Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yao Ge Date

Advanced Sparse Concept Detection and Recognition in Biomedical Texts via Few-Shot Learning Algorithms

By

Yao Ge
Doctor of Philosophy

Computer Science and Informatics

Abeed Sarker, Ph.D.
Advisor

Joyce C. Ho, Ph.D.
Committee Member

J. Lucas McKay, Ph.D.
Committee Member

Mohammed Ali Al-Garadi, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

Date

Advanced Sparse Concept Detection and Recognition in Biomedical Texts via Few-Shot Learning Algorithms

By

Yao Ge
B.S., Hainan University, Haikou, Hainan Province, China, 2016
M.S., Shandong University, Jinan, Shandong Province, China, 2019

Advisor: Abeed Sarker, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Abstract

Advanced Sparse Concept Detection and Recognition in Biomedical Texts via Few-Shot Learning Algorithms
By Yao Ge

Many natural language processing (NLP) problems involving biomedical texts have limited annotated data available. Traditional supervised machine learning and deep learning algorithms require large volumes of annotated data and underperform with small annotated datasets. Few-shot learning (FSL) methods aim to enable effective learning in the absence of large annotated datasets, but the performances of FSL-based NLP methods are suboptimal, particularly for biomedical texts, limiting their application in real-world settings. The overarching objective of this thesis is to rigorously validate the current state-of-the-art in FSL methods for named entity recognition (NER) from biomedical texts and to propose novel FSL approaches that can improve upon the state-of-the-art methods.

Given the emerging interest and early-stage development of FSL approaches in biomedical NLP, we conducted a systematic review and benchmarking of existing methods, revealing their underperformance on most biomedical datasets. To address data sparsity problems in FSL, we proposed a novel method combining data augmentation with a nearest neighbor classifier (DANN). We extended this method by adding a synthetic data generation module (HILGEN) that leverages hierarchical information of the Unified Medical Language System (UMLS) and information generated by large language models (LLMs). Finally, building on progress made in recent times, we further enhanced NER performance by leveraging LLMs with prompt engineering and a dynamic prompting strategy involving retrieval-augmented generation (RAG).

These methods improved NER performance across multiple datasets in FSL settings, including MIMIC III, NCBI disease, BC5CDR, and a dataset (Reddit-Impacts) specifically created as part of this research. For example, on MIMIC III in a 5-shot setting, BERT's near-zero F1 score improved to 19.69 with our DANN model, 58.68 with HILGEN-generated synthetic data, and 76.24 using RAG-based dynamic prompting. Similar gains were observed across other datasets. Our research demonstrates that combining enriched data representation, domain knowledge, synthetic data, and context-aware prompting effectively addresses data sparsity, enhancing biomedical NER in FSL settings. These advancements mark significant progress toward operationalizing FSL-based NER systems for biomedical applications.

Advanced Sparse Concept Detection and Recognition in Biomedical Texts via Few-Shot Learning Algorithms

By

Yao Ge
B.S., Hainan University, Haikou, Hainan Province, China, 2016
M.S., Shandong University, Jinan, Shandong Province, China, 2019

Advisor: Abeed Sarker, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Abeed Sarker. He has always been supportive, and give me constructive suggestions on my research. I always thought I am very lucky and very happy to be his first Ph.D. He's been really encouraging both in work and in life. Thank you for your patience, encouragement, and guidance. I am truly honored to have been part of your lab.

I would also like to express my sincerely appriaciate to all my committee members, Dr. Joyce C. Ho, Dr. J. Lucas McKay, Dr. Mohammed Ali Al-Garadi. From my qualifying exam to my prospectus and finally to my defense, you have provided invaluable guidance, support to me, as well as understanding. Your feedback and encouragement at every stage have been essential to my growth, and I am deeply grateful for your contributions to my journey.

To all the members of our lab and our collaborators, thank you for creating such an inspiring and supportive environment. I truly love the atmosphere of our lab and feel so fortunate to have met everyone over the years. A special thank-you goes to Dr. Sudeshna Das and Yuting Guo, who have been like two little suns shining brightly in my life. Sudeshna, you have been more than just a colleague—you're practically my co-mentor. Your incredible support as a postdoc, from revising my CV and papers to helping me prepare for presentations, has been invaluable. And to my dear Yuting, I owe you so much—not just for your help with experiments and your care in daily life, but also for recommending me to join this group in the first place. Your influence played a huge role in shaping my research direction, and I'm so grateful for it. I love this field, and I have you to thank for inspiring me to pursue it!

To my friends in the lab, thank you for making every day brighter. Sahithi and Swati, thank you for the daily chats, the shared laughter, and for always joining us for Chinese food! Jeanne, your incredible desserts and thoughtful parties brought us all closer together. Jamor, your energy and positivity have always kept the lab lively

and warm. It has been a joy to see our lab grow from just three members to twelve—I feel so lucky to have met each and every one of you.

To my friends, both old and new, thank you for your constant support and joy. Whether we met at Emory, at conferences, or elsewhere, every conversation with you has been both enriching and heartwarming. A special thank-you to my roommate Sa Suo, a passionate and sincere friend who listen to me a lot and accompany with me a lot during the past five years. I've known Sa since orientation, and her influence is woven into nearly every aspect of my five years here. Another special thank-you to my dearest friend Jun Li, who has been my confidant and emotional anchor for over a decade. Despite the 12-hour time difference, our daily chats and shared stories have been a constant source of joy.

To my cat, Oolong, thank you for being a little patch of light in my world. He is a really good companion. I'm truly happy to have been able to accompany him as he grows. Having him by my side has brought so much joy to my life.

I would also like to extend a special thanks to the exceptional Chinese swimmer, Shun Wang. His continued success in standing on the podium again at the Paris Olympics at the age of 30, have been a constant source of inspiration for me over the past six months. His unwavering dedication and embodiment of the Olympic spirit have motivated me to pursue my own goals with determination and optimism. I am grateful to him for letting us know the phrase "Per aspera ad astra"—through hardships to the stars—which has resonated deeply with me during this year.

Most importantly, to my parents and family, thank you for your unconditional love and support. Thank you for always standing firmly behind me, supporting every decision I've made, and giving me strength in my moments of doubt, confusion, and fear. I know I've grown up in a world filled with your love. I love you so much!

Contents

PUBLICATIONS **viii**

1 Introduction **1**

1.1 Overview 1

1.2 Few-shot Learning for Biomedical Named Entity Recognition 3

1.2.1 Few-shot Learning 4

1.2.2 FSL for Biomedical NER 4

1.2.3 Early Approaches to FSL for NER 6

1.2.4 LLMs on FSL for Biomedical NER 9

1.3 Research Questions 11

1.4 Thesis Outline 14

2 Literature Review **17**

2.1 Search strategy 17

2.2 Study selection and exclusion criteria 19

2.3 Data abstraction and synthesis 20

2.4 Results 21

2.4.1 Data collection results 21

2.4.2 Dimensions of characterization 22

2.4.3 Data characteristics 23

2.4.4 A summary of methodologies 24

2.4.5 Performance ranges 26
2.5 Discussion 26

3 Datasets **28**
3.1 Publicly Available Datasets 28
3.2 REDDIT-IMPACTS Dataset 31
3.2.1 Data collection 32
3.2.2 Annotation 33
3.2.3 Dataset creation 35

4 Few-shot Learning for Biomedical NER: Benchmarking Studies **37**
4.1 Traditional and FSL NER Models 38
4.1.1 Traditional NER Models 38
4.1.2 Few-shot Learning NER Models 40
4.2 Data Collection and Preparation 41
4.3 Experimental Setup 42
4.4 Results 42
4.5 Discussion 44
4.6 Conclusion 47

5 Data Augmentation with Nearest Neighbor Classifier **48**
5.1 Proposed Approach 49
5.1.1 Different Distance Methods 52
5.2 Results and Discussion 53
5.3 Conclusion 55

6 HILGEN: Hierarchically-Informed Data Generation for Biomedical NER Using Knowledge Bases and LLMs **57**
6.1 Background 58

6.1.1 UMLS in Biomedical Natural Language . . . 58
6.1.2 Synthetic Data Generation . . . 59
6.2 Proposed Approach . . . 59
6.2.1 Hierarchical Information and Semantic Network in UMLS . . . 60
6.2.2 UMLS-Based Data Generation . . . 62
6.2.3 GPT-Based Data Generation . . . 63
6.2.4 Fine-Tuning with Transformer-Based and Few-Shot Learning Models . . . 64
6.2.5 Ensemble Method . . . 65
6.2.6 Comparison with ZEROGEN . . . 65
6.3 Datasets and Experiment Setup . . . 65
6.4 Results . . . 66
6.4.1 Experimental Results . . . 66
6.4.2 Comparison with ZEROGEN . . . 67
6.4.3 Ensemble Approach . . . 68
6.5 Discussion . . . 69
6.5.1 Challenges of Zero-Shot Data Generation Approaches . . . 69
6.5.2 Impact of Ensemble Learning on Model Generalization . . . 70
6.6 Limitations . . . 71
6.7 Conclusion . . . 71

7 From Static to Dynamic: RAG-based Dynamic Prompting for Few-shot Learning 73

7.1 Background . . . 74
7.1.1 Retrieval-Augmented Generation . . . 74
7.2 Proposed Approach . . . 75
7.2.1 Static Prompt Engineering . . . 75
7.2.2 Dynamic Prompt Engineering . . . 79

7.3 Experimental Setup 82
7.4 Results 83
7.4.1 Task-specific Static Prompting 83
7.4.2 Dynamic Prompting with RAG 86
7.5 Discussion 92
7.5.1 Analysis of Different LLMs 92
7.5.2 Performance Improvements via RAG-based Prompting 92
7.5.3 Variability in the Impact of Shot Size 93
7.6 Limitations 94
7.7 Conclusion 94

8 Conclusion **96**
8.1 Future Work 97
8.1.1 Advancing Biomedical NER with Technical Innovations 97
8.1.2 Applications of LLMs in Few-shot BioNER 98

Appendix A Tables for Literature Review **101**

Appendix B Detailed Task-specific Static Prompts **113**

Appendix C Averaged Performance of the Baseline Dynamic Prompt Model **117**

Appendix D Results of 95% CIs for Each Metric **121**

Bibliography **124**

List of Figures

1.1 Architectures of Popular FSL Methodologies 10
1.2 Large Language Model structure 12
1.3 Research Overview 14

2.1 PRISMA Flow Diagram 21

3.1 Entity types and the frequency of each entity type. 34
3.2 Sample posts in the REDDIT-IMPACTS dataset. 36

4.1 Architectures of Two Traditional NER Models 39

5.1 Overall Architecture of DANN 50

6.1 Overall Architecture of HILGEN 60
6.2 Hierarchical Structure of UMLS 61
6.3 Semantic Network of UMLS 62
6.4 Prompts Used in HILGEN 64

7.1 Overview of Task-specific Static Prompting 76
7.2 Overview of Retrieval-based Dynamic Prompting 80
7.3 Performance Distribution of Prompting Strategies Across Datasets . . 86
7.4 Comparison of Average F_1-scores 89
7.5 F_1-score Distribution Across Retrieval Methods. 90

List of Tables

3.1 Statistics of publicly available datasets 30
3.2 Statistics of REDDIT-IMPACTS dataset 35

4.1 Performance on Benchmarking 43
4.2 Performance on ProtoNER model 45

5.1 Performance on DANN model 54

6.1 Performance comparison of various synthetic data generation strategies 67
6.2 Comparison of ZEROGEN and HILGEN approaches 68
6.3 Enhanced performance of ensemble with predictions from GPT-3.5 . . 68

7.1 Performance comparison of various prompting strategies 84
7.2 Performance of dynamic prompting strategies 87

A.1 Reviewed Articles for few-shot learning on medical data 106
A.2 A summary table showing primary few-shot approaches and evaluation methodologies 112

B.1 Specific static prompts for each component we used for the REDDIT-IMPACTS dataset. 114
B.2 Specific static prompts for each component we used for the BC5CDR dataset. 114

B.3 Specific static prompts for each component we used for the MIMIC III dataset. 115
B.4 Specific static prompts for each component we used for the NCBI dataset. 116
B.5 Specific static prompts for each component we used for the Med-Mentions dataset. 116

C.1 Averaged performance of the baseline dynamic prompt model on the REDDIT-IMPACTS dataset across different shot settings. 118
C.2 Averaged performance of the baseline dynamic prompt model on the BC5CDR dataset across different shot settings. 118
C.3 Averaged performance of the baseline dynamic prompt model on the MIMIC III dataset across different shot settings. 119
C.4 Averaged performance of the baseline dynamic prompt model on the NCBI dataset across different shot settings. 119
C.5 Averaged performance of the baseline dynamic prompt model on the Med-Mentions dataset across different shot settings. 120

D.1 95% CIs of static prompting strategies 122
D.2 95% CIs of dynamic prompting strategies 123

PUBLICATIONS

1. **Yao Ge**, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, Abeed Sarker. "Few-shot learning for biomedical text: A review of advances, trends, and opportunities" (2022). Journal of Biomedical Informatics, vol 144, pages: 144458. 2023. doi: 10.1016/j.jbi.2023.104458. PMID: 37488023 PMCID: PMC10940971.

2. **Yao Ge**, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, Abeed Sarker. "A comparison of few-shot and traditional named entity recognition models for biomedical text" Proceedings of the 10th IEEE International Conference on Healthcare Informatics (ICHI). 2022. pages: 84-89, doi: 10.1109/ichi54592.2022.00024. PMID: 37641590 PMCID: PMC10462421

3. **Yao Ge**, Mohammed Ali Al-Garadi, Abeed Sarker. "Data Augmentation with Nearest Neighbor Classifier for Few-Shot Named Entity Recognition." MEDINFO 2023—The Future Is Accessible, pages: 690-694. 2023. doi: 10.3233/SHTI231053. PMID: 38269897.

4. **Yao Ge**, Sudeshna Das, Karen O'Connor, Mohammed Ali Al-Garadi, Graciela Gonzalez-Hernandez, Abeed Sarker. "Reddit-Impacts: A Named Entity Recognition Dataset for Analyzing Clinical and Social Effects of Substance Use Derived from Social Media". 2024. arXiv preprint arXiv:2405.06145.

5. **Yao Ge**, Yuting Guo, Sudeshna Das, Swati Rajwal, Selen Bozkurt, Abeed Sarker. "HILGEN: Hierarchically-Informed Data Generation for Biomedical

NER Using Knowledgebases and LLMs". 2024.

6. **Yao Ge**, Sudeshna Das, Yuting Guo, Abeed Sarker. "From Static to Dynamic: Retrieval-based Dynamic Prompting for Few-shot Biomedical NER". 2024.

7. Sudeshna Das*, **Yao Ge***, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, Matthew Reyna, Reza Sameni, Yunyu Xiao, Sangmi Kim, Rasheeta Chandler, Natalie Hernandez, Danielle Mowery, Rachel Wightman, Jennifer Love, Anthony Spadaro, Jeanmarie Perrone, Abeed Sarker. "Two-layer retrieval augmented generation framework for low-resource biomedical question-answering: proof of concept using Reddit data." 2024. arXiv preprint arXiv:2405.19519.

8. Abeed Sarker, **Yao Ge**. "Mining long-COVID symptoms from Reddit: characterizing post-COVID syndrome from patient reports". JAMIA Open. 2021 Sep 2;4(3):ooab075. doi: 10.1093/jamiaopen/ooab075. PMID: 34485849; PMCID: PMC8411371. PMID: 34485849 PMCID: PMC8411371

9. Yuting Guo, **Yao Ge**, Abeed Sarker. "Detection of Medication Mentions and Medication Change Events in Clinical Notes Using Transformer-Based Models". Studies in Health Technology and Informatics, vol 310, pages: 685-689. 2024. doi: 10.3233/shti231052. PMID: 38269896.

10. Yuting Guo, Seyedeh Somayyeh Mousavi, **Yao Ge**, Madhumita Baskaran, Reza Sameni, Abeed Sarker. "Leveraging Few-Shot Learning and Large Language Models for Analyzing Blood Pressure Variations Across Biological Sex from Scientific Literature". 2024. (Under review by the IEEE Journal of Biomedical and Health Informatics (JBHI))

11. Abeed Sarker, Mohammed Ali Al-Garadi, **Yao Ge**, Nisha Nataraj, Christopher

*equal contribution

M Jones, Steven A Sumner. "Signals of increasing co-use of stimulants and opioids from online drug forum data", 2022:19 (51). Harm Reduction Journal. doi: 10.1186/s12954-022-00628-2. PMID: 35614501 PMCID: PMC9131693.

12. Abeed Sarker, Mohammed Ali Al-Garadi, **Yao Ge**, Nisha Nataraj, Londell McGlone, Christopher M Jones, Steven A Sumner. "Evidence of the emergence of illicit benzodiazepines from online drug forums." European Journal of Public Health 32, no. 6 (2022): 939-941. doi: 10.1093/eurpub/ckac161. PMID: 36342855 PMCID: PMC971342.

13. Cui, Hejie, Jiaying Lu, **Yao Ge**, and Carl Yang. "How Can Graph Neural Networks Help Document Retrieval: A Case Study on CORD19 with Concept Map Generation." Advances in Information Retrieval. ECIR 2022. Lecture Notes in Computer Science, vol 13186. Springer, Cham. doi: 10.1007/978-3-030-99739-7_9.

14. Shaina Raza, Brian Schwartz, Sahithi Lakamana, **Yao Ge**, Abeed Sarker. "A framework for multi-faceted content analysis of social media chatter regarding non-medical use of prescription medications.", BMC digital health 1, no. 1 (2023): 29. doi: 10.1109/TCBB.2023.3318209. PMID: 37680768 PMCID: PMC10483682.

15. Abeed Sarker, Sahithi Lakamana, Yuting Guo, **Yao Ge**, Abimbola Leslie, Omolola Okunromade, Elena Gonzalez-Polledo, Jeanmarie Perrone, Anne Marie McKenzie-Brown. "#ChronicPain: automated building of a chronic pain cohort from Twitter using machine learning", Health data science 3 (2023): 0078. doi: 10.34133/hds.0078. PMID: 38333075 PMCID: PMC10852024.

Chapter 1

Introduction

1.1 Overview

The task of named entity recognition (NER) aims to identify entity names in unstructured texts, and classify them into pre-defined entity types. In biomedical domain-specific datasets, common entity types include drug names, genes, adverse drug events (ADEs), indications, and symptoms, to name a few [22, 78, 159]. In recent years, deep neural network based methods (*a.k.a.*, deep learning) have achieved significant success in NER tasks when large labeled datasets are available [15, 89, 106], especially using self-supervised pre-trained language models (PLMs), such as BERT [35] and RoBERTa [111].

There are, however, still many open challenges in NER, especially for biomedical domain-specific texts and when the number of annotated instances is small [66]. In supervised learning settings with limited training instances, the application of traditional NER methods typically leads to overfitting (*i.e.*, the learner is incapable of generalizing the characteristics of the training data) [36, 103]. Within the biomedical domain, text-based datasets are often small (*e.g.*, for rare or novel diseases), and the availability of labeled data is limited. Even when large labeled datasets are created

for targeted tasks, due to restrictions associated with data privacy and patient security, it can be difficult or impossible to release or share them if they originate from biomedical sources, such as electronic health records (EHRs). Oftentimes, there is just not enough data to annotate, and even when data is available, manually annotating them can be time-consuming, error-prone and/or costly, and require high-skilled annotators [48].

The paradigm of few-shot learning (FSL) presents viable approaches to address the issue of learning from datasets where labeled data is sparse. Early FSL research progress in natural language processing (NLP) has been notably slower, primarily due to greater difficulties posed by natural language data and the lack of unified benchmarks in few-shot NLP [69]. Achieving high machine learning performances has also been challenging in few-shot settings. Text-based data often contain ambiguities and connotations that make generalization complicated. The presence of domain-specific terminologies, expressions, and associations in biomedical texts further exacerbates the difficulties of FSL [66]. Due to the potential utility of FSL in biomedical NLP, research on the topic is receiving growing attention, and progress has primarily occurred by building on a small set of related but distinct promising categories of approaches.

During the course of the research associated with this thesis, research within the field of FSL has undergone a significant shift. The widespread recognition and utilization of large language models (LLMs) such as the Generative Pre-trained Transformer (GPT [13, 135]) series have opened up unprecedented opportunities to explore and evaluate their potential in FSL settings [13]. These models, known for their ability to generate human-like text, have demonstrated remarkable proficiency in NLP tasks with relatively small training data. By leveraging the vast knowledge acquired during pre-training, LLMs can often effectively generalize from a few examples, and offer a promising approach to tackle unmet challenges in NLP and beyond. Designing effective prompts that guide the model to understand and perform the task correctly is

also a critical aspect of leveraging LLMs for FSL [105, 186]. Consequently, the increasing popularity of LLMs presents an exciting avenue for research and application in the space of FSL, providing a platform to investigate the extent of their capabilities and optimize their performance for distinct tasks.

In this thesis, we address the problem of sparse annotated data for NER in biomedical texts, a critical challenge exacerbated by the complexities of biomedical terminologies, privacy restrictions, and the resource-intensive nature of manual annotation. Our focus is on advancing the capabilities of FSL to overcome some of these challenges, and move the field forward towards robust and scalable biomedical NER systems. Specifically, we propose innovative approaches that leverage semantic augmentation, synthetic data generation through domain-specific knowledge bases like UMLS, and dynamic prompting within a RAG framework to enhance contextual understanding and performance, even in low-resource settings. This work seeks to bridge the gap between the current limitations of FSL-based NER methods and their potential for impactful applications in biomedical informatics.

1.2 Few-shot Learning for Biomedical Named Entity Recognition

In this section, we introduce some of the concepts of NLP that are associated with FSL and NER, emphasizing their importance and challenges in biomedical applications. Specifically, we explain how the problem of data sparsity in biomedical text processing necessitates the development of innovative FSL approaches. These methods aim to enable effective learning from limited annotated data, addressing challenges such as the sparsity of domain-specific entity types, the variability of biomedical terminologies, and the difficulties in generalizing to new entity classes. By exploring the interplay between FSL methodologies, NER tasks, and advancements in lever-

aging LLMs, we outline the current state of research and the strategies proposed to overcome the unique challenges in biomedical informatics.

1.2.1 Few-shot Learning

Few-shot learning is a machine learning paradigm designed to enable models to learn and perform specific tasks with only a limited amount of labeled training data. Unlike traditional supervised learning approaches that require large annotated datasets to achieve high performance, FSL aims to generalize from a small number of examples, often as few as one or five per class [130, 154]. This capability is particularly critical in domains where data annotation is labor-intensive, expensive, or constrained by domain expertise, such as biomedical informatics, where specialized knowledge is required to label data accurately [82].

The significance of FSL lies in its potential to address the limitations of annotated data sparsity without compromising performance [41]. By enabling effective learning in low data scenarios, FSL opens avenues for applying machine learning techniques in settings where traditional methods fail [171]. However, the inherent challenges of FSL make it a non-trivial problem. The limited availability of labeled data restricts the model's ability to capture diverse patterns, leading to issues such as insufficient coverage of the feature space [80, 88]. Moreover, models often struggle to generalize to unseen examples, as the paucity of training data leads to overfitting [144]. These challenges necessitate the development of algorithms and architectures specifically tailored to the unique demands of FSL.

1.2.2 FSL for Biomedical NER

Named Entity Recognition is a core task in NLP that aims to identify and classify entities within text into predefined categories such as names of people, organizations, locations, or domain-specific terms [125]. The task involves two primary steps: locat-

ing entity mentions within unstructured text and assigning them to the appropriate category. NER is foundational for numerous downstream applications, serving as a critical building block for tasks such as information extraction, question answering, and text summarization [100]. Despite its straightforward definition, NER is a non-trivial problem due to the inherent complexity of human language, which includes variations in syntax, ambiguity, and the presence of out-of-vocabulary terms [136]. The development of robust NER systems often requires advanced algorithms capable of understanding linguistic nuances and incorporating contextual information to resolve ambiguities effectively [1, 39, 100]. As a result, NER remains an active area of research in NLP, with methods evolving to address its diverse challenges across different languages and domains.

In the biomedical domain, NER involves identifying and classifying specific entities, such as names of diseases, drugs, anatomical terms, or procedures, within unstructured text [60]. It plays a critical role in extracting meaningful information from clinical narratives, scientific literature [96], and other lexical resources [184], enabling downstream applications such as information retrieval [65, 76], knowledge graph construction [21, 62], and decision support systems [72]. Unlike general NER tasks, biomedical NER is particularly challenging due to its dual objectives of entity detection (locating entity mentions in text) and entity classification (assigning these mentions to predefined categories) [170]. These objectives are inherently complex, as biomedical NER relies heavily on understanding contextual information to disambiguate terms that often appear in highly variable, domain-specific language [54].

FSL for NER magnifies these challenges due to the limited number of annotated examples typically available, particularly in low-resource settings [44]. Biomedical NER, in particular, faces significant difficulties because of the sparse distribution of entity labels. Annotating clinical narratives or scientific literature also often requires domain expertise [2], making large-scale labeling infeasible. In few-shot scenarios,

the available training data may cover only a small fraction of possible entity types, exacerbating the problem of generalizing to unseen categories during inference.

Another key difficulty in biomedical NER under FSL lies in the ability to generalize to new entity classes that are absent from the training set [149]. For example, a model trained on mentions of drugs and diseases may need to identify entirely new terms related to anatomical structures or procedures, which are semantically distinct. This highlights the need for models that can leverage domain knowledge and contextual cues effectively to bridge the gap between limited training data and the broader range of biomedical entities encountered in practice [57, 185]. These challenges underscore the importance of developing innovative FSL approaches that integrate contextual understanding and domain adaptation for biomedical NER tasks.

1.2.3 Early Approaches to FSL for NER

Early FSL research primarily focused on the field of computer vision, particularly with the goal of replicating how children learn to distinguish objects with minimal or no supervision [51, 137, 167]. FSL research progress in NLP has been notably slower, primarily due to greater difficulties posed by natural language data and the lack of unified benchmarks in few-shot NLP [69]. Attaining high machine learning performances has also been challenging in few-shot settings. Unlike images, text-based data often contain ambiguities and connotations that make generalization complicated. The presence of domain-specific terminologies, expressions, and associations in biomedical texts further exacerbates the difficulties of FSL [66]. Due to the potential utility of FSL in biomedical NLP, research on the topic is receiving growing attention, and progress has primarily occurred by building on a small set of related but distinct promising categories of approaches.

As only small numbers of labeled examples are available in the training data, prior knowledge, which is the knowledge the learner has before training, plays an

indispensable role in FSL [148]. Using prior knowledge, FSL can potentially generalize to new tasks in an effective manner as the small number of training instances are sufficient for fine-tuning the models for the task [171]. Wang et al. [171] divides FSL methods into three categories based on how prior knowledge is used: (i) data, which use prior knowledge to augment training data; (ii) model, which use prior knowledge to constrain hypothesis space; and (iii) algorithm, which use prior knowledge to guide how parameters are obtained.

One set of promising FSL approaches involves *meta-learning* (*a.k.a.*, "*learning to learn*" [67]). Meta-learning has perhaps been the most common framework for FSL, and is a branch of *metacognition*, which is concerned with learning about one's own learning and learning processes [147]. In the classical machine learning framework, training data is used to optimize a model for a specific task, and a separate set is used to evaluate the performance of the trained model. In the meta-learning framework, a model is trained using a set of training *tasks*, not data, and model performance is evaluated on a set of test tasks. In the experimental setting, the learner obtains prior knowledge by incorporating generic knowledge across different tasks (*i.e.*, algorithm level prior knowledge). The small number of labeled instances for the target task is then used to fine-tune the model. Figure 1.1a illustrates the meta-learning framework using a simple example—an entity recognition model is trained using different tasks involving news and music data, and is evaluated on a biomedical task.

Several additional classes of FSL methods have evolved over the years, some building on meta-learning. Ravi and Larochelle [137] presented a long-short term memory (LSTM) based meta-learner that is trained and customized separately for mini-batches of training data (referred to as *episodes*), rather than as a single model over all the mini-batches. Separately, *matching networks* were recently proposed, and they attempt to use two embedding functions (*i.e.*, functions that project data into vector space while capturing relevant semantics)—one for the training sets and

one for the test sets—to imitate how humans generalize the knowledge learned from examples. The framework attempts to optimize the two embedding functions from the training (*support sets*) and the validation examples (*query sets*), and attempts to measure how well the trained model can generalize [7, 167]. Figure 1.1b illustrates the functionality of matching networks in a simplified manner. A variant of matching networks utilizes active learning by adding a sample selection step that augments the training data by labeling the most beneficial unlabeled sample (*i.e.*, model level prior knowledge).

Another related class of FSL approaches known as *metric learning* employs distance-based metrics (*e.g.*, nearest neighbor). Given a support set, metric learning methods typically produce weighted nearest neighbor classifiers via non-linear transformations in an embedding space, and the examples in the support set close to the query example (based on the metric applied) are used to make classification decisions, imitating how humans use similar examples or analogies to learn. *Prototypical networks* [153], yet another similar class of approaches, particularly attempt to address the issue of overfitting due to small training samples by generating prototype representations of classes from the training samples, similar to how humans summarize knowledge learned from examples. Prediction of unknown data samples can be performed by computing distances to the class prototypes (*e.g.*, support set means), and choosing the nearest one as the predicted label. Figure 1.1c visually illustrates the functionality of a prototypical network. A semi-supervised variant of prototypical networks applies *soft assignment* on unlabeled samples, and incorporates these as prior knowledge (*i.e.*, data level prior knowledge. Transfer learning, a commonly used approach in FSL, also incorporates prior knowledge at the data level as knowledge learned from data in prior tasks are *transferred* to new few-shot tasks [129].

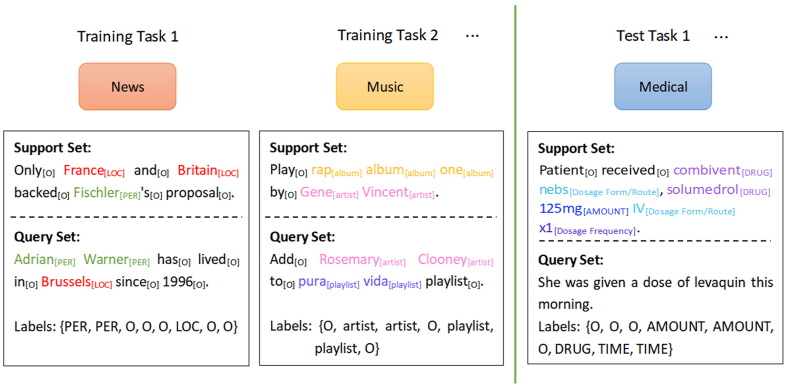
The problems that these and other FSL methods attempt to solve are closely aligned with the practical challenges faced by many biomedical NLP tasks. While a

number of FSL strategies have been explored for biomedical texts by distinct research communities (*e.g.*, health informatics, computational linguistics), there is currently no review that compares the performances of these strategies or summarizes the current state of the art. There is also no study that has compiled the reported performances of FSL methods on distinct biomedical NLP data/tasks. We attempt to address these gaps in this systematic review. Specifically, we review FSL methods for biomedical NLP tasks, and characterize each reviewed article in terms of type of task (*e.g.*, text classification, NER), primary aim(s), dataset(s), evaluation metrics, and other relevant aspects. We summarize our findings about FSL methods for biomedical NLP, and discuss challenges, limitations, opportunities and necessary future efforts for progressing research on the topic.

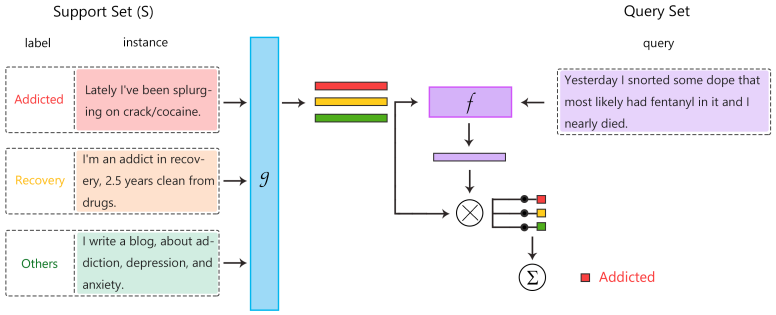
1.2.4 LLMs on FSL for Biomedical NER

Recent advancements highlight the potential of large language models (LLMs) such as Generative Pre-trained Transformer (GPT), for few-shot NER, especially when combined with domain-specific knowledge bases [168, 169]. The capabilities of LLMs offer an opportunity to evaluate their capabilities in few-shot scenarios by generating human-like texts and providing external knowledge with limited examples [13]. LLMs excel in generating natural language across domains and tasks, and their adaptability is enhanced by prompt-based strategies, which can significantly improve accuracy [105, 186].

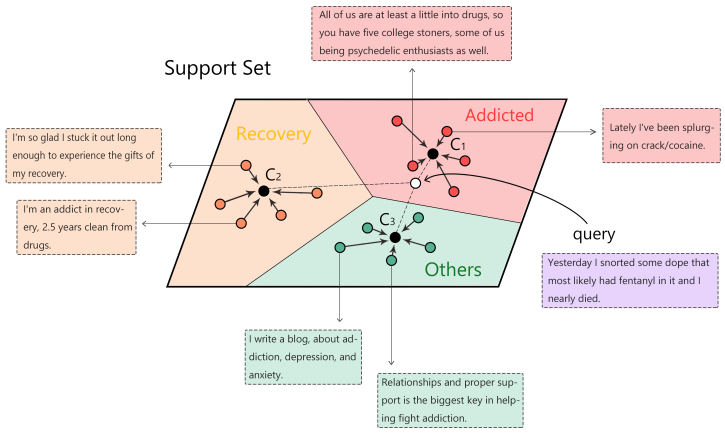
Information extraction from biomedical text involves deriving valuable insights from sources such as biomedical records, which often contain scarce, sensitive, and imbalanced data. The ability of LLMs to generate coherent and contextually relevant text offers new opportunities to address the intricacies of NLP tasks involving biomedical data. By generating synthetic texts that closely mimic real-world biomedical text, LLMs can provide additional training data that enhances the performance



(a) Meta-learning: each task mimics the few-shot scenario and can be completely non-overlapping. Support sets are used to train; query sets are used to evaluate the model.



(b) Matching networks: a small support set contains some instances with their labels (one instance per label in the figure). Given a query, the goal is to calculate a value that indicates if the instance is an example of a given class. Two embedding functions *g()* and *f()* are applied to transform the inputs.



(c) Prototypical network: a class's prototype is the mean of its support set in the embedding space. Given a query, its distance to each class's prototype is computed to decide its label.

Figure 1.1: Architectures of three popular few-shot learning methodologies. (a) Meta-learning. (b) Matching networks. (c) Prototypical network. Note: (b) and (c) use the DASH 2020 Drug Data [52].

of downstream tasks, such as NER and other critical applications in healthcare [152]. This ability to augment existing datasets with high-quality, contextually relevant text could significantly improve the accuracy and reliability of biomedical information extraction models. However, LLMs like GPTs might face problems like hallucination [6] and homogenization [5] when dealing with specialized biomedical concepts.

The advent of LLMs has shifted the focus towards prompt-based learning, which has shown promise in few-shot NLP [110, 133]. Figure 1.2 shows the shifts from three popular few-shot learning models to LLMs. The potential of LLMs and prompt-based strategies in few-shot settings is demonstrated by techniques like LM-BFF [104], which fine-tunes models using prompts, and PPT [55], which enhances prompt effectiveness through unsupervised pre-training. Incorporating biomedical knowledge bases like UMLS has also been explored [4, 121], demonstrating improvements over general-purpose models. Leveraging knowledge from both domain-specific knowledge bases and in-context information extracted by LLMs, however, is still relatively new. Thus, this presents a research gap that may have significant utility for challenging NLP tasks. We explore this potential utility on the task of few-shot NER using multiple biomedical datasets.

1.3 Research Questions

Biomedical NER remains a challenging task, especially in low data settings. As discussed, FSL offers a promising approach to addressing the data sparsity problem by enabling models to learn effectively from limited labeled examples. However, significant gaps persist in this area. Existing FSL approaches for open-domain NER often fail to generalize well to biomedical datasets due to the unique linguistic and domain-specific challenges, such as the variability of biomedical terminology and the need for contextual disambiguation. Moreover, the lack of standardized few-shot

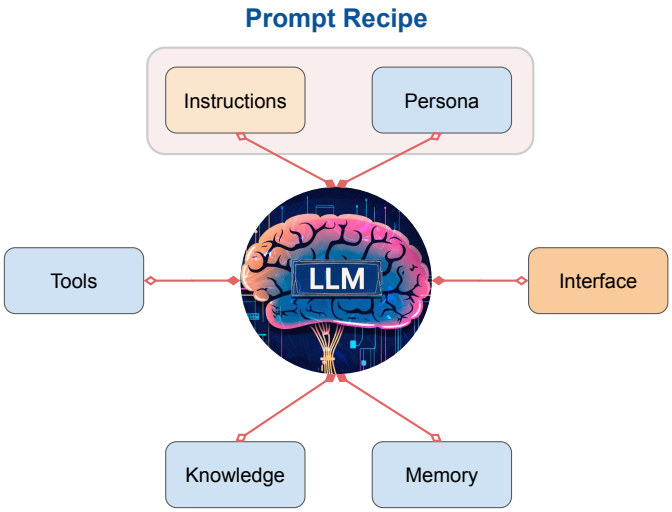


Figure 1.2: The average and standard deviation of critical parameters for Large Language Models.

datasets and benchmarking practices for biomedical NER hinders progress in the field.

To overcome these limitations, it is critical to explore innovative solutions that address data sparsity and adapt to the complex linguistic patterns of biomedical text. This thesis aims to fill these gaps by answering the following research questions, which collectively address the challenges and opportunities for advancing few-shot learning methods in biomedical NER:

- **RQ1:** How do existing FSL approaches for open-domain NER perform on biomedical datasets?
- **RQ2:** How can we address the challenge of data sparsity via data augmentation strategies?
- **RQ3:** How can we leverage knowledge from the Unified Medical Language System (UMLS) and LLMs to generate synthetic training examples for effectively expanding a few-shot dataset?

- **RQ4:** How can we employ effective techniques to transform static prompts into dynamic prompts for improving few-shot NER with retrieval augmented generation (RAG)?

Our contributions may be summarized as follows:

- **C1**: We conducted an in-depth review of FSL methods for biomedical NLP tasks [48]. Our findings revealed the lack of standardized few-shot datasets and benchmarking work for biomedical NER. We also proposed possible future research directions for few-shot biomedical NER.

- **C2**: We performed extensive benchmarking experiments to conduct head-to-head relative performance comparisons of FSL systems on public NER datasets, which demonstrated their severe underperformance [47].

- **C3**: We introduced Reddit-Impacts, a challenging NER dataset representing clinical and social impacts of substance use mentioned on social media, which is naturally suitable for FSL research due to the sparse occurrence of these concepts [50].

- **C4**: We proposed a novel method for FSL-based NER that uses data augmentation combined with a nearest neighbor classifier to address the data sparsity problem, and we also explored the influences of different distance metrics [49].

- **C5**: We explored knowledge augmentation methods based on the UMLS for improving NER in few-shot settings. We further leveraged LLMs for generating synthetic data to supplement UMLS knowledge to boost NER performance in the biomedical domain.

- **C6**: We explored the viability of employing LLMs for the extraction of named entities by utilizing task-specific static prompt engineering techniques, then en-

hanced it by employing RAG-based dynamic prompting, which further improves biomedical NER in few-shot settings.

While contributions C1-C3 are aimed at answering RQ1, C4 and C5 are motivated by RQ2 and RQ3, respectively, and C6 attempt to answer RQ4.

Figure 1.3 visually illustrates the research questions addressed and contributions made by this thesis.

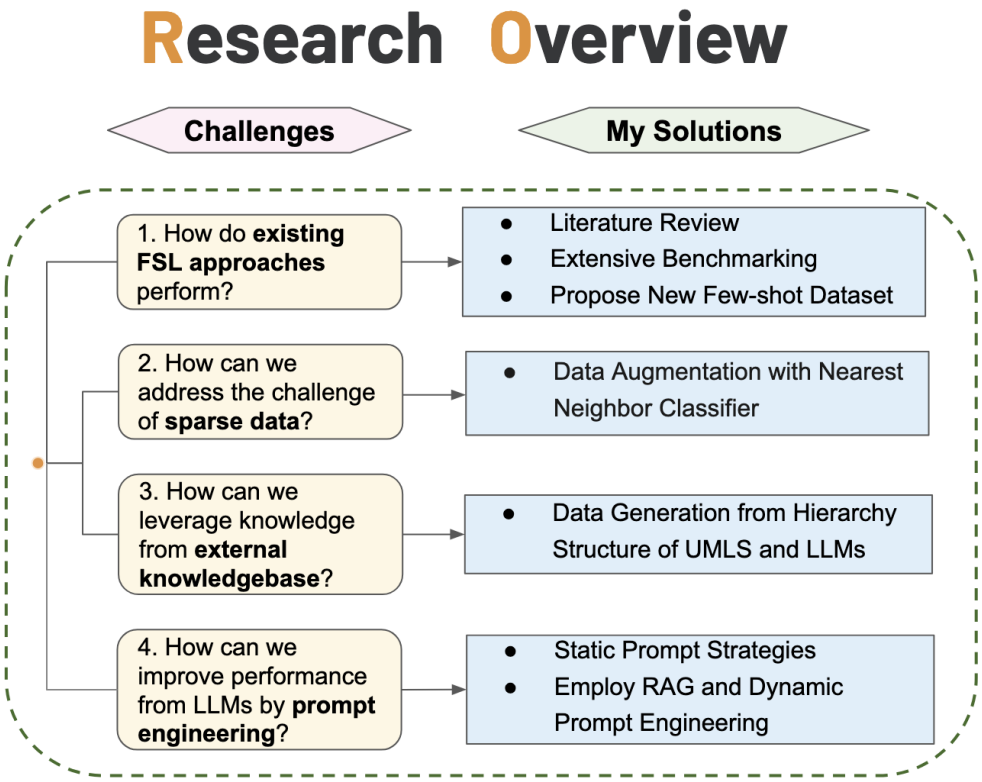


Figure 1.3: Overview of my research in FSL, including my contributions on literature review, benchmarking, proposing a new dataset, data augmentation method, synthetic data generation method and RAG-based dynamic prompting techniques to improve performance on inference from LLMs.

1.4 Thesis Outline

Chapter 2 provides a detailed overview of relevant literature on FSL for NER, with a focus on biomedical applications. It identifies gaps in existing approaches, includ-

ing the lack of standardized datasets, inconsistent evaluation strategies, and limited adaptability to the complexities of biomedical text. By synthesizing findings from prior studies, this chapter highlights the need for tailored methods and resources to address the unique challenges of biomedical NER in few-shot settings, forming the foundation for the research contributions in this thesis.

Chapter 3 presents a detailed explanation of the data we use. In particular, it discusses our corpus, which is specialized for few-shot NER. It also presents various corpus statistics, and the annotation process which was carried out as part of this research.

Chapter 4 describes our benchmarking work, which compares the performance of traditional NER models (e.g., BERT-Linear Classifier, BERT-CRF, SANER) and FSL-based models (e.g., StructShot, NNShot, Few-Shot Slot Tagging, ProtoNER) across five biomedical text datasets. The results demonstrate that while traditional models perform well with sufficient training data, all models exhibit poor performance in low data scenarios, highlighting the need for further advancements in FSL methods.

Chapter 5 details our DANN model, which combines semantic augmentation with a nearest neighbor classifier to address data sparsity in few-shot biomedical NER. Evaluated across five biomedical datasets, DANN demonstrates improved performance over baselines in several tasks, with Manhattan and 3-norm distances performing best under specific settings. However, challenges remain in noisy datasets like social media-based biomedical texts, highlighting the need for further domain-specific optimizations.

In Chapter 6, we identify possible approaches for generating synthetic data to address the challenges of data sparsity in few-shot biomedical NER. The proposed framework—HILGEN—leverages UMLS hierarchical knowledge and GPT-3.5 to create diverse and contextually rich training examples. By incorporating related concepts, parent-child relationships, and synthetically generated sentences, HILGEN en-

hances model performance across multiple datasets. The ensemble method, combining UMLS and GPT-generated outputs, further boosts precision and recall, showcasing the synergy between structured domain knowledge and generative models for tackling data limitations in biomedical NER.

Chapter 7 explores the transition from static to dynamic prompting. It begins by addressing the limitations of static prompts, which rely on predefined templates, and demonstrates their effectiveness as a baseline. The chapter then introduces a RAG-based dynamic prompting framework, highlighting its ability to adapt to contextual relevance, thereby significantly enhancing performance in low-resource scenarios. Comparative evaluations underscore the advantages of dynamic prompting over static methods in improving model flexibility and accuracy.

Finally, in Chapter 8, we conclude with a summary of the thesis, outlining future directions and possible applications of the work.

Chapter 2

Literature Review

Few-shot learning, also referred to as low-shot learning, is a machine learning paradigm where models learn to make predictions on a new class with only a small number of examples [153, 162]. This contrasts with traditional deep learning models that require large amounts of data [36, 103]. The goal of FSL is to train a model that can generalize to new classes with only a few examples, which makes it particularly useful for fine-grained classification tasks such as NER tasks, where obtaining large amounts of data for each class can be challenging, although it is conceptually possible to accomplish the tasks with small numbers of examples [90].

2.1 Search strategy

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) protocol to conduct this review [123]. FSL for NLP is a relatively recent research topic, so we concentrated on a short time range for our literature search—January 2016 to August 2021. We searched the following bibliographic databases to identify relevant papers: (1) PubMed/Medline, (2) Embase, (3) IEEE Xplore Digital Library, (4) ACL Anthology, and (5) Google Scholar, the latter being a meta-search engine, not a database. We included ACL Anthology (the primary source for the

latest NLP research) and IEEE Xplore, in addition to EMBASE and PubMed/Medline, because much of the methodological progress in FSL has been published in non-medical journals and conference proceedings. At the time of searching (September 2021), ACL Anthology hosted 71,290, and IEEE Xplore hosted over 5.4 million articles, although most articles in the latter did not focus on NLP or medicine. Over recent years, preprint servers have emerged as major sources of the latest information regarding research progress in computer science and NLP, and we used Google Scholar primarily as a medium for searching these preprint servers or published papers from other sources. Note that we also searched the ACM Digital Library*, but discovered no additional articles. Hence, we do not report it as a data source for our review.

We applied marginally different search strategies depending on the database to account for the differences in their contents. We used three types of queries:

1. Queries focusing on the technical field of research (phrases included: 'natural language processing', 'text mining', 'text classification', 'named entity recognition', and 'concept extraction');

2. Queries focusing on the learning strategy (phrases included: 'few-shot', 'low-shot', 'one-shot', and 'zero-shot'); and

3. Queries focusing on the domain of interest (phrases included: 'medical', 'clinical', 'biomedical', 'health', 'health-related').

All articles on PubMed and Embase fall within the broader biomedical domain, so we used combinations of the phrases in 1 and 2 above for searching these two databases, leaving out the phrases in 3. All articles in the ACL Anthology involve NLP, so we used phrases from 2 and 3 for this source. For IEEE Xplore and Google Scholar, the articles can be from any domain and on any topic, so we used combinations of all three sets of phrases for searching. PubMed, Embase, and IEEE only

*https://dl.acm.org/

returned articles that entirely matched the queries. However, ACL Anthology and Google Scholar retrieved larger sets of articles and ranked them by relevance. For ACL Anthology, the articles retrieved were reviewed sequentially in decreasing order of relevance. For each query combination, we continued reviewing candidate articles until we came across at least two pages (about 20 articles) of no relevant articles, at which point we decided that no relevant articles would be found in the following pages. Since FSL is a relatively new research area, we anticipated that there would be some relevant research papers that are not yet indexed in PubMed, Embase, or ACL Anthology. Specifically, preprint servers such as arXiv, biorXiv, and medrXiv are very popular among machine learning and NLP researchers as they enable the publication of the latest research progress early. We used Google Scholar as an auxiliary search engine to identify potentially relevant articles indexed in such preprint servers or other sources (*e.g.*, Open Review[†]). Google Scholar, like ACL Anthology, sorts returned articles by relevance, but the total number of articles returned is much larger. For this search engine, therefore, we reviewed the top 40 articles returned by each query combination, excluding those that were retrieved from the other databases.

2.2 Study selection and exclusion criteria

All articles shortlisted from initial searches were screened for eligibility by two authors of the manuscript (YGe and AS). We removed duplicate articles and those that either did not include at least one dataset from the biomedical domain or did not involve NLP. While it was always possible to identify the technical field/topic (NLP or not) from the titles and abstracts, to determine domain, we had to review full articles because a subset of papers included multiple datasets, and only some of these datasets were from the biomedical domain. We excluded papers if none of the datasets were related to medicine/health, or did not explicitly focus on few/low-shot settings, and

[†]`https://openreview.net/`

reviewed the remaining articles.

2.3 Data abstraction and synthesis

We abstracted the following details from each article, if available: publication year, data source, primary research aim(s), training set size(s), number of entities/classes, entity type for training, entity type for evaluation/testing, primary method(s), and evaluation methodology. For studies including data from multiple sources, we only abstracted those related to health/medicine. In terms of primary aim(s), some studies reported multiple objectives, and we abstracted all the NLP-oriented ones (*e.g.*, text classification, concept extraction). With respect to training set sizes, we abstracted information about the number of instances that were used for training, and, if applicable, how larger datasets were *reconstructed* to create few-shot samples. We also extracted the number of labels for each study/task; for NER/concept extraction methods, we identified the number of entities/concepts, and for classification, we identified the type of classification (*i.e.*, multi-label or multi-class) along with the number of classes. We also noted down the training domain(s) and test/evaluation domain(s) for each few-shot method, when applicable. Abstracting primary approach(es) and evaluation methodology was more challenging due to the complexities of some of the model implementations, and we reviewed and summarized the descriptions provided in each paper. For evaluation, we abstracted evaluation strategies and reported performances.

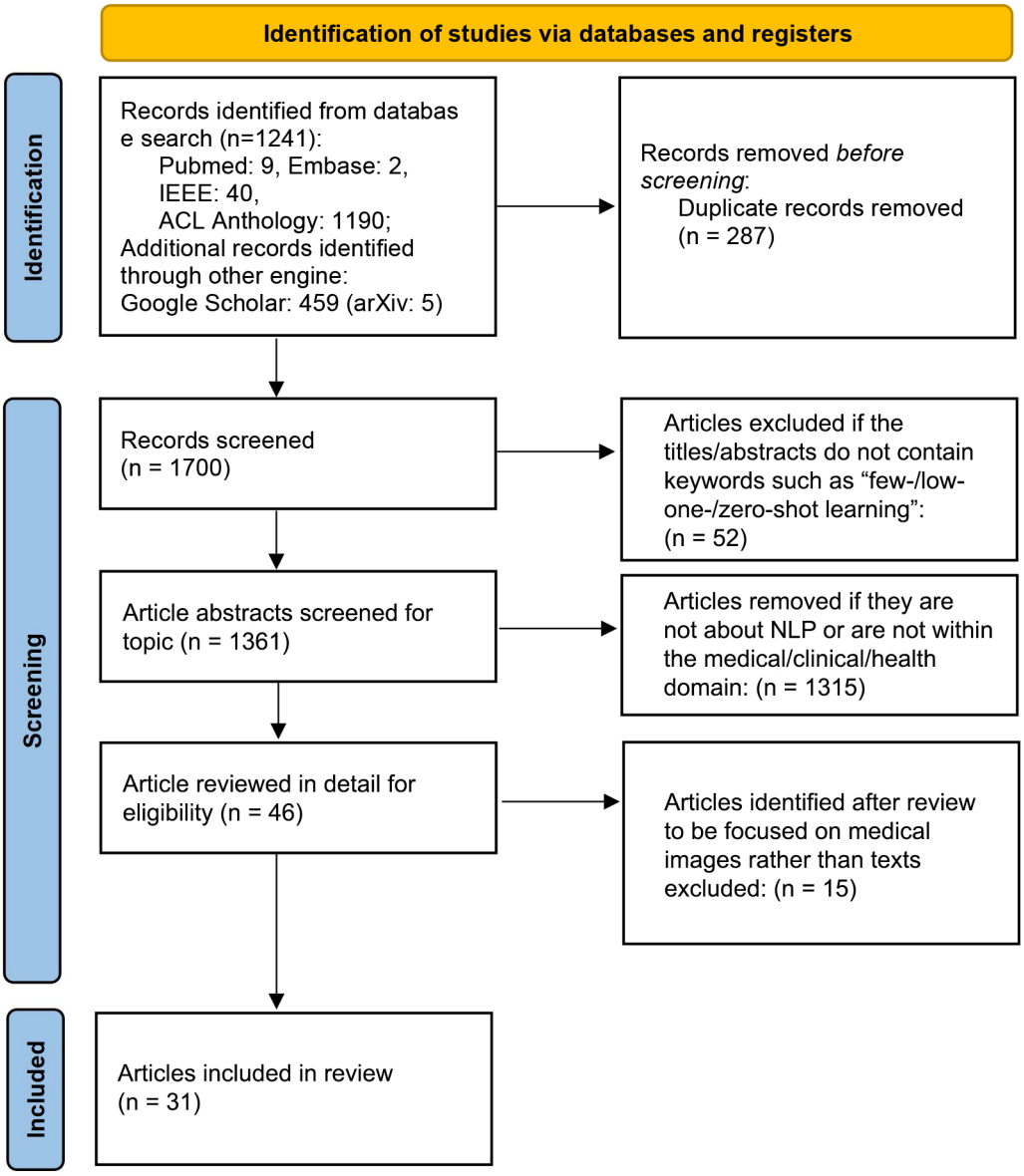


Figure 2.1: PRISMA flow diagram for the process of depicting the number of articles at each stage of collection and filtering.

2.4 Results

2.4.1 Data collection results

Our inclusion criteria were met by 31 studies. Initial searches retrieved 1241 articles from PubMed, Embase, IEEE Xplore, and ACL Anthology, and an additional 459 from Google Scholar. Figure 2.1 presents the screening procedures and numbers at

each stage. After initial filtering, we reviewed 46 full-text articles for eligibility, and excluded 15 from the final review. The first included study was from 2018, and most articles (22/31; 71%) were from 2020 and 2021, although for the latter year, only studies published prior to August 31 were included.

2.4.2 Dimensions of characterization

Table A.1 in Appendix A summarizes some fundamental characteristics of each study (authors, year, data source, retrieval search engine, and number of entities/classes); further abstracted statistics described in this paragraph are provided in Table S1 of the supplement (research aims, training set sizes, and training and evaluation entity types). In terms of training data sizes, 7/31 (23%) studies included zero-shot scenarios (*i.e.*, prediction without any labeled instances) into their research scope, including two on zero-shot learning only. 1-shot, 5-shot, and 10-shot were the most common '*shot*' settings, representing 12/31 (39%) of the reviewed studies. 6/31 (19%) reviewed studies used samples of larger datasets for training, often specified in percentages (*e.g.*, 25%, 50%). 3/31 (10%) studies did not explicitly specify shot values. 2 studies did not perform experiments in accordance with traditional few-shot scenarios, and divided all labels into three categories according to the frequency of occurrences (frequent group contained all labels occurring more than 5 times; few-shot group contained labels occurring between 1 and 5 times; and the zero-shot group included labels that never occurred in the training dataset), causing some labels to have large numbers of annotated samples. 11/13 (85%) few-shot NER tasks explicitly mentioned the number of entity types. For few-shot classification, 50% (5/10) specified the approximate number of classes. 7/31 (23%) studies involved cross-domain transfer, with different domains of training and test/evaluation data. In most cases, however, the training sets and test sets used were from the same domain.

Table A.2 in Appendix A provides summaries of the methods proposed, and the

evaluation strategies. Variants of neural network based (deep learning) algorithms, such as Siamese Convolutional Neural Networks [178], were the most common. Only 3/31 (10%) articles proposed new datasets, and 2/31 (7%) presented benchmarks for comparing multiple few-shot methods. Evaluation strategies had considerably less diversity. Almost all evaluation methodologies for classification tasks involved standard metrics such as accuracy, precision, recall, and F_1-scores, and NER tasks mainly relied on F_1-scores only.

2.4.3 Data characteristics

We grouped the datasets used into three categories: (i) publicly downloadable (de-identified) data; (ii) datasets from shared tasks; and (iii) new datasets specifically created for the target tasks. We found that datasets belonging to (ii) and (iii) were particularly difficult to obtain—shared task data are often difficult to obtain after their completion, and specialized datasets are often not made public, particularly if they contain protected health information (PHI). Studies using datasets from category (i) often reported performances on multiple datasets, consequently making the evaluations more comparable. Overlap of datasets among different studies was relatively low, making comparative analyses difficult. The MIMIC-III (Medical Information Mart for Intensive Care) dataset [78], was the most frequently used across studies (7/31; 23%), particularly for few-shot classification and NER tasks. This was likely due to the public availability of the dataset and the presence of many labels in it (7000) [141]. 6 papers used datasets from shared tasks, of which 4 were from BioNLP [11, 127], one from the Social Media Mining for Health Applications (SMM4H) [174], and one from the Medical Document Anonymization (MEDDOCAN) shared task [120]. Only 3 papers constructed new datasets, reflecting the paucity of corpora built to support FSL for biomedical NLP.

Reconstruction of datasets

There are 19/31 (61%) reviewed studies reconstructed existing datasets for conducting experiments in few-shot settings (*i.e.*, subsets of labeled instances were extracted from larger datasets). For multi-label text classification tasks, especially when the number of labels is very large, and for few-shot NER tasks, reconstructing datasets can be complex. A common way to represent data is *K-Shot-N-Way*, meaning that each of N classes or entities contains K labeled samples, as well as several queries from each class for each test batch. However, for multi-label classification tasks, each instance may have more than one class, often making it difficult to ensure that the reconstructed datasets included only K labeled samples for each class. Similar challenges exist for NER tasks, as each text segment may have overlapping entities. 39% (12/31) of the studies did not construct special datasets to represent few-shot settings. 16% (5/31) used existing datasets with high class imbalances, and the few-shot algorithms were focused on sparsely-occurring labels.

2.4.4 A summary of methodologies

23/31 (74%) studies addressed text classification or NER/concept extraction tasks while only 8 (26%) studies focused on others.

Few-shot text classification

10/31 studies (33%) focused on few-shot classification, with half of them involving multi-label text classification. Multi-label classification is a popular task because the associated datasets generally contain some very low-frequency classes. 7/10 (70%) classification papers proposed deep learning algorithms, and 3/10 (30%) were inspired by label-wise attention mechanisms. 2/10 (20%) combined few-shot tasks with graphs, such as similarity or co-occurrence graphs, or hierarchical structures that encode relationships between labels for knowledge aggregation. While convolutional neural

networks have been popular for FSL, transformer-based models such as BERT [35] and RoBERTa [111] rarely appeared in these articles. Only 1 paper [19] mentioned applying BERT to generate instance embeddings, and then passing top-level output representations into a label-wise attention mechanism.

Few-shot NER or concept extraction

8 reviewed papers were described as NER; 5 as concept extraction. Generally, studies described as concept extraction had fewer commonalities in their methods and involved task-specific configurations based on the characteristics of the data and/or extraction objectives. 63% (5/8) of the studies described as NER employed transfer learning, with training and testing data from different domains. Studies commonly used the BIO (beginning, inside, outside) or IO tagging schemes. 2 papers investigated both BIO and IO tagging schemes, concluding that systems trained using IO schemes outperform those trained using BIO schemes. Studies reported that the O (outside) tag was often ill-defined, as specific entities (*e.g.*, time entities such as 'today', 'tomorrow') would be tagged as O if they were not the primary focus of the dataset. 5 papers used BIO schemes, while 1 considered only the entity names without any tagging schemes. The NLP/machine learning strategies employed varied significantly, and included, for example, the application of fusion layers for combining features [180], biological semantic and positional features [56], prototypical representations and nearest neighbor classifiers [179], transition scorers for modeling transition probabilities between abstract labels [68, 71, 179], self-supervised methods [71, 86, 116], noise networks for auxiliary training [74, 86], and LSTM cells for encoding multiple entity type sequences [74].

Overview of other methods

6/31 (19%) studies applied meta-learning strategies, and 12/31 (39%) articles demonstrated the advantages of attention mechanisms in few-shot scenarios, such as handling the difficulty of recognizing multiple unseen labels. Among the latter, 5/12 used self-attention-related methods, and 4/12 used label-wise attention mechanisms. 8/31 (26%) studies reproduced prototypical networks, and/or added enhancements to them. Only 1 article used matching networks, and 2 studies included them as baselines.

2.4.5 Performance ranges

9/31 (29%) studies used *accuracy*, and the reported values varied considerably, between 44% and 97%. Two-thirds (6/9) reported accuracies higher than 70%. For the 17/31 (55%) studies that reported F_1-score, performance variations were even larger—from 11.7% to 95.7% (median: 68.6%). We were unable to determine in most cases if the performance differences were due to the effectiveness of the FSL methods, or if the dataset characteristics were primarily responsible.

2.5 Discussion

In this review, we systematically collected and compared 31 studies that lie at the intersection of FSL, NLP, and health. We generally found it difficult to perform head-to-head comparisons of the proposed methods due to the use of different evaluation strategies, training/test data, and experimental settings. For example, Chalkidis et al. [19] used 50 or less instances in their few-shot setting, while Rios and Kavuluru [140] used 5 or less. It was also often impossible to objectively compare performances with those reported in prior literature, as few-shot methods were expected to underperform compared to methods trained and evaluated on the same domain

and/or larger training sets. However, the review led to several observations that were relatively consistent across studies: (i) under the same experimental parameters, the performances reported on biomedical data were worse than those reported on data from other domains [71, 178, 179]; and (ii) creating specialized datasets for transfer learning typically produced better results than low-quality datasets (such as datasets lacking completeness, not specifically designed for FSL, or with unclear specifications) [68, 116, 179].

K-Shot-N-Way datasets were commonly reported for simulating few-shot scenarios for evaluations. In such synthetically created datasets, the number of instances for training is predetermined. Such consistency in characteristics is almost never the case with real-world text-based biomedical data. Though this design attempts to make direct comparisons between different methods or tasks easier, only speculative estimates can be made about how the proposed methods may perform if deployed in real-world settings. There is a need to evaluate systems on naturally distributed biomedical text data so that the deployment performances can be estimated—an aspect that future research should consider.

Few articles created new datasets specialized for FSL, or provided benchmarks that future studies could use for comparison. Considering the fact that FSL is still a relatively new field, such datasets and benchmarks are essential for promoting future development. The lack of standardized datasets, and the consequent need to reconstruct datasets for simulating few-shot scenarios is a notable obstacle to research in this space. Reconstructed datasets often use randomly sampled subsets for evaluation, making direct comparisons between systems difficult (since the specific training and test instances may not be known), and increasing the potential for biased performance estimates.

Chapter 3

Datasets

In this thesis, we used nine biomedical text datasets for building benchmarks and proposing new approaches. The datasets included eight existing publicly available datasets and the Reddit-Impacts dataset, which we collected from Reddit and annotated as part of the research described in this thesis.

3.1 Publicly Available Datasets

MIMIC III [79] is a large, single-center database that contains information relating to patients admitted to critical care units at a large tertiary care hospital and is publicly available. Data includes medications, laboratory measurements, observations and notes charted by care providers, diagnostic codes, imaging reports, hospital length of stay, survival data, etc. MIMIC III was one of the most frequently used datasets for few-shot classification and NER tasks.

N2C2 2018 [64] focuses on adverse drug events (ADE) and medication mentions. The dataset is used for tasks related to identifying and classifying medication-related entities and their associated adverse effects within electronic health records (EHRs).

I2B2 2014 [159] focuses on de-identification of longitudinal biomedical records. The primary task associated with this dataset is to identify and remove protected health information (PHI) from clinical narratives to ensure patient privacy. This involves detecting various types of PHI such as names, dates, locations, and other personal identifiers in biomedical records.

BioNLP 2016 [17] focuses on descriptions of genetic and molecular mechanisms from scientific articles. This dataset is used for tasks related to identifying various biological events and entities within the context of molecular biology, such as gene expression, protein interactions, and regulatory relationships. The goal is to facilitate the automatic extraction of detailed and structured information from the biomedical literature, which can support various applications in bioinformatics and computational biology.

SMM4H 2021 [174] focuses on ADE mentions in social media data. This dataset is part of a shared task series that aims to leverage social media platforms like Twitter (X) to identify and analyze health-related information, specifically ADEs. The challenge involves developing models that can accurately detect and classify mentions of adverse drug reactions in the noisy and informal text found in social media posts.

BC5CDR [99] is a resource specifically designed for relationships between chemicals and diseases from scientific literature. This corpus consists of biomedical articles annotated for mentions of chemicals, diseases, and the interactions between them; the primary goal of this dataset is to enable the development and evaluation of systems that can automatically identify these entities, which are crucial for various applications in biomedical research, including drug discovery, toxicology, and understanding disease mechanisms.

Med-Mentions [122] is a large biomedical corpus annotated with UMLS concepts. This dataset consists of scientific articles from PubMed, annotated for a wide range of biomedical entities linked to UMLS concepts. The annotations cover various types of biomedical information, including diseases, chemicals, genes, and anatomical terms. Med-Mentions supports tasks such as information extraction, literature mining, and knowledge base construction in the biomedical domain.

NCBI Disease [37] consists of a collection of PubMed abstracts annotated with disease names, linking these mentions to standardized concepts in the Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) databases. This corpus is used to train and evaluate models for the tasks of recognizing disease names in biomedical texts and normalizing these mentions to a consistent set of biomedical concepts.

Table 3.1: Statistics of the eight standardized biomedical datasets we used, including the source and aim of their tasks, training and test sizes (number of tokens), the number of entity types and the number of entities in each dataset.

Datasets	Training Size	Test Size	Entity Types	Entities
N2C2 2018 track 2 (adverse drug events and medication extraction)	611.0k	411.0k	9	76.9k
I2B2 2014 (de-identification of longitudinal medical records)	490.3k	206.1k	23	20.7k
MIMIC III (information relating to patients)	36.4k	6.4k	12	8.7k
BioNLP 2016 (Genetic and molecular mechanisms)	515.4k	148.5k	1	28.6k
SMM4H 2021 task 1b (distinguishing adverse effect mentions)	30.6k	4.5k	1	1.3k
BC5CDR (extracting relationships between chemicals and diseases)	228.8k	122.2k	2	28.8k
Med-Mentions (annotated with UMLS concepts)	847.9k	593.6k	1	340.9k
NCBI Disease (PubMed abstracts annotated with disease names)	134.0k	20.5k	4	6.3k

Table 3.1 presents relevant statistics for all publicly available datasets we used, presenting the source and aim of each NER task, training and test set sizes, the number of entity types, and the number of entities in each dataset.

3.2 Reddit-Impacts Dataset

In our literature review, we found that there is a paucity of datasets that are naturally suited specifically for FSL, but such datasets are essential for promoting future development. Therefore, we collected and created Reddit-Impacts [50], a challenging NER dataset curated from subreddits dedicated to discussions on prescription and illicit opioids, as well as medications for opioid use disorder.

Substance use disorders represent a critical challenge in public health, with both clinical and social consequences impacting individuals and communities worldwide [92, 114]. The pervasive nature of substance use, encompassing both prescription and illicit drugs, necessitates a deeper understanding of its impacts to inform more effective interventions and preventative measures [26, 31]. We introduce the Reddit-Impacts dataset, a unique corpus derived from Reddit, a platform known for its rich, anonymized discussions among diverse groups, including individuals who use drugs [58, 139, 145]. The dataset includes posts from 14 opioid-related subreddits, capturing a broad spectrum of experiences and discussions related to substance use.

Our research specifically focuses on the clinical and social impacts of nonmedical substance use. These impacts are critical yet under-represented in the available data, making them an ideal focus for applying FSL techniques to improve NER. The clinical impacts encompass the direct effects on an individual's health, while social impacts involve the broader consequences on relationships, communities, and societal structures. While such information is abundant on Reddit, they are embedded in vast volumes of other unrelated information, making it extremely challenging to detect them automatically with high accuracy from naturally distributed data. In this section, we detail the creation of the Reddit-Impacts dataset, describe our annotation process, and provide data statistics.

3.2.1 Data collection

Reddit is popular in the broader community of people who use drugs as it offers anonymity, and Reddit has seen rapid growth in its user base over the last several years. Reddit communities have also been found to serve as a means of social support for people who use drugs. We chose Reddit over other social networks or web-based forums such as Twitter, Bluelight, and Discord for several reasons. While all these sources contain information about substance use, the substance use community of Reddit is much larger and has been extensively used in peer-reviewed research related to substance use and emerging substance use trends. Additionally, Reddit threads are also heavily moderated, and posts must follow community-specific rules. Consequently, while these rules restrict some types of information from being posted, they also ensure that the data are reflective of the topical areas and the volume of spam, posts from bots, or irrelevant content is thereby lower. The existence of standard application programming interfaces (APIs) also makes data collection from Reddit relatively straightforward.

To identify potential Redditors (Reddit subscribers) who self-report opioid usage on Reddit, 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 (PRAW).*

The choice of these subreddits was based on their topical relevance and high levels of community discussion and engagement. Collection of data from these subreddits was not keyword-based. Instead, the API allowed the retrieval of all publicly posted threads and the associated comments. After retrieving all available posts of the 47,327 Redditors who had posted on the selected subreddits, we selected a random sample of these Redditors (N=13,812) and collected each of their past public posts across all subreddits (i.e., their longitudinal timelines) between November 2006 (corresponding

*`https://praw.readthedocs.io/en/latest/`

to the earliest post available) and March 2019 (corresponding to the last date of data collection).

3.2.2 Annotation

From the 13,812 public timelines we collected, we randomly selected 40 Redditors' timelines (i.e., all their posts in different subreddits) for manual review and annotation. This process finally yielded 26,126 posts for annotation. The annotation process was iterative and involved several steps. The posts were manually analyzed to develop the annotation guidelines, and then preliminary rounds of annotation were performed. We then discussed the disagreements, and updated the annotation guidelines for further clarity, and the final annotation was performed on a total of 91,601 sentences (2,500,489 tokens).

Due to the complexity of the annotation task, involving many entity types, and large numbers of posts that contained no entities at all, rather than annotating separately and then computing inter-annotator agreement, the data was first annotated by the lead annotator based on annotation guidelines and reviewed by two members of the study team. Following the annotation of all posts by two annotators, the annotations were reviewed by the full team, disagreements were resolved via discussion, and the annotation guideline was updated. Subsequent annotations were carried out in the same manner, adhering to the annotation guideline. All disagreements were resolved via discussion.

Based on the annotation guidelines we annotated lexical expressions in posts into 30 entity types that are independent of each other. Among them, 10 entity types belong to the basic personal information category, such as Age, Gender, Marital status, Location, Income, etc. 20 entity types related to medication information, such as Medicine intake, Illegal drug use, NMPDU, Method of intake, etc. Figure 3.1 shows all 30 entity types and their statistics in the annotated dataset.

Entity Types	
Advice to Others	273
Age	107
Alcohol: Co-ingestion or Amount or Frequency	21
Amount	535
Clinical Impacts	246
Co-ingestion	19
Country of Residence	6
Education Level	17
Ethnicity	8
Gender	70
Household Income	10
IDU: Switch From or Instead of Prescription or In Addition ..	30
Illegal Drug Use	664
Location	193
Marital Status	149
Medical Condition	388
Medicine Intake	1,372
Method of Intake	301
Nonmedical Prescription Drug Use	412
Occupation	49
Relapse	67
Social Impacts	72
Source of Drug	20
Supplements	162
Tobacco Use	19
Transition From Use to Abuse/misuse	13
Transition to IDU	3
Vape Flavor	2
Vape Use	34

Figure 3.1: Entity types and the frequency of each entity type.

The annotation process of our extensive dataset highlighted the prevalence of readily identifiable concepts such as medicine intake and illegal drug use. It also revealed that instances of clinical and social impacts—central to our study—are notably scarce. This scarcity poses significant challenges for research, as these impacts are crucial for understanding the broader consequences of nonmedical substance use on individual health and societal dynamics. To address these challenges and align with our objective of developing more effective public health strategies, we have concentrated our efforts on these two underrepresented entity types, thereby creating the specialized Reddit-Impacts dataset. This focused approach aims to enhance our ability to detect and study these rare but critical impacts in the discourse surrounding

substance use.

3.2.3 Dataset creation

From the total of 26,126 posts, only 318 posts (approximately 1.22%) were annotated as having clinical or social impacts. This extremely low occurrence rate underscores the sparsity of relevant data within the larger dataset. Due to the vast size and sparse nature of the original dataset, we opted to randomly select a subset of 1,380 posts for our experiments. We divided the annotated data into 3 sets: 60% for training, 20% for validation, and 20% for testing/evaluation. In summary, Reddit-Impacts comprises 843 posts for training, 259 for validation, and 278 for testing.

Table 3.2: Statistics of Reddit-Impacts dataset, including training and test sizes, the number of entity types and the number of entities in the dataset.

Datasets	**Entity Types**	**Training Size**	**Test Size**	**Entities**
Reddit-Impacts	Clinical Impacts,	30k tokens	6k tokens	0.2k tokens
	Social Impacts	1,102 posts	278 posts	318 posts

This refined dataset formation was pivotal for our experiments and subsequent release of the Reddit-Impacts dataset for the SMM4H 2024 shared task, aiming to provide a resource that is both concentrated and rich in the entities of interest—clinical impacts and social impacts. The number of instances of our Reddit-Impacts dataset is also shown in Table 3.2. In addition, Figure 3.2 presents an example of posts and their labels.

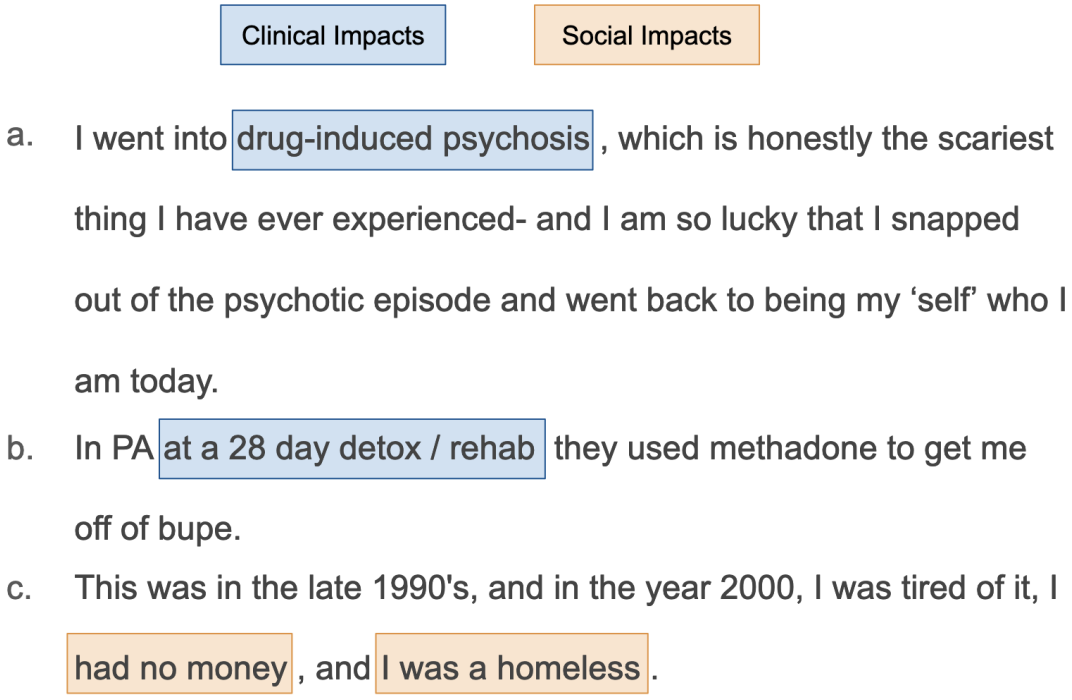


Figure 3.2: Sample posts in the REDDIT-IMPACTS dataset.

Chapter 4

Few-shot Learning for Biomedical NER: Benchmarking Studies

FSL for biomedical NER is an emerging research topic, and so there is a lack of benchmarks that allow the assessment of how well different approaches perform on the same data. To the best of our knowledge, no past research attempted to benchmark different FSL-based NER approaches on biomedical texts, and, at the same time, compared their performances to traditional NER models on the same datasets. In past related studies (*e.g.* [71]), only 1-2 biomedical text datasets were benchmarked on the same datasets. We attempt to address this gap in research. Specifically, we make the following contributions:

1. We benchmark several few-shot NER approaches on five standard biomedical text datasets.

2. We compare the performances of six models including three traditional NER models: BERT–Linear Classifier (BLC)*, BERT–CRF (BC)† and SANER‡, and

*https://github.com/smitkiri/ehr-relation-extraction
†https://github.com/kyzhouhzau/BERT-NER
‡https://github.com/cuhksz-nlp/SANER

three few-shot learning NER models: StructShot & NNShot[§], Few-Shot Slot Tagging (FS-ST)[¶] and ProtoNER[‖] on five biomedical text datasets.

3. We present a discussion of current research challenges for few-shot NER in the biomedical domain, and summarize important future research directions.

4.1 Traditional and FSL NER Models

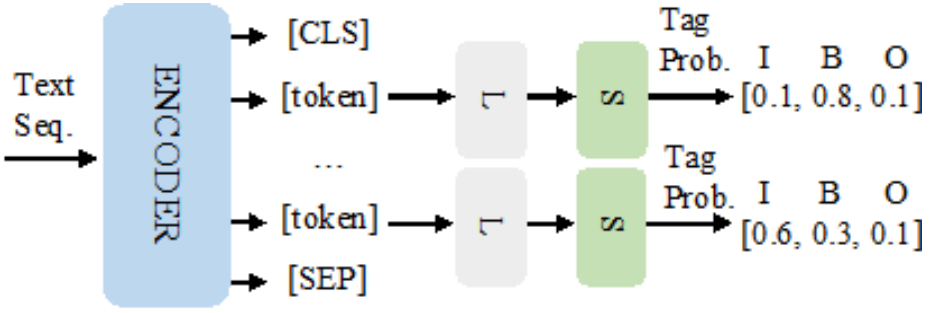
4.1.1 Traditional NER Models

We employed the base BERT model followed by a linear classifier as our first traditional NER model, and the model architecture is shown in Figure 4.1a. Taking advantage of self-supervised PLMs [35], we decided to use BERT to extract the contextualized representation of each token. The output of each token from BERT is passed through a fully connected neural network with a linear layer and a softmax layer at the end that projects tokens into entities. This architecture is typical for NER, and has been shown to achieve state-of-the-art performance on various datasets when sufficient training data is available. [89, 155]

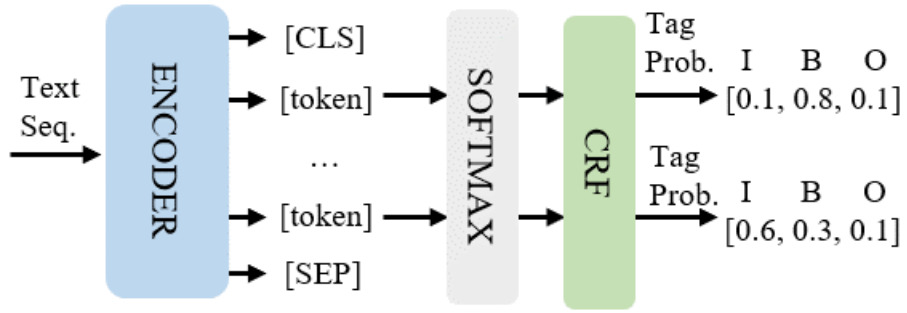
The second traditional NER scheme we used is also based on BERT but with a conditional random fields (CRF) layer after the softmax layer instead of a linear layer like the previous system. Figure 4.1b shows the basic architecture of this model. The linear classifier leads to the conditional independence of each classification decision, and thus, it is necessary to design a transition matrix with context relevance. The CRF layer can explicitly model the dependencies between entities as a table with transition scores between all pairs of entity types and add some constraints to ensure that the final prediction result is valid. These constraints can be automatically learned by the CRF layer during the training phase. For NER, texts are usually encoded in

[§]https://github.com/asappresearch/structshot
[¶]https://github.com/AtmaHou/FewShotTagging
[‖]https://github.com/Fritz449/ProtoNER

BIO or IO format ("B" represents "the beginning of the entity", "I" represents "the inside of the entity", and "O" represents "outside"). CRFs are effective in capturing dependencies between entities (*e.g.,* I-Drug cannot follow O, it must always follow B-Drug). In addition to the transition matrix, a CRF learner also includes an emission matrix. This emission matrix can be trained with Bi-LSTM, or it can be initialized randomly, but the performance is typically not as good as BERT. Consequently, we chose the BERT-CRF model as one of the traditional NER models in our experiments.



(a) BERT-Linear Classifier (BLC)



(b) BERT-CRF (BC)

Figure 4.1: Architectures of two traditional NER models based on BERT. (a) BERT-Linear Classifier (BLC): The output of each token from BERT is passed through a fully connected neural network with a linear layer and a softmax layer at the end that projects tokens into entities. (b) BERT-CRF (BC): The output of each token from BERT is fed into the CRF layer, which explicitly models dependencies between entities and adds constraints to ensure valid predictions.

The last traditional NER model we employed is SANER [128]. SANER is a neural-based NER method for social media text that utilizes augmented semantics to improve performance. The system contains a semantic enhancement module and a

gate module to encode and aggregate information separately so as to solve the problems of data sparsity when dealing with short and informal texts. We included this model since biomedical texts invariably contain domain-specific formal or informal language and abbreviations.

4.1.2 Few-shot Learning NER Models

The first two FSL-based NER models we explored were StructShot and NNShot [179]. StructShot uses contextual representations for each token in the support (training) set, and then uses a nearest neighbor (NN) and a Viterbi decoder to capture label dependencies. The authors of this model use a standard and reproducible evaluation setup for the few-shot NER task by using standard test sets and development sets from several domains. NNShot is a simpler variant of StructShot, which computes a similarity score between a token in the test example and all tokens in the support set without using the Viterbi decoder. The performance of StructShot was shown to be better than that of NNShot.

The second model we included is the Few-Shot Slot Tagging model (FS-ST) [68], which also includes a CRF layer. Since CRF considers both the transition score and the emission score to find the global optimal label sequence for each input, the framework in this paper includes two components: Transition Scorer and Emission Scorer. The transition scorer component captures the dependencies between labels. The authors introduce a collapsed dependency transfer mechanism into the CRF to transfer abstract label dependency patterns as transition scores. Specifically, they collapse specific labels into three abstract labels: O, B and I and modeled the transition from B and I to the same B (sB), a different B (dB), the same I (sI) and a different I (dI). To calculate the label transition probability for a new domain, the authors evenly distribute the abstract transition probabilities into corresponding target transitions. Then, the similarity between the word and each entity type is used

as the CRF emission score. In order to calculate this similarity, the paper proposes a label-enhanced task-adaptive projection network (L-TapNet) based on the few-shot classification model TapNet, which represents labels by using label name semantics. Unlike StructShot model, FS-ST model uses the more popular meta-learning framework for training and evaluation.

The third few-shot NER model is the ProtoNER model [44]. Metric learning methods, such as prototypical networks [153], which use prototypes (the average embeddings of support instances of each class) as the representations of each class, then compare the similarities between query instances and prototypes of each class based on certain distance metrics, showed state-of-the-art results in FSL for image classification tasks. Despite its success in image processing, metric learning has not been widely used in NLP tasks. Instead, in FSL settings for NLP, transfer learning is a more popular approach. Therefore, trying to adapt prototype-based methods such as prototypical networks for few-shot NER tasks naturally becomes another way of solving this problem. The ProtoNER model explored this possibility.

4.2 Data Collection and Preparation

Datasets are often reconstructed from existing ones to fit FSL scenarios. Since the most common way to represent data is *K-Shot-N-Way*, in this study, we conducted four sets of experiments using various proportions of the training data for the three traditional NER modes: 1-shot, 5-shot, 10%, and 100%. For the FSL models (StructShot & NNShot and FS-ST), we conducted experiments using 1-shot, 5-shot, and 15-shot settings. For the 10% setting, we randomly sampled 10% of the training set, and for 100% setting, we used the full training set.

For few-shot NER tasks, reconstructing datasets can be complex, as each text segment may contain more than one entity, often making it difficult to ensure that the

reconstructed datasets include only K labeled samples for each entity type. Therefore, we followed the construction method proposed by Yang et al. [179], and used the greedy sampling strategy to construct the training sets (*support sets*). In particular, we sampled entity sentences in increasing order relative to their frequency. Take 5-shot setting (K=5) as an example. We first extracted the entity type with the lowest frequencies, and after collecting 5 text segments containing this entity, we considered the entity type with the second lowest frequency and checked whether it appeared in the support set less than 5 times, as selected text segments may contain multiple entities. If it did occur less than 5 times, then we added more segments until it occurred 5 times. We followed these steps until all the entity types were included. The result of the greedy sampling strategy is to ensure that all entity types appear in the support set at least K times. If any `(instance, entity)` pair is deleted from the support set, at least one entity type appears in the support set less than K times.

4.3 Experimental Setup

Due to the structural differences in the included models, we could not use uniform parameter settings for all of them. Thus, we implemented experiments according to the parameter settings described in the publication associated with each model, including the number of training epochs, batch size, learning rate, and random seed numbers.

4.4 Results

Table 4.1 presents the F_1-scores of six NER models on each biomedical dataset. The table shows that all traditional NER models have relatively good performances when using full training data during training phases, especially on the relatively high-quality datasets (*i.e.*, datasets which were collected and analyzed using a strict set of guide-

Table 4.1: F_1-scores of six NER models on five biomedical datasets. The best performance in 5-shot settings and 1-shot settings has been highlighted in bold and underlined.

Models	Training Size	N2C2 2018	I2B2 2014	MIMIC III	BioNLP 2016	SMM4H 2021
SANER (traditional NER model)	**Whole training data**	80.63	90.62	66.57	81.78	44.56
	10% training data	79.27	80.67	46.68	70.50	23.4
	5-shot	10.27	**36.38**	**21.25**	23.14	0.00
	1-shot	7.92	**31.14**	7.07	4.32	0.00
BERT + Classifier (traditional NER model)	**Whole training data**	59.47	76.47	65.71	81.23	47.30
	10% training data	42.32	34.69	30.29	58.44	25.83
	5-shot	3.27	0.00	0.00	1.71	0.00
	1-shot	0.00	0.21	0.57	0.15	0.00
BERT + CRF (traditional NER model)	**Whole training data**	82.79	80.63	59.58	77.62	45.45
	10% training data	64.09	27.84	20.5	61.35	2.74
	5-shot	0.00	0.00	0.00	0.00	0.00
	1-shot	0.00	0.00	0.00	0.00	0.00
StructShot (few-shot model)	**5-shot**	**25.44**	20.30	3.18	0.03	0.00
	1-shot	17.59	20.26	0.63	0.00	0.00
NNShot (few-shot model)	**5-shot**	25.29	19.73	19.71	**28.88**	0.00
	1-shot	**16.70**	16.35	**15.37**	6.42	0.00
FewShot-Tagging (few-shot model)	**5-shot**	0.94	0.27	0.60	3.32	0.32
	1-shot	4.59	0.14	5.17	**6.81**	**0.35**

lines that ensure consistency and accuracy (low ambiguity), such as the N2C2 2018 and I2B2 2014 datasets). Even on the relatively noisy SMM4H dataset (social media), which has a small training set size, only one entity type, and relatively ambiguous annotations, their performances have been shown to be quite good [56]. SANER outperformed the other two traditional NER models on most datasets and most settings. In the few-shot scenarios, however, the F_1-scores for the BLC and BC models are mostly 0.00, suggesting that it is difficult for these models to generalize the characteristics from such small training data.

From the table, we can also see that NNShot outperforms most other models in few-shot settings. This finding contrasts the performances reported previously in the literature. We suspect that compared with StructShot, NNShot might have a better ability to extract and generalize features from biomedical texts. Another somewhat surprising result comes from the FS-ST model. In their original work, FS-ST model used the Ontonotes 5.0 dataset [173], WikiGold dataset [59], and several domains of

the SNIPS dataset [27] for training, and the remaining domains of SNIPS were used as the test set for evaluating. The reported performances of the FS-ST model on these datasets are far better than those for the biomedical domain data that we used. This suggests that both the similarity of texts and the overlap of entity types between these three datasets are higher than those of the biomedical datasets.

Table 4.2 shows the results of ProtoNER model. This model does not reconstruct datasets for satisfying few-shot settings. Instead, it randomly selected N sentences for training from the original dataset whose entity types are not evenly distributed. Then, it conducts separate experiments for each entity type. Therefore, we were unable to obtain the F_1-score for the entire dataset, and thus did not compare its results with other models. However, the values in table 4.2 show that, when the number of instances of an entity is obviously insufficient, the F_1-score is also very low, even going down to zero on occasions. In contrast to the success of the prototypical network in the field of image classification, its performance on few-shot NER tasks is not as competitive. The performances shown in the table are not high enough for application in real-life settings.

4.5 Discussion

In the benchmarking results shown in Table 4.1 and Table 4.2, the most important observation is perhaps that in few-shot biomedical NLP settings, all the models perform relatively poorly. The F_1-scores almost invariably fall below 30%, which renders them unsuitable for practical applications. More research is clearly required to develop FSL methods that are applicable in practical settings. This is particularly true for NER tasks involving biomedical data. Some of their performances in low-shot settings were, however, higher than the performances of traditional NER systems, which suggests that there is some promise for FSL NER methods.

Table 4.2: F_1-scores of ProtoNER model. The results obtained according to each entity type of each dataset. The entity types with less than 10 instances and their performance have been highlighted in bold and underlined.

Datasets	**Labels**	**Instances of label**	**F_1-score**
N2C2 2018	**Drug**	12510	63.92
	Strength	5519	83.45
	Form	5398	87.93
	Frequency	4062	63.57
	Dosage	3280	75.12
	Route	4672	85.72
	Duration	461	0.00
	Reason	2962	40.81
	ADE	692	24.24
I2B2 2014	**PATIENT**	903	61.81
	DOCTOR	1986	63.37
	USERNAME	60	93.88
	PROFESSION	161	66.67
	HOSPITAL	945	55.03
	ORGANIZATION	88	19.99
	STREET	162	96.87
	CITY	293	69.99
	STATE	250	76.71
	COUNTRY	61	85.71
	ZIP	164	90.14
	LOCATION-OTHER	**4**	**0.00**
	AGE	874	79.99
	DATE	5087	69.31
	PHONE	175	81.97
	FAX	**5**	**0.00**
	EMAIL	**2**	**0.00**
	URL	**6**	**0.00**
	HEALTHPLAN	**1**	**0.00**
	MEDICALRECORD	337	78.57
	IDNUM	78	54.05
	DEVICE	**7**	**0.00**
	BIOID	**1**	**0.00**
MIMIC III	**CONDITION/SYMPTOM**	2365	40.01
	DRUG	690	65.24
	AMOUNT	403	50.01
	TIME	326	40.63
	MEASUREMENT	665	49.85
	LOCATION	618	47.31
	EVENT	757	36.86
	FREQUENCY	62	0.00
	ORGANIZATION	114	28.62
	DATE	**2**	**0.00**
	AGE	44	95.25
	GENDER	36	99.98
BioNLP 2016	**GENE**	18258	27.87
SMM4H 2021	**ADE**	1124	7.84

We found that the quality of (*e.g.*, in terms of ambiguity) or the amount of noise in the datasets also plays a very important role in the performance of models on them. Although it is not shown in the tables, the performances of many models on the high-quality CoNLL 2003 dataset [143] are much better than that on other datasets (StructShot 1-shot on CoNLL 2003: 74.82%, FS-ST 1-shot on CoNLL 2003: 43.25%). It can also be seen in the table that almost all models have F_1-scores of 0.00 on the SMM4H dataset when the labeled data is very few. The SMM4H dataset is the only one that involves data from social media. Past research has shown that social media based biomedical NLP datasets are more difficult to obtain high performances on compared to biomedical datasets from other sources [146]. This is because social media data has specific characteristics that make NLP challenging, such as the presence of misspellings, colloquial expressions and noise. For example, "nosleep", as a symptom after taking drugs is marked as "adverse drug event" in one tweet, but not in another tweet, which might be due to the subtle differences in the contexts in which they are mentioned (*i.e.*, it can be an adverse event in some contexts and symptom in others, and it is not a standard biomedical term for either and therefore is unlikely to occur in other biomedical datasets).

The overarching aim of FSL is to enable systems to learn from few examples, as humans are often capable of doing [12]. Improving the intrinsic evaluation performances of FSL methods, especially on the datasets that have been explored, will still be one of the most important works in the future. Perhaps the other most influential work can be the creation of standardized publicly available datasets that will replicate real-world scenarios, and present actual FSL challenges. Currently, there is a paucity of such datasets, resulting in the need to reconstruct existing datasets to represent few-shot settings. Reconstructed datasets often do not accurately capture real-world scenarios. Specialized datasets representing few-shot scenarios will facilitate the thorough comparison of different FSL NER strategies, as well as the comparison of FSL

NER methods with traditional NER methods. Furthermore, there is currently no specialized dataset for FSL-based biomedical NLP, and contributions in this space are necessary to move the state-of-the-art in FSL-based NER for biomedical text forward. Future shared tasks should consider designing problems relevant to FSL-based NLP approaches.

4.6 Conclusion

In this chapter, we addressed the gap in benchmarking FSL approaches for biomedical NER by evaluating a variety of traditional and FSL-based models across multiple biomedical datasets. Our results demonstrated that traditional NER models perform well when ample labeled data is available, but their performance drastically declines in low-resource scenarios. In contrast, FSL models exhibited varying levels of success, with some outperforming traditional models in few-shot settings, but their overall performance remains far from practical applicability.

Our findings highlight several challenges in FSL-based biomedical NER, including the impact of dataset quality, noise, and the difficulty in generalizing across diverse biomedical domains. The results suggest a clear need for developing more robust FSL methods tailored to the biomedical domain. Additionally, the lack of standardized datasets specifically designed for FSL further complicates the evaluation and comparison of models, underscoring the importance of creating publicly available benchmark datasets that replicate real-world few-shot scenarios.

In conclusion, while current FSL approaches for biomedical NER show promise, significant improvements are needed to make them viable for real-world applications. Future work should focus on advancing FSL methodologies, improving dataset quality, and fostering the development of standardized benchmarks to enable rigorous evaluation and progress in this critical area of research.

Chapter 5

Data Augmentation with Nearest Neighbor Classifier

Data sparsity being the major issue in applying few-shot methods to biomedical NER, we explore data augmentation as a potential solution [29]. In the biomedical domain, the availability of labeled data is often limited, making it challenging to train robust models using few-shot learning techniques [36, 103].

One promising approach is semantic augmentation [24, 172]. Semantic augmentation adds semantic information such as synonyms, contextual data, or broader semantic relationships to the original data [102]. This makes the data more comprehensive, reducing the problems caused by information gaps due to data sparsity.

We also consider nearest neighbor (NN) classifier [87, 131] at inference, which is a simple and intuitive algorithm that makes predictions based on the closest training examples in the feature space. This approach helps preserve the local context of the data, which is crucial for biomedical NER, where the meaning and classification of entities often depend heavily on the surrounding context [28]. In scenarios with limited labeled data, the goal is often to classify data points into previously unseen classes based on a few examples. NN can seamlessly incorporate these new classes

into its decision-making process by simply including the new examples in its reference dataset [151].

5.1 Proposed Approach

The overarching aim of FSL-based NER systems is to learn from few examples to label names of entities of interest in text documents. In the following subsections, we first introduce the encoding procedure for augmenting semantic information, and then we present different distance metrics to explore the influences of methods for calculating similarities.

Rich semantic information is implicitly preserved in pre-trained word embeddings, making them potentially ideal resources for semantic augmentation. In order to generate contextual representations for all input tokens, we used a NER model with semantic augmentation [128] trained on the source domain as a token embedder to generate contextual representations of all tokens. The architecture of this data augmentation method combined with the nearest neighbor classifier (DANN) is shown in Figure 5.1. Considering a popular neural architecture for supervised NER models: a BERT-based NER model. For training these models on the source domain, we will follow the setting from Nie's paper [128]. After we obtain the pre-trained embeddings from the BERT-based NER model, for each token in the input sentence, we extract the most similar words of the token according to their pre-trained embeddings. Specifically, for each token $x_i \in \mathcal{X}$, we try to use pre-trained word embeddings from GloVe to extract the top m words that are most similar to x_i based on cosine similarities and denote them as:

$$C_i = \{c_{i,1}, c_{i,2}, \cdots, c_{i,j}, \cdots, c_{i,m}\} \tag{5.1}$$

Then use the BERT-based NER model again to get the embeddings $e_{i,j}$ of the ex-

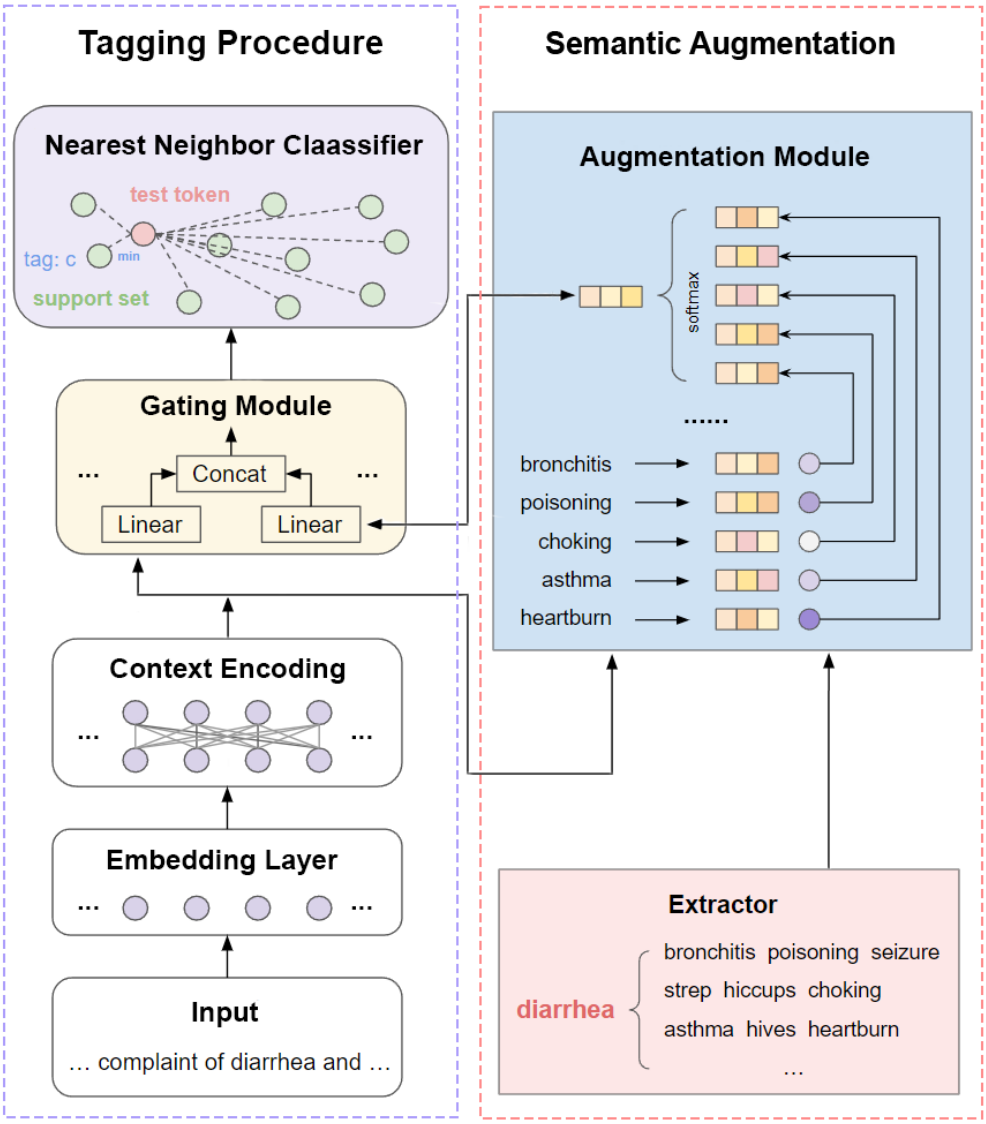


Figure 5.1: The overall architecture of our proposed model DANN: data augmentation method combined with Nearest Neighbor Classifier. An example sentence is given, where the augmented semantic information for the word "diarrhea" is also illustrated with the processing through the augmentation module and the gate module. After the gate module, the nearest neighbor classifier computes a similarity score between each token in the test set and all tokens in the support set, and it assigns the test token a tag c corresponding to the most similar token in the support set.

tracted words $c_{i,j}$. Since not all extracted words are helpful, afterwards, the augmentation module is used with an attention mechanism to weigh the semantic information carried by the extracted words. Specifically, for each token x_i, the augmentation module assigns a weight to each word $c_{i,j} \in C_i$ by:

$$p_{i,j} = \frac{\exp\left(\mathbf{h}_i \cdot \mathbf{e}_{i,j}\right)}{\sum_{j=i}^{m} \exp\left(\mathbf{h}_i \cdot \mathbf{e}_{i,j}\right)} \tag{5.2}$$

where h_i is the hidden vector for x_i obtained from the context encoder with its dimension matching that of the embedding (i.e., $e_{i,j}$) of $c_{i,j}$. Then, applying the weight $p_{i,j}$ to the word $c_{i,j}$ to compute the final augmented semantic representation by:

$$\mathbf{v}_i = \sum_{j=1}^{m} p_{i,j} \mathbf{e}_{i,j} \tag{5.3}$$

Therefore, the augmentation module ensures that the augmented semantic information is weighted based on their contributions. After the semantic augmentation module, a gate module [128] will be applied since the contribution of the obtained augmented semantic information to the NER task varied in different contexts. Particularly, we will use a RESET gate to control the information flow by:

$$\mathbf{g} = \sigma\left(\mathbf{W}_1 \cdot \mathbf{h}_i + \mathbf{W}_2 \cdot \mathbf{v}_i + \mathbf{b}_g\right) \tag{5.4}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are trainable matrices and $\mathbf{b}_g$ is the corresponding bias term. Afterwards, we use:

$$\mathbf{u}_i = [\mathbf{g} \circ \mathbf{h}_i] \oplus [(\mathbf{1} - \mathbf{g}) \circ \mathbf{v}_i] \tag{5.5}$$

to balance the information from the context encoder and the augmentation module, where $\mathbf{u}_i$ is the derived output of the gate module; $\circ$ represents the element-wise

multiplication operation and **1** is a 1-vector with its all elements equal to 1.

At inference, given a test example $\mathbf{x} = \{x_t\}_1^T$ and a K-shot entity support set $\mathcal{S} = \left\{\left(\mathbf{x}_n^{(sup)}, \mathbf{y}_n^{(sup)}\right\}_{n=1}^N$ comprising N sentences, we employed a token embedder $f_\theta(x) = \hat{x}$ to obtain contextual representations for all tokens in their respective sentences. Next, different distance metrics are used for computing similarities between tokens in the nearest neighbor classification.

5.1.1 Different Distance Methods

To improve the performance, we proposed two methods: replacing the representation of embeddings and changing the distance methods. For distance methods, we experimented with five approaches: squared Euclidean distance, cosine similarity, manhattan distance, infinity norm distance, and 3-norm distance, which are both commonly used measures of distance or dissimilarity.

The Euclidean distance between two points in Euclidean space is the length of a line segment between the two points.* The formula of Squared Euclidean Distance is:

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i - p_i)^2 \tag{5.6}$$

where p, q are two points in Euclidean n-space, q_i, p_i are Euclidean vectors, starting from the origin of the space (initial point), and n represents the n-space.

In the Euclidean space, Euclidean distance (2-norm distance) is usually used to compute the distance between two points. Other distances, based on other norms, are sometimes used instead.† For a point $(x_1, x_2, \ldots, x_n)$ and a point $(y_1, y_2, \ldots, y_n)$, the Minkowski distance of order p (p-norm distance) is defined as:

*https://en.wikipedia.org/wiki/Euclidean_distance

†https://en.wikipedia.org/wiki/Distance

$$
\begin{aligned}
\text{1-norm distance} &= \sum_{i=1}^{n} |x_i - y_i| \,| \\
\text{2-norm distance} &= \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2} \\
\text{p-norm distance} &= \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \\
\text{infinity norm distance} &= \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \\
&= \max \left(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n| \right)
\end{aligned}
\tag{5.7}
$$

The 2-norm distance is the Euclidean distance, and the 1-norm distance is more colorfully called Manhattan distance because it is the distance a car would drive in a city laid out in square blocks (if there are no one-way streets). The infinity norm distance is also called the Chebyshev distance. The p-norm is rarely used for values of p other than 1, 2, and infinity, so in our experiment, we only tried 3-norm.

Cosine similarity is a metric helpful in determining how similar the data objects are, irrespective of their size. In cosine similarity, data objects in a dataset are treated as a vector. The formula to find the cosine similarity between two vectors is:

$$
\text{cosine similarity } = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.8}
$$

where A and B are two given vectors, A_i and B_i are components of vector A and B respectively.

5.2 Results and Discussion

Table 5.1 shows the F_1-scores of our proposed model DANN with different distance metrics on five biomedical datasets, which we used in our benchmark. From the

Table 5.1: F1-scores of our proposed DANN model with four different distance metrics on five biomedical datasets compared with NNShot. The best performance of our models in 5-shot settings and 1-shot settings has been highlighted in bold and underlined.

Models	Training Size	N2C2 2018	I2B2 2014	MIMIC III	BioNLP 2016	SMM4H 2021
NNShot (few-shot model)	5-shot	**25.29**	19.73	19.51	**<u>28.88</u>**	0.00
	1-shot	**<u>16.70</u>**	16.35	**<u>15.37</u>**	6.42	0.00
DANN + Squared euclidean distance	5-shot	0.21	25.18	19.34	24.02	0.00
	1-shot	2.25	11.95	9.55	22.68	0.00
DANN + Manhattan distance	5-shot	0.16	**27.29**	**19.68**	24.21	0.00
	1-shot	1.95	10.81	5.38	22.97	0.00
DANN + Infinity norm distance	5-shot	0.13	16.99	16.99	23.97	0.00
	1-shot	1.68	**<u>16.99</u>**	9.70	22.84	0.00
DANN + 3-norm distance	5-shot	0.14	22.87	18.98	23.93	0.00
	1-shot	2.25	13.92	10.54	**<u>23.16</u>**	0.00

table, we see that our experimental results outperform the comparison model on most tasks, and our model's performance is the best one when conducting experiments on the I2B2 2014 dataset. On other datasets with different settings, the performance of our model is also already close to that of the baseline model, except for the N2C2 2018 dataset. We found that the experimental results on this dataset show that it is difficult for our model to extract useful semantic information from a few examples of this dataset. This is probably because the size of the N2C2 2018 dataset is the largest. Hence, the numbers of "O" entity types in the dataset are much higher than in other datasets, thus leading to the introduction of more noise. The table also shows that for the social media dataset (SMM4H 2021), none of the models are able to make accurate predictions with few samples. Previous research has shown that social media based biomedical NLP datasets are more difficult to obtain high performances as social media data has specific characteristics that make NLP challenging, such as the presence of misspellings and colloquial expressions.

For the same settings of the DANN model, we can horizontally compare five methods for calculating similarity. From Table 5.1, we see that Manhattan distance performs relatively well in the 5-shot setting, slightly outperforming other distance

metrics on three datasets. Meanwhile, the best-performing distance method in the 1-shot setting is the least frequently used 3-norm distance metric, which performs the best on three datasets.

The essence of our method is to change the input from the simple embedding generated by the BERT model to a more complex generation method. Specifically, we use a data augmentation module based on the nearest neighbor classifier. In this experiment, we used a BERT-based NER model to generate encodings, and used GloVe to select words that are similar to the input tokens. These are both good mechanisms for obtaining word vectors, but they have no unique advantages for biomedical data. Therefore, we also experimented with more domain-specific models, such as BioBERT and ClinicalBERT, to try to obtain the representations of tokens that are learned from biomedical or scientific data. These experiments, however, did not produce results better than other approaches.

5.3 Conclusion

In this chapter, we introduced the DANN (Data Augmentation with Nearest Neighbor) model, which combines semantic data augmentation with a nearest neighbor classifier to address data sparsity in few-shot biomedical NER tasks. By leveraging pre-trained embeddings to enhance semantic representation and incorporating a gating mechanism to balance augmented information with contextual features, DANN effectively preserves the local context critical for accurate biomedical NER.

Our experimental results demonstrate that the DANN model outperforms baseline methods on several biomedical datasets, particularly in low-resource scenarios such as 1-shot and 5-shot settings. We found that the choice of distance metric plays a significant role in model performance, with Manhattan and 3-norm distances performing best in different settings. However, the model faced challenges in extracting useful se-

mantic information from datasets with high noise or imbalanced entity distributions, such as the N2C2 2018 dataset.

While DANN shows promise in addressing data sparsity and improving few-shot NER, its performance on noisy social media datasets like SMM4H highlights the need for further advancements. Future work should explore more domain-specific embedding models and refined augmentation techniques tailored to biomedical text characteristics. Additionally, addressing dataset imbalance and noise remains a key challenge for achieving robust and generalizable NER systems in real-world applications. Overall, DANN provides a foundation for advancing data-efficient methods in biomedical NLP and paves the way for more effective few-shot learning solutions.

Chapter 6

HILGEN: Hierarchically-Informed Data Generation for Biomedical NER Using Knowledge Bases and LLMs

While the data augmentation method described in the previous chapter improved NER performance in biomedical text, there is room for further improvement in model performance. Our experiments and error analyses revealed that the NER performances vary substantially depending on the instances chosen for the training set. Therefore, our intuition was that expanding a given few-shot training set with synthetic data may help boost performance. In this chapter, we explore two synthetic data generation strategies that leverage knowledge encoded in the UMLS and LLMs. The UMLS has been curated and maintained for over two decades now, and consequently, it encapsulates a large volume of biomedical domain knowledge. Meanwhile, in recent years, the emergence of generative LLMs that can generate contextually relevant texts has opened up novel opportunities for generating problem-specific synthetic data. However, generic LLMs may face challenges when dealing with specialized biomedical concepts, as they may lack domain-specific knowledge and may generate

context-irrelevant or incorrect biomedical information [183].

To leverage encoded knowledge for generating synthetic data, we propose HILGEN (Hierarchically-Informed Data Generation for Biomedical NER Using Knowledge bases and LLMs), which infuses domain knowledge and hierarchical information from the UMLS [9], with synthetic data generated by LLMs. Our approach aims to enrich the representations of sparsely occurring biomedical concepts, thereby enhancing performance in few-shot learning for biomedical NER tasks—a relatively underexplored research space.

6.1 Background

6.1.1 UMLS in Biomedical Natural Language

The UMLS is widely used in biomedical NLP for its comprehensive repository of biomedical terminologies, concepts, and relationships, serving as a critical resource for tasks like NER, information extraction, and text classification. It provides a structured repository containing over two million concepts, including synonyms, hierarchical relationships, and semantic types, making it an invaluable resource for disambiguating, and standardizing biomedical terms, facilitating NLP systems in processing clinical and research data more accurately. By leveraging UMLS in this study, we ground LLM-generated examples in accurate biomedical contexts, ensuring that the representations of biomedical entities remain semantically coherent and clinically relevant. This integration enables the dynamic generation of enriched examples informed by UMLS, enhancing the model's understanding of rare biomedical terms and its ability to generalize across diverse biomedical datasets. Furthermore, the use of UMLS helps mitigate potential biases inherent in the training data of LLMs by providing a more balanced perspective on biomedical knowledge, ultimately resulting in more reliable outcomes in biomedical NER tasks.

6.1.2 Synthetic Data Generation

Data augmentation and transfer learning are widely used techniques in machine learning to address data sparsity by generating or utilizing additional data, such as synthetic or noisy data, to improve data representation and diversity. However, synthetic datasets often struggle to capture the naturalness and realism of human-written texts, particularly in biomedical domains, and they may introduce biases that affect the validity of downstream tasks. LLMs have been explored for generating biomedical text, leveraging their capacity to store and produce health-related information. Recent studies have demonstrated the potential of LLMs in data augmentation for clinical tasks, employing techniques such as the label-to-data method to mitigate the scarcity and sensitivity of biomedical data. Traditional data augmentation approaches using pretrained language models often involve fine-tuning on existing datasets to generate synthetic data. More recent methods focus on generating synthetic data with minimal supervision, using carefully crafted prompts or reverse tasks to produce high-quality data points. Unlike these approaches, our work augments data by incorporating domain knowledge from UMLS alongside LLMs, leading to the generation of high-quality synthetic data that enhances the performance of biomedical NER tasks in FSL settings.

6.2 Proposed Approach

Our method, HILGEN, generates synthetic data by comparing and integrating linguistic structures from LLMs with information extracted from the UMLS. The overall architecture of the proposed approach is shown in Figure 6.1. Our approach leverages the 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. We employ the GPT model to generate additional ex-

amples based on few-shot training examples. These related concepts are added to the few-shot training data to create additional synthetic instances. The new synthetic instances are then added to the original few-shot training data, and the models are fine-tuned on the augmented data. While UMLS-based data generation helps us augment the data with domain-specific knowledge, GPT-based data generation allows us to leverage the vast amount of open-domain data. We now provide further details about the data generation strategies and resources leveraged.

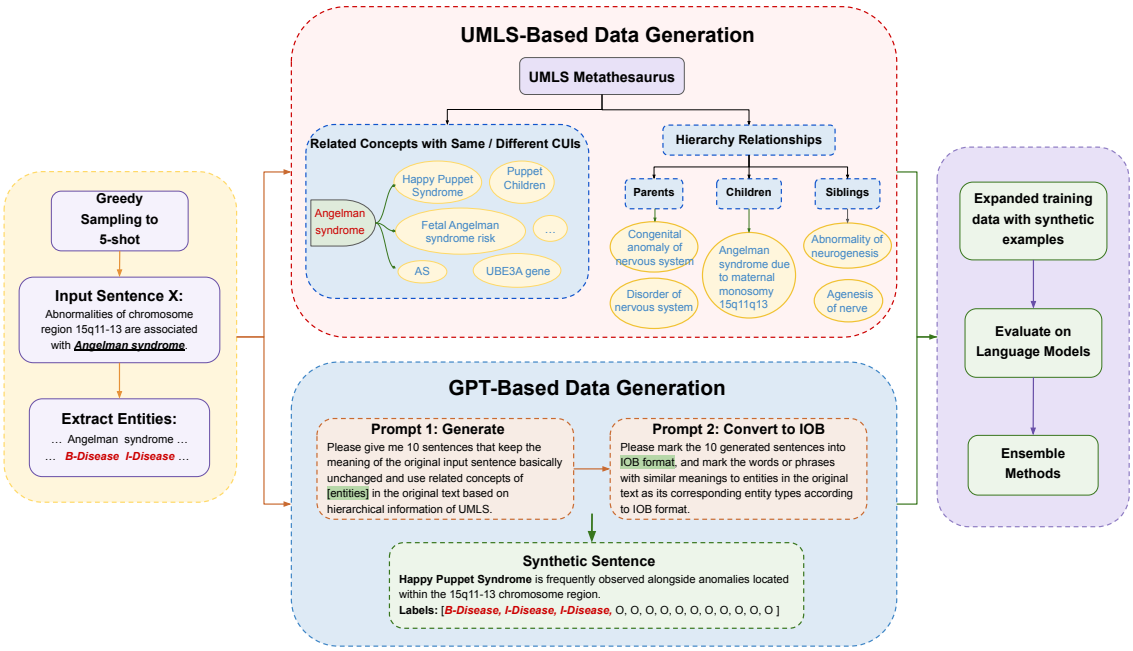


Figure 6.1: The overall architecture of the HILGEN model. UMLS- and GPT-based methods are first employed for synthetic training data generation. An ensemble of the two approaches is used for the final training data generation.

6.2.1 Hierarchical Information and Semantic Network in UMLS

One of the key features of the UMLS is its hierarchical organization of concepts, which represents the relationships between concepts in a hierarchical structure [9], similar to a tree. The hierarchy of information in the UMLS provides a way to access information about concepts organizationally related to a given concept. This

hierarchical structure allows for easy navigation of the UMLS and helps to organize and categorize concepts based on their relationships. The hierarchy includes several different types of relationships between concepts, including 'isa' (is a), 'has_parent', and 'has_child' relationships. Figure 6.2 shows an example of the tree-like structure of the UMLS.

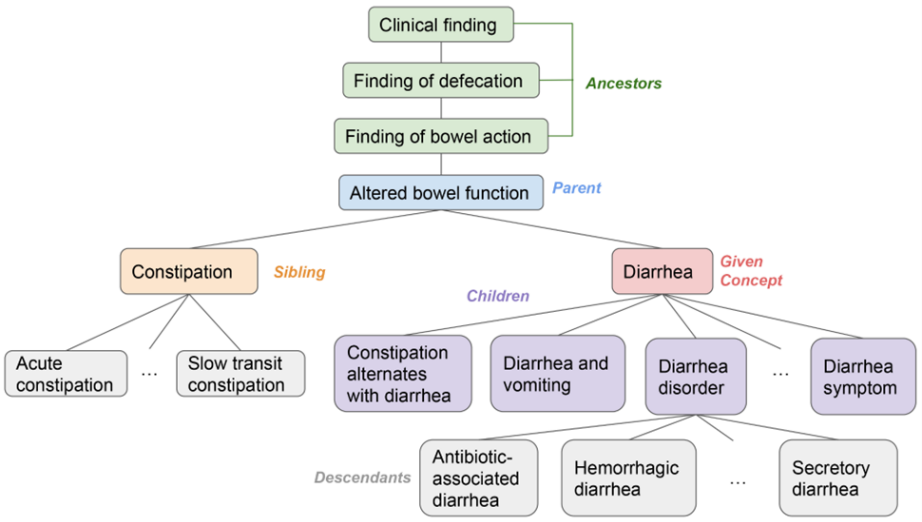


Figure 6.2: A subtree of the hierarchical structure of concept "diarrhea" in SNOMEDCT_US dictionary (Systematized Nomenclature of Medicine–Clinical Terms).

In addition to the hierarchy of concepts, the UMLS also includes a semantic network that describes the relationships between concepts in semantic space. The semantic network* 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 6.3. 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 [109]. Within each category, concepts are further organized based on their relationships to other concepts, such as 'isa' relationships or 'part_of' relationships.

Both the hierarchical information and the semantic network are important for

*https://www.ncbi.nlm.nih.gov/books/NBK9679/

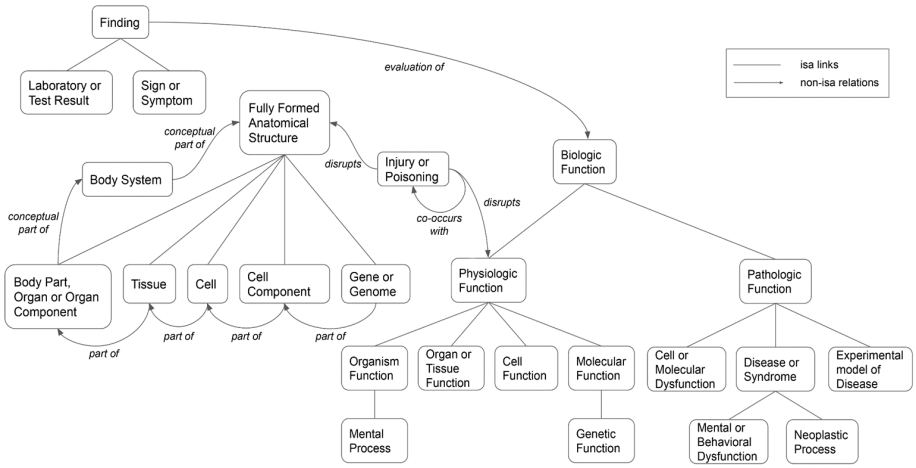


Figure 6.3: A portion of the UMLS semantic network. 'isa' links and 'non-isa' relations are represented in the figure, respectively.

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 fairly comprehensive understanding of the relationships between concepts within the UMLS.

6.2.2 UMLS-Based Data Generation

Our approach utilizes the UMLS to generate 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.

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 (a 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 biomedical free texts. The third layer of augmentation considers the hierarchical associations in biomedical 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 terminology that is widely used in the healthcare industry [8].

6.2.3 GPT-Based Data Generation

Our approach to utilizing GPT for generating new examples involves providing the model with complete sentences. As illustrated in Figure 6.1, we begin with an input sentence and use a two-step prompt to generate varied, semantically similar sentences that use different expressions for the same entities and convert the sentences into IOB format for subsequent fine-tuning. This strategy enables GPT to leverage contextual information to enhance its comprehension of the given concepts, thereby facilitating the generation of semantically coherent examples. Furthermore, we control the number of generated examples to match the quantity extracted from the UMLS, ensuring that the results are not biased by discrepancies in the amount 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 prompts used to extract hierarchical information and convert generated sentences into IOB format for GPT-based generation in the HILGEN model are listed in Figure 6.4.

Listing (a): Prompt for extracting related concepts

> Please give me 10 sentences that keep the meaning of the original input sentence basically unchanged and use related concepts of [entities] in the original text based on hierarchical information of UMLS.

Listing (b): Prompt for extracting parents and children

> Based on your knowledge of hierarchical information of UMLS, please find the parents and children of [entities] in the input sentence by using SNOMEDCT_US dictionary. Then, please give me 10 sentences that keep the meaning of the original input sentence basically unchanged and use parents and children of [entities] in the original text.

Listing (c): Prompt for extracting siblings

> Based on your knowledge of hierarchical information of UMLS, please find the siblings of [entities] in the input sentence by using SNOMEDCT_US dictionary. Then, please give me 10 sentences that keep the meaning of the original input sentence basically unchanged and use siblings of [entities] in the original text.

Listing (d): Prompt for converting sentences to IOB format

> Please mark the 10 generated sentences into IOB format, and mark the words or phrases with similar meanings to entities in the original text as its corresponding entity types according to IOB format.

Figure 6.4: Prompts used in HILGEN model generation process.

6.2.4 Fine-Tuning with Transformer-Based and Few-Shot Learning Models

We fine-tuned a transformer-based model, BERT, and our previously proposed few-shot learning model, DANN, on the expanded synthetic training data across four biomedical text datasets. BERT is a pre-trained transformer model widely adopted in NLP and serves as a baseline for comparison. The DANN model incorporates a semantic augmentation module with a nearest neighbor classifier, which enriches

the diversity and representativeness of the training data by selecting examples most similar to the target concepts. This approach provides additional context, improving the model's ability to generalize to unseen examples.

6.2.5 Ensemble Method

To further improve the robustness and accuracy of our biomedical NER models, we employ several ensemble approaches, including weighted voting and intersection. These ensembles combine models trained on synthetic data generated from both UMLS and GPT-3.5. The ensemble methods involve aggregating the predictions from multiple models to produce a final prediction. By leveraging the strengths of both data sources, the ensemble model enhances the overall performance, reducing the impact of any single model's weaknesses.

6.2.6 Comparison with ZEROGEN

We further evaluated the impact of synthetic text generation by the HILGEN approach against the ZEROGEN [181] system, a zero-shot learning framework that leverages large pre-trained language models (PLMs) to generate synthetic datasets for training smaller task-specific models. We specifically used GPT-3.5 for generating synthetic data to ensure a fair comparison and consistency in data generation at both the sentence and entity levels.

6.3 Datasets and Experiment Setup

We utilized four biomedical text datasets (MIMIC III, BC5CDR, NCBI-Disease, and Med-Mentions) as benchmarks to evaluate the performance of our approaches. These datasets provide a diverse range of clinical narratives and biomedical information, allowing for a comprehensive assessment of our methods. We conducted experiments

in the few-shot settings with 5 examples available for each label and used the metrics precision (P), recall (R), and F_1-score (F_1) for evaluation.

6.4 Results

In this section, we present the results of our experiments on four biomedical datasets using the HILGEN and other approaches. All approaches were trained and evaluated on the same data.

6.4.1 Experimental Results

The results in Table 6.1 demonstrate the effectiveness of HILGEN in generating synthetic data by incorporating prior knowledge through hierarchical information from UMLS and GPT-3.5. Leveraging both UMLS and GPT-3.5 for data generation, we observed significant improvements across all datasets. Incorporating knowledge from related concepts, as well as parent and child relationships from UMLS, often resulted in higher precision and F_1-scores, indicating that hierarchical and semantic relationships provide valuable context closely matching the target entities. The performance when using sibling relationships was somewhat mixed, with improvements in certain datasets but not consistently outperforming the other methods. Improvements were particularly noticeable in difficult cases where baseline models struggled to make accurate predictions.

When comparing GPT-3.5 to the incorporation of UMLS, both approaches showed improvements over baseline models. GPT-3.5 generally performed better across most datasets, suggesting its strength in generating diverse, contextually rich examples and 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 biomedical vocabularies, and the hi-

Table 6.1: Performance comparison of various synthetic data generation strategies for Biomedical NER Tasks. The table shows precision (P), recall (R), and F_1-score (F_1) for models trained on synthetic data generated by HILGEN using hierarchical information from UMLS and GPT-3.5. For each dataset, we compare the performance of the original 5-shot model, models using synthetic data generated with related concepts, parent-child relationships, and sibling relationships, and the best ensemble model.

Dataset		**MIMIC III**			**BC5CDR**			**NCBI-Disease**			**Med-Mentions**		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BERT-Large													
Original 5-shot	N/A	0.74	0.08	0.14	5.14	0.37	0.69	6.57	0.76	1.36	9.34	26.88	13.86
HILGEN: Generated by hierarchical information from UMLS	with related concepts	37.33	59.65	45.92	49.69	66.42	56.85	34.86	34.56	**34.71**	27.62	60.13	37.85
	with parents and children	40.41	57.05	47.30	46.01	59.93	52.05	36.39	28.14	31.74	27.83	60.56	38.14
	with siblings	36.11	56.78	44.14	48.81	55.08	51.76	37.52	29.19	32.83	26.81	59.38	36.94
HILGEN: Generated by GPT-3.5	with related concepts	38.95	60.54	47.41	51.10	52.83	51.95	30.32	32.34	31.30	27.94	60.29	38.18
	with parents and children	41.95	62.08	**50.06**	46.29	62.87	53.32	28.81	30.14	29.46	28.02	62.12	**38.62**
	with siblings	34.44	63.06	44.54	49.26	68.12	**57.18**	30.99	33.64	32.26	27.32	60.34	37.61
HILGEN: Best-Ensemble	N/A	43.72	60.16	**50.63**	53.17	63.97	**58.06**	37.79	34.51	**36.07**	29.36	64.97	**40.44**
DANN Model													
Original 5-shot	N/A	19.22	21.40	19.68	27.66	50.52	35.75	18.67	27.93	22.38	48.05	57.62	52.40
HILGEN: Generated by hierarchical information from UMLS	with related concepts	52.16	58.11	54.97	52.41	73.76	**61.27**	33.65	46.04	**38.88**	60.79	67.86	64.13
	with parents and children	51.09	56.34	53.59	51.98	72.33	60.49	35.78	35.78	35.78	60.03	67.57	63.58
	with siblings	53.95	60.26	**56.93**	50.63	65.83	57.23	34.87	40.68	37.55	60.13	68.12	63.88
HILGEN: Generated by GPT-3.5	with related concepts	46.87	62.34	53.51	53.72	69.53	60.61	35.21	40.86	37.82	61.08	68.92	**64.76**
	with parents and children	46.22	58.91	51.80	47.08	62.57	53.73	34.31	39.72	36.82	60.44	68.76	64.33
	with siblings	41.54	56.64	47.94	53.11	69.30	60.13	35.24	41.01	37.91	60.28	67.81	63.83
HILGEN: Best-Ensemble	N/A	52.79	64.60	**58.68**	60.52	73.85	**65.09**	37.10	42.99	**39.83**	63.49	70.28	**66.72**

erarchical information from UMLS contributed to more accurate and contextually relevant synthetic data, highlighting its usefulness in providing structured biomedical knowledge.

6.4.2 Comparison with ZEROGEN

Table 6.2 provides a detailed comparison of the performance metrics (precision, recall, and F_1-score) between the ZEROGEN and HILGEN approaches, which clearly illustrates the superior performance of HILGEN compared to ZEROGEN in all evaluated

Table 6.2: Comparison of ZEROGEN and HILGEN approaches using BERT-Large and DANN Models on biomedical datasets, demonstrating HILGEN's superior performance across all metrics and datasets.

		MIMIC III			**BC5CDR**			**NCBI-Disease**			**Med-Mentions**		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BERT-Large	ZeroGen	10.17	3.34	4.62	35.74	21.64	26.96	21.22	4.25	7.82	14.70	20.30	17.06
	HILGEN	43.72	60.16	**50.63**	53.17	63.97	**58.06**	37.79	34.51	**36.07**	29.36	64.97	**40.44**
DANN Model	ZeroGen	17.32	7.80	10.75	47.95	34.05	39.82	17.13	10.78	13.24	45.42	18.46	26.25
	HILGEN	52.79	64.60	**58.68**	60.52	73.85	**65.09**	37.10	42.99	**39.83**	63.49	70.28	**66.72**

scenarios.

HILGEN achieved up to a 28.19% higher F_1-score on the BC5CDR dataset on average. In contrast, ZEROGEN's zero-shot approach, though efficient, often generated more generic and less domain-specific data, resulting in lower precision, particularly in datasets like MIMIC III and Med-Mentions. HILGEN's incorporation of UMLS also led to more consistent improvements across datasets, demonstrating its ability to more accurately reflect the complexity and specificity of biomedical language compared to ZEROGEN.

6.4.3 Ensemble Approach

Table 6.3: Enhanced performance of ensemble with predictions from GPT-3.5 on biomedical datasets. Despite HILGEN's competitive results, the ensemble method, which combines HILGEN and GPT-3.5 outputs, improves the overall performance.

	MIMIC III			**BC5CDR**			**NCBI-Disease**			**Med-Mentions**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
GPT-3.5	62.99	64.10	63.54	56.81	83.61	67.66	27.79	45.10	34.39	25.66	37.81	30.57
HILGEN	52.79	64.60	58.68	60.52	73.85	65.09	37.10	42.99	39.83	63.49	70.28	**66.72**
Ensemble	59.82	70.22	**64.60**	72.03	71.65	**71.84**	38.85	43.39	**40.99**	53.16	65.66	58.75

The results in Table 6.3 present the performances of the ensemble approaches in improving performance, depicting notable improvements in precision, recall, and

F_1-scores. For the MIMIC III dataset and BC5CDR dataset, the ensemble method outperforms both the results from GPT-3.5 inference and the use of HILGEN for generating synthetic training data. This improvement demonstrates the robustness of the ensemble strategy in handling diverse biomedical texts and achieving higher accuracy in entity extraction.

HILGEN outperforms GPT-3.5 on the NCBI-Disease and Med-Mentions datasets, achieving an F_1-score of 66.72% on Med-Mentions, compared to GPT-3.5's 30.57%. The ensemble method further improves performance by combining the strengths of both models, resulting in a more robust and accurate NER system, as reflected by higher F_1-scores.

6.5 Discussion

Our findings underscore the value of incorporating structured domain knowledge, such as that found in UMLS, into synthetic data generation. By leveraging hierarchical relationships, HILGEN consistently produced semantically coherent examples, enhancing the performance of NER tasks, particularly in FSL scenarios. The improvements in precision and F_1-scores suggest that the hierarchical and semantic relationships embedded in UMLS provide valuable context for identifying and categorizing biomedical entities.

6.5.1 Challenges of Zero-Shot Data Generation Approaches

Zero-shot approaches such as ZEROGEN, while eliminating the need for extensive manual annotation, face certain limitations. Firstly, ZEROGEN uses generic prompts with minimal domain-specific constraints, often generating synthetic data that lacks the specific context found in biomedical texts, leading to overly generic or irrelevant content for NER tasks. Secondly, the generated data may exhibit inconsistencies in

style and structure relative to the original datasets, failing to capture the language patterns present in actual biomedical texts. Even when the required entity types are provided, ZEROGEN's synthetic datasets often have repetitive sentence structures, failing to capture the linguistic diversity of biomedical texts, which reduces the effectiveness of NER models. Thirdly, generating sentences with multiple entities that resemble original clinical or biomedical dataset structures is challenging. This is compounded by the fact that, although the entity type distribution may match, the generated text often fails to capture the nuanced context and relationships between entities, leading to a significant drop in model performance.

By incorporating UMLS, HILGEN benefits from a comprehensive and structured biomedical knowledge base, ensuring that the generated synthetic examples are semantically coherent and closely aligned with the domain-specific context of biomedical texts. This meticulous approach to maintaining sentence-level and entity-level consistency is crucial, as it allows the synthetic data to accurately reflect the intricate structures and relationships present in the original datasets, thereby improving HILGEN's ability to mimic them and significantly enhancing model performance.

6.5.2 Impact of Ensemble Learning on Model Generalization

The ensemble approach, combining models trained on synthetic data generated from both UMLS and GPT-3.5, consistently achieved the highest performance metrics across all datasets. This approach leverages the complementary strengths of UMLS's hierarchical domain-specific knowledge and GPT-3.5's diverse, contextually rich examples. By integrating the outputs of models trained on different synthetic data sources, the ensemble approach achieved balanced improvements in both precision and recall. Also, it largely mitigates the issue of data sparsity in FSL scenarios by effectively utilizing the diverse examples generated from UMLS and GPT-3.5. This results in more comprehensive training data, enabling the model to generalize more

effectively to unseen instances while maintaining accuracy. Our results highlight the complementary benefits of combining domain-specific knowledge from UMLS with the generative capabilities of LLMs.

6.6 Limitations

While HILGEN presents a robust approach for generating high-quality synthetic data based on few-shot scenarios for biomedical NER tasks, several limitations must be acknowledged. First, the scope of our current data generation and expansion is somewhat limited. Specifically, we identified and used only the top 10 related concepts for each entity, and our expansion and generation process relied on a 5-shot setting. It is plausible that utilizing a higher number of annotated examples, such as 10 or 20, and incorporating a wider array of related concepts could potentially yield superior results. Our primary objective in this study was to establish the feasibility and effectiveness of the HILGEN approach. We hypothesize that further expanding the synthetic dataset would result in improved model performance. Nonetheless, this expansion would also entail additional computational and resource costs.

Another limitation is our focus on methodology over prompt engineering. The prompts we used were relatively basic. Based on the evolving space of LLMs and their increasing capabilities, we believe that more sophisticated prompt engineering may lead to better results. In the next chapter, we describe our pursuit of this promising avenue and subsequent findings.

6.7 Conclusion

Our experiments showed that the HILGEN model, which combines synthetic data from UMLS and LLMs, performs significantly better than other baseline approaches in few-shot settings. Our motivation for using UMLS and GPT-3.5 for biomedical

data generation was twofold. First, UMLS, with its domain-specific knowledge, complements FSL by providing critical information that is not present in the training data, countering the belief that LLMs alone can replace expert-curated knowledge. This proves essential for improving predictions on unseen data. Second, GPT models offer a context-aware understanding of entities, enriching entity recognition and expanding training data without the need for additional manual annotations, especially for rare or complex cases. Overall, using information from the hierarchical structure of UMLS and LLMs as external knowledge bases can generate high-quality synthetic datasets to address key challenges in FSL with biomedical text datasets, including limited training data and the need for domain-specific knowledge. This approach can enhance FSL models across various biomedical applications, showcasing a valuable use case for long-established knowledge sources supporting biomedical NLP research.

Chapter 7

From Static to Dynamic: RAG-based Dynamic Prompting for Few-shot Learning

The research, development and evaluation of the HILGEN approach suggested that one possible course for further improving biomedical FSL NER performance is to improve the prompting and in-context learning strategies. The emergence of LLMs has transformed NLP capabilities, including for biomedical NER in few-shot settings. Methods for effectively leveraging LLMs, particularly in few-shot, restricted-domain settings are largely underexplored, although some past studies have suggested possible pathways. For example, the potential of inference from LLMs and prompt-based strategies in overcoming the challenges posed by few-shot settings has been demonstrated with techniques like LM-BFF [104], which utilizes prompts to fine-tune models on limited data. Additionally, approaches like PPT [55] enhance prompt effectiveness by pre-training prompt token representations with unsupervised data. Building on past works, in this chapter, we systematically explore prompting strategies for improving performance, and we introduce a novel in-context learning strategy that

employs RAG to dynamically update the prompt based on the input text.

7.1 Background

The most common and straightforward method currently is to use pre-defined, *static* prompts. Static, in this context, refers to the use of the same, consistent prompt for every instance in a dataset. This means that regardless of the input content, the model applies the same fixed prompt and in-context examples for processing. However, static prompts lack flexibility and cannot adjust to specific input data, leading to sub-optimal performance when dealing with diverse and complex datasets. Their fixed format restricts performance potential, as they do not adjust based on context, even when more suitable annotated examples are present in the training data. Consequently, systems employing static prompts exhibit high variance depending on the relevance of the in-context examples to the input unlabeled texts.

7.1.1 Retrieval-Augmented Generation

Inspired by the RT framework [101], which combines Retrieval-Augmented Generation (RAG) and Chain-of-Thought (CoT) prompting to enhance few-shot biomedical NER performance. RAG is an advanced technique that enhances the capabilities of LLMs by combining retrieval with generative text modeling [98]. Unlike traditional generation methods that rely solely on the model's pre-existing knowledge, RAG introduces an additional step: retrieving relevant information from a corpus or database. This retrieval process is typically guided by similarity measures [107], such as cosine similarity between embeddings, which helps the model access contextually relevant examples or documents tailored to the input query. Once these relevant texts are retrieved, they are integrated into the prompt or used as additional context for the model's response generation, creating a more informed output.

The motivation behind RAG is to address the limitations of LLMs in handling tasks that require specialized or up-to-date information [46]. Even with extensive training data, LLMs may struggle with domain-specific concepts or recent developments due to knowledge cutoffs and lack of domain specificity in training corpora [77]. By introducing contextually relevant information at inference time, RAG can significantly improve performance in specialized applications [176], such as biomedical text analysis, where precision and relevance are critical.

In biomedical NER tasks, RAG may improve a model's adaptability by retrieving examples or contexts that closely resemble the input text, thereby enabling a more accurate identification of entities [176]. In FSL settings, RAG architectures have the potential to reduce reliance on large annotated datasets by dynamically selecting relevant data [73], making it particularly useful for domains where annotated data is limited. Additionally, RAG complements prompting techniques, like CoT prompting, by enabling stepwise reasoning based on retrieved information [142], which may lead to better precision and recall for complex, sparse entities.

In an attempt to address the inherent limitations of static prompts in FSL settings, we explore dynamic prompt updating techniques, which involve automatically retrieving training examples and adjusting prompts based on contextual similarity. Following the optimization of prompts, we evaluate the effectiveness of two types of prompting—static and dynamic—using multiple LLMs including GPT-3.5, GPT-4, and LLaMA 3 (open source), on multiple datasets.

7.2 Proposed Approach

7.2.1 Static Prompt Engineering

Figure 7.1 presents the components of the static prompts used for LLMs. We systematically designed task-specific static prompts for use with LLMs, which comprise

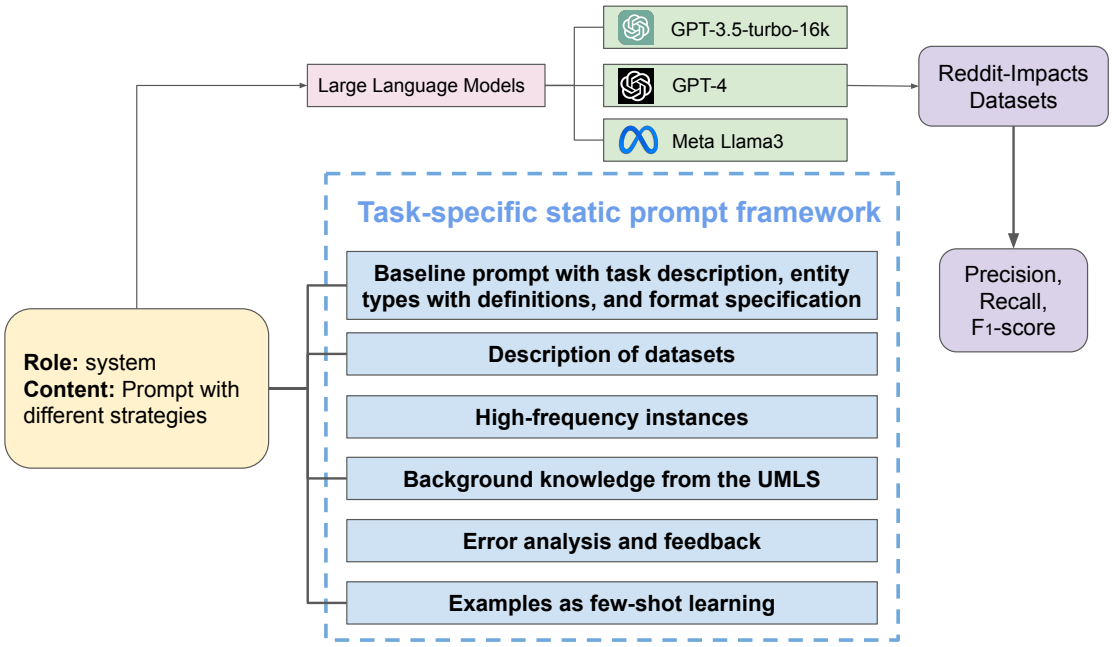


Figure 7.1: An overview of the NER strategy based on static prompting on three LLMs. Static prompts containing different information are provided to the LLMs, which, in turn, generate predictions for evaluation.

the following components:

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 their definitions. Entity definitions provide detailed and unequivocal explanations of an entity in the context of a specific task, crucially guiding the LLM toward accurately pinpointing entities within texts. Also, we provided the input, and instructions regarding the output format, which is a crucial step in ensuring the successful completion of the task. For generative LLMs, NER presents greater challenges, relative to classification, as it is essentially a sequence-to-sequence problem, where each token is assigned a corresponding label. However, when a prompt includes a sentence as is, we found that it can be difficult for LLMs to directly and accurately assign labels to each

token, resulting in mismatches in the number of input and output tokens. This issue is exacerbated by the fact that LLMs have their own tokenization mechanisms, which may differ from the tokenization in the annotated data. If the input and labels are provided in the BIO format instead, it often results in degraded performance due to the LLM's inability to fully understand the text.

One input approach is to provide a text and indicate the entities within it [175]. For example, in the sentence 'I was a codeine addict,' the phrase 'codeine addict' is identified as an entity and is annotated as 'Clinical Impacts'. However, this format can become ambiguous when faced with long sentences that contain the same word or phrase multiple times, each with different contextual meanings, not all of which may be labeled as the relevant entity. Another input method involves providing spans corresponding to the entities [70], but this also causes mismatches between spans and entities frequently when generative LLMs are used.

To address these challenges, we adopt a new format for constructing the input and output for the LLMs. We provide LLMs with a list of tokens that have already been tokenized. For the output, we instruct the model to return each token concatenated with its corresponding label. 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 example:

Input: ['I', 'was', 'a', 'codeine', 'addict.']

Output: ['I-O', 'was-O', 'a-O', 'codeine-B-Clinical_Impacts', 'addict.I-Clinical_Impacts']

To minimize the potential loss of sentence context caused using only tokens, we also explored the effectiveness of using the untokenized sentences as input, and tagged

tokens as output:

Input: ['I was a codeine addict.']
Output: ['I-O', 'was-O', 'a-O', 'codeine-B-Clinical_Impacts', 'addict.I-Clinical_Impacts']

2. Description of datasets: By describing a dataset's origin, content, and themes, we aim to provide LLMs with a basic understanding of the dataset. For example, for the REDDIT-IMPACTS dataset, we described that it focuses on individuals who use opioids, and we are interested in the impact of opioid use on their health and lives.

3. High-frequency instances: Some entities do not have clear definitions, and the determination is more ambiguous. Therefore, we provide the most frequently occurring words or phrases in each entity type within the training dataset to assist LLMs in understanding the potential distribution of entities and the theme of the text for this task. This is akin to providing a LLM with a lexicon of the concepts of interest.

4. Incorporation of background knowledge from the UMLS: We provide LLMs with comprehensive and structured information we obtained from the UMLS. Our intuition, based on the findings reported in the previous chapter, was that this knowledge could enhance the understanding and interpretation of biomedical concepts, relationships, and terminologies.

5. Error analysis and feedback: To improve the model's accuracy and address prediction errors, we provide an error analysis and feedback mechanism. After an initial set of predictions was made by LLMs on unseen training set instances, we manually reviewed the errors by comparing the model's predictions with the gold

standard annotations. For each incorrect prediction, we analyze the type and cause of the error, such as misclassification, missed entities, or spurious entities. Based on this analysis, we provide a summarization of feedback to the model. This feedback includes only general descriptions of errors without any examples. While this element of the prompt requires preliminary explorations of the dataset, common possible errors can be identified easily using a small set of training examples (*e.g.,* 5-shot), and this enables a mechanism of incorporating expert feedback into the process.

6. Annotated samples: We provide k annotated instances within the prompt for in-context learning. Samples are randomly selected and formatted according to the task description and entity markup guide.

We compared the effectiveness of different components of static prompting by incrementally incorporating descriptions of datasets, high-frequency instances, background knowledge from the UMLS, error analysis and feedback, and varying k-shot annotated samples. Detailed prompts used for each dataset are provided in Appendix B.

7.2.2 Dynamic Prompt Engineering

In prompt-based strategies using LLMs for in-context learning, the common approach has been to provide the model with a static prompt to guide its predictions. These prompts often include example instances, and CoT prompting. However, a significant limitation of this approach is that the provided examples may differ substantially from the texts from which the model is expected to extract named entities. Note that even in the presence of additional annotated samples, the LLMs context window size may limit the number of instances that can be embedded in a prompt for in-context learning. A static prompt, thus, does not generalize well, leading to high variance in performance.

To address this issue, we attempted to improve upon static prompting, and adopted a dynamic approach involving RAG. In our proposed approach, a retrieval engine is first indexed with the annotated examples from the training set. Upon receiving an input sentence, the system first retrieves the top n annotated examples using the retrieval engine. The retrieved examples are then embedded into the prompt, which is then passed to the LLM along with the input text. Figure 7.2 presents an overview of the system architecture.

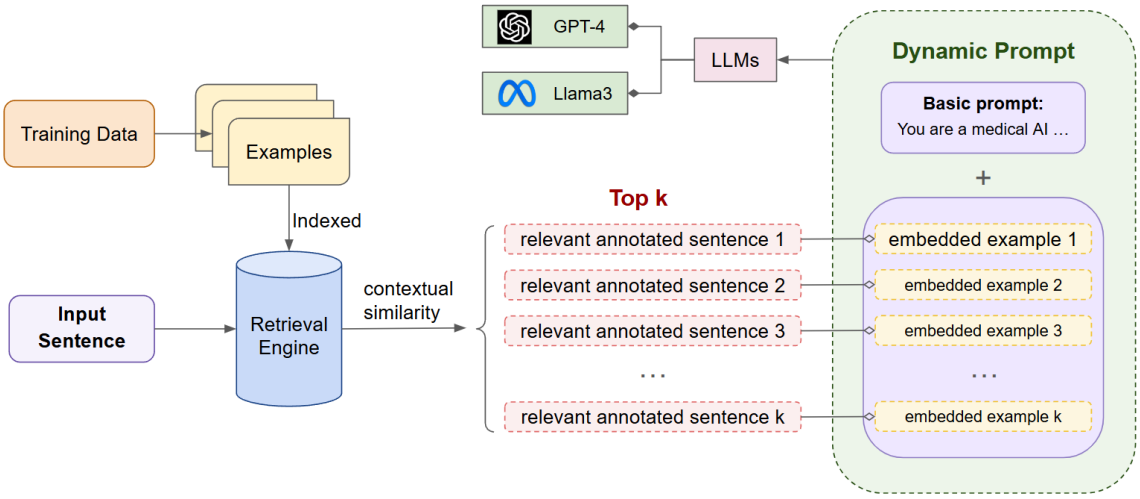


Figure 7.2: Overview of Retrieval-based Dynamic Prompting model. First the training data are provided to the retrieval engine for indexing. During inference, the system first ranks all training examples based on contextual similarity with the input text. Finally, the top n retrieved instances are embedded in the prompt, which is passed to the LLM (e.g., GPT-4, LLaMA 3).

Retrieval Engines

Selecting an effective retrieval engine is crucial since the examples embedded in the prompt influence the model's performance. We considered several retrieval methods, each chosen for its unique strengths in handling diverse biomedical texts, and applicability in FSL settings. The engines we selected are: TF-IDF [156], Sentence-BERT (SBERT) [138], ColBERT [84], and Dense Passage Retrieval (DPR) [83]. These search mechanisms offer a range of capabilities, from efficient keyword matching to advanced

deep-learning-based retrieval. We provide further details below.

1. TF-IDF: Term Frequency-Inverse Document Frequency (TF-IDF) scores the relevance of documents based on the frequency of terms. We included TF-IDF due to its efficiency and simplicity, which allows for rapid retrieval of relevant examples based on keyword overlap. While it lacks semantic understanding, it serves as a strong baseline, particularly when the input contains well-defined biomedical terminologies.

2. Sentence-BERT (SBERT): SBERT leverages a pre-trained BERT model fine-tuned for semantic similarity tasks. By encoding input sentences into dense embeddings, SBERT can capture the semantic relationships between sentences, making it well-suited for identifying contextually similar examples even when the input phrasing differs from the training data. This capability is particularly advantageous in the biomedical domain, where synonymous terms and varied expressions are common.

3. ColBERT: ColBERT (Contextualized Late Interaction over BERT) enhances retrieval performance by focusing on contextualized token representations. It uses a late-interaction mechanism that allows for more nuanced matching of query and document tokens. We selected ColBERT for its ability to capture fine-grained semantic details, which is essential for handling complex biomedical texts with diverse and context-dependent entity mentions.

4. Dense Passage Retrieval (DPR): DPR employs a dual-encoder architecture, where separate encoders are used for queries and documents. It uses deep neural networks to learn dense embeddings, optimizing for maximum similarity between relevant query-document pairs. DPR's strength lies in its ability to handle open-domain retrieval tasks effectively, making it a powerful choice for dynamically selecting annotated examples that are highly relevant to the input text, thus improving the

contextual adaptability of our dynamic prompts.

In our experiments, we evaluated the performance of each retrieval method, assessing their impact on few-shot NER across multiple biomedical datasets.

7.3 Experimental Setup

Below, we report our experimental setup for the two prompting strategies—Static Prompting and Dynamic Prompting.

1. Static Prompting: For static prompting, we evaluated three language models: GPT-3.5, GPT-4, and LLaMA 3. We used prompts containing five examples per label to provide context and guide the models' predictions. For GPT-3.5, we used the OpenAI API version "2023-07-01-preview", and for GPT-4, we used the version "2024-02-15-preview". Both models were configured with the following settings: temperature = 0.2, top_p = 0.1, frequency_penalty = 0, presence_penalty = 0, and no stop tokens specified.

For LLaMA 3, we used the Meta-Llama-3-70B-Instruct model, with a temperature of 0.5 and top_p of 0.95. Preliminary experiments (reported later in this chapter) revealed that GPT-3.5 consistently performed significantly worse compared to GPT-4. Hence, we excluded GPT-3.5 from further experiments in the dynamic prompting phase to limit API usage costs. To ensure robustness in the static prompting phase, the few-shot examples were randomly selected four times, and the reported results are the average of these four random selections.

2. Dynamic Prompting: In the dynamic prompting phase, we focused on evaluating GPT-4 and LLaMA 3 on multiple datasets. We conducted experiments using three different in-context settings: 5-shot, 10-shot, and 20-shot, to assess the impact of increasing the number of examples on the model's performance. The baseline

prompts in this phase also used randomly selected examples, with the results averaged over four random runs.

The evaluations were conducted on five biomedical datasets: MIMIC-III (clinical notes dataset), BC5CDR (disease and chemical entity recognition), NCBI-Disease (disease annotations from PubMed abstracts), Med-Mentions (large-scale UMLS concepts dataset), and our REDDIT-IMPACTS dataset (annotated for clinical and social impacts entity extraction). Further details about these datasets are provided in Chapter 3. We used precision (P), recall (R), and F_1-score (F_1) as evaluation metrics to comprehensively asses the models' performance across different datasets. In addition, to account for the variability in performance across different experimental runs, we include 95% confidence intervals (CIs) [124] for each metric, providing a measure of the statistical robustness of the results. The confidence intervals were computed via bootstrap resampling [38] with 1000 samples with replacement.

7.4 Results

7.4.1 Task-specific Static Prompting

The results, as presented in Table 7.1, demonstrate consistent performance improvements across the five biomedical datasets when all components of the static prompting strategy are combined for all three LLMs. Compared to the basic prompt (baseline), the integration of additional task-specific components, such as dataset descriptions, high-frequency instances, error analysis, and few-shot learning, led to significant improvements in performance metrics across all datasets and evaluation criteria. GPT-4 showed the largest improvements when the full structured prompt is used. For GPT-4, the average F_1-score increased by 0.09 across datasets, ranging from 0.08 (Med-Mentions) to 0.12 (BC5CDR). GPT-3.5 obtained an average F_1-score increase of 0.07, with gains ranging from 0.05 (NCBI) to 0.11 (BC5CDR). LLaMA 3-70B, which

Table 7.1: Performance comparison of various prompting strategies across different datasets in terms of F_1-score (F_1), Precision (P), and Recall (R). The row "BP + All components" represents the combination of all strategies, with the best performance across datasets highlighted in bold. The red bold font indicates the best F_1 score achieved by an individual component, while black bold font highlights the highest Precision, and underlined text denotes the best Recall for a single component. Additionally, green bold font is used to mark F_1 scores that are lower than the baseline performance (BP).

	Reddit_Impacts			*BC5CDR*			*MIMIC III*			*NCBI*			*Med-Mentions*		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
GPT-3.5															
Basic Prompt (BP)	10.37	43.26	16.73	55.64	76.88	64.56	55.31	54.11	54.70	18.28	51.33	26.96	8.55	10.12	9.27
BP + Description of datasets	13.25	52.38	21.15	59.25	81.47	68.61	59.54	54.18	56.73	26.24	50.25	34.48	16.52	10.33	12.71
BP + High-frequency instances	13.40	50.13	21.15	59.08	82.96	69.01	59.93	55.66	57.72	26.54	55.68	35.95	20.16	15.03	17.22
BP + UMLS knowledge	10.17	42.86	**16.44**	54.24	80.57	64.83	47.16	54.50	**50.57**	23.93	43.02	30.75	10.26	11.58	10.88
BP + Error analysis	12.02	48.21	19.24	57.36	82.49	67.67	64.63	55.16	59.52	25.23	48.31	33.15	18.74	13.24	15.52
BP + 5-shot learning with sentences	12.33	44.42	19.30	59.30	82.04	68.84	53.09	53.37	57.03	37.09	43.78	40.16	17.28	25.54	20.61
BP + 5-shot learning with tokens	**13.63**	53.11	**21.69**	**62.27**	82.02	**70.79**	**67.50**	55.97	**61.21**	**40.15**	46.32	**43.01**	**20.54**	30.57	**24.57**
BP + All above	**15.36**	**53.92**	**23.91**	**63.64**	**84.86**	**72.73**	**67.77**	**57.10**	**61.99**	**42.73**	**48.07**	**45.24**	**22.15**	**55.32**	**31.63**
GPT-4															
Basic Prompt (BP)	12.75	48.15	20.16	59.56	83.22	69.43	57.57	55.72	56.63	25.13	50.48	33.56	18.27	11.12	13.83
BP + Description of datasets	15.12	52.94	23.52	60.66	84.58	70.65	63.35	56.42	59.68	26.43	55.22	35.75	21.23	11.96	15.30
BP + High-frequency instances	15.98	53.75	24.64	63.89	84.06	72.60	**64.61**	56.14	60.08	35.02	41.44	37.96	21.72	17.69	19.50
BP + UMLS knowledge	12.85	50.14	20.46	59.48	84.63	69.86	55.37	54.90	**55.13**	22.80	47.92	**30.90**	18.72	11.83	14.50
BP + Error analysis	14.87	52.04	23.13	67.92	82.75	74.61	63.93	56.72	60.11	34.86	41.38	37.84	20.28	16.59	18.25
BP + 5-shot learning with sentences	14.71	51.48	22.88	65.04	83.18	73.00	58.49	55.03	58.25	36.96	45.67	40.86	27.42	30.33	28.80
BP + 5-shot learning with tokens	**17.23**	52.57	**25.95**	**68.10**	87.66	**76.65**	63.40	62.49	**62.94**	**40.72**	48.42	**44.24**	**27.71**	41.41	**33.20**
BP + All above	**18.87**	**52.01**	**27.60**	**68.62**	**90.32**	**78.03**	63.06	**64.12**	**63.58**	**45.02**	49.02	**46.93**	27.26	**60.06**	**37.49**
Llama3-70B															
Basic Prompt (BP)	9.93	36.42	15.61	52.52	76.04	62.13	46.57	55.64	50.70	15.63	24.37	19.15	19.59	23.17	21.23
BP + Description of datasets	13.27	35.26	19.28	58.53	80.22	67.68	56.64	55.81	56.22	17.30	36.43	21.44	23.59	19.87	21.57
BP + High-frequency instances	**14.52**	34.53	**20.44**	60.85	78.05	68.39	55.65	56.47	56.06	21.11	42.97	26.62	**23.99**	31.20	27.12
BP + UMLS knowledge	7.87	35.88	**12.91**	57.11	74.63	64.71	46.78	51.63	**48.92**	14.95	34.72	20.91	22.09	25.51	23.68
BP + Error analysis	12.46	38.86	18.87	58.96	80.52	68.07	63.18	55.20	58.92	16.84	44.65	24.46	22.64	29.92	25.78
BP + 5-shot learning with sentences	11.35	39.71	17.65	65.16	77.28	70.70	62.90	51.61	56.85	21.97	49.94	30.52	23.87	64.66	34.87
BP + 5-shot learning with tokens	13.33	40.32	20.04	**66.03**	78.57	**71.76**	**63.89**	60.19	**61.98**	**34.49**	32.41	**33.42**	23.72	68.45	**35.23**
BP + All above	13.16	**57.86**	**21.43**	**68.97**	78.36	**73.32**	59.30	**67.27**	**62.94**	**35.81**	**34.71**	**34.80**	**25.89**	**67.05**	**37.26**

started with the lowest baseline performance, showed an average F_1-score increase of 0.08, with its largest improvement observed in the REDDIT-IMPACTS dataset, with an increase of 37%.

Table 7.1 also shows that GPT-4 consistently outperformed GPT-3.5 and LLaMA 3-70B in all configurations, benefiting more from the integration of task-specific components, particularly in datasets such as BC5CDR and Med-Mentions, where it achieved the highest F_1-score. GPT-3.5, while achieving slightly lower overall performance still exhibited performance improvements relative to the baseline. This is evident in datasets such as REDDIT-IMPACTS, where its F_1-score exhibited significant growth with the integration of additional components. LLaMA 3-70B, despite its initially lower baseline performance, achieved competitive results when task-specific components were applied.

As illustrated in the Table 7.1, high-frequency instances, and dataset descriptions had the most notable impact on recall. For example, in the Med-Mentions dataset, adding high-frequency instances improved recall for GPT-4 by 0.05 (from 0.41 to 0.46). FSL at the token level provided the most significant increase in precision across models. For instance, precision in the NCBI dataset increased by 0.06 for GPT-3.5 (from 0.40 to 0.46) and by 0.08 for GPT-4 (from 0.36 to 0.44).

The box plot in Figure 7.3 further highlights the performance variability of different prompting strategies across datasets. The figure illustrates that the integration of UMLS knowledge yielded mixed outcomes across the datasets. While it improved recall in certain datasets such as BC5CDR, it underperformed compared to the basic prompt in datasets like REDDIT-IMPACTS and NCBI. This component aimed to provide foundational biomedical knowledge by introducing descriptions and context derived from UMLS. However, this approach may have introduced noise, particularly in datasets that are not strongly aligned with UMLS's predefined biomedical concepts. For example, in the REDDIT-IMPACTS dataset, GPT-3.5's F_1-score decreased

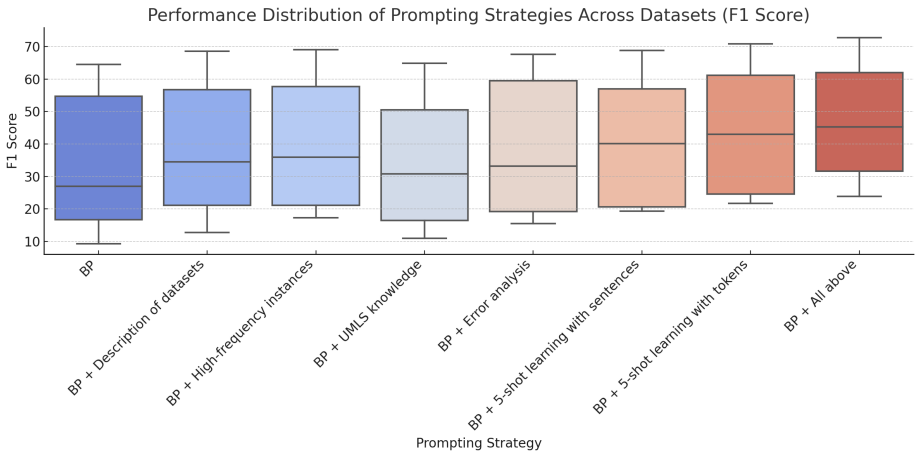


Figure 7.3: Performance distribution of prompting strategies across datasets (F_1-score). The box plots depict the performance of various prompting strategies applied to five biomedical datasets, highlighting the range, median, and distribution of F_1-scores for each strategy.

slightly from 16.73 to 16.44, suggesting that the background information from UMLS diluted the model's ability to capture task-specific cues.

Furthermore, 95% confidence intervals (CIs) for each metric are provided in Table D.1 in Appendix D.

7.4.2 Dynamic Prompting with RAG

The results in Table 7.2 demonstrate the effectiveness of dynamic prompting in different FSL settings (5-shot, 10-shot, and 20-shot) for GPT-4 and LLaMA 3 across five biomedical datasets. As mentioned in the Experimental Setup subjection, the baseline prompts used randomly selected examples, and the results were averaged over four random runs to ensure reliability. Detailed results for each random run, along with the averaged results, are presented in Appendix C. 95% confidence intervals (CIs) for each metric are provided in Table D.2 in Appendix D.

For GPT-4, somewhat surprisingly, TF-IDF retrieval consistently outperforms other methods in most cases. For example, on the BC5CDR dataset, TF-IDF achieves the highest F_1-score of 85.88% in the 5-shot setting, 86.64% in the 10-shot setting,

Table 7.2: Evaluation of dynamic prompting strategies (5-shot, 10-shot, and 20-shot) using GPT-4 and Llama 3 across five biomedical datasets. The table presents F_1-score, precision, and recall for each retrieval method: Base Prompt, TF-IDF, SBERT, ColBERT, and DPR. The row "Base" represents using static prompts we proposed in the former section.

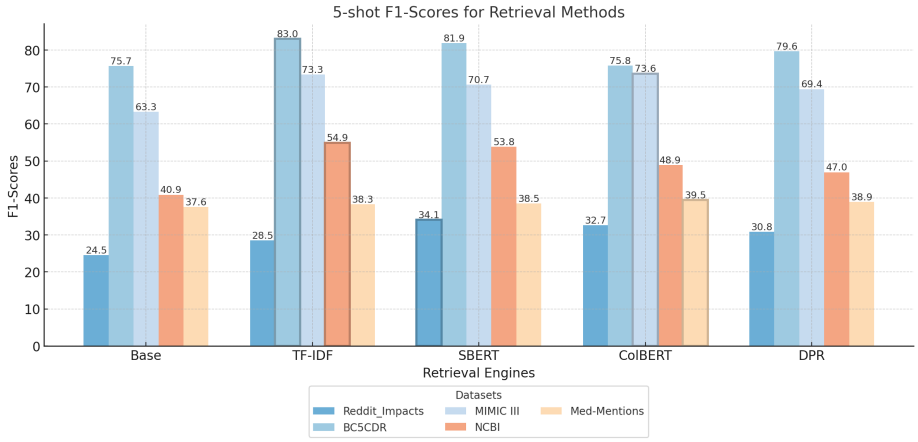
		Reddit_Impacts			*BC5CDR*			*MIMIC III*			*NCBI*			*Med-Mentions*		
		P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
GPT-4																
5-shot	**Base**	18.87	52.01	27.60	68.62	90.32	78.03	63.06	64.12	63.58	45.02	49.02	46.93	27.26	60.06	37.49
	TF-IDF	19.71	51.25	28.47	**82.31**	89.76	**85.88**	**74.43**	78.14	**76.24**	**56.86**	63.68	**60.08**	27.22	62.68	37.96
	SBERT	**24.31**	55.00	**33.72**	76.63	91.41	83.37	72.63	74.27	73.44	55.05	60.30	57.56	28.05	64.65	39.12
	ColBERT	22.66	56.79	32.39	78.64	81.03	79.82	74.14	77.02	75.56	50.43	54.48	52.38	**28.14**	68.69	**39.93**
	DPR	22.60	58.79	32.64	79.39	88.24	83.58	69.77	70.00	69.89	46.67	52.39	49.37	27.90	65.49	39.13
10-shot	**Base**	22.25	56.66	31.92	75.33	88.31	81.27	66.38	74.24	70.09	53.23	52.13	52.67	26.67	59.20	36.74
	TF-IDF	21.53	56.25	31.14	83.81	89.67	**86.64**	73.85	77.29	75.53	**58.81**	65.66	**62.05**	28.14	71.42	40.37
	SBERT	**25.41**	58.75	**35.47**	83.94	87.99	85.92	72.73	75.08	73.89	58.79	63.02	60.83	**28.32**	70.26	40.37
	ColBERT	23.86	58.02	33.81	83.49	88.05	85.71	**74.69**	78.06	**76.34**	55.12	59.56	57.25	28.15	71.99	**40.48**
	DPR	22.96	56.25	32.61	**85.16**	84.42	84.79	71.84	72.42	72.13	56.82	60.72	58.70	28.25	70.04	40.25
20-shot	**Base**	27.74	58.75	37.67	74.57	89.18	81.15	70.65	71.32	70.98	51.68	52.29	51.98	28.10	60.78	38.39
	TF-IDF	27.72	62.20	38.35	85.41	88.98	87.16	**75.81**	79.61	**77.66**	**61.80**	67.13	**64.36**	**28.20**	77.30	**41.32**
	SBERT	28.44	59.50	38.22	85.37	89.57	**87.42**	73.79	76.54	75.14	60.89	63.59	62.21	26.81	74.09	39.37
	ColBERT	**31.19**	66.67	**42.49**	82.09	83.94	83.00	75.27	78.19	76.70	56.13	59.35	57.69	27.70	75.47	40.53
	DPR	28.55	60.75	38.84	**85.81**	85.40	85.60	71.82	72.74	72.28	59.00	61.74	60.34	27.16	69.37	39.23
Llama3-70B																
5-shot	**Base**	13.16	57.86	21.43	68.97	78.36	73.32	59.30	67.27	62.94	35.81	34.71	34.80	25.89	67.05	37.26
	TF-IDF	18.89	58.62	28.57	**78.49**	81.78	80.11	66.48	74.84	70.41	48.93	50.70	49.80	26.46	72.06	38.68
	SBERT	**23.20**	66.67	**34.42**	77.26	83.79	**80.39**	64.04	72.21	67.88	**50.66**	49.59	**50.12**	26.15	68.92	37.91
	ColBERT	22.05	65.12	32.94	71.21	72.33	71.76	**68.37**	75.32	**71.68**	44.93	46.08	45.50	**26.68**	72.38	**38.99**
	DPR	19.20	59.26	29.00	74.47	76.91	75.67	65.74	72.54	68.97	41.06	48.66	44.54	26.51	71.38	38.66
10-shot	**Base**	22.37	59.94	32.50	72.56	77.91	75.15	59.13	71.63	63.77	39.67	31.49	35.60	25.57	64.33	36.50
	TF-IDF	23.53	62.65	34.21	80.82	80.32	80.57	55.79	55.34	55.56	49.59	49.41	49.50	24.03	68.00	35.51
	SBERT	22.27	59.76	32.45	77.72	84.94	81.17	67.67	76.09	71.63	**52.84**	49.94	**51.35**	**27.61**	66.88	**39.08**
	ColBERT	22.58	60.50	32.89	78.40	82.37	80.34	**69.65**	76.37	**72.85**	38.72	38.81	38.77	26.49	67.58	38.06
	DPR	**24.37**	57.83	**34.29**	**85.16**	84.42	**84.79**	65.85	73.68	69.54	47.60	45.04	46.28	25.80	70.97	37.85
20-shot	**Base**	24.52	53.81	33.67	**75.42**	75.58	75.50	62.01	62.12	62.05	40.71	42.58	41.62	26.57	64.79	37.67
	TF-IDF	27.62	66.95	39.11	74.64	82.47	**78.36**	55.95	58.51	57.66	45.39	49.83	47.50	**27.80**	64.39	**38.83**
	SBERT	**29.93**	68.06	**41.43**	75.04	80.75	76.85	**65.90**	64.77	65.35	42.09	46.40	44.14	25.48	61.36	36.01
	ColBERT	23.57	65.48	34.66	73.74	70.70	72.19	58.25	57.03	57.63	**47.08**	49.88	**48.44**	25.41	66.98	36.85
	DPR	26.15	65.04	37.30	72.58	77.15	74.80	62.72	69.19	**65.80**	37.18	44.13	40.36	26.10	62.88	36.89

and 87.16% in the 20-shot setting. Similarly, for the MIMIC III dataset, TF-IDF achieves the top F_1-score of 76.24% in the 5-shot setting and 77.66% in the 20-shot setting. In contrast, SBERT exhibits strong performance on the REDDIT-IMPACTS dataset, where it achieves the highest F_1-scores of 33.72% (5-shot) and 35.47% (10-shot). Moreover, SBERT achieves an F_1-score of 41.43% on REDDIT-IMPACTS in the 20-shot setting, outperforming TF-IDF by a margin of 3.08%. For LLaMA 3, DPR retrieval achieves competitive results, particularly on BC5CDR, where it achieves the highest F_1-scores of 84.79% (10-shot) and 74.80% (20-shot). SBERT also performs strongly on the REDDIT-IMPACTS dataset, achieving F_1-scores of 34.42% (5-shot) and 41.43% (20-shot).

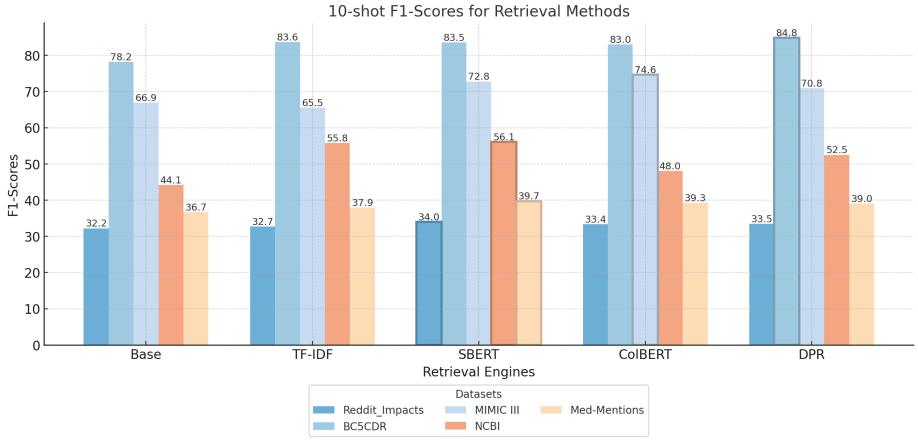
Both models benefit significantly from retrieval-augmented methods, as evident in the figures. TF-IDF and ColBERT frequently produce the highest F_1-scores for both models, demonstrating their effectiveness. SBERT also demonstrates consistent improvement over the base method, especially for GPT-4.

Figure 7.4 presents the F_1-scores of retrieval methods for different shot settings: 5-shot, 10-shot, and 20-shot. The results are averaged across evaluations conducted using GPT-4 and LLaMA 3 models. Across all settings, retrieval methods consistently improve performance compared to the Base prompt.

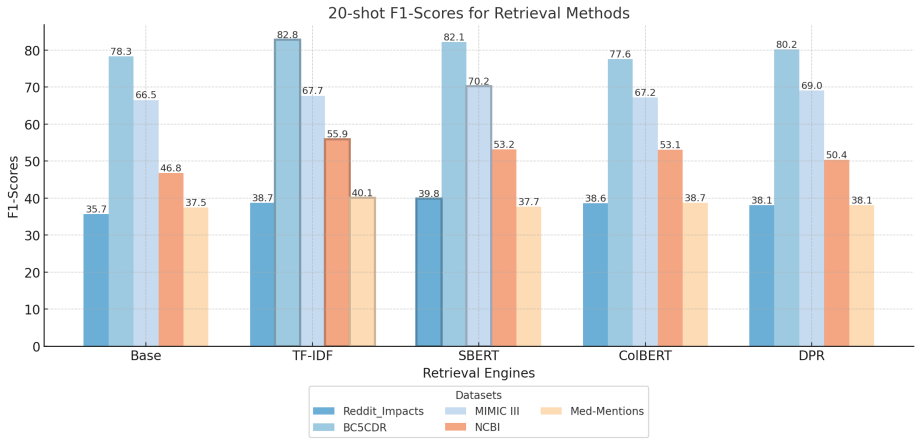
1. 5-shot Analysis: In the 5-shot setting, the SBERT retrieval engine achieves the highest average F_1-score for the REDDIT-IMPACTS dataset (34.1%), while TF-IDF performs best on BC5CDR (83.0%) and NCBI (54.9%). For MIMIC III, ColBERT leads with an F_1-score of 73.6%, and on Med-Mentions, ColBERT also stands out with a top score of 39.5%. These results highlight the dataset-specific strengths of different retrieval methods, with TF-IDF showing strong performance on entity-rich datasets like BC5CDR and NCBI.



(a) 5-shot F_1-scores for Retrieval Methods.



(b) 10-shot F_1-scores for Retrieval Methods.



(c) 20-shot F_1-scores for Retrieval Methods.

Figure 7.4: Comparison of average F_1-scores across different retrieval methods, for GPT-4 and LLaMA 3 models under varying shot settings.

2. 10-shot Analysis: In the 10-shot setting, SBERT stands out as the best-performing retrieval method overall, achieving the highest F_1-scores on three datasets: REDDIT-IMPACTS (34.0%), Med-Mentions (39.7%), and NCBI (56.1%). DPR achieves the top score on BC5CDR (84.8%), while ColBERT performs best on MIMIC III with an F_1-score of 74.6%. These results highlight a departure from the 5-shot setting, where TF-IDF dominated, indicating that SBERT is better suited for slightly larger data scenarios.

3. 20-shot Analysis: In the 20-shot setting, TF-IDF once again demonstrates strong performance, achieving the highest F_1-scores on three datasets: BC5CDR (82.8%), NCBI (55.9%), and Med-Mentions (40.1%). SBERT leads on REDDIT-IMPACTS with a top score of 39.8%, while it also performs best on MIMIC III with an F_1-score of 70.2%. These results highlight TF-IDF's and SBERT's consistent robustness across multiple datasets as the top-performing retrieval method.

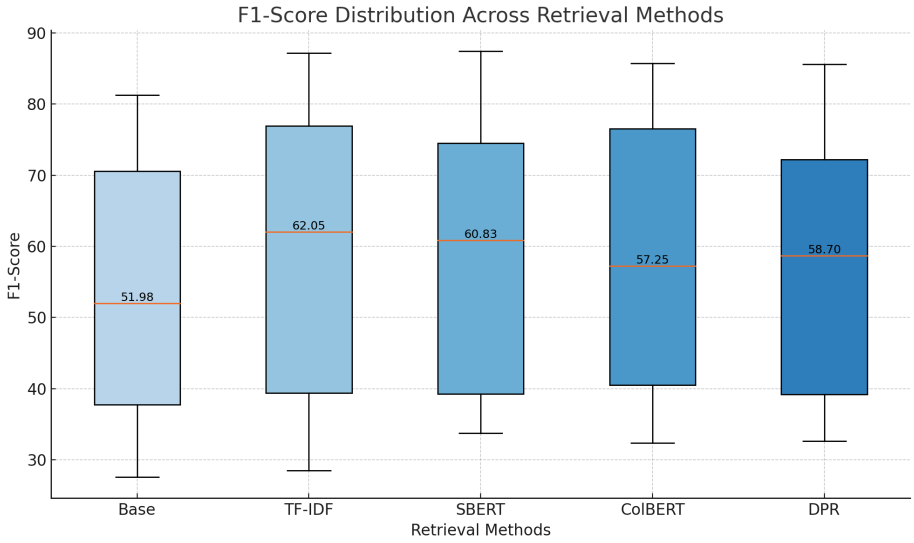


Figure 7.5: F_1-score Distribution Across Retrieval Methods.

Figure 7.5 shows the distribution of F_1-scores across different retrieval methods, providing an overview of their variability and effectiveness. All retrieval-based methods show significant improvements, demonstrating the benefit of incorporating re-

trieval methods into the prompting strategy. TF-IDF achieves the highest median F_1-score of 62.05%, indicating consistent performance across datasets, followed closely by SBERT with a median F_1-score of 60.83%. The variability is smallest for TF-IDF, as indicated by the narrow interquartile range, while methods like Base and DPR display higher variability, suggesting greater sensitivity to dataset characteristics.

Overall, GPT-4 consistently achieves higher F_1-scores compared to LLaMA 3 across most datasets and retrieval methods, particularly in 5-shot setting and 10-shot setting. This benefit becomes even more significant in datasets with sparse or noisy data, where retrieval-augmented methods play a critical role. LLaMA 3 shows comparable performance in 20-shot setting, but struggles to close the gap with GPT-4 in scenarios with fewer examples or more noisy data. This highlights GPT-4's robustness in leveraging limited training data.

Across all datasets and shot settings on GPT 4, larger training sizes (20-shot) tend to yield higher F_1-scores, precision, and recall. This suggests that larger training datasets, perhaps including synthetically generated data, can lead to further improvements. However, on LLaMA 3, this increase is less consistent, with the best performance often observed at the 10-shot setting across multiple datasets. The combination of effective retrieval methods and larger shot sizes contributes significantly to the overall improvements observed in model performance across all datasets. Based on the experimental results described in the previous chapter, combining RAG-based dynamic prompting with synthetic data generation methods may lead to further performance improvements.

7.5 Discussion

7.5.1 Analysis of Different LLMs

GPT-4 consistently outperforms GPT-3.5 and LLaMA 3-70B across datasets and configurations, demonstrating its robustness in understanding nuanced biomedical information. There might be several reasons. First, GPT-4 has significantly more parameters compared to GPT-3.5 and LLaMA 3-70B, enabling it to capture finer-grained contextual nuances, especially in complex and domain-specific tasks. The increased capacity allows GPT-4 to better model relationships between terms and concepts, particularly in structured datasets such as BC5CDR and NCBI. Second, GPT-4 is trained on a broader and more diverse corpus, which likely includes a richer representation of biomedical and scientific texts. This extensive exposure perhaps enhances its understanding of specialized terminologies and complex sentence structures, making it particularly effective in tasks like entity recognition and relationship extraction. Third, in datasets with sparse or ambiguous annotations, such as Reddit-Impacts or Med-Mentions, GPT-4 achieves higher recall, indicating its ability to identify relevant entities and relationships more comprehensively.

7.5.2 Performance Improvements via RAG-based Prompting

Retrieval engines improve performance by providing task-relevant context that enhances the model's understanding of the input, effectively bridging the gap between the model's general pretraining knowledge and the specific requirements of the task. By retrieving contextually relevant examples or background information, these engines reduce ambiguity and help the model focus on critical patterns, which is particularly beneficial for specialized domains like biomedicine. Our results broadly show that TF-IDF based retrieval works well for datasets that have low noise and limited out-of-vocabulary expressions. In contrast, engines like SBERT perform better

on linguistically diverse datasets by leveraging semantic embeddings, which capture nuanced relationships between words and phrases.

More advanced retrieval methods like ColBERT and DPR often underperform compared to TF-IDF and SBERT in biomedical tasks. This may be due to several reasons. ColBERT and DPR rely on dense representations, which, while powerful for general-purpose semantic matching, may fail to capture the precise, domain-specific distinctions critical in biomedical datasets. Furthermore, their reliance on dense embeddings can sometimes be overfit to irrelevant semantic similarities, retrieving documents that are semantically related but not contextually relevant to the query.

7.5.3 Variability in the Impact of Shot Size

The effect of shot size on performance is not uniform, as observed in the results across datasets. While increasing the shot size from 5 to 20 generally improves F_1-scores, the extent of improvement is dataset-dependent. Datasets with formal texts, like BC5CDR, which already benefit from retrieval engines aligning with predefined terms, exhibit marginal gains with additional examples. In contrast, noisy datasets like Reddit-Impacts are more sensitive to shot size, as more examples help the model adapt to diverse linguistic patterns and reduce misclassifications.

20-shot does not always yield the best results. One reason is diminishing returns: as the number of examples increases, redundancy or noise may be introduced, especially in datasets where retrieval engines already provide strong task-specific context. Another potential reason arises from the inherent constraints of LLMs, such as input token limits. As the shot size grows, the available space for processing task-specific context diminishes, potentially diluting the effectiveness of the prompt or truncating important information.

7.6 Limitations

This study demonstrates the effectiveness of retrieval-augmented prompting strategies in improving the performance of LLMs across diverse biomedical datasets. Despite the promising results achieved in this study, several limitations warrant discussion:

1. Lack of Biomedical-Specific Retrieval Methods While the study evaluates general-purpose retrieval engines, it does not incorporate biomedical-specific retrieval methods tailored to the biomedical domain, such as MedCPT [76]. Retrieval methods fine-tuned on biomedical corpora could potentially provide a better alignment with the linguistic and structural complexities of biomedical texts.

2. Dependence on Retrieved Results Our results rely on the quality and quantity of the retrieved examples. If the retrieval engine fails to retrieve sufficient or relevant results, such as ColBERT, model performance may be negatively impacted.

3. Indexing Only Sentences with Entities To ensure the retrieved examples satisfy the k-shot requirement (providing sufficient examples for each entity type), this study indexed only sentences containing entities. While this approach ensured prompt quality, it limited the scope of the retrieval process and may have excluded other relevant examples that could improve task generalization.

By addressing these limitations in future research, the study's findings can be further extended to explore the full potential of retrieval-augmented prompting strategies, particularly for specialized tasks in the biomedical domain.

7.7 Conclusion

In this work, we demonstrated the effectiveness of transitioning from static to dynamic prompting strategies, using RAG, for few-shot biomedical NER. Static prompts en-

riched with task-specific components improved performance, with GPT-4 showing the best results. Dynamic prompting further enhanced adaptability by retrieving contextually relevant examples, with methods like TF-IDF and SBERT outperforming others in most cases. While increasing shot size generally improved performance, diminishing returns were observed in some cases.

Dynamic prompting proved to be a robust approach for tackling data-sparsity challenges, but limitations such as computational overhead and dependency on retrieval quality remain. Future work should explore domain-specific retrieval methods and optimize prompting strategies to maximize efficiency and scalability in biomedical NLP.

Chapter 8

Conclusion

FSL approaches have substantial promise for NLP in the biomedical domain as many biomedical datasets naturally have low numbers of annotated instances. In this proposed thesis, we addressed the challenge of sparse data in NER task for biomedical texts through the application of few-shot learning algorithms and data augmentation techniques. Our research demonstrated that integrating domain-specific knowledge from the UMLS and leveraging the generative capabilities of LLMs can significantly improve NER performance in biomedical texts. We developed a novel method that combines data augmentation with nearest neighbor classifiers, which showed promising results in benchmark experiments. Additionally, the creation of the REDDIT-IMPACTS dataset provided a valuable resource for analyzing clinical and social impacts from social media data. Our findings underscore the potential of few-shot learning and synthetic data generation in enhancing biomedical NLP tasks, paving the way for more effective and scalable solutions in the field. Future work will focus on refining the methods of prompt engineering and exploring applications of LLMs in broader biomedical contexts.

This thesis provides significant advancements in sparse concept detection and recognition in biomedical texts using few-shot learning techniques. The study sys-

tematically benchmarked existing FSL models and highlighted their limitations in low-resource biomedical NER tasks. To address these challenges, three novel contributions were introduced: semantic data augmentation via nearest neighbor classifiers, synthetic training data generation leveraging the UMLS and LLMs, and dynamic prompting within a retrieval-augmented generation framework. The results demonstrated substantial performance improvements across multiple biomedical datasets, showcasing the potential of combining enriched data representation, domain knowledge, and adaptive prompting techniques to mitigate data sparsity issues.

The contributions of this thesis extend beyond performance gains; we provide a roadmap for integrating domain-specific knowledge with cutting-edge NLP methodologies, enhancing the practical applicability of FSL systems in biomedical contexts. These findings bridge gaps in existing research and lay a foundation for scalable, robust solutions to address the complexities of biomedical NER in real-world low-resource settings.

8.1 Future Work

Few-shot learning with LLMs presents significant opportunities for advancing biomedical NER. To address challenges related to data scarcity, domain specificity, and real-world applicability, future research directions should focus on technical innovations, practical applications, and overcoming current limitations.

8.1.1 Advancing Biomedical NER with Technical Innovations

Our future work will focus on enhancing the performance of LLMs in sparse concept detection and biomedical NER tasks by leveraging advanced fine-tuning methods tailored for low-resource settings. Biomedical-specific LLMs based on transformer architectures, such as BioGPT [115] and BioMedLM [10], a domain-tuned GPT variant. To

efficiently adapt these models to domain-specific tasks, we plan to explore lightweight fine-tuning techniques, like Quantized Low-Rank Adaptation (QLoRA) [34]. This approach allows efficient adaptation of base LLMs to domain-specific tasks, addressing challenges posed by limited annotated data and high computational costs. By utilizing QLoRA's innovative double quantization technique, we aim to achieve significant reductions in memory usage while maintaining or improving model performance. Specifically, we will explore how QLoRA can align LLMs with target outputs in scenarios with sparse data, evaluating its effectiveness across various biomedical NER tasks. In addition, we plan to investigate the trade-offs between memory efficiency, computational cost, and task accuracy, providing insights into its scalability and practicality for low-resource settings.

To further improve the reasoning capabilities of LLMs in biomedical NER tasks, our future work will also explore the application of Chain-of-Thought prompting. This technique facilitates step-by-step reasoning, enabling models to handle complex challenges such as nested or discontinuous entities more effectively. By explicitly structuring the reasoning process, CoT not only enhances accuracy in entity boundary detection but also improves the interpretability of model predictions, making it highly applicable to intricate biomedical contexts.

8.1.2 Applications of LLMs in Few-shot BioNER

Our current work does not deeply explore how entity types vary across datasets or investigate the implications of these variations for real-world applications, particularly in the context of few-shot NER in biomedical domain. For example, understanding the overlap or divergence in entity types across datasets could inform tasks such as identifying novel therapeutic applications for existing drugs, disease-specific information extraction, or the development of specialized clinical decision support tools. Furthermore, we have not yet conducted targeted experiments on individual entity

types, such as diseases, medications, or procedures, which could reveal their unique challenges and better support their use in specific biomedical contexts. In the future, we aim to analyze the distribution and characteristics of entity types across datasets and conduct targeted experiments to align the performance of LLMs with practical applications in the biomedical domain, particularly by leveraging few-shot learning techniques to address the scarcity of annotated biomedical data in BioNER tasks.

Looking ahead, we envision applying advancements in integrating LLMs with domain-specific knowledgebases like UMLS in real-world biomedical and research settings, where sparse data and complex biomedical terminology often limit the applicability of traditional models. Given these challenges, we believe synthetic data generation remains a valuable approach to augment limited datasets and improve model performance in such contexts. To this end, we considered two approaches to enhancing synthetic sentence generation using UMLS: encoding API access methods directly into the prompt to dynamically query UMLS [77] or feeding the results retrieved from UMLS back into the prompt to guide the generation process. These strategies leverage UMLS's hierarchical structure and semantic relationships, aligning with a chain-of-thought prompting approach to improve contextual relevance and accuracy in generated outputs. While such strategies require substantial research and development, these methods hold promise for improving tasks like decision-making, personalized treatment planning, and biomedical knowledge discovery.

However, synthetic data generation also presents challenges that require further investigation. One significant issue is hallucination, where LLMs may produce incorrect or non-existent entities, potentially compromising the quality and reliability of the synthetic data. Furthermore, evaluating the generated text poses another critical challenge, particularly in ensuring it aligns with domain-specific requirements and accurately reflects real-world biomedical contexts. Developing robust evaluation metrics and methods to measure the accuracy of generated text will also be one of the a

key focuses of future work. Addressing these issues will be essential for maximizing the impact of synthetic data generation in biomedical applications and ensuring its effectiveness for tasks such as few-shot NER.

Appendix A

Tables for Literature Review

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Rios et al. [140]	2018	MIMIC II [81] and MIMIC III [78]	Multi-label Text Classification	Frequent group (all labels that occur >5 times), the few-shot group (labels that occur between 1 and 5 times), and the zero-shot group (labels that never occur in the training dataset), reconstructed	Not mentioned	Medical, discharge summaries annotated with a set of ICD-9 diagnosis and procedure labels	Medical, discharge summaries annotated with a set of ICD-9 diagnosis and procedure labels
Rios et al. [141]	2018	MIMIC II [81] and MIMIC III [78]	Multi-label Text Classification	Original dataset, with no reconstruction	Not mentioned	Medical	Medical
Hofer et al. [66]	2018	i2b2 2009[188], 2010[189], 2012[161]; CoNLL-2003[143]; BioNLP-2016[25]; MIMIC-III[78]; UK CRIS[14, 158]	NER	10-shot	Not mentioned	Medical and nonmedical (e.g, news)	Medical
Pham et al. [132]	2018	The Europarl datasets[85], and IWSLT17[16] for English→Spanish; the UFAL Medical Corpus and HIML2017 dataset ‡ for German→English	Neural Machine Translation (NMT)	One-Shot	N/A †	German→English: medical; English→Spanish, the proceedings of the European Parliament and data from TED	German→English: medical; English→Spanish: data from TED
Yan et al. [178]	2018	Multigames dataset[177], HCR dataset[157] and SS-Tweet dataset[163], SemEval-2013 Dataset (SemEval b)[126]	Text Classification	Few-shot, but reconstructed	Multigames : 3 Hcr: 5 SS-Tweet: 3 SemEval-2013 Dataset: 3	Tweets about sentiment and games, and health Care Reform (HCR) data	Tweets about sentiment and games, and Health Care Reform (HCR) data
Manousogiannis et al.[118]	2019	Tweets (provided by SMM4H 2019)[174]	NER	Original dataset, with no reconstruction	1, ADR with 319 MEDDRA codes	Medical (ADR)	Medical (ADR)
Gao et al. [45]	2019	Based on FewRel dataset[61], this paper propose FewRel 2.0 by constructing a new test set from the biomedical domain	Relation Classification	5-Way 1-Shot / 5-Way 5-Shot / 10-Way 1-Shot / 10-Way 5-Shot	25	Wikipedia corpus and Wikidata knowledge bases	Biomedical literature with UMLS, a large-scale knowledge base in the biomedical sciences
Lara-Clares et al.[93]	2019	MEDDOCAN shared task dataset[120]	NER	500 clinical cases, with no reconstruction	29	Clinical	Clinical

Table A.1 continued from previous page

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Ferré et al. [40]	2019	BB-norm dataset from the Bacteria Biotope 2019 Task[11]	Entity Normalization	Original dataset with no reconstruction and zero-shot	Not mentioned	Biological	Biological
Hou et al. [68]	2020	Snips dataset[27]	Slot Tagging (NER)	1-shot and 5-shot	7	Six of Weather, Music, PlayList, Book (including biomedical), Search Screen (including biomedical), Restaurant and Creative Work.	The remaining one
Sharaf et al. [150]	2020	ten different datasets collected from the Open Parallel Corpus (OPUS)[165]	Neural Machine Translation (NMT)	Sizes ranging from 4k to 64k training words (200 to 3200 sentences), but reconstructed	N/A †	Bible, European Central Bank, KDE, Quran, WMT news test sets, Books, European Medicines Agency (EMEA), Global Voices, Medical (ufal-Med), TED talks	Bible, European Central Bank, KDE, Quran, WMT news test sets, Books, European Medicines Agency (EMEA), Global Voices, Medical (ufal-Med), TED talks
Lu et al. [113]	2020	MIMIC II[81] and MIMIC III[78], and EU legislation dataset[18]	Multi-label Text Classification	5-shot for MIMIC II and III, 50-shot for EU legislation	MIMIC II: 9 MIMIC III: 15 EU legislation: 5	Medical	Medical
Jia et al. [74]	2020	BioNLP13PC and BioNLP-13CG[127], CoNLL-2003 English dataset[143], Broad Twitter dataset[32], Twitter dataset[112] and CBS SciTech News dataset[75]	NER	Four few-shot (reconstructed) and zero-shot	CoNLL: 4 Broad Twitter: 3 Twitter: 4 BioNLP13PC: >=3 BioNLP13CG: >=3 CBS News: 4	For the BioNLP dataset, BioNLP13PC as the source domain dataset; In the Broad Twitter dataset, the CoNLL-2003 as the source domain dataset; In the Twitter dataset,the CoNLL-2003 as thesource domain dataset	for the BioNLP dataset, BioNLP13CG is used as the target domain dataset; In the Broad Twitter dataset, Broad Twitter is used as the target domain dataset; In the Twitter dataset, Twitter is used as the target domain dataset

Table A.1 continued from previous page

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Chalkidis et al.[19]	2020	EURLEX57K[18], MIMIC III[78] and AMAZON13K[97]	Multi-label Text Classification	The labels are divided into frequent (>50), few-shot ($\leq$ 50), and zero-shot	Not mentioned	English legislative documents, English discharge summaries from US hospitals, English product descriptions from Amazon	English legislative documents, English discharge summaries from US hospitals, English product descriptions from Amazon
Lwowski et al. [116]	2020	Tweets about COVID-19[91]	Text Classification	100 tweets, with no reconstructed	4	Tweets about COVID-19	Tweets about COVID-19
Hou et al. [69]	2020	dialogue utterances from the AIUI open dialogue platform of iFlytek[§]	Dialogue Language Understanding: includes two sub-tasks: Intent Detection (classification) and Slot Tagging (sequence labeling)	1-shot, 3-shot, 5-shot and 10-shot	Train Domains: 45 Dev Domains: 5 Test Domains: 9	General dialogue (including health domain)	General dialogue (including virus-Search domain)
Chen et al. [23]	2020	WIKIBIO dataset[94]	Natural Language Generation (NLG)	Dataset sizes: 50, 100, 200 and 500, with no reconstruction	N/A [†]	Books, Songs and Human domain (including biomedical)	Books, Songs and Human domain (including biomedical)
Vaci et al. [166]	2020	UK-CRIS system that provides a means of searching and analysing deidentified clinical case records from 12 National Health Service Mental Health Trusts[14, 158]	NER	Original dataset, with no reconstruction	7	Clinical	Clinical
Huang et al. [71]	2020	10 public datasets	NER	5-shot, 10%, 100%	CoNLL: 4 Onto: 18 WikiGold: 4 WNUT: 6 Movie: 12 Restaurant: 8 SNIPS: 53 ATIS: 79 Multiwoz: 14 I2B2: 23	10 public datasets, different domains	10 public datasets, different domains

Table A.1 continued from previous page

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Chen et al. [20]	2020	MRI image dataset and MRI text reports¶	Text Classificaition	Original dataset, with no reconstruction	Not mentioned	MRI data	MRI data
Yin et al. [182]	2020	MLEE [134] and BioNLP'13-GE [127]	Sequence Tagging (NER)	5-way-10-shot, 5-way-15-shot and 5-way-20-shot	5	Biological event	Biological event
Goodwin et al. [53]	2020	TensorFlow DataSets catalogue‖	Abstractive Summarization	Zero-shot and 10-shot	N/A †	3 general domain & 1 consumer health	3 general domain & 1 consumer health
Yang et al. [179]	2020	OntoNotes 5.0[173], CoNLL-2003[143], I2B2 2014[159], WNUT 2017[33]	NER	1-shot and 5-shot	Onto: 18 CoNLL: 4 I2B2-14: 23 WNUT: 6	Three of general, news, medical and social media	The remaining one of general, news, medical and social media
Hartmann et al. [63]	2021	The IULA dataset (Spanish)[119], The NUBES dataset (Spanish)[108], The FRENCH dataset[30]; all from the hospitals. And other Negation Scope Resolution datasets	NER	Zero-shot, with no reconstruction	1, Negation	No training data for the clinical datasets	clinical
Fivez et al.[43]	2021	SNOMED-CT1** disorder names as biomedical synonym sets and ICD-10	Name Normalization	zero-shot, with no reconstructed	N/A †	Biomedical	Biomedical
Lu et al. [180]	2021	construct and share a novel dataset †† based on Weibo for the research of few-shot rumor detection, and use PHEME dataset[187]	Rumor Detection (NER)	For the Weibo dataset: 2-way 3-event 5-shot 9-query; for PHEME dataset: 2-way 2-event 5-shot 9-query	Weibo: 14 PHEME: 5	Source posts and comments from Sina Weibo related to COVID-19	source posts and comments from Sina Weibo related to COVID-19
Ma et al. [117]	2021	CCLE, CERES-corrected CRISPR gene disruption scores, GDSC1000 dataset, PDTC dataset and PDX dataset‡‡	Drug-response Predictions	1-shot, 2-shot, 5-shot and 10-shot	N/A †	Biomedical	Biomedical
Kormilitzin et al. [86]	2021	MIMIC-III[78] and UK-CRIS datasets[14, 158]	NER	25%, 50%, 75% and 100% of the training set, with no reconstruction	7	Electronic health record	Electronic health record

Table A.1 continued from previous page

Study	Year	Data source	Research aim	Size of training set	Number of entities / classes	Entity type of training domain	Entity type of test domain
Guo et al. [56]	2021	abstracts of biomedical literatures (from relation extraction task of BioNLP Shared Task 2011 and 2019[11]) and structured biological datasets	NER	100%, 75%, 50%, 25%, 0% of training set, with no reconstructed	Not mentioned	Biomedical entities	Biomedical entities
Lee et al. [95]	2021	COVID19-Scientific[160], COVID19-Social[3] (fact-checked by journalists from a website called Politi-fact.com), FEVER[164] (Fact Extractionand Verification, generated by altering sentences extracted from Wiki-pedia to promote research onfact-checking systems)	Fact-Checking (close to Text Classification)	2-shot, 10-shot and 50-shot	Not mentioned	Facts about COVID-19	Facts about COVID-19
Fivez et al.[42]	2021	extract sets of high-level concepts and their constituent names from 2 large-scale hierarchies of disorder concepts, ICD-10 and SNOMED-CT**	Name Normalization	15-shot	N/A †	Biomedical	Biomedical

Table A.1: Articles published for few-shot learning on medical data, their publication years, data sources, search engine, which downstream tasks the literature focus, size of the training set (number of the shot), number of the entities (for few-shot NER tasks), number of the classes (for few-shot classification tasks), type of data in training domain, type of data in test domain.

* The research aim of this paper is text classification or NER, but the size of training set is not mentioned in the paper.

† The research aim of this paper is neither text classification nor NER.

‡ UFAL Medical Corpus v.1.0 and HIML2017 dataset: http://aiui.xfyun.cn/index-aiui. Last accessed November 22, 2021.

§ iFlytek: http://aiui.xfyun.cn/index-aiui. Last accessed November 22, 2021.

¶ Those datasets are not released.

‖ TensorFlow DataSets: https://www.tensorflow.org/datasets. Last accessed November 22, 2021.

** SNOMED-CT1: https://www.snomed.org. Last accessed November 22, 2021.

†† A novel dataset proposed by this paper: https://github.com/jncsnlp/Sina-Weibo-Rumors-for-few-shot-learning-research. Last accessed November 22, 2021.

‡‡ Links are provided in the original paper.

Study	Research aim	Primary approach(es)	Evaluation methodology
Rios et al. [140]	Multi-label Text Classification	Propose and evaluate a neural architecture suitable for handling few- and zero-shot labels in the multi-label setting where the output label space satisfies two constraints: (1). the labels are connected forming a DAG and (2). each label has a brief ♮*languagedescriptor*.	R@5 and R@10 (Recall), P@10 (Precision), Macro-F_1 scores
Rios et al. [141]	Multi-label Text Classification	Propose a novel semi-parametric neural matching network for diagnosis/procedure code predictionfrom EMR narratives.	Precision, Recall, F_1-scores, AUC (PR), AUC (ROC), P@k, R@k
Hofer et al. [66]	NER	Five improvements on NER tasks when only 10 annotated examples are available: 1.Layer-wise initialization with pre-trained weights (single pre-training); 2.Hyperparameter tuning; 3.Combining pre-training data; 4.Custom word embeddings; 5. Optimizing out-of-vocabulary words.	F_1-score
Pham et al. [132]	Neural Machine Translation (NMT)	Present a generic approach to use phrase-based models to simulate Experts to complement neural machine translation models show that the model can be trained to copy the annotations into the output consistently.	BLEU score, SUGGESTION (SUG) and SUGGESTION ACCURACY (SAC)
Yan et al. [178]	Text Classification	Propose a short text classification framework based on Siamese CNNs and FSL, which will learn the discriminative text encoding so as to help classifiers distinguish those obscure or informal sentence. The different sentence structures and different descriptions of a topic will be learned by FSL strategy to improve the classifier's generalization.	Accuracy
Manousogiannis et al. [118]	Concept Extraction	Propose a simple Few-Shot learning approach, based on pre-trained word embeddings and data from the UMLS, combined with the provided training data.	Relaxed and strict Precision/Recall /F_1-scores
Gao et al. [45]	Relation Classification	Propose FewRel 2.0, a new task containing two real-world issues that FewRel ignores: (1) few-shot domain adaptation, and (2) few-shot none-of-the-above detection.	Accuracy
Lara-Clares et al. [93]	NER	This work is based in the Few-shot Learning Model to learn high level features. propose a hybrid Bi-LSTM CNN model adding a Part-of-Speech (POS) tagging layer, that is, information about multi-word entities. And use wikipedia2vec to automatically extract and classify keywords.	F_1-score

Table A.2 continued from previous page

Study	Research aim	Primary approach(es)	Evaluation methodology
Ferré et al. [40]	Entity Normalization	Propose C-Norm, a new neural approach which synergistically combines standard, weak supervision, ontological knowledge integration and distributional semantics.	The offcial evaluation tool of the BB-norm task: a similarity score and a strict exact match score.
Hou et al. [68]	Slot Tagging (NER)	Proposed Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network 1.A collapsed dependency transfer mechanism into CRF to transfer abstract label dependency patterns as transition scores 2.The emission score of CRF: word's similarity to the representation of each label. 3. A Label-enhanced Task-Adaptive Projection Network (L-TapNet) based on TapNet, by leveraging label name semantics in representing labels.	1.Cross-validate the models on different domains. One target domain for testing, one domain for development, and rest domains as source domains for training. 2. Evaluate F_1-scores within each few-shot episode, and average 100 F_1-scores from all 100 episodes as the final result
Sharaf et al. [150]	Neural Machine Translation (NMT)	Frame the adaptation of NMT systems as a meta-learning problem, where can learn to adapt to new unseen domains based on simulated offline meta-training domain adaptation tasks.	Use BLEU, measure case-sensitive de-tokenized BLEU with SacreBLEU.
Lu et al. [113]	Multi-label Text Classification	Present a simple multi-graph aggregation model that fuses knowledge from multiple label graphs encoding different semantic label relationships in order to study how the aggregated knowledge can benefit multi-label zero/few-shot document classification.	Recall@K and nDCG@K. K was set to 10 for MiMIC-II/III and 5 for URLEX57K.
Jia et al. [74]	NER	The method creates distinct feature distributions for each entity type across domains, which can give better transfer learning power compared to representation networks that do not explicitly differentiate entity types.	F_1-score
Chalkidis et al. [19]	Multi-label Text Classification	1. Hierarchical methods based on Probabilistic Label Trees (PLTs); 2. Combines BERT with LWAN; 3. Investigate the use of structural information from the label hierarchy in LWAN. Leverage the label hierarchy to improve few and zero-shot learning.	R-Precision@K (RP@K), a top-K version of R-Precision of each document, and nDCG@K
Lwowski et al. [116]	Text Classification	Propose a self-supervised learning algorithm to monitor COVID-19 Twitter using an autoencoder to learn the latent representations and then transfer the knowledge to COVID-19 Infection classifier by fine-tuning the Multi-Layer Perceptron (MLP) using few-shot learning.	Accuracy, Precision, Recall and F_1-score

Table A.2 continued from previous page

Study	Research aim	Primary approach(es)	Evaluation methodology
Hou et al. [69]	Dialogue Language Understanding: includes two sub-tasks: Intent Detection (classification) and Slot Tagging (sequence labeling)	Present FewJoint, a novel FSL benchmark for NLP. This benchmark introduces few-shot joint dialogue language understanding, which additionally covers the structure prediction and multi-task reliance problems.	Intent Accuracy, Slot F_1-score, Sentence Accuracy
Chen et al. [23]	Natural Language Generation (NLG)	The design of the model architecture is based on two aspects: content selection from input data and language modeling to compose coherent sentences, which can be acquired from prior knowledge.	BLEU-4, ROUGE-4 (F-measure) follow the same trend with BLEU-4
Vaci et al. [166]	Concept Extraction	Used a combination of methods to extract salient information from electronic health records. First, clinical experts define the information of interest and subsequently build the training and testing corpora for statistical models. Second, built and fine-tuned the statistical models using active learning procedures.	Precision, Recall and F_1-score
Huang et al. [71]	NER	Present the first systematic study for few-shot NER, a problem that is previously little explored in the literature. Three distinctive schemes and their combinations are investigated; perform comprehensive comparisons of these schemes on 10 public NER datasets from different domains; Compared with existing methods on few-shot and training-free NER settings, the proposed schemes achieve SoTA performance despite their simplicity.	F_1-score
Chen et al. [20]	Classificaition	Propose a classification and diagnosis method for Alzheimer's patients based on multi-modal feature fusion and small sample learning. And then the compressed interactive network is used to explicitly fuse the extracted features at the vector level. Finally, the KNN attention pooling layer and the convolutional network are used to constract a small sample learning network to classify the patient diagnosis data.	Accuracy and F_1-score
Yin et al. [182]	Sequence Tagging (NER)	Mainly adopt the prototypical network, and use the relation module as the distance measurement function to model the task of biomedical event trigger identification. In addition, in order to make full use of the external knowledge base to learn the complex biological context, we introduced a self-attention mechanism.	F_1-score

Table A.2 continued from previous page

Study	Research aim	Primary approach(es)	Evaluation methodology
Goodwin et al. [53]	Abstractive Summarization	Compare the summarization quality produced by three SOTA transformer-based models: BART, T5, and PEGASUS.	ROUGE-1, ROUGE-2, and ROUGE-L F_1-scores, BLEU-4 and Repetition Rate
Yang et al.[179]	NER	Propose STRUCTSHOT. 1.Use contextual representation to represent each token, uses a nearest neighbor (NN) classifier and a Viterbi decoder for prediction. 2. Test systems on both identifying new types of entities in the source domain as well as identifying new types of entities in various target domains in one-shot and five-shot settings.	F_1-score
Hartmann et al. [63]	Concept Extraction	Present a universal approach to multi-lingual negation scope resolution, and study an approach for zero-shot cross-lingual transfer for negation scope resolution in clinical text, exploiting data from disparate sources by data concatenation, or in an MTL setup.	Two widely used evaluation metrics for negation scope prediction: Percentage of correct spans (PCS) and F_1-score over scope tokens
Fivez et al. [43]	Name Normalization	Take a next step towards truly robust representations, which capture more domain-specific semantics while remaining universally applicable across different biomedical corpora and domains. Use conceptual grounding constraints which more effectively align encoded names to pretrained embeddings of their concept identifiers.	For synonym retrieval: Mean average precision (mAP) over all synonyms. For concept mapping, Accuracy (Acc) and Mean reciprocal rank (MRR) of the highest ranked correct synonym.
Lu et al. [180]	Rumor Detection	Collect and contribute a publicly available rumor dataset that is suitable for few-shot learning from Sina Weibo. And introduce a FSL-based multi-modality fusion model named COMFUSE for COVID-19 rumor detection, including text embeddings modules with pre-trained BERT model, feature extraction module with multilayer Bi-GRUs, multi-modality feature fusion module with a fusion layer, and meta-learning based few-shot learning paradigm for rumor detection.	Accuracy

Table A.2 continued from previous page

Study	Research aim	Primary approach(es)	Evaluation methodology
Ma et al. [117]	Drug-response Predictions	Applied the few-shot learning paradigm to three context-transfer challenges: (1) transfer of a predictive model learned in one tissue type to the distinct contexts of other tissues; (2) transfer of a predictive model learned in tumor cell lines to patient-derived tumor cell (PDTC) cultures in vitro; and (3) transfer of a predictive model learned in tumor cell lines to the context of patient-derived tumor xenografts (PDXs) in mice in vivo.	Accuracy, Pearson's correlation, AUC
Kormilitzin et al. [86]	NER	First, the underlying deep neural network language model was pre-trained in a self-supervised manner using the cloze-style approach. Second, using the weak-supervision method, developed synthetic training data with noisy labels. Lastly, incorporated all ingredients into an active learning approach.	Accuracy, Precision, Recall and F_1-score
Guo et al. [56]	Extract Entity Relations	Proposes BioGraphSAGE model, a Siamese graph neural network with structured databases as domain knowledge to extract biological entity relations from literatures.	Precision (P-value), Recall (R-value) and F_1-score
Lee et al. [95]	Fact-Checking (close to Text Classification)	Propose a novel way of leveraging the perplexity score from LMs for the few-shot fact-checking task and demonstrate the effectiveness of the perplexity-based approach in the few-shot setting.	Accuracy and the Macro-F_1-score
Fivez et al. [42]	Name Normalization	Explore a scalable few-shot learning approach for robust biomedical name representations which is orthogonal to this paradigm. And use more general higher-level concepts which span a large range of fine-grained concepts.	Spearman's rank correlation coefficient between human judgments and similarity scores of name embeddings, reported on semantic similarity (sim) and relatedness (rel) benchmarks.

Table A.2: A summary table showing primary few-shot approaches and evaluation methodologies

Appendix B

Detailed Task-specific Static Prompts

Prompt Strategies	*Reddit-Impacts*
Basic Prompt	**[Task Description]:** You are a medical AI trained to identify and classify tokens into three categories: Clinical Impacts, Social Impacts, and Outside ('O'). Your task is to extract and classify the clinical and social impacts from this dataset, considering your knowledge of the lifestyle of this population and the potential clinical and social impacts they might experience. **[Entity Types with Definitions]:** 'Clinical Impacts' refer to tokens describing the effects, consequences, or impacts of substance use on individual health or well-being, as defined in UMLS. 'Social Impacts' describe the societal, interpersonal, or community-level effects, also based on UMLS definitions. Any token not falling into these categories should be labeled as 'O'. **[Format Specification]:** For example, the sentence 'I was a codeine addict.' is tokenized and labeled as follows: ['I', 'was', 'a', 'codeine', 'addict', '.'] with labels ['O', 'O', 'O', 'Clinical Impacts', 'Clinical Impacts', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should be tokens with their labels: ['I-O', 'was-O', 'a-O', 'codeine-Clinical Impacts', 'addict-Clinical Impacts', '.-O'].
Description of datasets	The data you are working with has been collected from 14 forums on Reddit (subreddits) that focused on prescription and illicit opioids, and medications for opioid use disorder. This dataset represents a social media context, coming from individuals who may use prescription and illicit opioids and stimulants.
High-frequency instances	In this dataset, high-frequency clinical impacts include 'withdrawal', 'rehab', 'addicted', 'detox', 'overdosed', and 'rehabs'. High-frequency social impacts include 'lost', 'homeless', 'charged', 'streets', 'jail', and 'disorderly'.
UMLS knowledge	The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.
Error analysis	Possible analysis of prediction errors: If a sentence describes the background information of an event, facility, or project, then even if it mentions keywords related to social impact like 'at jail', it still cannot be determined as describing a patient being in jail. It is essential to clearly determine whether the sentence is describing the patient's condition. Second, if the sentence is about the usage, operation, or introduction of a drug or medicine, it does not belong to the patient's clinical impacts, even if it mentions some symptoms. Pay attention to whether the sentence contains words like 'if' that indicate conditions.

Table B.1: Specific static prompts for each component we used for the REDDIT-IMPACTS dataset.

Prompt Strategies	*BC5CDR*
Basic Prompt	**[Task Description]:** You are a medical AI trained to identify and classify tokens into three categories: 'Disease', 'Chemical' and Outside ('O'). Your task is to extract and classify the Disease and Chemical related concepts from this dataset. **[Entity Types with Definitions]:** 'Disease' is a particular abnormal condition that adversely affects the structure or function of all or part of an organism and is not immediately due to any external injury. Diseases are often known to be medical conditions that are associated with specific signs and symptoms. A disease may be caused by external factors such as pathogens or by internal dysfunctions. For example, internal dysfunctions of the immune system can produce a variety of different diseases, including various forms of immunodeficiency, hypersensitivity, allergies, and autoimmune disorders. 'Chemical' in this context refers to substances or compounds with specific chemical properties and structures. These can include drugs, neurotransmitters, elements or ions, vitamins, and other medically relevant chemicals. Any token not falling into Disease categories should be labeled as 'O'. **[Format Specification]:** For example, the sentence 'The hypotensive effect of 100 mg / kg alpha-methyldopa was also partially reversed by naloxone .' is tokenized and labeled as follows: ['The', 'hypotensive', 'effect', 'of', '100', 'mg', '/', 'kg', 'alpha-methyldopa', 'was', 'also', 'partially', 'reversed', 'by', 'naloxone', '.'] with labels ['O', 'Disease', 'O', 'O', 'O', 'O', 'O', 'O', 'Chemical', 'O', 'O', 'O', 'O', 'O', 'Chemical', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ['The-O', 'hypotensive-Disease', 'effect-O', 'of-O', '100-O', 'mg-O', '/-O', 'kg-O', 'alpha-methyldopa-Chemical', 'was-O', 'also-O', 'partially-O', 'reversed-O', 'by-O', 'naloxone-Chemical', '.-O'].
Description of datasets	The data you are working with is BC5CDR dataset, a benchmark dataset for biomedical natural language processing, created from PubMed abstracts. It includes annotations for two entity types—chemicals and diseases—and their relationships, specifically chemical-induced disease interactions. The dataset is widely used for tasks such as named entity recognition and relation extraction, supporting research in biomedical text mining and information extraction.
High-frequency instances	In this dataset, high frequency of 'Disease' include 'pain', 'toxicity', 'renal', 'failure', 'disease', 'hypotension'; high frequency of 'Chemical' include 'cocaine', 'acid', 'dopamine', 'nicotine', 'morphine', 'lithium'.
UMLS knowledge	The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.
Error analysis	Possible analysis of prediction errors: The prediction errors mainly stem from challenges in distinguishing between entity boundaries and contextual usage. For instance, multi-token entities were partially labeled, causing boundary mismatches. Additionally, certain terms such as "receptor" or "antagonist" were incorrectly labeled as 'O,' despite being part of chemical-related entities. Misclassification also occurred in sentences with conditional phrases or background information, where the relation between entities was not accurately captured. Furthermore, entities mentioned in descriptive or abstract contexts, were sometimes overlooked. These errors highlight difficulties in handling complex sentence structures, context-specific classification, and multi-token entity recognition.

Table B.2: Specific static prompts for each component we used for the BC5CDR dataset.

Prompt Strategies	*MIMIC III*
Basic Prompt	**[Task Description]:** You are a medical AI trained to identify and classify tokens into 13 categories: 'CONDITION/SYMPTOM', 'DRUG', 'AMOUNT', 'TIME', 'MEASUREMENT', 'LOCATION', 'EVENT', 'FREQUENCY', 'ORGANIZATION', 'DATE', 'AGE', 'GENDER' and Outside ('O'). Your task is to extract and classify the concepts from this dataset. **[Entity Types with Definitions]:** 'ORGANIZATION' refers to entities or groups associated with healthcare or emergency medical services. These could be specific departments, teams, or services within a medical or emergency response organization. 'DATE' in this context refers to specific calendar dates. These dates are typically used to mark particular events, appointments, or deadlines. 'AGE' in this context refers to the length of time that a person has lived or the number of years since their birth. It can be expressed in various formats, including numerical values, abbreviated forms, or written out in words. 'GENDER' in this context refers to the socially constructed roles, behaviors, activities, and attributes that a given society considers appropriate for men and women. It encompasses the identities of 'male' and 'female,' which are often associated with biological sex but are also shaped by cultural and social factors. 'FREQUENCY' in this context refers to the rate or regularity at which an event or phenomenon occurs. It can describe how often something happens, ranging from sporadic or irregular occurrences to more regular or constant patterns. 'EVENT' in this context refers to specific occurrences or actions that take place, particularly in a medical or clinical setting. These can include procedures, assessments, or other significant incidents. 'LOCATION' in this context refers to specific places or areas, particularly within a healthcare or medical setting. These can include types of facilities, specific locations within a facility, or other relevant places. 'MEASUREMENT' in this context refers to quantitative assessments or values used to evaluate specific physiological or medical parameters. These can include vital signs, laboratory test results, numerical values, or other metrics related to patient health. 'TIME' in this context refers to specific points or periods in the temporal continuum, particularly as they relate to healthcare or medical events. These can include general time references, specific durations, or events tied to time. 'AMOUNT' in this context refers to specific quantities or dosages, particularly in a medical or pharmaceutical setting. These can include measurements of medication, frequency or number of administrations, and methods of delivery. 'DRUG' in this context refers to specific medications or pharmaceutical substances used in the treatment, prevention, or diagnosis of diseases. These can include brand names, generic names, or forms of administration. 'CONDITION/SYMPTOM' in this context refers to physical or subjective signs that indicate a medical condition or disease. These can include sensations of discomfort, specific types of pain or discomfort, respiratory issues, or gastrointestinal symptoms. Any token not falling into categories above should be labeled as 'O'. **[Format Specification]:** For example, the sentence 'The patient was readmitted to the hospital on 2195-6-6 due to fevers to 103 at the rehabilitation facility despite being on intravenous antibiotics HISTORY OF PRESENT ILLNESS 55 year-old female presents with 2/5 week history of non-bloody diarrhea' is tokenized and labeled as follows: ['The', 'patient', 'was', 'readmitted', 'to', 'the', 'hospital', 'on', '2195-6-6', 'due', 'to', 'fevers', 'to', '103', 'at', 'the', 'rehabilitation', 'facility', 'despite', 'being', 'on', 'intravenous', 'antibiotics', 'HISTORY', 'OF', 'PRESENT', 'ILLNESS', '55', 'year-old', 'female', 'presents', 'with', '2/5', 'week', 'history', 'of', 'non-bloody', 'diarrhea'] with labels ['O', 'O', 'O', 'EVENT', 'O', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'O', 'MEASUREMENT', 'MEASUREMENT', 'MEASUREMENT', 'O', 'LOCATION', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'DRUG', 'DRUG', 'O', 'O', 'O', 'O', 'AGE', 'O', 'GENDER', 'O', 'O', 'AMOUNT', 'AMOUNT', 'O', 'O', 'CONDITION/SYMPTOM', 'CONDITION/SYMPTOM']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ['The-O', 'patient-O', 'was-O', 'readmitted-EVENT', 'to-O', 'the-LOCATION', 'hospital-LOCATION', 'on-O', '2195-6-6-O', 'due-O', 'to-O', 'fevers-MEASUREMENT', 'to-MEASUREMENT', '103-MEASUREMENT', 'at-O', 'the-LOCATION', 'rehabilitation-LOCATION', 'facility-LOCATION', 'despite-O', 'being-O', 'on-O', 'intravenous-DRUG', 'antibiotics-DRUG', 'HISTORY-O', 'OF-O', 'PRESENT-O', 'ILLNESS-O', '55-AGE', 'year-old-O', 'female-GENDER', 'presents-O', 'with-O', '2/5-AMOUNT', 'week-AMOUNT', 'history-O', 'of-O', 'non-bloody-CONDITION/SYMPTOM', 'diarrhea-CONDITION/SYMPTOM'].
Description of datasets	The data you are working with is MIMIC-III (Medical Information Mart for Intensive Care) dataset, a large, publicly available database containing de-identified health data from critical care patients at the Beth Israel Deaconess Medical Center. It includes structured data, such as demographics, lab results, and vital signs, as well as unstructured data, such as clinical notes and discharge summaries. The dataset is widely used for research in machine learning, natural language processing, and clinical decision support to improve healthcare outcomes.
High-frequency instances	In this dataset, high frequency of 'CONDITION/SYMPTOM' include 'pain', 'chest', 'cough', 'breath', 'nausea', 'abdominal'; high frequency of 'DRUG' include 'iv', 'lasix', 'ceftriaxone', 'oxygen', 'ns', 'coumadin'; high frequency of 'AMOUNT' include 'iv', '2', '1', 'mg', 'days', 'one'; high frequency of 'TIME' include 'day', 'admission', 'prior', 'last', 'ago', 'morning'; high frequency of 'MEASUREMENT' include 'bp', 'hr', 'pressure', 'blood', 'rr', 'rate', 'heart'; high frequency of 'LOCATION' include 'hospital', 'right', 'home', 'floor', 'emergency', 'micu'; high frequency of 'EVENT' include 'ct', 'placed', 'cxr', 'intubated', 'exam', 'review'; high frequency of 'FREQUENCY' include 'chronic', 'intermittent', 'daily', 'occasionally', 'frequent', 'intermittently'; high frequency of 'ORGANIZATION' include 'ems', 'service', 'surgery', 'pcp', 'emergency', 'neuro', 'medicine'; high frequency of 'DATE' include '2171114', '21491117'; high frequency of 'AGE' include '60', '80yo', '78', '61', 'seventyeightyearold', '69'; high frequency of 'GENDER' include 'man', 'woman', 'f', 'male', 'female', 'm'.
UMLS knowledge	The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.
Error analysis	The prediction errors stem from several factors. Entity boundary recognition issues were common, particularly with multi-token entities like "shortness of breath" or "paroxysmal nocturnal dyspnea," where some tokens were missed or incorrectly labeled as 'O.' Additionally, the model struggled with entity type confusion, such as distinguishing "pain" as a symptom versus its contextual use related to location. Context-dependent misinterpretations also contributed to errors, especially in handling negations like "denies chest pain" or temporal references such as "last few months." Overlapping entities posed further challenges, where closely related terms (e.g., "MI" and "CABG") interfered with accurate classification. Finally, rare or unseen entities in the training data led to occasional misclassifications, highlighting gaps in the model's ability to generalize.

Table B.3: Specific static prompts for each component we used for the MIMIC III dataset.

Prompt Strategies	*NCBI*
Basic Prompt	**[Task Description]:** You are a medical AI trained to identify and classify tokens into five categories: DiseaseClass, SpecificDisease, Modifier, CompositeMention and Outside ('O'). Your task is to extract and classify the DiseaseClass, SpecificDisease, Modifier and CompositeMention from this dataset.
	[Entity Types with Definitions]: 'DiseaseClass' refers to a classification system or category used to group various medical conditions or diseases based on certain characteristics, such as their nature, affected biological systems, or underlying causes. 'SpecificDisease' appears to describe particular diseases that are identified and classified based on their specific clinical features, genetic origins, or biochemical abnormalities. 'Modifier' refers to specific attributes or variations or conditions that can modify or influence the presentation, progression, or characteristics of a disease, alter the manifestation or course of a disease, potentially affecting its diagnosis, treatment, and prognosis. 'CompositeMention' describes medical conditions or characteristics that are composed of several elements or features, often involving multiple tissues, organs, or systems. Any token not falling into these categories should be labeled as 'O'.
	[Format Specification]: For example, the sentence 'Histidinemia. Classical and atypical form in siblings.' is tokenized and labeled as follows: ['Histidinemia.', 'Classical', 'and', 'atypical', 'form', 'in', 'siblings.'] with labels ['SpecificDisease', 'O', 'O', 'O', 'O', 'O', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ['Histidinemia.-SpecificDisease', 'Classical-O', 'and-O', 'atypical-O', 'form-O', 'in-O', 'siblings.-O'].
Description of datasets	The data you are working with is NCBI disease corpus, a collection of 793 PubMed abstracts fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community. Each PubMed abstract was manually annotated by two annotators with disease mentions and their corresponding concepts in Medical Subject Headings (MeSH) or Online Mendelian Inheritance in Man (OMIM). The public release of the NCBI disease corpus contains 6892 disease mentions, which are mapped to 790 unique disease concepts. Of these, 88 percent link to a MeSH identifier, while the rest contain an OMIM identifier. We were able to link 91 percent of the mentions to a single disease concept, while the rest are described as a combination of concepts.
High-frequency instances	In this dataset, high-frequency 'DiseaseClass' include 'disorder', 'abnormalities', 'tumors', 'mental', 'disorders', 'retardation'. High-frequency 'SpecificDisease' include 'deficiency', 'syndrome', 'dystrophy', 'familial', 'myotonic', 'colorectal'. High-frequency 'Modifier' include 'tumor', 'tumour', 'APC', 'choroideremia', 'DM', 'DMD'. High-frequency 'CompositeMention' include 'breast', 'ovarian', 'cancer', 'muscular', 'andor', 'becker'.
UMLS knowledge	The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.
Error analysis	The prediction errors in the NCBI dataset primarily stem from challenges in distinguishing composite mentions and modifiers within complex biomedical contexts. For instance, entities like "BRCA1 gene" were incorrectly segmented, with "BRCA1" labeled as a modifier instead of being part of the composite mention. Additionally, multi-token composite mentions such as "breast and ovarian cancer" were not consistently labeled, with individual tokens occasionally missed or misclassified. Contextual ambiguity, such as distinguishing between mentions of general biological terms (e.g., "tumor") and their specific functional roles (e.g., "tumor suppressor"), also contributed to errors.

Table B.4: Specific static prompts for each component we used for the NCBI dataset.

Prompt Strategies	*Med-Mentions*
Basic Prompt	**[Task Description]:** You are a medical AI trained to identify and classify tokens into two categories: Disease and Outside ('O'). Your task is to extract and classify the Disease related concepts from this dataset.
	[Entity Types with Definitions]: 'Disease' is a particular abnormal condition that adversely affects the structure or function of all or part of an organism and is not immediately due to any external injury. Diseases are often known to be medical conditions that are associated with specific signs and symptoms. A disease may be caused by external factors such as pathogens or by internal dysfunctions. For example, internal dysfunctions of the immune system can produce a variety of different diseases, including various forms of immunodeficiency, hypersensitivity, allergies, and autoimmune disorders. Any token not falling into Disease categories should be labeled as 'O'.
	[Format Specification]: For example, the sentence 'A total of 200 children and adolescents with type 1 diabetes, ages 9-18 years, completed the DEPS-R Turkish version.' is tokenized and labeled as follows: ['A', 'total', 'of', '200', 'children', 'and', 'adolescents', 'with', 'type', '1', 'diabetes,', 'ages', '9-18', 'years,', 'completed', 'the', 'DEPS-R', 'Turkish', 'version.'] with labels ['O', 'O', 'O', 'O', 'Disease', 'O', 'Disease', 'O', 'Disease', 'Disease', 'Disease', 'Disease', 'O', 'Disease', 'O', 'O', 'Disease', 'Disease', 'Disease']. The output format should include tokens with their labels: ['A-O', 'total-O', 'of-O', '200-O', 'children-Disease', 'and-O', 'adolescents-Disease', 'with-O', 'type-Disease', '1-Disease', 'diabetes,-Disease', 'ages-Disease', '9-18-O', 'years,-Disease', 'completed-O', 'the-O', 'DEPS-R-Disease', 'Turkish-Disease', 'version.-Disease'].
Description of datasets	The data you are working with is Med-Mentions, a new manually annotated resource for the recognition of biomedical concepts. What distinguishes Med-Mentions from other annotated biomedical corpora is its size (over 4,000 abstracts and over 350,000 linked mentions), as well as the size of the concept ontology (over 3 million concepts from UMLS 2017) and its broad coverage of biomedical disciplines.
High-frequency instances	In this dataset, high frequency 'Disease' related entities include 'patients', 'cells', 'treatment', 'cancer', 'analysis', 'disease', 'clinical'.
UMLS knowledge	The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.
Error analysis	The prediction errors in the Med-Mentions dataset are primarily due to challenges in identifying complex and overlapping disease mentions, as well as distinguishing between general biomedical terms and specific disease entities. Multi-token entities such as "renal pedicle occlusion" or "intention-to-treat analyses" were often partially labeled, with some tokens being misclassified or excluded. Additionally, the presence of nested or overlapping mentions, such as "prostate cancer" and its relationship to broader contexts like "treatment disparities," led to inconsistent labeling. The model also struggled with domain-specific terminology, misclassifying general terms like "maternal genotype" or "outcome" as disease mentions. These errors highlight limitations in handling nuanced biomedical language, especially when entities span multiple tokens or overlap with related terms.

Table B.5: Specific static prompts for each component we used for the Med-Mentions dataset.

Appendix C

Averaged Performance of the Baseline Dynamic Prompt Model

		Precision	Recall	F_1-score			Precision	Recall	F_1-score
GPT 4	**5-shot**	16.40	50.00	24.70	**Llama3**	**5-shot**	12.56	56.32	20.55
		17.67	50.62	26.20			13.68	55.81	21.97
		17.48	53.09	26.30			12.14	59.52	20.16
		23.91	54.32	33.20			14.25	59.79	23.02
	AVG	18.865	52.0075	27.60		**AVG**	13.1575	57.86	21.425
	10-shot	25.41	58.75	35.47		**10-shot**	24.51	59.52	34.72
		21.67	54.32	30.99			19.63	62.35	29.86
		21.43	59.26	31.48			23.96	57.44	33.81
		20.47	54.32	29.73			21.38	60.43	31.59
	AVG	22.245	56.6625	31.9175		**AVG**	22.37	59.935	32.495
	20-shot	28.74	58.63	38.57		**20-shot**	27.22	57.65	36.98
		27.03	57.54	36.78			23.24	52.44	32.21
		26.52	59.79	36.74			22.12	52.08	31.06
		28.65	59.02	38.57			25.49	53.06	34.44
	AVG	27.735	58.745	37.665		**AVG**	24.5175	53.8075	33.6725

Table C.1: Averaged performance of the baseline dynamic prompt model on the REDDIT-IMPACTS dataset across different shot settings.

		Precision	Recall	F_1-score			Precision	Recall	F_1-score
GPT 4	**5-shot**	69.97	91.43	79.27	**Llama3**	**5-shot**	68.09	84.35	75.35
		68.32	88.3	77.29			69.70	76.67	73.02
		64.58	90.15	75.25			71.00	79.15	74.86
		71.59	91.39	80.29			67.10	73.25	70.04
	AVG	68.615	90.3175	78.025		**AVG**	68.9725	78.355	73.3175
	10-shot	75.11	90.84	82.23		**10-shot**	72.38	79.18	75.63
		74.41	90.32	81.60			73.31	74.24	73.77
		74.13	85.71	79.50			71.15	77.05	73.98
		77.65	86.35	81.73			73.40	81.17	77.2
	AVG	75.325	88.305	81.265		**AVG**	72.56	77.91	75.145
	20-shot	75.36	88.33	81.33		**20-shot**	75.12	74.84	74.98
		73.13	91.74	81.38			72.03	71.88	71.96
		71.84	91.1	80.33			76.88	77.20	77.04
		77.94	85.54	81.57			77.64	78.40	78.02
	AVG	74.5675	89.1775	81.1525		**AVG**	75.4175	75.58	75.5

Table C.2: Averaged performance of the baseline dynamic prompt model on the BC5CDR dataset across different shot settings.

		Precision	Recall	F_1-score			Precision	Recall	F_1-score
GPT 4	**5-shot**	62.38	63.78	63.07	**Llama3**	**5-shot**	57.69	68.39	62.58
		61.88	62.19	62.03			61.83	61.33	61.58
		65.26	68.14	66.67			58.24	68.06	62.77
		62.71	62.36	62.54			59.44	71.29	64.82
	AVG	63.0575	64.1175	63.5775		**AVG**	59.3	67.2675	62.9375
	10-shot	66.38	74.24	70.09		**10-shot**	59.13	71.63	64.78
		68.37	74.86	71.47			62.68	63.8	63.23
		66.72	75.33	70.77			61.71	67.58	64.51
		66.46	73.34	69.73			62.24	62.84	62.54
	AVG	66.9825	74.4425	70.515		**AVG**	61.44	66.4625	63.765
	20-shot	70.41	69.84	70.12		**20-shot**	63.89	62.04	62.95
		70.11	71.24	70.67			63.96	62.83	63.39
		71.45	73.39	72.41			59.22	61.71	60.44
		70.64	70.80	70.72			60.96	61.88	61.42
	AVG	70.6525	71.3175	70.98		**AVG**	62.0075	62.115	62.05

Table C.3: Averaged performance of the baseline dynamic prompt model on the MIMIC III dataset across different shot settings.

		Precision	Recall	F_1-score			Precision	Recall	F_1-score
GPT 4	**5-shot**	46.17	50.52	48.25	**Llama3**	**5-shot**	33.88	40.67	36.97
		45.69	49.07	47.32			32.26	39.14	35.37
		40.24	43.65	41.88			36.38	28.49	31.95
		47.96	52.83	50.28			40.73	30.53	34.9
	AVG	45.015	49.0175	46.9325		**AVG**	35.8125	34.7075	34.7975
	10-shot	52.98	52.67	52.82		**10-shot**	42.79	31.08	35.66
		53.71	53.36	53.54			40.18	28.73	36.27
		52.99	50.35	51.64			35.75	32.05	33.76
		53.85	52.3	53.06			39.95	34.11	36.71
	AVG	53.3825	52.17	52.765		**AVG**	39.6675	31.4925	35.6
	20-shot	54.57	54.73	54.65		**20-shot**	40.59	42.07	41.32
		44.86	45.43	45.14			40.87	42.8	41.81
		51.6	51.75	51.68			41.48	43.02	42.24
		55.7	57.24	56.46			39.88	42.41	41.1
	AVG	51.6825	52.2875	51.9825		**AVG**	40.705	42.575	41.6175

Table C.4: Averaged performance of the baseline dynamic prompt model on the NCBI dataset across different shot settings.

		Precision	Recall	F_1-score			Precision	Recall	F_1-score
GPT 4	**5-shot**	27.24	62.73	37.99	**Llama3**	**5-shot**	24.94	73.03	37.18
		26.71	58.58	36.69			25.95	61.96	36.58
		29.23	60.47	39.41			26.11	73.57	38.54
		25.86	58.41	35.85			26.55	59.62	36.74
	AVG	27.26	60.0475	37.485		**AVG**	25.8875	67.045	37.26
	10-shot	26.92	58.26	36.82		**10-shot**	25.67	70.61	37.65
		26.1	59.15	36.22			25.1	59.19	35.15
		24.41	58.56	34.46			25.76	70.06	37.67
		29.19	60.82	39.45			25.73	57.46	35.54
	AVG	26.655	59.1975	36.7375		**AVG**	25.565	64.33	36.5025
	20-shot	27.86	59.12	37.87		**20-shot**	26.18	63.59	37.09
		25.02	60.64	35.42			26.63	61.14	37.1
		29.33	61.71	39.76			25.93	63.57	36.84
		30.18	61.66	40.52			27.54	70.85	39.66
	AVG	28.0975	60.7825	38.3925		**AVG**	26.57	64.7875	37.6725

Table C.5: Averaged performance of the baseline dynamic prompt model on the MedMentions dataset across different shot settings.

Appendix D

Results of 95% CIs for Each Metric

	Reddit_Impacts	BC5CDR	MIMIC III	NCBI	Med-Mentions
GPT-3.5					
Basic Prompt (BP)	16.73 [11.53, 22.83]	64.56 [61.55, 67.73]	54.70 [49.60, 58.73]	26.96 [24.43, 30.98]	9.27 [7.81, 12.22]
BP + Description of datasets	21.15 [14.88, 26.64]	68.61 [66.74, 70.72]	56.73 [52.58, 61.22]	34.48 [31.08, 39.25]	12.71 [10.93, 15.65]
BP + High-frequency instances	21.15 [15.75, 27.40]	69.01 [66.24, 70.98]	57.72 [52.75, 62.26]	35.95 [33.36, 38.44]	17.22 [14.47, 19.80]
BP + UMLS knowledge	16.44 [8.43, 23.07]	64.83 [61.83, 66.41]	50.57 [46.17, 55.04]	30.75 [27.73, 33.26]	10.88 [8.81, 12.29]
BP + Error analysis	19.24 [12.91, 26.17]	67.67 [65.53, 70.32]	59.52 [54.96, 64.47]	33.15 [31.24, 38.87]	15.52 [13.14, 17.20]
BP + 5-shot learning with sentences	19.30 [12.26, 25.78]	68.84 [67.25, 70.49]	57.03 [53.06, 62.85]	40.16 [38.78, 46.45]	20.61 [17.58, 22.29]
BP + 5-shot learning with tokens	**21.69 [15.92, 28.89]**	**70.79 [68.87, 73.15]**	**61.21 [56.81, 66.05]**	**43.01 [41.43, 48.21]**	**24.57 [22.88, 26.64]**
BP + All above	**23.91 [15.87, 30.97]**	**72.73 [70.32, 74.86]**	**61.99 [57.24, 66.38]**	**45.24 [42.64, 50.58]**	**31.63 [29.36, 34.74]**
GPT-4					
Basic Prompt (BP)	20.16 [13.29, 26.54]	69.43 [66.28, 72.44]	56.63 [51.27, 60.83]	33.56 [31.59, 37.25]	13.83 [11.85, 15.09]
BP + Description of datasets	23.52 [16.46, 30.84]	70.65 [67.47, 72.72]	59.68 [55.18, 64.09]	35.75 [33.54, 40.58]	15.30 [13.61, 17.15]
BP + High-frequency instances	24.64 [17.72, 31.11]	72.60 [71.17, 74.28]	60.08 [56.33, 65.37]	37.96 [36.95, 41.73]	19.50 [17.11, 22.97]
BP + UMLS knowledge	20.46 [13.84, 27.07]	69.86 [66.05, 72.62]	55.13 [50.20, 60.29]	30.90 [28.68, 34.30]	14.50 [12.57, 16.46]
BP + Error analysis	23.13 [16.65, 30.69]	74.61 [71.44, 77.29]	60.11 [55.44, 64.72]	37.84 [34.13, 42.71]	18.25 [15.06, 20.43]
BP + 5-shot learning with sentences	22.88 [16.23, 30.59]	73.00 [71.26, 76.22]	58.25 [53.28, 63.95]	40.86 [39.37, 45.36]	28.80 [26.71, 30.20]
BP + 5-shot learning with tokens	**25.95 [18.50, 32.07]**	**76.65 [74.15, 77.92]**	**62.94 [57.56, 66.87]**	**44.24 [42.93, 48.28]**	**33.20 [31.64, 35.70]**
BP + All above	**27.60 [19.43, 33.80]**	**78.03 [75.51, 80.02]**	**63.58 [58.73, 67.18]**	**46.93 [44.85, 51.58]**	**37.95 [35.88, 39.90]**
Llama3					
Basic Prompt (BP)	15.61 [8.20, 22.12]	62.13 [59.24, 63.58]	50.70 [45.93, 54.19]	19.15 [15.21, 21.38]	21.23 [19.24, 23.42]
BP + Description of datasets	19.28 [11.71, 25.96]	67.68 [64.86, 69.10]	56.22 [52.77, 60.25]	21.44 [20.80, 24.65]	21.57 [19.30, 24.76]
BP + High-frequency instances	**20.44 [13.79, 27.51]**	68.39 [66.48, 70.35]	56.06 [52.62, 61.42]	26.62 [22.16, 28.31]	27.12 [26.37, 29.35]
BP + UMLS knowledge	12.91 [7.40, 18.71]	64.71 [61.44, 67.01]	48.92 [44.75, 53.37]	20.91 [17.07, 22.61]	23.68 [20.59, 25.17]
BP + Error analysis	18.87 [13.34, 25.13]	68.07 [65.41, 70.58]	58.92 [53.90, 63.84]	24.46 [20.97, 25.20]	25.78 [23.48, 27.56]
BP + 5-shot learning with sentences	17.65 [13.62, 24.69]	70.70 [69.36, 72.83]	56.85 [52.32, 61.33]	30.52 [26.50, 33.96]	34.87 [32.18, 37.25]
BP + 5-shot learning with tokens	20.04 [14.81, 27.29]	**71.76 [69.58, 73.51]**	**61.98 [56.59, 65.18]**	**33.42 [28.72, 35.12]**	**35.23 [33.17, 37.08]**
BP + All above	**21.43 [14.24, 28.80]**	**73.32 [72.27, 74.26]**	**62.94 [57.07, 65.79]**	**34.80 [28.57, 35.44]**	**37.26 [35.45, 39.08]**

Table D.1: Evaluation of static prompting strategies using GPT-3.5, GPT-4 and Llama 3 across five biomedical datasets. The table presents F_1-score with 95% confidence intervals reported for each metric to indicate the statistical reliability of the results.

		Reddit_Impacts	*BC5CDR*	*MIMIC III*	*NCBI*	*Med-Mentions*
GPT-4						
5-shot	**Base**	27.60 [19.43, 33.80]	78.03 [75.51, 80.02]	63.58 [58.73, 67.18]	46.93 [44.85, 51.58]	37.95 [35.88, 39.90]
	TF-IDF	28.47 [21.78, 35.47]	**85.88 [84.53, 86.42]**	**76.24 [72.98, 79.63]**	**60.08 [56.70, 63.32]**	37.96 [35.90, 39.84]
	SBERT	**33.72 [26.28, 42.20]**	83.37 [82.51, 84.22]	73.44 [69.91, 76.81]	57.56 [54.05, 60.73]	39.12 [36.84, 41.34]
	ColBERT	32.39 [25.10, 39.85]	79.82 [78.24, 80.98]	75.56 [72.06, 78.94]	52.38 [49.06, 55.55]	**39.93 [37.93, 41.73]**
	DPR	32.64 [25.42, 40.17]	83.58 [82.30, 84.88]	69.89 [65.75, 73.63]	49.37 [45.37, 52.94]	39.13 [34.44, 41.35]
10-shot	**Base**	31.92 [23.77, 38.44]	81.27 [80.81, 82.37]	70.52 [66.10, 73.81]	52.67 [49.36, 56.76]	36.74 [32.29, 38.83]
	TF-IDF	31.14 [24.33, 38.13]	**86.64 [85.15, 88.09]**	75.53 [72.18, 79.10]	**62.05 [58.79, 65.11]**	40.37 [38.23, 42.43]
	SBERT	**35.47 [27.17, 43.21]**	85.92 [83.09, 87.27]	73.89 [70.22, 77.80]	60.83 [57.47, 64.03]	40.37 [38.23, 42.39]
	ColBERT	33.81 [26.24, 41.55]	85.71 [84.42, 86.07]	**76.34 [73.01, 79.68]**	57.25 [53.75, 60.72]	**40.48 [38.13, 42.54]**
	DPR	32.61 [24.50, 40.33]	84.79 [82.96, 86.78]	72.13 [68.06, 75.85]	58.70 [54.99, 61.91]	40.25 [30.83, 50.75]
20-shot	**Base**	37.67 [30.04, 43.44]	81.15 [80.40, 82.24]	70.98 [65.77, 73.82]	51.98 [50.33, 58.84]	38.39 [35.26, 40.29]
	TF-IDF	38.35 [30.77, 46.28]	87.16 [85.77, 88.62]	**77.66 [71.91, 78.88]**	**64.36 [61.18, 67.87]**	**41.32 [39.21, 43.26]**
	SBERT	38.22 [28.57, 44.90]	**87.42 [85.26, 89.12]**	75.14 [71.77, 78.75]	62.21 [59.01, 65.18]	39.37 [35.05, 40.39]
	ColBERT	**42.49 [32.52, 48.33]**	83.00 [81.39, 84.40]	76.70 [73.11, 79.89]	57.69 [54.22, 61.18]	40.53 [37.61, 43.26]
	DPR	38.84 [29.01, 44.44]	85.60 [84.28, 86.93]	72.28 [68.56, 75.95]	60.34 [56.54, 63.72]	39.23 [34.22, 41.56]
Llama3						
5-shot	**Base**	21.43 [14.24, 28.80]	73.32 [72.27, 74.26]	62.94 [57.07, 65.79]	34.80 [28.57, 35.44]	37.26 [35.45, 39.08]
	TF-IDF	28.57 [21.74, 36.06]	80.11 [79.25, 81.00]	70.41 [66.87, 73.76]	49.80 [46.38, 53.03]	38.68 [35.67, 40.81]
	SBERT	**34.42 [26.28, 41.52]**	**80.39 [79.50, 81.33]**	67.88 [64.09, 71.69]	**50.12 [46.89, 53.66]**	37.91 [36.02, 39.81]
	ColBERT	32.94 [25.00, 39.84]	71.76 [70.75, 72.69]	**71.68 [68.08, 75.21]**	45.50 [41.95, 49.49]	**38.99 [36.15, 41.34]**
	DPR	29.00 [22.86, 36.36]	75.67 [74.67, 76.70]	68.97 [65.05, 72.70]	44.54 [41.24, 48.25]	38.66 [36.78, 40.50]
10-shot	**Base**	32.50 [26.94, 42.26]	75.15 [74.65, 76.67]	63.77 [58.59, 67.75]	35.60 [32.17, 39.12]	36.50 [35.73, 39.57]
	TF-IDF	34.21 [27.24, 42.03]	**80.57 [79.65, 81.47]**	55.56 [53.11, 60.44]	49.50 [46.05, 52.92]	35.51 [34.75, 37.45]
	SBERT	32.45 [25.33, 39.63]	81.17 [80.26, 82.03]	71.63 [67.75, 75.15]	**51.35 [47.49, 55.16]**	**39.08 [36.39, 41.38]**
	ColBERT	32.89 [20.35, 35.05]	80.34 [79.53, 81.24]	**72.85 [69.46, 76.49]**	38.77 [34.91, 42.29]	38.06 [35.52, 40.71]
	DPR	**34.29 [26.11, 41.98]**	74.72 [73.77, 75.73]	69.54 [65.61, 73.17]	46.28 [42.77, 49.65]	37.85 [36.06, 39.69]
20-shot	**Base**	33.67 [24.09, 40.88]	75.50 [73.57, 76.36]	62.05 [58.23, 67.15]	41.62 [38.83, 45.71]	37.67 [35.22, 40.57]
	TF-IDF	39.11 [31.34, 47.70]	**78.36 [77.42, 79.30]**	57.66 [51.19, 59.80]	47.50 [43.87, 50.84]	**38.83 [37.54, 39.11]**
	SBERT	**41.43 [31.58, 48.98]**	76.85 [74.86, 78.96]	65.35 [60.44, 70.40]	44.14 [40.57, 47.86]	36.01 [34.16, 37.75]
	ColBERT	34.66 [24.07, 36.31]	72.19 [71.17, 73.20]	57.63 [53.19, 61.93]	**48.44 [45.07, 51.78]**	36.85 [34.10, 39.29]
	DPR	37.30 [27.13, 44.76]	74.80 [72.49, 76.36]	**65.80 [61.82, 69.69]**	40.36 [36.96, 43.96]	36.89 [34.46, 38.84]

Table D.2: Evaluation of dynamic prompting strategies (5-shot, 10-shot, and 20-shot) using GPT-4 and Llama 3 across five biomedical datasets. The table presents F_1-score for each retrieval method: Base Prompt, TF-IDF, SBERT, ColBERT, and DPR, with 95% confidence intervals reported for each metric to indicate the statistical reliability of the results.

Bibliography

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649, 2018.

[2] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930, 2013.

[3] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90, 2018.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL `https://aclanthology.org/W19-1909`.

[5] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogeniza-

tion effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, page 413–425, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704857. doi: 10.1145/3635636.3656204. URL https://doi.org/10.1145/3635636.3656204.

[6] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120, 2023.

[7] Philip Bachman, Alessandro Sordoni, and Adam Trischler. Learning algorithms for active learning. In *international conference on machine learning*, pages 301–310. PMLR, 2017.

[8] Tim Benson. *Principles of health interoperability HL7 and SNOMED*. Springer Science & Business Media, Berkshire, UK, 2012. doi: 10.1007/978-1-4471-2801-4.

[9] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[10] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

[11] Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. Bacteria biotope at bionlp open shared tasks 2019. *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 121–131, 2019. doi: 10.18653/v1/D19-5719.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] Felicity Callard, Matthew Broadbent, Mike Denis, Matthew Hotopf, Murat Soncul, Til Wykes, Simon Lovestone, and Robert Stewart. Developing a new model for patient recruitment in mental health services: a cohort study using electronic health records. *BMJ open*, 4(12):e005654, 2014. doi: 10.1136/bmjopen-2014-005654.

[15] Alberto Cetoli, Stefano Bragaglia, Andrew D O'Harney, and Marc Sloan. Graph convolutional networks for named entity recognition. *arXiv preprint arXiv:1709.10053*, 2017.

[16] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation*, pages 261–268, 2012.

[17] Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessieres, Loic Lepiniec, et al. Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In *BioNLP Shared Task*, page 113. The Association for Computational Linguistics, 2016.

[18] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019.

[19] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. pages 7503–7515. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[20] DeHua Chen, Li Zhang, and Chuang Ma. A multimodal diagnosis predictive model of alzheimer's disease with few-shot learning. In *2020 International Conference on Public Health and Data Science (ICPHDS)*, pages 273–277. IEEE, 2020.

[21] Jiaoyan Chen, Yuxia Geng, Zhuo Chen, Jeff Z Pan, Yuan He, Wen Zhang, Ian Horrocks, and Huajun Chen. Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *Proceedings of the IEEE*, 111(6):653–685, 2023.

[22] Yao Chen, Changjiang Zhou, Tianxin Li, Hong Wu, Xia Zhao, Kai Ye, and Jun Liao. Named entity recognition from chinese adverse drug event reports with lexical feature based bilstm-crf and tri-training. *Journal of biomedical informatics*, 96:103252, 2019.

[23] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. Few-shot nlg with pre-trained language model. *arXiv preprint arXiv:1904.09521*, 2019.

[24] Zitian Chen, Yanwei Fu, Yinda Zhang, Yu-Gang Jiang, Xiangyang Xue, and Leonid Sigal. Multi-level semantic feature augmentation for one-shot learning. *IEEE Transactions on Image Processing*, 28(9):4594–4605, 2019.

[25] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train

good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174, 2016.

[26] C Brendan Clark, Cosmas M Zyambo, Ye Li, and Karen L Cropsey. The impact of non-concordant self-report of substance use in clinical trials research. *Addictive behaviors*, 58:74–79, 2016.

[27] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

[28] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers: (with python examples). *arXiv preprint arXiv:2004.04523*, 2020.

[29] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. *arXiv preprint arXiv:2010.11683*, 2020.

[30] Clément Dalloux, Natalia Grabar, and Vincent Claveau. Détection de la négation: corpus français et apprentissage supervisé. *Revue des Sciences et Technologies de l'Information-Série TSI: Technique et Science Informatiques*, pages 1–21, 2019.

[31] Louisa Degenhardt and Wayne Hall. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *The Lancet*, 379(9810): 55–70, 2012.

[32] Leon Derczynski, Kalina Bontcheva, and Ian Roberts. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, 2016.

[33] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the wnut2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, 2017.

[34] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

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

[37] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2013.12.006. URL `https://www.sciencedirect.com/science/article/pii/S1532046413001974`.

[38] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

[39] Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in bioinformatics*, 22(3):bbaa110, 2021.

[40] Arnaud Ferré, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, and Claire Nédellec. C-norm: a neural approach to few-shot entity normalization. *BMC Bioinformatics*, 21:579, 2020. URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03886-8`.

[41] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[42] Pieter Fivez, Simon Suster, and Walter Daelemans. Scalable few-shot learning of robust biomedical name representations. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 23–29, 2021.

[43] Pieter Fivez, Simon Suster, and Walter Daelemans. Conceptual grounding constraints for truly robust biomedical name representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2440–2450, 2021.

[44] Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, page 993–1000, 2019. doi: https://doi.org/10.1145/3297280.3297378.

[45] Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fewrel 2.0: Towards more challenging few-shot relation classification. *arXiv preprint arXiv:1910.07124*, 2019.

[46] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[47] Yao Ge, Yuting Guo, Yuan-Chi Yang, Mohammed Ali Al-Garadi, and Abeed Sarker. A comparison of few-shot and traditional named entity recognition models for medical text. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 84–89. IEEE, 2022.

[48] Yao Ge, Yuting Guo, Sudeshna Das, Mohammed Ali Al-Garadi, and Abeed Sarker. Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, page 104458, 2023.

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

[50] Yao Ge, Sudeshna Das, Karen O'Connor, Mohammed Ali Al-Garadi, Graciela Gonzalez-Hernandez, and Abeed Sarker. Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media. *arXiv preprint arXiv:2405.06145*, 2024.

[51] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. Induction networks for few-shot text classification. *arXiv preprint arXiv:1902.10482*, 2019. URL `https://arxiv.org/abs/1902.10482`.

[52] Shalmoli Ghosh, Janardan Misra, Saptarshi Ghosh, and Sanjay Podder. Utilizing social media for identifying drug addiction and recovery intervention. In *Proceedings of Workshop on Data Analytics for Smart Health (DASH 2020), co-located with IEEE BigData*, 2020. URL `https://ieeexplore.ieee.org/abstract/document/9378092`.

[53] Travis R Goodwin, Max E Savery, and Dina Demner-Fushman. Flight of the pegasus? comparing transformers on few-shot and zero-shot multi-document

abstractive summarization. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2020, page 5640. NIH Public Access, 2020.

[54] Mourad Gridach. Character-level neural network for biomedical named entity recognition. *Journal of biomedical informatics*, 70:85–91, 2017.

[55] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. PPT: Pre-trained prompt tuning for few-shot learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.576. URL https://aclanthology.org/2022.acl-long.576.

[56] Shuyu Guo, Lan Huang, Gang Yao, Ye Wang, Haotian Guan, and Tian Bai. Extracting biomedical entity relations using biological interaction knowledge. *Interdisciplinary Sciences: Computational Life Sciences*, 13:312–320, 2021. URL https://link.springer.com/article/10.1007/s12539-021-00425-8.

[57] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 124–141. Springer, 2020.

[58] Yuting Guo, Swati Rajwal, Sahithi Lakamana, Chia-Chun Chiang, Paul C Menell, Adnan H Shahid, Yi-Chieh Chen, D Pharm, Nikita Chhabra, Wan-Ju Chao, et al. Generalizable natural language processing framework for mi-

graine reporting from social media. *AMIA Summits on Translational Science Proceedings*, 2023:261, 2023.

[59] Iryna Gurevych and Torsten Zesch. Proceedings of the 2009 workshop on the people's web meets nlp: Collaboratively constructed semantic resources (people's web). In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, 2009.

[60] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[61] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*, 2018.

[62] Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and Bouchra El Asri. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042, 2021.

[63] Mareike Hartmann and Anders Søgaard. Multilingual negation scope resolution for clinical text. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 7–18, 2021.

[64] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 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, 2020.

[65] William Hersh. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media, 2008.

[66] Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo Nevado-Holgado. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468*, 2018. URL `https://arxiv.org/abs/1811.05468`.

[67] Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3079209.

[68] Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*, 2020.

[69] Yutai Hou, Jiafeng Mao, Yongkui Lai, Cheng Chen, Wanxiang Che, Zhigang Chen, and Ting Liu. Fewjoint: a few-shot learning benchmark for joint language understanding. *arXiv preprint arXiv:2009.08138*, 2020.

[70] 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. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259, 01 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad259. URL `https://doi.org/10.1093/jamia/ocad259`.

[71] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*, 2020.

[72] Qinghua Huang, Fan Zhang, and Xuelong Li. Few-shot decision tree for diag-

nosis of ultrasound breast tumor using bi-rads features. *Multimedia Tools and Applications*, 77:29905–29918, 2018.

[73] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*, 2024.

[74] Chen Jia and Yue Zhang. Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5906–5917, 2020. doi: 10.18653/v1/2020.acl-main.524.

[75] Chen Jia, Xiaobo Liang, and Yue Zhang. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, 2019.

[76] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

[77] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

[78] 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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. doi: 10.1038/sdata.2016.35.

[79] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo An-

thony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. doi: 10.1038/sdata.2016.35.

[80] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54, 2019.

[81] Vianney Jouhet, Georges Defossez, Anita Burgun, Pierre le Beux, P. Levillain, Pierre Ingrand, and Vincent Claveau. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(03):242–251, 2012. doi: 10.3414/ME11-01-0005.

[82] Jelena Jovanović and Ebrahim Bagheri. Semantic annotation in biomedicine: the current landscape. *Journal of biomedical semantics*, 8:1–18, 2017.

[83] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[84] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[85] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[86] Andrey Kormilitzin, Nemanja Vaci, QiangLiu, and Alejo Nevado-Holgado. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086, 2021. doi: 10.1016/j.artmed.2021.102086.

[87] Oliver Kramer and Oliver Kramer. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23, 2013.

[88] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in artificial intelligence*, 5(4):221–232, 2016.

[89] Kai Labusch, Preußischer Kulturbesitz, Clemens Neudecker, and David Zellhöfer. Bert for named entity recognition in contemporary and historical german. In *Proceedings of the 15th Conference on Natural Language Processing*, pages 9–11, 2019.

[90] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. One-shot learning by inverting a compositional causal process. *Advances in neural information processing systems*, 26, 2013.

[91] Rabindra Lamsal. Design and analysis of a large-scale covid-19 tweets dataset. *Applied Intelligence*, 51(5):2790–2804, 2021.

[92] Laura Lander, Janie Howsare, and Marilyn Byrne. The impact of substance use disorders on families and children: from theory to practice. *Social work in public health*, 28(3-4):194–205, 2013.

[93] Alicia Lara-Clares and Ana Garcia-Serrano. Key phrases annotation in medical documents: Meddocan 2019 anonymization task. In *IberLEF@ SEPLN*, pages 755–760, 2019.

[94] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*, 2016.

[95] Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung.

Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*, 2021.

[96] Ulf Leser and Jörg Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369, 2005.

[97] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[98] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[99] 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. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL `https://doi.org/10.1093/database/baw068`.

[100] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

[101] Mingchen Li, Huixue Zhou, Han Yang, and Rui Zhang. Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association*, page ocae095, 2024.

[102] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11516–11525, 2021.

[103] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7260–7268, 2019.

[104] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL `https://aclanthology.org/2021.acl-long.353`.

[105] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL `https://aclanthology.org/2021.acl-long.353`.

[106] Yinghao Li, Pranav Shetty, Lucas Liu, Chao Zhang, and Le Song. Bertifying the hidden markov model for multi-source weakly supervised named entity recognition. *arXiv preprint arXiv:2105.12848*, 2021.

[107] T Warren Liao, Zhiming Zhang, and Claude R Mount. Similarity measures for retrieval in case-based reasoning systems. *Applied Artificial Intelligence*, 12(4): 267–288, 1998.

[108] Salvador Lima-López, Naiara Pérez, Montse Cuadros, and German Rigau. Nubes: A corpus of negation and uncertainty in spanish clinical texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5772–5781, 2020.

[109] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Yearbook of medical informatics*, 2(01):41–51, 1993. doi: 10.1055/s-0038-1634945.

[110] Juhua Liu, Qihuang Zhong, Liang Ding, Hua Jin, Bo Du, and Dacheng Tao. Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2629–2642, 2023. doi: 10.1109/TASLP.2023.3290431.

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

[112] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, 2018.

[113] Jueqing Lu, Lan Du, Ming Liu, and Joanna Dipnall. Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs. *arXiv preprint arXiv:2010.07459*, 2020.

[114] Monica Luciana, James M Bjork, Bonnie J Nagel, Deanna M Barch, Raul Gonzalez, Sara Jo Nixon, and Marie T Banich. Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery. *Developmental cognitive neuroscience*, 32:67–79, 2018.

[115] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022.

[116] Brandon Lwowski and Peyman Najafirad. Covid-19 surveillance through twitter using self-supervised and few shot learning. In *EMNLP 2020 Workshop NLP-COVID*, 2020. URL `https://openreview.net/forum?id=OOq31Bnxx5I`.

[117] Jianzhu Ma, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F. A. Wessels, Marc Hafner, Roded Sharan, Jian Peng, and Trey Ideker. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2:233–244, 2021. URL `https://www.nature.com/articles/s43018-020-00169-2`.

[118] Manolis Manousogiannis, Sepideh Mesbah, Selene Baez Santamaria, Alessandro Bozzon, and Robert-Jan Sips. Give it a shot: Few-shot learning to normalize adr mentions in social media posts. In *Proceedings of the Fourth Social Media Mining for Health Applications (SMM4H) Workshop and Shared Task*, pages 114–116, 2019.

[119] Montserrat Marimon, Jorge Vivaldi, and Núria Bel Rafecas. Annotation of negation in the iula spanish clinical record corpus. In Eduardo Blanco, Roser Morante, and Roser Saurí, editors, *SemBEaR 2017. Computational Semantics*

Beyond Events and Roles, pages 43–52, Valencia, Spain, 2017. Association for Computational Linguistics (ACL).

[120] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638, 2019.

[121] George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.139. URL `https://aclanthology.org/2021.naacl-main.139`.

[122] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with {umls} concepts. In *Automated Knowledge Base Construction (AKBC)*, 2019. URL `https://openreview.net/forum?id=SylxCx5pTQ`.

[123] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS medicine*, 6(7):1549–1676, 2009. doi: https://doi.org/10.1371/journal.pmed.1000097.

[124] Richard D Morey et al. Confidence intervals from normalized data: A correction

to cousineau (2005). *Tutorials in quantitative methods for psychology*, 4(2):61–64, 2008.

[125] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[126] Preslav Nakov, Sara Rosenthal, Svetlana Kiritchenko, Saif M. Mohammad, Zornitsa Kozareva, Alan Ritter, Veselin Stoyanov, and Xiaodan Zhu. Developing a successful semeval task in sentiment analysis of twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65, 2016. doi: 10.1007/s10579-015-9328-1.

[127] Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop*, pages 1–7, 2013.

[128] Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. Named entity recognition for social media texts with semantic augmentation. *arXiv preprint arXiv:2010.15458*, 2020.

[129] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. doi: 10.1109/TKDE.2009.191.

[130] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[131] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[132] Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. Towards one-shot learning for rare-word translation with external experts. *arXiv preprint arXiv:1809.03182*, 2018.

[133] Gabriele Prato, Ella Charlaix, and Mehdi Rezagholizadeh. Fully quantized transformer for machine translation. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1–14, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.1. URL https://aclanthology.org/2020.findings-emnlp.1.

[134] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun'ichi Tsujii, and Sophia Ananiadou. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581, 2012.

[135] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. URL https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[136] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the thirteenth conference on computational natural language learning (CoNLL-2009)*, pages 147–155, 2009.

[137] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *International Conference on Learning Representations*, 2016. URL https://openreview.net/forum?id=rJY0-Kcll.

[138] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[139] Kayla B Rhidenour, Kate Blackburn, Ashley K Barrett, and Savanna Taylor. Mediating medical marijuana: exploring how veterans discuss their stigmatized substance use on reddit. *Health Communication*, 37(10):1305–1315, 2022.

[140] Anthony Rios and Ramakanth Kavuluru. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empir-*

ical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2018, page 3132. NIH Public Access, 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6375489/.

[141] Anthony Rios and Ramakanth Kavuluru. Emr coding with semi–parametric multi–head matching networks. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2081. NIH Public Access, 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7720861/.

[142] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[143] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003. URL https://arxiv.org/abs/cs/0306050.

[144] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

[145] Abeed Sarker. Social media mining for toxicovigilance of prescription medications: End-to-end pipeline, challenges and future work. *arXiv preprint arXiv:2211.10443*, 2022.

[146] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*, 54:202–212, 2015.

[147] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May 1987. URL `http://www.idsia.ch/~juergen/diploma.html`.

[148] Hiemke Katharina Schmidt, Martin Rothgangel, and Dietmar Grube. Prior knowledge in recalling arguments in bioethical dilemmas. *Frontiers in psychology*, 6:1292, 2015. doi: https://doi.org/10.3389/fpsyg.2015.01292.

[149] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8247–8255, 2019.

[150] Amr Sharaf, Hany Hassan, and Hal Daumé III. Meta-learning for few-shot nmt adaptation. *arXiv preprint arXiv:2004.02745*, 2020.

[151] Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. knn-prompt: Nearest neighbor zero-shot inference. *arXiv preprint arXiv:2205.13792*, 2022.

[152] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023.

[153] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. URL `https://arxiv.org/abs/1703.05175`.

[154] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash

Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023.

[155] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[156] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[157] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63, 2011.

[158] Robert Stewart, Mishael Soremekun, Gayan Perera, Matthew Broadbent, Felicity Callard, Mike Denis, Matthew Hotopf, Graham Thornicroft, and Simon Lovestone. The south london and maudsley nhs foundation trust biomedical research centre (slam brc) case register: development and descriptive data. *BMC psychiatry*, 9(1):1–12, 2009. doi: 10.1186/1471-244X-9-51.

[159] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

[160] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. Caire-covid: a question answering and query-focused multi-document summarization system for covid-19 scholarly information management. *arXiv preprint arXiv:2005.03975*, 2020.

[161] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 2013. doi: 10.1136/amiajnl-2013-001628.

[162] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[163] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012. doi: 10.1002/asi.21662.

[164] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[165] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer, 2012.

[166] Nemanja Vaci, Qiang Liu, Andrey Kormilitzin, Franco De Crescenzo, Ayse Kurtulmus, Jade Harvey, Bessie O'Dell, Simeon Innocent, Anneka Tomlinson, Andrea Cipriani, and Alejo Nevado-Holgado. Natural language processing for structuring clinical text data on depression using uk-cris. *Evidence-based mental health*, 23:21–26, 2020. doi: 10.1136/ebmental-2019-300134.

[167] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, and Daan Wierstra. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016. doi: /10.5555/3157382.3157504.

[168] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[169] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stick-

ier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[170] Xu Wang, Chen Yang, and Renchu Guan. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, 9:373–382, 2018.

[171] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020. doi: https://doi.org/10.1145/3386252.

[172] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[173] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, and Michelle Franchini. Ontonotes release 5.0. Linguistic Data Consortium, Philadelphia, PA, 2013.

[174] Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30, 2019.

[175] Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. Empirical study of zero-shot NER with ChatGPT. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore, Decem-

ber 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.493. URL https://aclanthology.org/2023.emnlp-main.493.

[176] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.

[177] Leiming Yan, Wenying Zheng, Huajie (Harry) Zhang, Hao Tao, and Ming He. Learning discriminative sentiment chunk vectors for twitter sentiment analysis. *J Inf Technol*, 18(7):1605–1613, 2017. doi: 10.6138/JIT.2017.18.7.20170410.

[178] Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 7:29799–29810, 2018. URL https://link.springer.com/article/10.1007/s11042-018-5772-4.

[179] Yi Yang and Arzoo Katiyar. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*, 2020.

[180] Heng yang Lu, Chenyou Fan, xiaoning Song, and Wei Fang. A novel few-shot learning based multi-modality fusion model for covid-19 rumor detection from online social media. *PeerJ Computer Science*, 7:2376–5992, 2021. URL https://peerj.com/articles/cs-688/.

[181] 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.

[182] Shujuan Yin, Weizhong Zhao, Xingpeng Jiang, and Tingting He. Knowledge-aware few-shot learning framework for biomedical event trigger identification. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 375–380. IEEE, 2020.

[183] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023. doi: https://doi.org/10.48550/arXiv.2303.14070.

[184] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098, 2013.

[185] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021.

[186] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/zhao21c.html`.

[187] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Learning reporting dynamics during breaking news for rumour detection in social media. *arXiv preprint arXiv:1610.07363*, 2016.

[188] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010. doi: 10.1136/jamia.2010.003947.

[189] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the*

American Medical Informatics Association, 18(5):552–556, 2011. doi: 10.1136/amiajnl-2011-000203.