team; contains administrative territorial entity; head of government; country of citizenship; original network; heritage designation; performer; participant of; position held; has part; location of formation; located on terrain feature; architect; country of origin; publisher; director; father; developer; military branch; mouth of the watercourse; nominated for; movement; successful candidate; followed by; manufacturer; instance of; after a work by; member of political party; licensed to broadcast to; headquarters location; sibling; instrument; country; occupation; residence; work location; subsidiary; participant; operator; characters; occupant; genre; operating system; owned by; platform; tributary; winner; said to be the same as; composer; league; record label; distributor; screenwriter; sports season of league or competition; taxon rank; location; field of work; language of work or name; applies to jurisdiction; notable work; located in the administrative territorial entity;

**MMerpati** flight 106 departed Jakarta ( CGK ) on a domestic flight to Tanjung Pandan ( TJQ ) . <head> TJQ <tail> Tanjung Pandan <relation> place served by transport hub
**It is** approximately 8 km away from Mount Korbu , the tallest mountain of the Titiwangsa Mountains . <head> Mount Korbu <tail> Titiwangsa Mountains <relation> mountain range
...
[unlabeled data] <head> [head entity] <tail> [tail entity] <relation>: [label]

Figure 28: Prompt for PGDA4 used for FewRel Experiemtns.

**Relation:** place served by transport hub
**Relation Definition:** territorial entity or entities served by this transport hub (airport, train station, etc.)
**Relation:** mountain range
**Relation Definition:** range or subrange to which the geographical item belongs
...
**Identify the relation between the head entity and the tail entity in the given sentence.**
Sentence: Merpati flight 106 departed Jakarta ( CGK ) on a domestic flight to Tanjung Pandan ( TJQ ) . Head Entity: TJQ; Tail Entity: Tanjung Pandan
**Relation:** place served by transport hub
**Sentence1:** [unlabeled data]
**Head Entity1:** [head entity]; Tail Entity1:[tail entity]
**Relation:** [label]

Figure 29: Prompt for PGDA5 used for FewRel Experiemtns.

### A.11 Discussion on Annotation of More Complex Tasks

The primary aim of this paper is to provide assistance to the majority of individuals and small and medium-sized enterprises (SMEs) engaged in data annotation. To achieve this objective, we have specifically chosen the most frequently employed tasks that possess broad downstream applications. In instances where more intricate tasks are encountered, previous research endeavors (Drozdov et al., 2023; Liu et al., 2023; Qin et al., 2023b) have demonstrated the remarkable capabilities of LLMs in complex NLP tasks, encompassing reasoning, semantic parsing, and dialogues, even in situations with minimal or no training data. Furthermore, there is a recent inclination to utilize LLMs for generating self-instruct data, exemplified by projects such as Alpaca[10] and Vicuna[11]. Consequently, the utilization of large language models for data annotation represents a nascent field with ample room for further research and enhancement.

[10] https://crfm.stanford.edu/2023/03/13/alpaca.html
[11] https://lmsys.org/blog/2023-03-30-vicuna/

### A.12 Domain-Specific Knowledge of GPT-3

Figure 30 shows that GPT-3 has memorized a large amount of domain-specific knowledge.

Generate the abstract for a paper related to stem cell.

This paper explores the potential of stem cell research and its implications for medical applications. It reviews the current research and findings on stem cell biology and regenerative medicine, including the use of stem cells in treating diseases such as cancer, Parkinson's, and heart disease. It also examines the ethical and regulatory concerns surrounding the use of stem cells in medicine, and explores the potential of stem cell research in developing new treatments and therapies. The paper concludes that stem cell research holds tremendous promise for the future of medical treatment and has the potential to revolutionize the way we think about and treat diseases.

Generate the abstract for a paper related to biology.

This paper examines the role of biology in human health and disease. In particular, we explore the relationship between genetics and the development and progression of various diseases. We review current research on the effects of genetic variation on disease risk and the potential implications for diagnosis, treatment, and prevention. We also discuss the ethical implications of genetic testing, and suggest directions for future research. Our findings suggest that genetics plays an important role in the development and progression of various diseases, and that further research is needed to better understand the implications of genetic testing.

Generate five Hydrophilic Chemical Organics.

1. Polyethylene glycol
2. Sorbitol
3. Glycerol
4. Polyvinyl alcohol
5. Sodium polyacrylate

Write code for Quicksort.

```
def quickSort(arr):
  if len(arr) <= 1:
    return arr
  else:
    pivot = arr[len(arr) // 2]
    left = [x for x in arr if x < pivot]
    middle = [x for x in arr if x == pivot]
    right = [x for x in arr if x > pivot]
    return quickSort(left) + middle + quickSort(right)
```

Figure 30: Examples showing that GPT-3 has memorized a large amount of domain-specific knowledge.

### A.13 Results for SST2 under zero-shot and 2-shot settings

| Settings | Approach | Number of Samples Annotated / Generated | Cost (USD) | Time (Mins) | Results |
|---|---|---|---|---|---|
| Zero-shot | PGDA | 3000 | 1.82 | 14† | 86.11 |
| | | 6000 | 3.65 | 27† | **87.31** |
| | PGDG | 3000 | 0.8 | 4† | 78.25 |
| | | 6000 | 1.61 | 8† | 80.15 |
| | DADG | 3000 | 3.10 | 13† | 73.53 |
| | | 6000 | 6.21 | 25† | 76.66 |
| 2-shot | PGDA | 3000 | 3.18 | 16 | 85.89 |
| | | 6000 | 6.36 | 32† | **89.07** |
| | PGDG | 3000 | 0.97 | 4† | 79.57 |
| | | 6000 | 1.94 | 9† | 79.24 |
| | DADG | 3000 | 3.68 | 15† | 75.34 |
| | | 6000 | 7.38 | 29† | 77.32 |

Table 6: Costs, time spending, and results of SST2 under zero-shot and 2-shot settings. †means multiprocessing (5 processes) is enabled. Time for manual labeling excludes the time spent on instruction preparation and training.

### A.14 Case Study of Multilingual Data Annotation

Figure 31 and 32 shows that GPT-3, ChatGPT and GPT-4 can be used to annotate data in non-English languages.

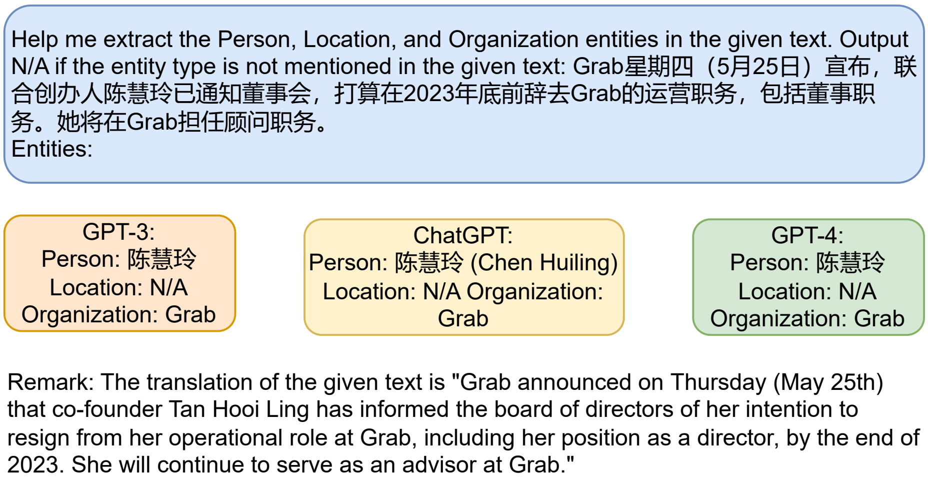


Figure 31: Illustrations of Annotating Chinese NER using GPT-3, ChatGPT and GPT-4.

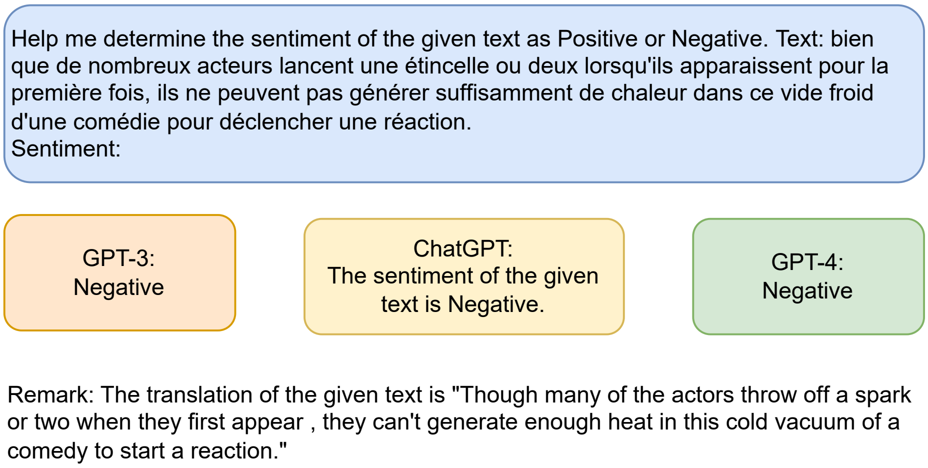


Figure 32: Illustrations of Annotating French Text Classification Data using GPT-3, ChatGPT and GPT-4.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 7 Limitations*

☑ A2. Did you discuss any potential risks of your work?
*Section Ethics Consideration*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract & Section 1 Introduction*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 4 Experiments*

☑ B1. Did you cite the creators of artifacts you used?
*Section 4 Experiments*

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Not applicable. All the datasets used in this pare are open-sourced datasets.*

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Not applicable. All the datasets used in this pare are open-sourced datasets.*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. All the datasets used in this pare are open-sourced datasets.*

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Not applicable. All the datasets used in this pare are open-sourced datasets.*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section 4 Experiments*

### C ☑ Did you run computational experiments?

*Section 4 Experiments*

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Not applicable. We followed the baseline codes.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4 Experiments*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 4 Experiments*

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Not applicable. We followed the baseline codes.*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*