LARGE LANGUAGE MODEL FOR GENERATING TWITTER DATA IN SENTIMENT ANALYSIS

Haorui Wang, Xianle Feng, Xinhan Yang & Haijiao Tao
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA

ABSTRACT

Data sparsity is a big challenge in the machine learning field. Recently, Large language models (LLMs) have shown strong capabilities across various tasks, since they embed extensive world knowledge within their parameters. The impressive performance of LLMs motivates a question: can LLMs be utilized to create synthetic data for specific applications? In this work, we explore the use of LLMs in generating Twitter data for sentiment analysis. Finding that the state-of-the-art method, Attrprompt, falls short in generating Twitter data, we introduce Fact-imagination Prompt, which asks LLMs to imagine potential reasons and events underlying tweets, and then use these imagined scenarios to produce new tweets. The experiment results over two datasets demonstrate that Fact-imagination Prompt outperforms simple class-conditional prompts in terms of diversity. In the comparative analysis, we find that while the Fact-Imagination Prompt increases data diversity and maintains label balance, it adversely impacts data accuracy, thereby reducing the finetuned model's performance in downstream prediction tasks. The results show the importance of balancing data diversity and accuracy in sentiment analysis. [1]

1 INTRODUCTION

Many challenges in machine learning are linked to insufficient data: either the available datasets are too small or even while it's relatively easy to capture unlabeled data, manually labeling it can be extremely costly (Nikolenko, 2021). One potential approach to address this issue is to utilize generative models to generate synthetic data (Patki et al., 2016; Assefa et al., 2020). Although generative models such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), Variational Autoencoders (VAEs) (Doersch, 2016), and diffusion models (Song et al., 2020) are capable of producing data that mimics the characteristics and statistical properties of original datasets, they need to be trained on the original datasets. Such training processes need to be implemented by users, necessitating them to understand the technical details of these generative models. In contrast, large language models (LLMs) have shown remarkable performance on various generative tasks without training by users, such as web operations (Zhou et al., 2023), human behavior generation (Park et al., 2023), and video games (Wang et al., 2023). Users just need to provide natural language prompts, as LLMs already have a vast range of world knowledge embedded in them from their training phase (Huang et al., 2022). This raises an intriguing question: can LLMs be effectively leveraged to generate synthetic training data using straightforward prompting techniques? In this work, we study how to utilize LLMs to generate synthetic Twitter data for sentiment analysis.

Several studies have LLMs as sources for generating training data, particularly in zero-shot settings for text classification tasks (Chen et al., 2023; Gao et al., 2023; Ye et al., 2022). Yu et al. (2023) is the most similar work with us. They introduced a prompting approach named Attrprompt, where LLMs first analyze the attributes of data, and then generate synthetic data based on the attribute diversity. During our initial exploration, we conducted experiments over the Covid-19 tweets dataset.

[1] The data and code is available at https://github.com/leoier/Fact-Imagination-Prompt

The results are shown in Fig. 1. Interestingly, the LLM only summarizes some vague and abstract attributes for Twitter data, which implies that there are no specific attributes for a tweet that could help to determine its sentiment. There are mainly two reasons for this: first, the Attrprompt was originally developed for generating news articles, which are typically longer and more structured than tweets. The tweets lack consistent elements such as style or location. Second, short phrases or single sentences in tweets might be highly ambiguous, so tweets often lack sufficient context for accurately determining the sentiment.

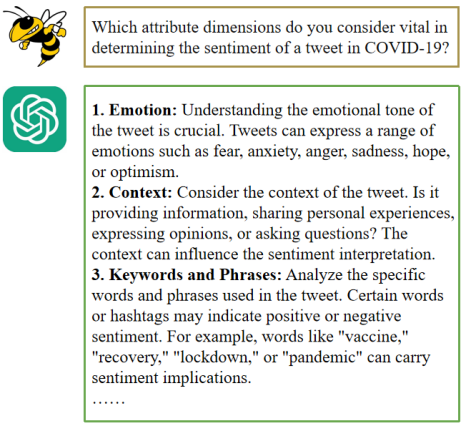

Figure 1: Innitial exploration using Attrprompt

To overcome Attrprompt's shortcomings in applying it to Twitter data, we propose a novel prompt method called Fact-imagination Prompt. This method contrasts with existing approaches like AttrPrompt by focusing on the analysis and reflection of tweet motivations. To evaluate its effectiveness, we compare it with baseline methods, namely the Simple Prompt and Label-Based Prompt. Our data augmentation process employs GPT-3.5 Turbo to generate synthetic samples for each dataset.

The experiment results show that the Fact-imagination Prompt are effective in generating data with high diversity and label balance. By contrast, using the Simple Prompt or the Label-Based Prompt would compromise the label balance or diversity of the generated dataset. However, the evaluation of the fine-tuned model shows that only Simple Prompt has a positive effect on sentiment analysis performance. The fact-imagination Prompt has degraded the performance of the model. And Fact-imagination guided by the LLM itself works better than the one guided by humans.

Our key contributions are: (1) We introduce the Fact-imagination Prompt method, a technique designed to enhance the generation of diverse Twitter data by LLMs. (2) In the experiment, we find that only increasing the diversity of data may not be sufficient for improving the fine-tuned model, as it overlooks data accuracy.

2 Related Work

LLMs as data generators. Several existing works utilize LLMs as data generators to generate various types of data, such as tabular data (Borisov et al., 2022), relation triplets (Chia et al., 2022) and sentence pairs (Schick & Schütze, 2021). For more complex types of data, Ho et al. (2022) let LLMs generate Question Answering chain-of-thought (CoT) reasoning steps, and utilize these CoT reasoning steps to finetune small language models. Wang et al. (2022) design a pipeline to generate instructions, input, and output samples from an LLM, then use them to finetune the original model. Furthermore, LLMs have been used to generate synthetic training data for conversational AI, enabling the creation of customized training examples for lightweight "student" models, thereby addressing the challenge of expensive and complex data collection and labeling (Rosenbaum et al., 2023). Among various settings, we anchor on synthetic data generation for sentiment analysis in the few-shot setting, mainly focusing on the diversity of generated data. In this direction, Chen et al. (2023) highlighted the use of class-conditional prompts during data generation and the potential limitations in diversity. Yu et al. (2023) proposed Attrprompt, where LLMs first analyze the attributes of data and then generate synthetic data based on the attribute diversity.

Sentiment analysis. Data sparsity is a significant challenge in sentiment analysis, particularly in the context of Twitter data, due to tweets' noisy and abbreviated nature. Several approaches have been proposed to alleviate data sparsity in sentiment analysis. Saif et al. (2012) proposed using two different sets of features to address data sparsity: the semantic feature set, which extracts semantically hidden concepts from tweets and incorporates them into classifier training through interpolation, and the sentiment-topic feature set. Additionally, Akhtar et al. (2018) proposed minimizing the

effect of data sparsity by leveraging bilingual word embeddings learned through a parallel corpus. Furthermore, the dynamic generation of stopword lists has been suggested as an optimal method for maintaining high classification performance while reducing data sparsity and shrinking the feature space (Saif et al., 2014). These approaches demonstrate ongoing efforts to mitigate the impact of data sparsity on sentiment analysis, particularly in the context of Twitter data.

Prompt design. The design of prompts in LLMs is a crucial aspect of leveraging their full potential. There are many famous prompting methods. For example, Chain-of-Thought (CoT) (Wei et al., 2022) provides a few examples of reasoning steps in the prompt to let LLMs imitate. Zero-shot Chain-of-Thought (Kojima et al., 2022) let LLMs output "Let's think step by step" as the first sentence in their outputs to elicit their reasoning abilities. More related to us, Zhou et al. (2022); Drozdov et al. (2022) prompt LLMs to break down complex tasks into simpler ones, tackling them sequentially. Their approach primarily targets the inference phase. In contrast, our fact-imagination method doesn't decompose the problem into simple ones. Instead, it mirrors the natural process of composing a tweet: encountering or observing something, followed by sharing thoughts on social media.

3 METHOD

As the introduction mentions, AttrPrompt only summarizes vague and abstract attributes for tweets. Hence, we need to design a new prompting method from a different perspective. In this section, we present the design of our proposed method, the Fact-Imagination Prompt and other Prompt methods used to compare the Fact-Imagination Prompt, such as simple Prompt and Label-Conditional Prompt.

Fact-Imagination Prompt. Fact-Imagination Prompt, unlike AttrPrompt, uses reasons and events behind tweets to augment data. It comprises two parts, and the detailed prompt is provided in Figure 2. First, we upload a randomly sampled subset of the original dataset to the LLM and then instruct the LLM to analyze and reflect on the reasons and events behind those tweets. For instance, some events behind negative tweets are personal loss, health issues, and work-related stress. In the second step, we instruct the LLM to generate new data by giving these reasons and events in the prompt.

It is essential to highlight that the first step of the Fact-Imagination Prompt can be conducted either through the LLM (designated as "Fact-Imagination Prompt by LLM") or via manual analysis (termed "Fact-Imagination Prompt by Human"). We utilize and analyze both methods in the experiment step.

Baseline Prompt method. To compare the performance of the Fact-Imagination Prompt, we use Simple Prompt and Label-Conditional Prompt (Figure 3) as our baseline prompt methods. The Simple Prompt is a straightforward method where we directly ask the LLM to generate new data without specifying any particular constraints or themes. On the other hand, the Label-Conditional Prompt narrows the focus, directing the LLM to generate new data specifically for a specific class or category.

Data Augmentation. Our data augmentation process employs `gpt-3.5-16k-turbo` via the OpenAI API. For each prompting method: Fact-Imagination, Simple, and Label-Conditional Prompts, we generate 500 samples for Emotion in Text Tweets dataset, and around 1000 samples for the Covid-19 dataset. The specific API parameters used for generating the data are summarized in Table 1.

Table 1: GPT API Parameter Setting

Temperature	Max Tokens	Top P	Frequency Penalty	Presence Penalty
1	2048	1	0	0

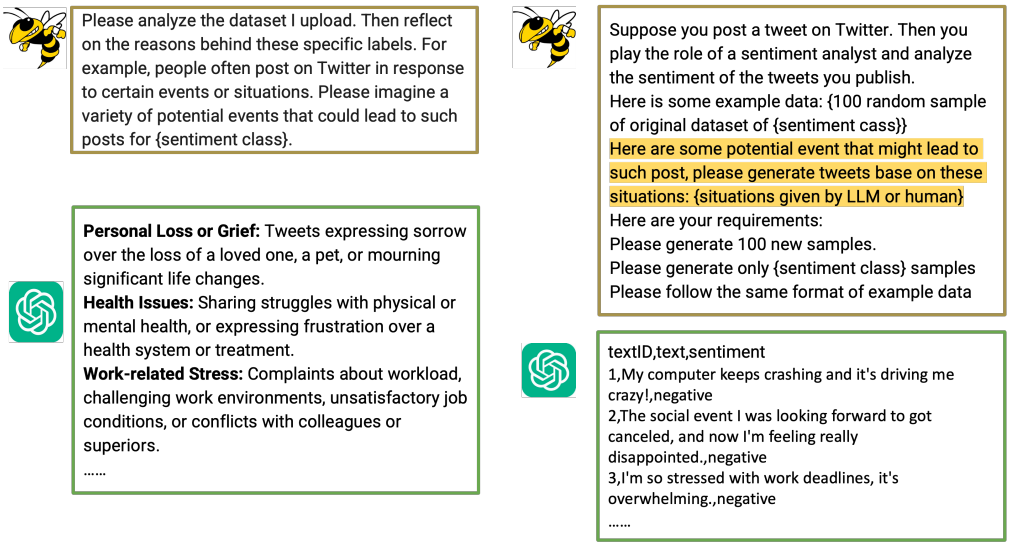

Figure 2: **Fact-Imagination Prompt**. We show the detailed prompt of the Fact-Imagination Prompt. The analyze and reflect step is shown on the left, and the second step, generating new data, is shown on the right. The highlighted part is the most important part, which makes the Fact-Imagination Prompt method different from other prompt methods.

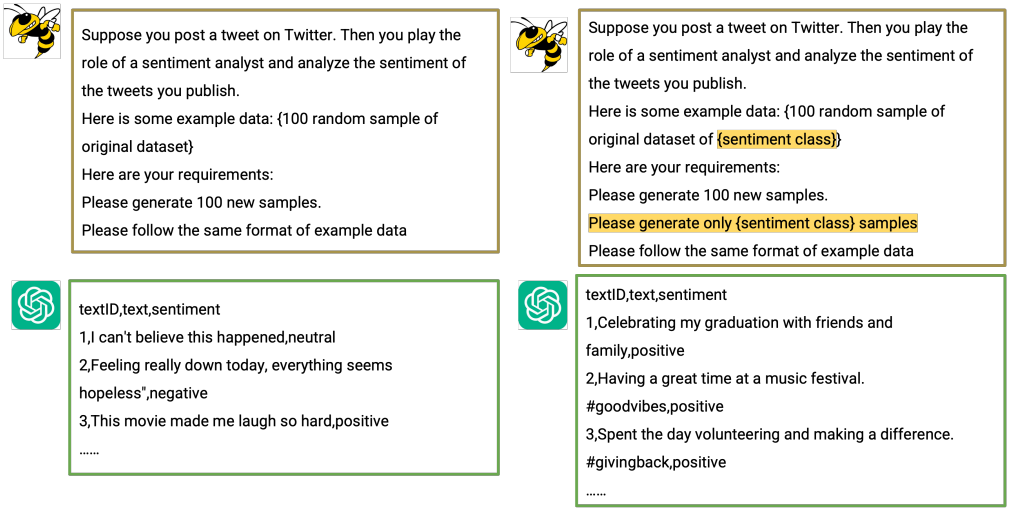

Figure 3: **Baseline Prompt methods**. We show the detailed prompts of (a) Simple Prompt, and (b) Label-Conditional Prompt. The part shows that the difference between the Simple Prompt and the Label-Conditional Prompt is that the Label-Conditional Prompt only generates data for one class each time.

Table 2: Parameters for Fine-Tuning

Learning Rate	Number of Epochs	Batch Size	Warmup Ratio	Weight Decay
0.0005	20	16	0.1	0.01

Model Fine-tuning. For the downstream sentiment analysis tasks, we employed `distilbert-base-uncased`[2], a light version of BERT known for fast training and inferencing with comparable performance, in fine-tuning. The model is trained on the cross-entropy loss. The hypermeters for training are listed in Table 2. 15% of the training data is used for validation. After running all the epochs, the model with the best performance in the validation set will be used for testing.

4 Experiments

4.1 Datasets

Our experiments are conducted based on the following two Twitter datasets for sentiment analysis tasks:

- **Emotion in Text Tweets**[3]: The Emotion in Text Tweets dataset contains publicly available ordinary tweets sourced from Twitter on a large diversity of topics. These tweets are labeled "negative," "neutral," and "positive."
- **Covid-19 Tweets**[4]: The Covid-19 dataset contains tweets related to covid-19 in 2020. The tweets are labeled "extremely negative," "negative," "neutral," "positive," and "extremely positive."

Note that we want to simulate the low availability of labeled data, so only subsets of these datasets are used to construct the training and validation sets. The sizes of the augmented data are roughly the same as the size of the original training set.

Table 3 summarizes the descriptive statistics of the datasets with and without data augmentation.

Table 3: Statistics of Datasets

Dataset	Aug. Method	Train + Valid Size	Test Size	Imbalance Ratio
Emotion in Text Tweets	None	500	1000	1.36
	Simple Prompt	1000	1000	1.67
	Label-Conditional Prompt	1000	1000	1.27
	Fact-Imagine by LLM	1000	1000	1.61
	Fact-Imagie by Human	992	1000	1.23
Covid-19 Tweets	None	1000	3798	1.92
	Simple Prompt	1920	3798	5.38
	Label-Conditional Prompt	1923	3798	1.88
	Fact-Imagine by LLM	1751	3798	1.50
	Fact-Imagine by Human	2000	3798	1.38

A metric of the label imbalance within each dataset was conducted by computing the imbalance ratio, defined as the occurrence of the most frequent label over the occurrence of the least frequent label. Data augmentation by Simple Prompt exacerbates label imbalance in the dataset. When the model generates data without guidance, it will generate a larger proportion of data for some specific labels. Further, as shown in Figure 4, the LLM tends to generate more data with non-'neutral' labels for Emotion in Text Tweets. The LLM tends to generate more data with the label 'extremely positive' than 'extremely negative' for Covid-19 Tweets. The tension in generating more positive data might be related to the human-value alignment for LLMs.

By contrast, data augmentation with the Fact-imagination Prompt guided by humans effectively mitigates label imbalance. Also, this method preserves the distribution of the labels in the real dataset.

[2] Sanh et al. (2020)

[3] https://www.kaggle.com/c/tweet-sentiment-extraction/data

[4] https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification

4.2 Diversity of Generated Data

To measure the diversity of the generated data, we define the similarity of each sample in the dataset as the cosine similarity of the embedding vectors by encoding each row with Sentence-BERT[5]. Then, we calculate the average pairwise similarity (APS) of the samples within each dataset. The lower the APS is, the higher the diversity of the data.

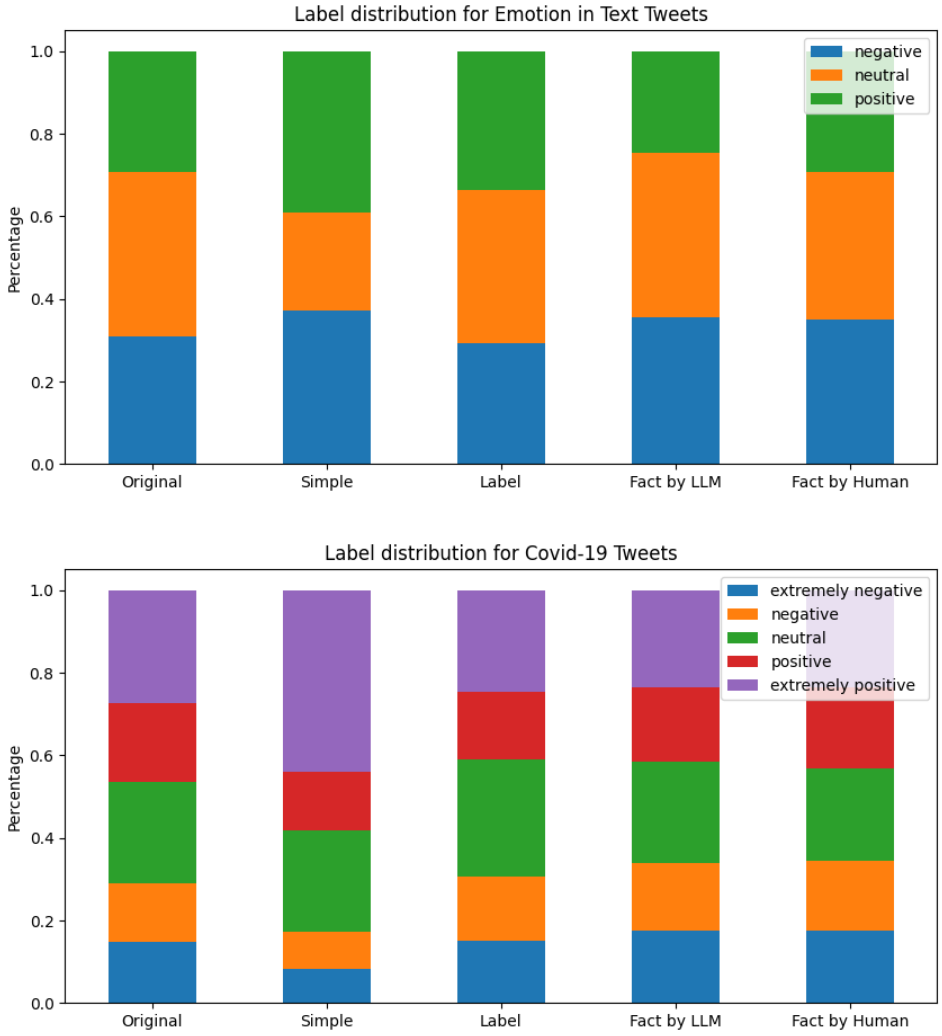

Figure 4: Label Distribution in each Dataset

As shown in Table 4, the original dataset without augmentation has the highest data diversity. The data generated by LLM has lower diversity than the actual data. As the model is autoregressive, when the output is sufficiently long, the model will inevitably generate repetitive content or rephrase what has been output before. The Label-Conditional Prompt yields data with the lowest diversity among the data generation prompting methods. Thus, when the LLM has the restriction of the label for generated data, its ability to be more diverse is limited. It is possible that given the context of a fixed label, the prediction branches easily converge to similar content, and thereby, the model would generate similar data in each round.

On the other hand, Fact-imagination Prompts are effective in increasing the diversity of the generated data. Adding an extra step to ask either LLM or humans to analyze the facts related to the data-

[5]Reimers & Gurevych (2019)

Table 4: Sample Diversity of the Datasets

Augmentation Method	Emotion in Text Tweets		Covid-19 Tweets	
	In-label APS	APS	In-label APS	APS
None	0.127	0.119	0.313	0.308
Simple Prompt	0.181	**0.141**	0.358	0.342
Label-Conditional Prompt	0.195	0.156	0.383	0.363
Fact-Imagination by LLM	**0.180**	0.145	0.353	0.335
Fact-Imagination by Human	0.185	0.142	**0.332**	**0.306**

generation process helps the model be more diverse. As shown by the increased diversity in fact-imagination guided by the LLM itself, it is beneficial to have an extra step that explicitly asks the model to analyze how to make the data more diverse. For fact-imagination guided by humans, especially in the Covid-19 dataset, humans are good at analyzing the facts behind the tweets written by the users based on their expertise or past experience. The additional information contributes to guiding the LLM to generate more diverse data.

4.3 Fine-tuned Model Performance

To evaluate whether data argumentation helps the Performance of downstream sentiment analysis tasks, we fine-tune the model described in the Method section on each original dataset and the ones with argumentation by each prompting method.

Figure 5 shows the F1 score on the test set after the model is fine-tuned on the datasets with and without augmentation. We can see that Simple Prompt has the best model performance. On the contrary, fine-tuning the data generated by Label-Conditional Prompt yields the worst Performance. Between the Fact-imagination Prompt guided by LLM and humans, the one by LLM itself works better than the one by humans. However, as shown in Table 5, only Simple Prompt positively impacts model performance. All other data augmentation methods degrade the Performance of the fine-tuned model.

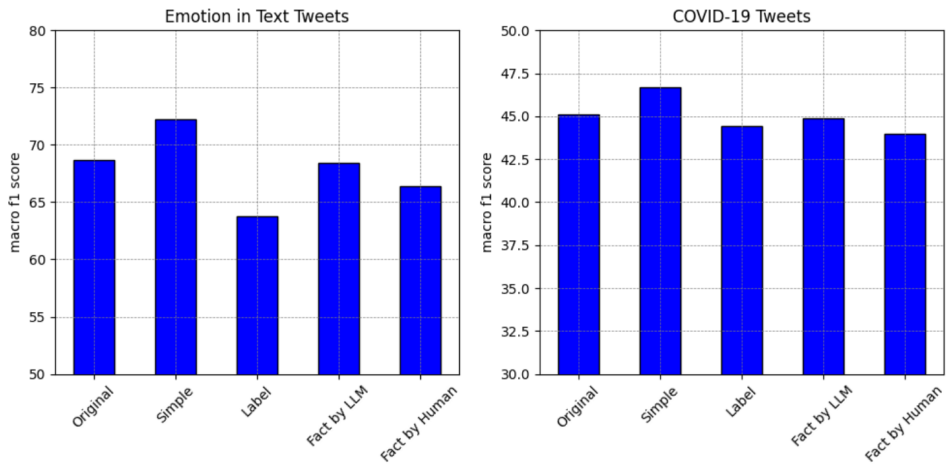

Figure 5: Comparison of macro F1 score for data augmentation methods in each dataset

The above results show that even though the data generated by Simple Prompt suffers from high label imbalance and moderately low sample diversity, it still pushes up the Performance of the model. By contrast, the data generated by the Fact-imagination Prompt did well in preserving the label distribution of the real dataset and achieved a high sample diversity. However, training on these data harms the Performance of the model. Thus, only focusing on mimicking the distribution of labels and forcing the model to generate more diverse data is not sufficient to benefit the Performance of the sentiment analysis task.

Table 5: Performance of the model on the augmented datasets

Augmentation Method	Emotion in Text Tweets		Covid-19 Tweets	
	Accuracy	F1	Accuracy	F1
Simple Prompt	71.90 +3.40	72.26 +3.58	46.29 -1.31	46.71 +1.59
Label-Conditional Prompt	63.40 -5.10	63.76 -4.92	45.47 -2.13	44.41 -0.71
Fact-Imagination by LLM	68.10 -0.40	68.39 -0.29	45.87 -1.73	44.90 -0.22
Fact-Imagination by Human	65.90 -2.60	66.37 -2.31	46.39 -1.21	43.98 -1.14

Intuitively, training on more diverse tweets would teach the model more information to perform well for the sentiment analysis tasks. However, the LLM may generate incorrect data that confuses the model. The LLM with alignment might be biased towards generating more positive text than it is supposed to be. Besides, the LLM also generates out-of-scope data. For example, we observe that LLM generates tweets that are way too formal and lengthy, which are rare in real life. These data are more diverse. However, they do not help the sentiment analysis of the real data or even introduce noises in the training set. Moreover, compared to Simple Prompt, for which the LLM might retrieve information from its knowledge, generating data by Fact-imagination Prompt is more difficult. In this case, the LLM is more likely to hallucinate or make mistakes. As a result, the accuracy of the generated data might be compromised.

5 Conclusion

Conclusion. In this work, we propose a prompting method named "Fact-Imagination Prompt" for generating synthetic Twitter data. The Fact-Imagination Prompt asks LLMs to imagine potential reasons and events underlying tweets and then use these imagined scenarios to produce new tweets. In the comparative analysis, we find that while the Fact-Imagination Prompt successfully increases the diversity of the generated data, it tends to compromise accuracy. This observation demonstrates that when implementing data augmentation using LLMs, it's crucial to balance the diversity and accuracy of the generated data.

Limitations. While Fact-Imagination Prompts increase data diversity, they often compromise accuracy. Data augmented by these prompts, despite being diverse, introduces noise and deviates from typical real-life tweet content. This leads to decreased model performance. In the future, Firstly, we plan to investigate how to evaluate the accuracy of the generated data for these prompting methods, either by humans or by a more advanced LLM like GPT-4, so that we can better understand the relationship between the quality of generated data and the performance drop of the fine-tuned model. Moreover, if there is a positive relationship between data accuracy and model performance, a natural question is how we could increase the accuracy of the generated data while preserving the favorable aspects such as label balances and diversity. A potential method is to use another LLM or train a machine learning model to help us classify and filter low-quality data. Using more sophisticated prompting methods might also help to model generate data with higher quality.

REFERENCES

Md S Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. Association for Computational Linguistics, 2018.

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. Mixture of soft prompts for controllable data generation. *arXiv preprint arXiv:2303.01580*, 2023.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *arXiv preprint arXiv:2203.09101*, 2022.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*, 2022.

Jiahui Gao, Renjie Pi, Lin Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. Self-guided noise-free data generation for efficient zero-shot learning. In *International Conference on Learning Representations (ICLR 2023)*, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pp. 9118–9147. PMLR, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Sergey I Nikolenko. *Synthetic data for deep learning*, volume 174. Springer, 2021.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410. IEEE, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Andy Rosenbaum, Pegah Kharazmi, Ershad Banijamali, Lu Zeng, Christopher DiPersio, Vivi Wei, Gokmen Oz, Clement Chung, Karolina Owczarzak, Fabian Triefenbach, and Wael Hamza. Calico: Conversational agent localization via synthetic data generation. In *NeurIPS 2023 Workshop on SyntheticData4ML*, 2023. URL https://www.amazon.science/publications/calico-conversational-agent-localization-via-synthetic-data-generation.

Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*, pp. 508–524. Springer, 2012.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 2020.

Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. *arXiv preprint arXiv:2104.07540*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. Zerogen: Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv:2202.07922*, 2022.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

A APPENDIX

Table 6: **Sample generated tweets for Emotion in Text Tweets dataset**. This table includes text ID, texts and their corresponding sentiments.

TextID	Text	Sentiment
1	A Second Head Thinking about what to have for lunch. Any suggestions?	neutral
2	Finished reading a really interesting article in the magazine. Highly recommend it.	neutral
3	Going to meet a friend for coffee later. Looking forward to catching up.	neutral
4	Just finished a workout. Feeling refreshed and energized.	neutral
5	Taking a break from work to watch a TV show. Excited to see what happens next.	neutral
6	Planning a weekend getaway to the beach. Can't wait for some relaxation.	neutral
7	Just got a new book. Looking forward to diving into it later.	neutral
8	Feeling so tired at work because I didn't sleep well. This headache won't go away.	negative
9	Feeling like a failure because I can't pass the driving test. It's so frustrating!	negative
10	Waiting for sleeping pills to kick in. Tomorrow's gonna be a tough day at work.	negative
11	This rainy weather is making me feel so tired. Can't wait for it to stop!	negative
12	Everyone is going out tonight, but I'm stuck at work until late. Feeling left out.	negative
13	I'm sorry if I disappointed you in the past few days. I didn't mean to.	negative
14	My knee hurts so bad after my coworker accidentally bumped into me. It's really painful.	negative
15	Finally finished writing my novel! So proud of myself for accomplishing this major goal.	positive
16	After months of hard work and dedication, I have lost 20 pounds! Feeling amazing and confident.	positive
17	Just received my diploma in the mail. Officially a college graduate! Celebrating this milestone in my life.	positive
18	I passed my driving test today! Can't believe I finally have my driver's license. Celebrating with a road trip.	positive
19	So excited for my upcoming vacation to Hawaii! Counting down the days until I can relax on the beach.	positive
20	Tickets to my favorite band's concert just arrived in the mail. Can't wait for the night of great music and unforgettable memories.	positive
21	Looking forward to a weekend getaway with friends. It's going to be a fun-filled adventure.	positive
22	Only a few more days until my birthday! Excited to celebrate with loved ones and enjoy some delicious cake.	positive
23	The new season of my favorite TV show is premiering tonight. Grabbing some popcorn and getting ready for a night of entertainment.	positive
24	Had an amazing time catching up with old friends over a delicious dinner. Love spending quality time with great company.	positive
25	Attended a surprise party for a dear friend today. Seeing their happy face made it all worth it.	positive