Data Augmentation via Large Language Models and UMLS for Few-shot Named Entity Recognition in Medical Texts

Anonymous ACL submission

Abstract

Few-shot learning with large language models holds substantial potential in the biomedical domain where obtaining extensive annotated data for specialized tasks can often be challenging. In the presence of small annotated datasets, incorporating domain knowledge from external sources is a common strategy. In this paper, we explore knowledge augmentation strategies for biomedical named entity recognition (NER) by incorporating information encapsulated in the Unified Medical Language System (UMLS). We leverage UMLS knowledge along with its hierarchical structure, and information from large language models (LLMs) to automatically generate new training examples in few-shot settings. We further explore the viability of employing GPT-3.5 for the extraction of biomedical named entities from Reddit data focused on prescription and illicit opioids. The results show an improvement of 13% on the F_1-score on average over five established NER datasets, and a 6% increase on the REDDIT-IMPACTS dataset after appropriate prompt engineering improvements. Our findings indicate that utilizing UMLS and LLMs as a joint source of prior knowledge can be a viable approach for improving the state of the art for few-shot learning-based NER in medical text.

1 Introduction

Few-shot learning in biomedical Natural Language Processing (NLP), contrasts with traditional lexicon-based approaches, which may struggle with lexical variants or ambiguous expressions in larger datasets, and deep learning models that require large amounts of data (Dong and Xing, 2018; Ge et al., 2022), makes it particularly useful for fine-grained classification tasks such as Named Entity Recognition (NER), where obtaining large amounts of data can be challenging. For instance, in medical diagnosis, few-shot learning has been used to develop models that can make accurate predictions with only a few examples (Sung et al., 2018; Lake et al., 2013), which is particularly beneficial for rare or new diseases (Jadon, 2021; Yoo et al., 2021).

Few-shot learning has shown promising results in NER tasks. The extensive popularity and adoption of Large Language Models (LLMs) like Generative Pre-trained Transformer (GPT (Radford et al., 2018; Brown et al., 2020)) recently, also offer a valuable chance to assess the capabilities of LLMs in few-shot learning scenarios (Brown et al., 2020), by generating human-like texts and providing external knowledge with limited examples. This adaptability is further enhanced by prompt-based strategies, where Well-designed prompts can significantly improve accuracy (Zhao et al., 2021; Li and Liang, 2021a). The strength of LLMs lies in their ability to generate text and provide contextually relevant responses to queries. However, they might face challenges when dealing with specialized medical concepts, as they may lack domain-specific knowledge and may generate unrealistic or incorrect medical information (Yunxiang et al., 2023). To address this, we propose incorporating domain knowledge from the Unified Medical Language System (UMLS) (Bodenreider, 2004), a comprehensive database of medical terminologies to enrich the representations of sparsely occurring medical concepts and enhance the performance in few-shot learning for biomedical NER tasks.

In this paper, our contributions are three-fold:

C1. We explore knowledge augmentation methods using UMLS for improving NER in few-shot settings. Specifically, we used the encapsulated knowledge of UMLS and its hierarchical structure to enhance the training data when dealing with rare entities.

C2. We leverage in-context information generated by GPT-3.5 to supplement UMLS knowledge to boost NER performance in the clinical domain.

C3. We introduce a new dataset, REDDIT-

IMPACTS, collected from subreddits (Reddit forums) focused on prescription and illicit opioids, and medications for opioid use disorder. We also explore the viability of employing GPT-3.5 for the extraction of named entities from REDDIT-IMPACTS dataset by utilizing prompt engineering techniques and integrating background knowledge from the UMLS.

The results show an improvement of 13% on the F_1 score on average over five NER datasets, as compared to the baseline model. Our findings indicate that utilizing UMLS and LLMs as a source of prior knowledge can be a viable approach for improving the state of the art for few-shot learning-based NER in medical text. Also, LLMs with appropriate prompts can significantly improve the performance where there are only few annotated samples.

2 Related Work

Early research in few-shot learning within biomedical NLP faced challenges due to the complexities of natural language data containing domain-specific terminologies and associations. Prior knowledge has been identified as crucial (Schmidt et al., 2015) to combat this issue. FSL methods based on the use of prior knowledge are categorized into data, model, and algorithm types (Wang et al., 2020). Initial approaches, such as meta-learning and metric learning, leveraged prior knowledge at different levels to generalize to new tasks with limited data (Schmidt et al., 2015). Meta-learning (Hospedales et al., 2021) has perhaps been the most common framework for early stage of FSL research, which is trained using a set of training tasks. Other approaches such as matching networks (Vinyals et al., 2016), which use embedding functions to generalize knowledge, prototypical networks (Snell et al., 2017), which generate prototype representations of classes to address overfitting issues, and transfer learning were considered mainstream directions in the field of few-shot learning (Pan and Yang, 2009), prior to the extensive exploration of LLMs in few-shot learning.

The advent of LLMs has shifted the focus towards prompt-based learning, which has shown promise in few-shot NLP (Prato et al., 2020; Liu et al., 2023). This approach relies on the ability of LLMs to generate text and provide contextually relevant responses with minimal training data, offering a promising solution to the challenges posed by few-shot settings, particularly in the clinical domain. The use of data from social media for biomedical tasks presents additional complexities in the form of noisy data and inconsistent language usage (Hu et al., 2024). The potential of LLMs and prompt-based strategies in overcoming the challenges posed by few-shot settings has been demonstrated with techniques like LM-BFF (Li and Liang, 2021b) utilizing prompts to fine-tune models on limited data. Additionally, approaches like PPT (Gu et al., 2022) enhance prompt effectiveness by pre-training prompt token representations with unsupervised data. Augmenting knowledge from biomedical and clinical knowledgebases such as UMLS have also been explored (Michalopoulos et al., 2021; Alsentzer et al., 2019) and found to outperform general purpose pre-trained language models. Leveraging knowledge from both domain-specific knowledgebases and in-context information extracted by LLMs, however, is relatively nascent.

3 Methods

First, our aim is to compare and contrast the dynamically generated linguistic structures from LLMs, with those systematically extracted from the UMLS. We explored the usage of hierarchical information and structured knowledge encapsulated in the UMLS and its semantic networks to automatically retrieve concepts related to the named entities in the few-shot training data. Considering the generative capabilities of GPT-3.5, which enable it to produce contextually relevant linguistic structures, we also utilize GPT-3.5 to provide external knowledge. These related concepts are added to the few-shot training data to create additional synthetic instances. The synthetic instances and the original few-shot training data are then used to train models and compare their performances.

Second, we explore the feasibility of employing GPT-3.5 for extracting named entities, specifically clinical and social impacts, from our REDDIT-IMPACTS datasets. We proposed a task-specific prompt to refine the prompt engineering process, and to evaluate GPT-3.5's performance in comparison to conventional few-shot learning models on biomedical-related social media datasets.

3.1 Hierarchy Information and Semantic Network in UMLS

One of the key features of the UMLS is its hierarchical organization of concepts, which provides a

way to access information about the parents, children, and siblings of a given concept. In addition to the hierarchy, the UMLS also includes a semantic network that describes the relationships between concepts based on their semantic similarity. The hierarchy information in the UMLS represents the relationships between concepts in a hierarchical structure (Mishra et al., 2018), similar to a tree. This hierarchical structure allows for easy navigation of the UMLS and helps to organize and categorize concepts based on their relationships. The hierarchy is based on a number of different types of relationships between concepts, including 'isa' (is a), 'has_parent' and 'has_child' relationships. Figure 1 shows an example of the tree-like structure of the UMLS.

The semantic network[1] in the UMLS, represents the relationships between concepts based on their semantic similarity, rather than their hierarchical relationships. A portion of the UMLS semantic network is shown in Figure 2. The semantic network is organized into a set of categories, such as 'Anatomy', 'Chemicals and Drugs', and 'Physiology', each of which represents a different area of biomedical and health-related concepts (Lindberg et al., 1993). Within each category, concepts are further organized based on their relationships to other concepts, such as 'isa' relationships or 'part_of' relationships.

Both the hierarchy information and the semantic network are important for understanding the relationships between concepts in the UMLS. The hierarchy allows for navigation and understanding of general relationships, while the semantic network provides insight into specific relationships based on semantic similarity. Together, these two approaches help to provide a comprehensive understanding of the relationships between concepts within the UMLS.

3.2 Data Augmentation by UMLS and GPT-3.5

Our approach utilizes the UMLS for generating new examples in several ways. When faced with entity types with small numbers of labeled samples, we use the knowledge encoded in the UMLS to expand the training data and add synthetic examples into the training set, so that the original few-shot training set is expanded to a larger one. Specifically, we incorporate knowledge in multiple layers.

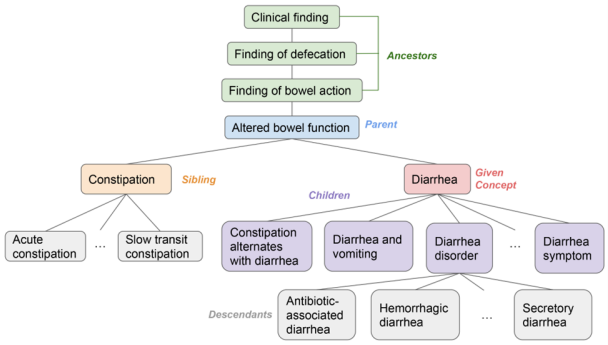

Figure 1: A subtree of the hierarchical structure of concept "diarrhea" in SNOMEDCT_US dictionary (Systematized Nomenclature of Medicine–Clinical Terms).

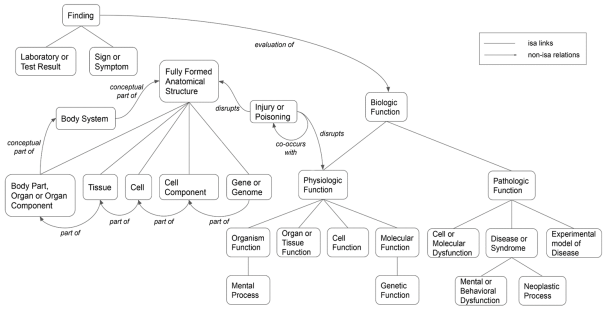

Figure 2: A portion of the UMLS semantic network. Isa links and non-isa relations are represented in the figure, respectively.

The first layer consists of lexical expressions with the same UMLS concept IDs (typically referred to as concept unique identifiers or CUIs), which are added to create synthetic examples. Thus, this layer of knowledge augmentation adds potential synonyms of the original named entities in the training data.

The second layer of expansion consists of augmenting the training data from the first layer with additional closely related CUIs that are under the same UMLS semantic type (broad category of concepts, such as pharmacological substances). This layer, thus, adds additional examples that are likely to be conceptually closely related to the entities in the training data, and thus, are likely to occur in similar contexts in medical free texts. The third layer of augmentation considers the hierarchical associations in medical concepts. Specifically, we utilize the parent-child relationships between concepts and extract parents, children, and siblings of given concepts based on the SNOMEDCT_US dictionary (Systematized Nomenclature of Medicine–Clinical Terms), which is a comprehensive clinical termi-

[1] https://www.ncbi.nlm.nih.gov/books/NBK9679/

nology that is widely used in the healthcare industry (Benson, 2012).

Contrary to the extraction of linguistic structures to a given concept from the UMLS, our approach in utilizing GPT-3.5 for generating new examples involves providing the model with complete sentences. This strategy enables GPT-3.5 to leverage contextual information to enhance its comprehension of the given concepts, thereby facilitating the generation of semantically coherent new examples. Furthermore, we strictly control the number of generated examples to match the quantity extracted from the UMLS, ensuring that the results are not influenced by discrepancies in the volume of training data. In the use of prompts, we adopt a fundamental prompt strategy, which involves providing the sentence itself, indicating its task and the expected output format, while mandating that it generates based on the knowledge from the UMLS. The overall architecture of our data augmentation approach is shown in Figure 3.

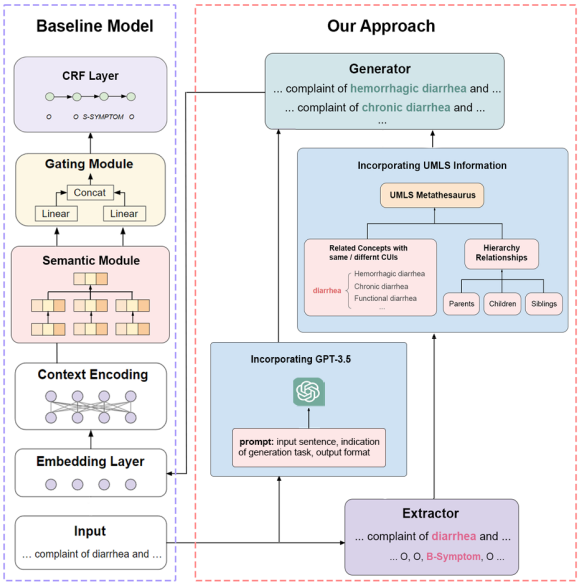

Figure 3: The overall architecture of our data augmentation approach. An example sentence is shown, where the sentence is directly fed into GPT-3.5, while the word "diarrhea" is extracted from the training data by the Extractor, then its related information is retrieved by using UMLS Metathesaurus, and a Generator module expands the training data. After that, new training data is put into the encoder for training.

3.3 GPT-3.5 with Prompt Engineering

3.3.1 Prompt Engineering

Figure 4 illustrates the components of the prompts used for GPT models and the main workflow of our experiment. We designed task-specific prompts for use with GPT-3.5 which comprises the following components:

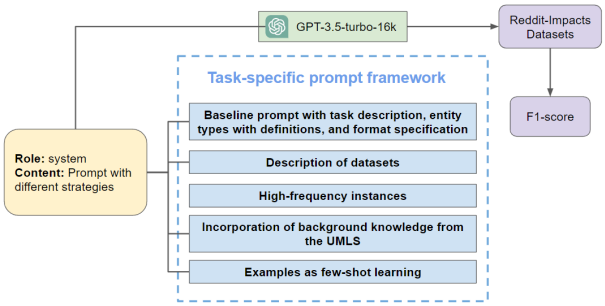

Figure 4: An overview of our study workflow. Prompt with different strategies are delivered to GPT-3.5-turbo-16k model, then the predictions are returned back from API for evaluation.

1. Baseline prompt with task description, entity types with definitions, and format specification: The baseline component provides the LLM with essential information regarding the basic aims of the task, which is extracting and classifying entities. The categories of labels present in the dataset along with elucidation of their definitions, as we only focused on these two entity types. Entity definitions provide detailed and unequivocal explanations of an entity in the context of a specific task, crucially guiding the LLM towards accurately pinpointing entities within textual documents. Also, we provided the input and output format in which the LLMs are expected to deliver results, which is a crucial step in ensuring the successful completion of the task, and it presents a greater challenge for NER problems.

For GPT-3.5, NER present more challenges as it is essentially a sequence-to-sequence problem, where each token is assigned a corresponding label. However, if our query only provides a sentence, it is difficult for GPT-3.5 to directly and accurately assign labels to each token. As it is challenging to ensure that the number of labels in the output matches the number of words in the input sentence. One issue is that GPT-3.5 has its own tokenization mechanisms which may differ from what we expected. Providing labels in the IOB format also adds to the complexity. Moreover, handling punctuation often presents difficulties in this context.

A common approach is to provide a sentence and indicate the entities within it (Xie et al., 2023). For example, in the sentence "I was a codeine addict," the phrase "codeine addict" is identified as an entity and is annotated as a 'Clinical Impacts' entity in this task. However, this format can become ambiguous when facing with long sentences contain the same word or phrase multiple times, each with different contextual meanings, not all of which may be labeled as entities. A subsequent method involves providing spans corresponding to the entities (Hu et al., 2024), but this adds significant challenges for GPT models, as mismatches between spans and entities can frequently occur.

To address these challenges, we adopt a new format for constructing the input and output for the GPT model. We provide GPT-3.5 with a list of tokens that have already been tokenized. For the output, we instruct the model to return each token along with its corresponding label and concatenate together. This method allows us to easily extract labels for evaluation, and it ensures a one-to-one correspondence between the predicted labels and tokens, with the number of labels always consistent with the number of tokens in the input sentence.

For instance,
Input: ['I', 'was', 'a', 'codeine', 'addict', '.']
Output: ['I-O', 'was-O', 'a-O', 'codeine-Clinical Impacts', 'addict-Clinical Impacts', '.-O']

2. Description of datasets: By describing the dataset's origin, content, and themes, we aim to provide GPT-3.5 with a basic understanding of the dataset, which can further refine the filtering of predicted entities. For example, the REDDIT-IMPACTS dataset focuses on individuals who use opioids, and we are interested in the impact of opioid use on their health and life.

3. High-frequency instances: Given that clinical impacts and social impacts are relatively abstract concepts, unlike entities such as medicines or symptoms which have clear definitions, the determination of clinical and social impacts is more ambiguous. Therefore, we provide the most frequently occurring words or phrases in these two entity types within the training dataset to assist GPT-3.5 in understanding the potential distribution of entities and the theme of the text for this task.

4. Incorporation of background knowledge from the UMLS: We provided GPT-3.5 with integrating the comprehensive and structured information provided by UMLS into the analysis or processing of medical data. We expect this knowledge can enhance the understanding and interpretation of medical concepts, relationships, and terminologies.

5. Annotated samples: To further assist the LLMs in understanding the task and generating accurate results, we provided a set of annotated samples to improve its performance in a few-shot learning setting. We randomly selected either 5 annotated examples for each entity types from the training set and formatted them according to the task description and entity markup guide.

In this work, we compare the effectiveness of different prompt components by incrementally incorporating description of datasets, high-frequency instances, background knowledge from the UMLS and annotated samples.

4 Datasets

We conduct our data augmentation experiments on five medical text datasets in 5-shot settings, which included: (i) MIMIC III (Medical Information Mart for Intensive Care) dataset (Johnson et al., 2016); (ii) the N2C2 2018 shared task track (Henry et al., 2019); (iii) the BC5CDR Disease dataset (Li et al., 2016); (iv) the MedMentions dataset (Mohan and Li, 2019); and (v) the NCBI Disease dataset (Doğan et al., 2014). In addition to our experiments on these five datasets, we also introduce the REDDIT-IMPACTS dataset, which we use to evaluate the performance of our proposed approach on biomedical data obtained from social media.

Reddit communities have been found to serve as a means of social support for people who use drugs, for both prescription and non-prescription usage. We identified 14 opioid-related subreddits spanning discussions on prescription and illicit opioids, and collected all retrievable posts using the Python-Reddit API Wrapper for Reddit[2]. We extracted unique users from the retrieved posts, resulting in a cohort of users who had posted on the selected sub-reddits. We selected a random sample of these Redditors (N=13,812) and collected each of their past public posts across all subreddits (i.e., timelines), between November 2006 (corresponding to the earliest post available) and March 2019 (corresponding to the last date of data collection). From this, we randomly selected 40 Redditors and manually reviewed and annotated 26,126 posts.

We applied natural language processing to generate lexical variants of all included prescription and

[2]https://praw.readthedocs.io/en/latest/

illicit opioids and stimulants, and detect mentions of them on the chosen subreddits. We developed annotation guidelines to categorize the words or phrases in posts into 33 entity types and manually annotated a set. The annotation process was iterative and involved several rounds. We first explored the posts, developed the annotation guidelines, and then annotated the texts. The authors then discussed the disagreements, updated the guideline, and re-annotated the posts.

However, the annotation of this large dataset revealed that some concepts, such as medicine intake, illicit drug use, were relatively easy to find and annotate. Therefore, for our few-shot learning research, we focused on two entity types in our dataset, which belonged to the category with the least number of samples: clinical impacts and social impacts, and named the dataset containing only these two entity types as REDDIT-IMPACTS dataset.

From this sparse dataset, we randomly extracted totaling 1,380 posts for our experiments, including 843 posts for training, 259 for validation, and 278 for testing. Among these, 27.8% of the posts contain words or phrases marked as clinical or social impacts, with 184 entities annotated as clinical impacts and 67 entities in social impacts.

5 Results and Discussion

5.1 Comparison of incorporating UMLS and GPT-3.5

Our primary findings were that incorporating prior knowledge via UMLS outperforms baseline few-shot models on all included datasets. The results in Table 1 show that using the UMLS and its hierarchical information helps the model utilize the parent-child relationships between concepts for NER. Similarly, models enhanced with GPT-3.5 information show improved performance, especially in the BC5CDR disease and MIMIC III datasets. The most significant improvements are observed when incorporating related concepts and parents and children, indicating the importance of hierarchical and semantic relationships in entity recognition. In comparison to the baseline models, our model achieved higher F_1-scores: on average, our results showed an improvement of around 13% compared to the baseline models. The improvement is particularly noticeable in difficult cases where the baseline models struggled to make accurate predictions.

Comparing the use of GPT-3.5 and the incorporation of UMLS, both approaches show improvements. However, GPT-3.5 tends to perform better in the most datasets, suggesting its strength in understanding complex clinical text. UMLS incorporation shows more consistent improvements across all datasets, as it provides a solid foundation for identifying and categorizing entities based on established medical vocabularies, indicating its usefulness in providing structured medical knowledge.

This approach significantly improves the performance of the model, especially for cases with only a few samples. In addition, these results demonstrate the effectiveness of our approach and its potential for use in real-world applications. Importantly, the augmented models consistently outperform the baseline model. The integration of external knowledge sources like UMLS and GPT models appears to be beneficial for NER in clinical texts.

Our motivation behind leveraging UMLS information and utilizing GPT-3.5 for data augmentation to improve the performance of few-shot learning in medical text datasets is manifold. First, the UMLS is a comprehensive ontology that contains a vast amount of knowledge about biomedical and health-related concepts. By using this information as data augmentation, we can provide our few-shot learning model with access to domain-specific knowledge that is not present in the training data. This can help to improve the model's ability to make accurate predictions on unseen data (Tian et al., 2020). Second, GPT-3.5 might offer a more context-aware understanding of entities, potentially leading to richer and more nuanced entity recognition. Therefore, we could effectively increase the amount of training data available to our model without the need for additional manual annotations. This can help to improve the model's performance on rare or complex cases that may not be well-represented in the original training data.

Overall, using information from UMLS and GPT-3.5 as data augmentation can help to address some of the key challenges in few-shot learning in medical text datasets, including limited training data and the need for domain-specific knowledge. This approach has the potential to improve the performance of few-shot learning models in a variety of medical applications. This is also a high-utility use case for knowledge sources that have been developed and maintained for decades for supporting biomedical NLP research.

Baseline Model	Incorporation	Models	N2C2 2018	MIMIC III	BC5CDR disease	Med-Mentions	NCBI disease
SANER (Nie et al., 2020)	**N/A**	**SANER (baseline model)**	10.27	21.25	37.25	52.12	22.11
	with UMLS	**Incorporate related concepts**	17.37	34.44	54.71	55.83	28.49
		Incorporate parents and children	**26.62**	33.79	51.12	53.29	26.31
		Incorporate siblings	23.98	32.65	54.82	53.94	27.19
	with GPT-3.5	**Incorporate related concepts**	18.54	28.95	55.96	54.57	**30.92**
		Incorporate parents and children	20.93	32.23	54.94	56.12	26.80
		Incorporate siblings	22.10	26.58	53.77	55.88	27.85
DANN (Ge et al., 2024)	**N/A**	**DANN (baseline model)**	0.21	19.68	35.75	52.40	22.38
	with UMLS	**Incorporate related concepts**	11.36	34.35	55.04	55.87	26.88
		Incorporate parents and children	16.58	32.33	50.86	54.77	25.04
		Incorporate siblings	17.53	30.77	55.31	52.62	28.68
	with GPT-3.5	**Incorporate related concepts**	13.76	28.19	**56.72**	**58.90**	30.16
		Incorporate parents and children	16.27	**34.63**	56.48	57.46	26.88
		Incorporate siblings	20.41	26.45	54.16	58.31	27.57

Table 1: F_1-scores of incorporating UMLS or GPT-3.5 compared with baseline on five medical datasets. All of the results used BERT (Devlin et al., 2019) as the embedding layer, and the best performances are highlighted in bold.

5.2 Performance on REDDIT-IMPACTS Dataset

The results in Table 2 show that the best performance we received so far on this new dataset is 54.36% by using DANN model with full training data. Both the SANER and DANN models struggle in the 5-shot setting, when incorporating related concepts from GPT-3.5, the performance improves slightly. As our datasets were collected from social media, which means the texts are often informal, unstructured, and contains abbreviations and typos.

In the few-shot settings, the performance of GPT-3.5 in making predictions is surprising and impressive. Compared to other few-shot learning models, GPT-3.5's performace significantly outperforms others when provided with only 5-shot training data. Although these results are lower than those obtained using the full training dataset, it is reasonable. The inability to utilize the full dataset stems from the current constraints of the OpenAI platform, which does not facilitate direct fine-tuning on our own datasets, coupled with restrictions on the lengths of prompts and outputs. Therefore, it might be prudent to provide GPT-3.5 with an augmented volume of training data, to explore the upper limit of GPT-3.5 on this new dataset.

Figure 5 shows the performance on GPT-3.5 with different prompt-based strategies. We found that, providing with description of the datasets, high-frequency instances from the dataset or more examples with annotations all improves the performance significantly, which indicates that aspects such as a clear description of the dataset's origin, content, themes, and high-frequency instances are very helpful for GPT-3.5 to understand the task and identify entities. Notably, the impacts of elucidating these elements are comparable to providing more annotated data. However, it is surprising that merely informing GPT-3.5 about UMLS-related knowledge in the prompt leads to worse results. A potential reason could be that simply stating the use of UMLS knowledge without effectively integrating it into the prompt may not provide GPT-3.5 with sufficient context to properly utilize this information, potentially leading to ambiguity in its predictions. Furthermore, it was observed that GPT-3.5 predicted a large number of spurious entities. "Spurious" here means that these entities were not annotated as such in the golden tests, but GPT-3.5 identified them as belonging to either clinical impacts or social impacts. One reason is that the language used in social media posts can be ambiguous, and without clear boundaries for what constitutes a clinical or social impact.

6 Conclusions and Future Work

Few-shot learning approaches have substantial promise for NLP in the medical domain, as many medical datasets naturally have low numbers of annotated instances. Our experimental results demonstrated that using UMLS and GPT-3.5 to incorporate prior knowledge is a possible solution for ex-

Conduct on GPT-3.5	Size of Training Data	Models	F_1-score
Few-shot learning models	**Full training data**	SANER	49.26
		DANN	**54.36**
	5-shot	SANER	0.00
		DANN	0.00
		Incorporate related concepts from GPT-3.5 based on SANER	5.41
		Incorporate related concepts from GPT-3.5 based on DANN	4.64
GPT-3.5 with different prompt-based strategies	**1-shot**	Basic Prompts	16.73
		Incorporate with description of datasets	21.15
		Incorporate with high-frequency instances	21.15
		Incorporate with background knowledge of UMLS	16.29
	5-shot	Incorporate with 5-shot examples	21.69
		Incorporate with all above	**22.90**

Table 2: Performance on Reddit-Impacts Datasets by using few-shot learning methods and GPT-3.5 with different prompts.

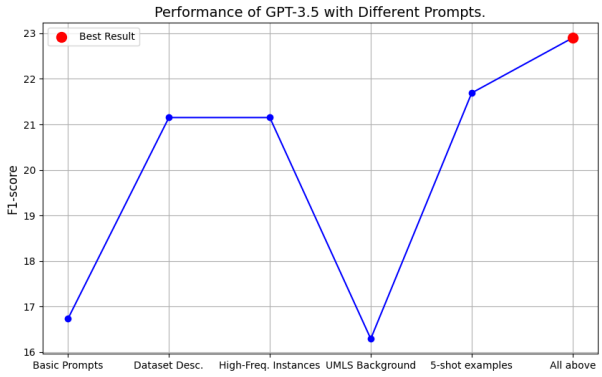

Figure 5: Performance of GPT-3.5 with Different Prompts.

ploring few-shot learning methods on medical text. Furthermore, we proposed a new dataset which was collected from Reddit, our experimental results reveal that GPT-3.5 significantly outperforms other models in few-shot settings. Additionally, by tuning the prompts, we have successfully achieved a notable improvement in GPT-3.5's performance on dataset. This indicates that appropriate prompt tuning strategies can significantly improve the performance of large language models on specific tasks, especially in scenarios with limited annotated samples.

Our future work will focus on exploring a wider variety of prompting techniques to augment specialized biomedical knowledgebases. We also plan to study the potential of small LLMs and quantization techniques in the context of few-shot learning. By investigating how these approaches can be applied to develop efficient and effective models, we aim to advance the state-of-the-art in few-shot natural language processing tasks while addressing the computational constraints associated with large-scale language models.

Limitations

While our model has shown significant improvement in results on medical text datasets, especially when annotated examples are scarce, its performance is still far from reaching the state-of-the-art level. Future research still has considerable room for enhancement. Another limitation stems from the study of prompt engineering; attempting to perfect prompts from the perspective of providing more useful information is challenging and task-specific. More general methods of generating prompts are already leading us to further research. Furthermore, due to the limitations of the API, we are unable to directly fine-tune GPT on other datasets. Therefore, the emergence of small LLMs and the development of quantization are also directions for our future research.

Acknowledgements

Redacted for anonymity.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and

Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tim Benson. 2012. *Principles of health interoperability HL7 and SNOMED*. Springer Science & Business Media, Berkshire, UK.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl-1):D267–D270.

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.

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.

Nanqing Dong and Eric P. Xing. 2018. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Data augmentation with nearest neighbor classifier for few-shot named entity recognition. In *MEDINFO 2023—The Future Is Accessible*, pages 690–694. IOS Press.

Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker. 2022. Few-shot learning for medical text: A systematic review. *arXiv preprint arXiv:2204.14081*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. 2021. Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Shruti Jadon. 2021. Covid-19 detection from scarce chest x-ray image data using few-shot deep learning approach. In *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, volume 11601, pages 161–170. SPIE.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2013. One-shot learning by inverting a compositional causal process. *Advances in Neural Information Processing Systems*, 26.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.

Xiang Lisa Li and Percy Liang. 2021a. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021b. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51.

Juhua Liu, Qihuang Zhong, Liang Ding, Hua Jin, Bo Du, and Dacheng Tao. 2023. Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2629–2642.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

A. Mishra, V. Verma, M. Reddy, A. S., P. Rai, and A. Mittal. 2018. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380, Los Alamitos, CA, USA. IEEE Computer Society.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with {umls} concepts. In *Automated Knowledge Base Construction (AKBC)*.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. 2020. Fully quantized transformer for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Hiemke Katharina Schmidt, Martin Rothgangel, and Dietmar Grube. 2015. Prior knowledge in recalling arguments in bioethical dilemmas. *Frontiers in psychology*, 6:1292.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1199–1208.

Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. 2020. Differentiable meta-learning model for few-shot semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12087–12094.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, pages 1–34.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Tae Keun Yoo, Joon Yul Choi, and Hong Kyu Kim. 2021. Feasibility study to improve deep learning in oct diagnosis of rare retinal diseases with few-shot classification. *Medical & Biological Engineering & Computing*, 59:401–415.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.