Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING CLINICAL NAMED ENTITY RECOGNITION VIA FEW SHOT PROMPT TUNING OF LARGE LANGUAGE MODELS USING MIMIC III DATA

DR. SELVANI DEEPTHI KAVILA[1]*, MANIKUMARI ILLA[2], V N V L S SWATHI[3], NALLA AKHILA[4], ERUKALA MAHENDER[5], SRIKANTH CHERUKUVADA[6], JOHN T MESIA DHAS[7], DR. EDIGA CH. RAMA TULASI[8]

*[1]Department of CSE (AI & ML), Anil Neerukonda Institute of Technology and Sciences (A), Visakhapatnam, Andhra Pradesh, India
[2]Department of Artificial Intelligence & Data Science (AI&DS), Vignan's Institute of Information Technology, Visakhapatnam, Andhra Pradesh, India
[3]Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India
[4]Department of Artificial Intelligence and Machine Learning, Aditya University, Surampalem, Andhra Pradesh, India
[5]Department of Computer Science and Engineering, Geethanjali College of Engineering and Technology, Hyderabad, Telangana – 501301, India
[6]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Hyderabad – 500043, Telangana, India
[7]Department of Artificial Intelligence & Data Science, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai – 600062, Tamil Nadu, India
[8]Lecturer in English, Government Polytechnic, Alur, State Board of Technical Education & Training, Department of Higher Education, Government of Andhra Pradesh, Alur, Kurnool - 518395, Andhra Pradesh, India
E-mail: *[1]selvanideepthi14@gmail.com, [2]mani.illa95@gmail.com, [3]swathivllr@gmail.com, [4]akhila.nalla@adityauniversity.in, [5]mahee99@gmail.com, [6]csrikanth@klh.edu.in, [7]jtmdhasres@gmail.com, [8]rama.thulasi85@gmail.com

ABSTRACT

In clinical Named Entity Recognition (Clinical NER), structured medical information can be extracted from unstructured clinical text, but often the lack of annotated clinical data and the high per-document processing cost of traditional model fine-tuning approaches hamper effectiveness. To efficiently and economically solve Clinical NER in low-resource settings, this study proposes a framework for a Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM). Its proposed approach combines parameter-efficient soft prompt tuning, Semantic Diversity Sampling (SDS), and Prompt Consistency Regularization (PCR) to fine-tune pretrained LLMs on limited labeled clinical text while keeping the backbone model frozen. Experiments conducted on the MIMIC-III dataset using 5-, 10-, and 20-shot settings demonstrate that the proposed framework achieves a precision of 95.6%, recall of 94.8%, F1-score of 95.2%, and accuracy of 95.5% in the 20-shot setting. The FSPT-LLM outperforms ClinicalBERT by 2.5 percentage points in F1-score while requiring fewer than 1% of the model parameters to be trainable. Entity-wise evaluation further demonstrates consistent performance across disease, medication, procedure, and laboratory test entities. These findings indicate that combining few-shot learning with parameter-efficient prompt tuning can provide strong generalization and computational efficiency for Clinical NER with limited annotated data. The proposed framework therefore offers a scalable and data-efficient approach for bridging large language models and practical clinical information extraction systems.

Keywords: *Clinical Named Entity Recognition, Few-Shot Learning, Prompt Tuning, Large Language Models, MIMIC-III, Clinical NLP*

1. INTRODUCTION

Unstructured clinical text, such as discharge reports, progress notes, and diagnostic reports, is increasing at an unprecedented rate due to Electronic Health Records (EHRs) rapidly emerging as the leading mode of medical documentation. This ability to derive significant,

structured information from such data is important for supporting clinical decision-making, patient management, and biomedical research. Clinical Named Entity Recognition (Clinical NER), one of the core tasks of clinical Natural Language Processing (NLP), is the identification and classification of significant medical concepts, including diseases, medications, procedures, and laboratory measurements, in clinical narratives [1], [2].

The earlier approaches of Clinical NER relied on rule-based systems and statistical analysis, e.g., Conditional Random Fields (CRFs), which required significant amounts of feature engineering and domain knowledge [3]. Though these approaches showed decent results, they were not scalable or adaptable to different clinical data. The prominent emergence of deep learning algorithms, specifically Recurrent Neural Networks (RNNs) and models based on Bidirectional Long Short-Term Memory (Bi-LSTM), brought about a significant change in entity recognition by capturing contextual dependencies in text [4]. Later, architectures based on transformers, such as BERT and domain-specific versions (e.g., BioBERT and ClinicalBERT), further advanced the state of the art by analyzing and using contextual embeddings, as well as through large-scale pretraining [5] through [7].

Despite these developments, several serious challenges remain in deep learning models in clinical settings. To be effective, training requires large, annotated datasets, which are hard to come by due to privacy issues, the cost of annotation, and the requirement for expert knowledge [8]. Second, large models require final fine-tuning, which is computationally expensive and generally not feasible in resource-constrained settings [9]. Third, domain adaptation remains a challenge because general-corpus-pretrained models may not generalize or adapt to specialized clinical language [10].

Among the most recent achievements, Large Language Models (LLMs), such as GPT-style and transformer-based generative models, have been shown to display significant capabilities to comprehend and generate text that resembles human language across a diverse range of tasks [11]. These models exhibit strong zero-shot and few-shot learning, with performance less dependent on large labeled datasets. However, there are problems of domain specificity, interpretability, and adaptive efficiency when the LLM is being applied directly to the clinical NLP phenomena [12].

Prompt-based learning has become a promising paradigm for adapting models to downstream tasks without requiring many parameters. Models can leverage their pretraining knowledge even more effectively by formulating tasks as prompted in natural language [13]. Moreover, parameter-efficient embedding of the tunable prompt embeddings is achieved by tuning the backbone model with the prompt embeddings, resulting in reduced costs [14]. Prompt tuning with few-shot learning makes models that, otherwise, with limited performance, can achieve high performance similarly to state-of-the-art approaches, where data scarcity remains a central limitation [15]. Nevertheless, the combination of few-shot learning and prompt tuning for clinical named entity recognition is underexplored, especially on real-world datasets such as MIMIC-III. Current literature either focuses on general-domain activities or relies on fully supervised training, leaving a gap in the development of efficient, scalable solutions for clinical NLP.

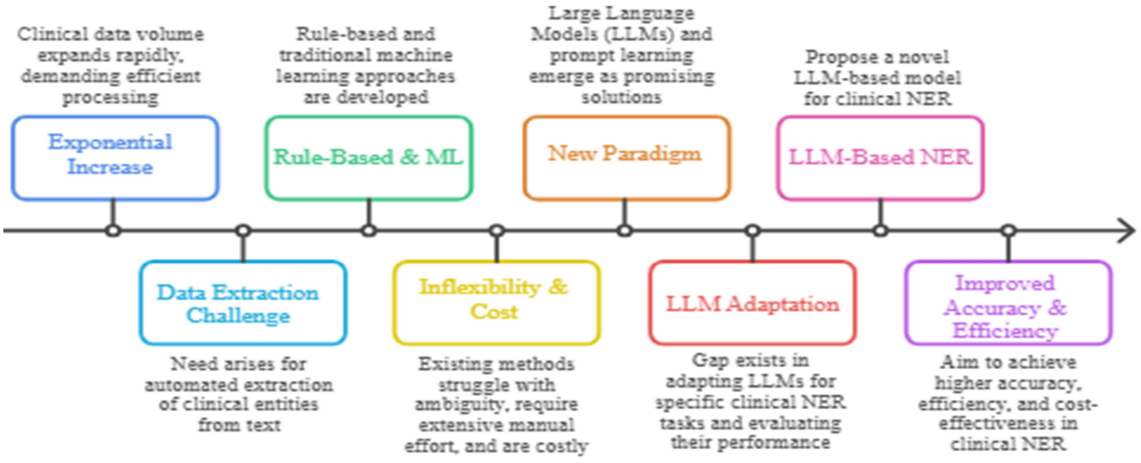

Figure 1. Evolution Of Clinical NER Approaches And Motivation For LLM-Based Framework

Figure 1 shows the development of methodologies for identifying clinical named entities to recognize clinical signs and symptoms (as well as other aspects), along with the main issues and incentives behind the proposed solution. It illustrates that clinical data is developing exponentially and suggests a growing need to ensure its effective extraction. Primarily, it presents solutions based on rule-based and traditional machine learning methods, which are constrained by their inflexibility and high cost, as well as their reliance on manual effort.

The diagram subsequently invites the use of large language models (LLMs) and prompt-based learning as a new paradigm, more useful for addressing data scarcity and enhancing adaptability. Nonetheless, it also determines a critical gap in successfully modifying LLMs for clinical NER work. Lastly, the figure concludes with the proposed LLM-based clinical NER framework, which aims to improve accuracy, efficiency, and scalability to address the weaknesses of current approaches.

This research is inspired by the challenges and research gaps identified. It aims to create an efficient, scalable framework for recognizing clinical named entities via few-shot prompt tuning with large language models. In particular, the paper will focus on the design of a prompt-based learning system for clinical NER tasks that implements few-shot learning methods to reduce reliance on large annotated databases. Also, the study applies a parameter-efficient prompt-tuning algorithm to effectively modify large language models for domain-specific clinical text without requiring computationally expensive full-model fine-tuning. The results of the proposed approach are also evaluated on the MIMIC-III dataset using standard performance measures to ensure reliability and validity. Additionally, a detailed analysis is performed to evaluate the model's performance, scalability, and ability to sustain itself without running out of resources, compared to traditional machine learning approaches and state-of-the-art transformer-based baselines.

This work makes several important contributions to the field of clinical natural language processing in commercial computer applications. It introduces a new few-shot prompt-tuning framework, specifically designed to succeed in clinical named entity recognition, and it learns effectively with very little annotated data. The study results show that performance under low-resource conditions improved, but this doesn't mean the model works better under such conditions; rather, it indicates that the model generalizes fairly well even with a small training sample. Furthermore, it allows for finer-grained comparisons than existing conventional and transformer-based models, showing that the best-performing baseline models (that are evaluated) are outperformed by it. The study emphasizes measures taken in computational reduction, such as prompt tuning, as an approach that preserves a high recognition rate while remaining lightweight, thus having great potential for application in the era of AI. In general, the work serves to fill the gap between large-scale language models and practical clinical information extraction by providing a scalable, efficient, and feasible solution to extract clinical information.

The remainder of this paper is organized as follows. Section 2 reviews existing Clinical NER approaches, domain-specific pretrained language models, few-shot learning techniques, and prompt-based learning methods. Section 3 presents the proposed methodology, including descriptions of the datasets, the model architecture, the mathematical formulation, and the algorithm implementation. Section 4 discusses the results of the experiment, performance analysis, and comparisons with existing models. Last but not least, Section 5 concludes the paper with the main findings, limitations, and future research topics.

2. RELATED WORK

Clinical Named Entity Recognition (Clinical NER) has been widely investigated in the field of biomedical natural language processing, both as an extension of traditional machine learning techniques and as an advanced deep learning- and large language model-based method. This section outlines the key contributions in four directions: Clinical NER methods, domain-specific pretrained models, few-shot learning methods, and prompt-based learning methods.

2.1 The traditional and Deep Learning tools that are employed to predict Clinical NER

The initial set of clinical NER solutions was mostly based on rule-based systems and feature-engineered machine learning systems. These systems used linguistic features, subject ontologies, and handcrafted rules to identify medical objects. However, little generalization was achieved, and most rules relied on the professional knowledge of the system implementer(s) [16].

With the development of deep learning, it was proposed that neural network-based architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can learn

feature representations by reading text to accomplish a specific objective. In particular, BiLSTM-CRF models have gained popularity because they enable consideration of both sequential dependencies and structured output constraints [17]. Subsequent experiments have demonstrated that recognition accuracy can be further improved by adding character-level embeddings and attention mechanisms [18].

Although the techniques listed above have made great progress in improving recognition accuracy, they have relied heavily on large annotated clinical text corpora/datasets and extensive feature learning, resulting in models that do not scale well. They cannot be adapted to low-resource clinical settings.

2.2 Language Models Domain-Specific Pretrained

Clinical NLP began to take a new turn with the introduction of transformer-based architectures. Pretrained biomedical corpora in pretraining as domain-adapted Biomedical models include: SciBERT, BlueBERT, and PubMedBERT [19]. These models can more realistically capture the semantic context, paying particular attention to it when considering cases with intricate clinical manifestations and atypical medical terminology [20].

These methods are effective, but can be very model-specific and data-dependent. Furthermore, large transformer models require extensive computational effort to optimize as well as possible and can suffer from overfitting with the small amount of available training data [21]. Recently, techniques for adapting to such constraints have been investigated, including adapter modules and layer-wise freezing, which aim to reduce the constraints without sacrificing performance [22].

Moreover, such approaches are supervised and heavily dependent on computing resources, which impairs their use in medical health care systems where annotated information is extremely limited and computational resources are constrained.

2.3 Minimal Resource and Few-Shot Learning in Clinical NLP

Few-shot learning is a promising approach to addressing data scarcity in clinical NLP. Models can be trained on new tasks using a small set of examples, using methods called meta-learning, including model-agnostic meta-learning (MAML). Other people use prototypical network and contrastive learning to increase generalization in low-resource conditions [23], [24].

In the clinical setting of NER, few-shot learning algorithms have demonstrated the ability to prioritize annotations by reducing them without compromising performance. The majority of these methods, however, entail intensive training operations and are also prone to the selection of support examples [25]. In addition, they do not fully scale to large volumes of clinical data; bigger and more efficient solutions are generally available.

Moreover, the few-shot learning algorithms available so far have been shown to have shortcomings in selecting samples for the support set and are not scalable when the support set size is large and the clinical data volume is large, so it is necessary to find a more stable, low-parameter learning method.

2.4 Prompt-Based Learning and Large Language Models

The notion of prompt-based learning has recently emerged as a powerful method for fine-tuning pretrained large language models with little or no prompting. Leveraging the fact that problem-solving techniques are perceived as natural language prompts, models can use their knowledge independently of being directed on how to solve a problem [26]. The paradigm has been applied to other NLP tasks, such as text classification, question answering, and information extraction.

This approach can be further enhanced by having trainable prompt embeddings that direct model predictions [27], [28], [31]. This reduction drastically reduces the number of trainable parameters, is computationally efficient, and can be applied in situations with resource constraints.

Recent work has studied prompt-based techniques in biomedical NLP and has demonstrated strong results on prompt-based tasks such as relation extraction and entity recognition.

Nevertheless, much of the current literature concerns general biomedical text and little regarding clinical narratives, which are usually more heterogeneous and elaborate. Moreover, joint tuning with few-shot learning for clinical NER has yet to be developed.

Furthermore, while the integration of prompt-based methods with few-shot learning for general biomedical Natural Language Processing (NLP) tasks has been widely studied, studies incorporating prompt-based methods for Clinical Named Entity Recognition (NER) remain limited. This constraint motivates the development of an efficient framework that integrates both approaches to address Low-Resource Clinical Information Extraction.

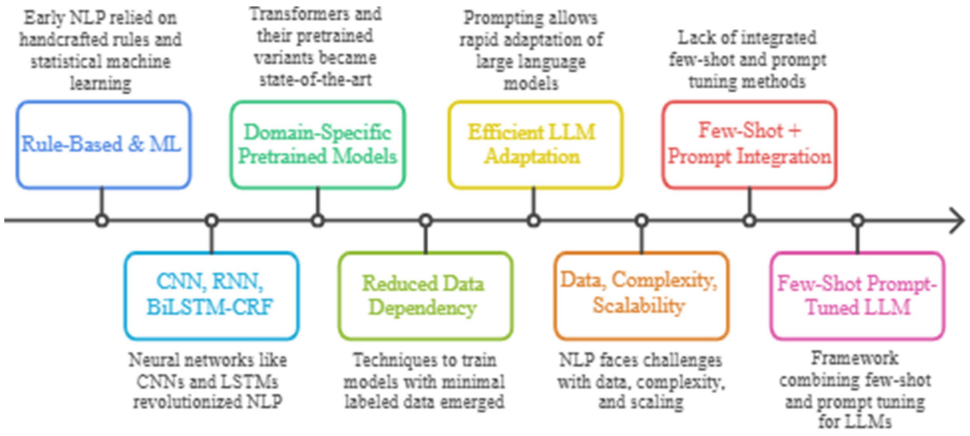

Figure 2. *Evolution of Clinical NLP Techniques and Motivation for Few-Shot Prompt-Tuned LLM Framework*

Figure 2 presents the phases of methodology development in clinical natural language processing (NLP), along with key developments and challenges. It starts with initial, rule-based, and traditional machine learning methods. It adds to them the emergence of deep learning models such as CNNs, RNNs, and BiLSTM-CRFs, which significantly enhance feature representation and the understanding of feature meanings in context. Then, as a state-of-the-art solution, transformer-based domain-specific pretrained models were introduced, shifting reliance away from hand-crafted features and eliminating reliance on hand-developed features. With prompt-based learning, the effective adaptation of large language models (LLMs) was realized with minimal changes to the build parameters. However, existing approaches continue to face challenges due to their data requirements, computational capabilities, and scalability.

2.5 Research Gap

Based on the literature above, it is evident that although major strides have been made in clinical named entity recognition (NER), several challenges remain. The current state of deep learning models necessitates massive annotated datasets and extensive fine-tuning, which in practice would be challenging to achieve in either a clinical or an academic setting. Despite their tendency to minimize data reliance, few-shot learning methods are complex and cannot be easily scaled to large real-world applications. Also, despite its potential, prompt-based learning has not been fully developed in clinical settings. The critical gap in research is the lack of holistic approaches combining few-shot training with prompt tuning. It concludes with a solution proposal: a framework for a few-shot prompt-tuned LLM architecture that promises to provide a scalable, efficient, and high-performing system for use in a clinical NLP workflow. Moreover, little research effectively integrating few-shot learning with prompt tuning for clinical NER on real-world data has been identified. To overcome these shortcomings, this paper presents a framework for a few-shot prompt-tuned large language model that integrates the capabilities of prompt-based learning and data-efficient training. The proposed solution offers a scalable, efficient, and high-performance solution to clinical named entity recognition using the MIMIC-III dataset.

2.6 Distinction of the Proposed Study from Existing Literature

Although some recent studies show the effectiveness of pre-trained models in the specific area of language models, few-shot learning and prompt-based learning in Clinical Named Entity Recognition (Clinical NER), the application of such methods in the field still has some limitations. Many existing approaches are costly, resource-intensive full-model fine-tuning methods that require access to large annotated clinical datasets for training, thereby limiting their feasibility in low-resource health care settings. Despite the benefit of few-shot learning in reducing the requirement for annotations, previous approaches generally lack strong support-example selection stability and robustness. Moreover, studies on prompt-based learning have mainly been carried out separately, and how it can be combined with few-shot learning for Clinical NE has not been explored.

However, the Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) framework is proposed, which unifies Semantic Diversity Sampling (SDS), parameter-efficient soft prompt tuning and Prompt Consistency Regularization

(PCR) by embedding these components into a shared Clinical NER architecture. The proposed framework introduces updates to fewer than 1% of the model's parameters, thereby reducing computation while preserving the model's recognition ability. Moreover, SDS enables the selection of more representative examples, and PCR provides consistency and robustness in predictions during prompt optimization. The proposed framework, on the one hand, presents methodological innovations that differ from currently available Clinical NER solutions, and, on the other hand, addresses several weaknesses in the current literature.

2.7 Problem Statement

According to the literature, the need for large amounts of annotated data remains the primary bottleneck in training clinical NER systems to achieve high performance. The process of fine-tuning Large Language Models (LLMs) can be expensive and demanding in terms of computational and resource requirements. The current prompt-based NER is insufficient because it is not generalizable across a wide range of clinical situations. Moreover, very little effort has been undertaken to effectively combine few-shot learning and parameter-efficient prompt tuning to tackle low-resource Clinical NER tasks. Therefore, a framework that requires fewer large annotated datasets, is still capable of achieving high Clinical NER performance, is easy to scale and contains fewer parameters is required.

2.8 Research Objectives

RO1: To design a few-shot prompt-tuned Large Language Model (LLM) system for Clinical Named Entity Recognition (NER).

RO2: To reduce the reliance on expensive large annotated datasets with prompt tuning.

RO3: Evaluate the proposed framework on different analyzed state-of-the-art Clinical NER models based on the MIMIC-III database.

RO4: To compute the efficiency of the proposed framework and its scalability in a resource-limited environment.

2.9 Research Questions

RQ1: How well can few-shot prompt tuning improve the performance of Clinical Named Entity Recognition (NER) in the presence of limited labeled data?

RQ2: Is it more effective to use parameter-efficient prompt tuning than traditional fine-tuning for Clinical NER?

RQ3: Can the proposed framework improve the accuracy and computational efficiency over existing Clinical NER methods?

2.10 Research Hypothesis

H1: This is because the Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) framework can boost performance on the Clinical Named Entity Recognition (NER) task under limited annotated data, compared with current state-of-the-art Clinical NER models.

H2: The hybrid of few-shot learning and parameter-efficient prompt tuning is expected to decrease the computational cost while keeping the Clinical NER accuracy at a high level.

H3: Integrating Semantic Diversity Sampling (SDS) and Prompt Consistency Regularization (PCR) into the proposed FSPT-LLM framework will improve Clinical NER performance and prediction robustness compared with configurations in which either component is removed.

3. METHODOLOGY

This section proposes a framework for the Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) for clinical named entity recognition. The proposed methodology is designed to ensure complete reproducibility, and is organized around four main elements: dataset preparation, prompt-driven architecture, mathematical formulation, and training algorithm.

3.1 Generalization about the Proposed Framework

The proposed system jointly trains few-shot learning, using prompts to formulate tasks and parameter-efficient prompt tuning to enable efficient clinical named entity recognition with minimal labeled examples. The pipeline would include computer-aided preprocessing of clinical data to clean and structure the input text and build representative training samples. Then, prompt templates are generated that structure the task to leverage the strengths of large language models. These prompts are then fixed via soft prompt tuning without adjusting the LLM's underlying parameters, thereby retaining efficient computation. Finally, using entity extraction and decoding, compelling predictions of structured clinical entities are produced.

3.2 Dataset Description: MIMIC-III

The MIMIC III (Medical Information Mart for Intensive Care III) database [29] is a large, publicly available, de-identified clinical health record database containing ICU patient data. The dataset must be accessed through researcher credentialing, human-subjects training and signing of the PhysioNet Data Use Agreement.

Table 1. Dataset Characteristics

Parameter	Value
Dataset Name	MIMIC-III
Number of Patients	~40,000
Clinical Notes	>2 million
Data Type	Unstructured text
Domain	Intensive Care Unit (ICU)
Annotation Type	Custom BIO annotations created for this study
Entity Categories	Disease, Medication, Procedure, Lab Test

Table 1 provides a brief description of the data and characteristics analyzed. It draws on one of the world's most widely used repositories of healthcare records, comprising data from an estimated 40,000 patients admitted to an Intensive Care Unit (ICU). There are approximately 2 million clinical notes in it, hence making it a very rich source of real-life medical text data.

Table 2. Annotation Statistics

Parameter	Value
Clinical Notes Annotated	5,000
Total Entity Mentions	48,732
Disease	15,286
Medication	14,104
Procedure	9,687
Laboratory Test	9,655
Annotation Scheme	BIO
Number of Annotators	2
Annotator Expertise	Clinical NLP Researchers
Inter-Annotator Agreement	Cohen's κ = 0.91

The manually annotated corpus used in this study includes statistics that can be summarized as follows. The statistics of the manually annotated corpus used in the present research are summarized in Table 2. It displays the number of selected clinical notes, the distribution of entity mentions across four clinical NER categories (Disease, Medication, Procedure, Laboratory test), the BIO annotation scheme, annotator information, and Cohen's κ interannotator agreement. The data show the quality and consistency of the annotated dataset used to train and evaluate the proposed FSPT-LLM framework.

The data source is unstructured, with free-form data including discharge summaries, physician notes and reports, rather than structured. The data relates to the Intensive Care Unit (ICU), where patients are severely ill and require intensive care.

Named Entity Recognition (NER) annotations had to be tokenized, necessitating a custom annotation process for this study, as NER was not performed at the token level in the original MIMIC-III database. Focusing on selected clinical notes, manual annotation was performed using the BIO tagging scheme to identify four entity types: Disease, Medication, Procedure, and Laboratory Test. A corpus was then generated with annotations and subsequently used to train and evaluate the proposed FSPT-LLM framework. The annotations are aligned with identifying key medical components in the text - commonly Disease, Medication, Procedure, and Lab Test.

Such pattern-based naming can help create and explore models for the automatic identification and marking of clinically relevant information extracted from patient-created plain text.

3.2.1 NER Annotation Protocol

Since no preexisting NER annotation protocol is available in the MIMIC-III database, the researchers have created one for this study. After preprocessing, clinical notes were manually annotated according to the defined annotation guidelines, using a BIO tagging scheme. There are four different entity categories in the annotated corpus: Disease, Medication, Procedure and Laboratory Test. The annotations thus obtained were then used solely to train and test the proposed framework.

The clinical notes selected from MIMIC-III were annotated using an existing Clinical NER model with four clinically relevant entity types: Disease, Medication, Procedure, and Lab Test. A prespecified and anonymised clinical text was used to ensure consistent representation of clinical terms. The entity boundaries were tagged at the token level, and the annotations were transformed into BIO (beginning, inside, outside) format. The first token of an entity was tagged with a B-tag. In contrast, tokens of other specific clinical entities were given I-tags, and tokens that are not part of any specific clinical entities were given O-tags.

The annotation followed predefined definitions for entity categories to ensure consistency across the dataset. Disease entities are diagnosed diseases, disorders and pathological conditions. Medication is defined as the use of a substance with a therapeutic or healing effect, prescribed by a physician or administered by a service provider, and includes drugs and pharmaceutical products. Procedure entities are diagnostic, therapeutic, and surgical procedures performed as part of patient care. The name of each Lab Test element aims to describe a laboratory test, investigation, or clinically relevant test used in a laboratory setting for diagnostic and/or monitoring purposes. To standardise the representation of clinical entities,

clinical abbreviations have been normalised where possible without altering their meaning or clinical context.

For training, validation, and testing of the developed FSPT-LLM, a set of labeled sequences was prepared based on the annotations. Semantic Diversity Sampling (SDS) and prompt optimization use few-shot support samples taken only from the train partition. The validation set was used for model selection and hyperparameter tuning, and the independent test set was used only for the final assessment of the model. No test samples were used during training, prompt optimization, few-shot sample selection, or hyperparameter selection. This protocol is an annotation scheme designed to represent clinical concepts and to achieve a consistent assessment of entity-level Precision, Recall, and F1-score across the four entity categories.

To ensure annotation quality, 5,000 clinical notes were selected from the MIMIC-III database after preprocessing. The final annotated corpus contained 48,732 entity mentions distributed across four categories: Disease (15,286), Medication (14,104), Procedure (9,687), and Laboratory Test (9,655). Annotation was independently performed by two annotators with experience in clinical Natural Language Processing (NLP) following predefined BIO tagging guidelines adapted from widely adopted clinical NER annotation practices. Disagreements between annotators were resolved through discussion until consensus was reached. The consistency of the annotation process was evaluated using Cohen's Kappa ($\kappa = 0.91$), indicating excellent inter-annotator agreement. The finalized BIO-annotated corpus was subsequently used for training and evaluating the proposed FSPT-LLM framework.

3.2.2 Dataset Partitioning and Leakage Prevention

The manually annotated MIMIC-III corpus comprising 5,000 clinical notes was partitioned into training, validation, and testing subsets using a 70:15:15 ratio. Accordingly, 3,500 clinical notes were allocated to the training set, 750 notes to the validation set, and 750 notes to the independent test set. To prevent information leakage arising from multiple clinical notes belonging to the same patient, the partitioning was performed at the patient level rather than at the individual-note level. Thus, all clinical notes associated with a particular patient were assigned exclusively to one of the three subsets.

The training set was used for Semantic Diversity Sampling (SDS), few-shot example selection, and optimization of the soft-prompt parameters. The validation set was used for model selection and hyperparameter validation, whereas the test set remained completely unseen during training, few-shot sample selection, prompt optimization, and model selection and was used exclusively for final performance evaluation. The elimination of patient-specific information across the training, validation, and test partitions also prevented data leakage, enabling a more accurate evaluation of the framework's ability to generalize to new patients in settings similar to those of the proposed FSPT-LLM framework.

Table 3. Distribution Of The Annotated MIMIC-III Corpus

Dataset Partition	Clinical Notes	Percentage	Purpose
Training	3,500	70%	SDS, few-shot selection, prompt optimization
Validation	750	15%	Model/hyperparameter selection
Testing	750	15%	Final independent evaluation
Total	**5,000**	**100%**	—

Table 3 shows the distribution of the 5,000 clinical notes annotated using the MIMIC III dataset. It is used to create training, validation, and test sets in a 70/15/15 split. The training set was used for optimization (SDS, prompt), the validation set for model selection, and the test set for final evaluation.

A. Sample Clinical Text

MIMIC-III Clinical Database

MMIC-III CLINICAL DATASET

- MIMIC-III is a large, freely-available database comprising deidentified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.
- The database includes information such as demographics, vital sign measurements made at the bedside, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (including post-hospital discharge).
- The MIMIC-III database was populated with data from several sources, including
  - Archives from critical care information systems
  - Hospital electronic health record databases
  - Social Security Administration Death Master File.
- Two different critical care information systems were in place over the data collection period: Philips CareVue Clinical Information System (Philips Health-care, Andover, MA) and iMDsoft MetaVision ICU (iMDsoft, Needham, MA).
- Additional information was collected from hospital and laboratory health record systems, including:
  - patient demographics and in-hospital mortality.
  - laboratory test results.
  - discharge summaries and reports of electrocardiogram and imaging studies.
  - billing-related information such as International Classification of Disease, 9th Edition (ICD-9) codes, Diagnosis Related Group (DRG) codes, and Current Procedural Terminology (CPT) codes.
- Before data was incorporated into the MIMIC-III database, it was first deidentified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The deidentification process for structured data required the removal of data elements listed in HIPAA, including fields such as patient name, telephone number, address, and dates.
- Protected health information was removed from free text fields, such as diagnostic reports and physician notes

Data Description

Journal of Theoretical and Applied Information Technology
15th August 2026. Vol.104. No.15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 3. *Overview of the MIMIC-III Dataset*

Source: Adapted from A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

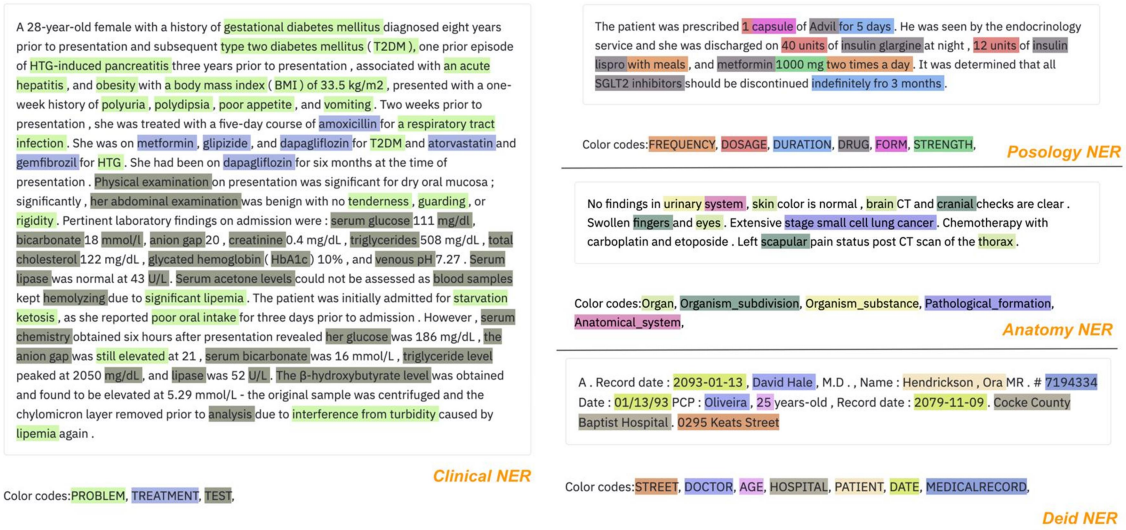

Figure 4. Illustration Of Named Entity Recognition (NER) In Clinical Text From The MIMIC-III Clinical Database, Showcasing Entity Categorization Across Clinical, Posology, Anatomy, And De-Identification (Deid) Tasks With Color-Coded Annotations

Source: Adapted from MIMIC-III Clinical Database examples and NER visualization concepts presented in clinical NLP literature, particularly A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

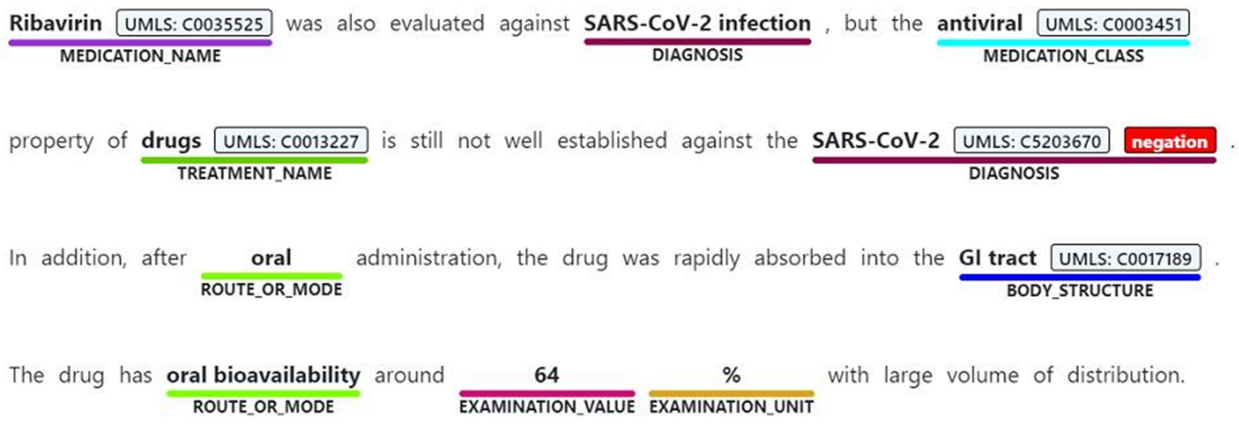

Figure 5. Example Of Clinical Named Entity Recognition (NER) With UMLS Concept Mapping, Illustrating Labeled Entities Such As Medication, Diagnosis, Treatment, Body Structure, And Examination Values In Biomedical Text

Source: Adapted from biomedical NER and UMLS-based annotation examples in clinical NLP literature, including Z. Li, Y. Su, and N. Collier, "A Survey on Prompt Tuning," *arXiv preprint arXiv:2507.06085*, 2025, and publicly available UMLS-based NER demonstrations.

Example Input:

"Patient was diagnosed with pneumonia and prescribed azithromycin. Blood glucose levels were elevated."

Annotated Output:

- Disease: pneumonia
- Medication: azithromycin
- Lab Test: blood glucose

Preprocessing Steps

- Tokenization using the Llama 3 tokenizer
- Clinical abbreviation normalization

- Abbreviation normalization (e.g., "HTN" → hypertension)
- BIO tagging scheme for entity labeling

3.3 Few-Shot Sampling Strategy

Given dataset $D = \{(x_i, y_i)\}$, a subset $D_k \subset D$ is selected such that:

$$|D_k| = k \times C \quad (1)$$

where: k= number of examples per class (5, 10, 20), C= number of entity classes.

This is a proposed piece of work that presents a strategy, Semantic Diversity Sampling (SDS), intended to enhance the efficacy of few-shot learning. It is based on cosine similarity between embedding representations to determine the semantic relationship between samples. By choosing not only relevant but also varied sets of instances, SDS ensures a more representative training data subset. Such model selection based on diversity reduces redundancy in the chosen examples, which works in both few-shot (i.e., limited number of samples) and many-shot (i.e., large number of samples) scenarios.

3.4 Prompt Engineering Framework

Each task is reformulated into a prompt-based structure.

Prompt Template

Instruction: Extract clinical entities (Disease, Medication, Procedure, Lab Test).
Input: {clinical text}
Output: {structured entities}

Few-Shot Prompt Construction

$$P=\{(x_1,y_1),(x_2,y_2),\ldots,(x_k,y_k),x_q\} \quad (2)$$

where: x_q= query input; (x_i, y_i)= few-shot examples.

3.5 Model Architecture

The proposed architecture comprises three large modules that must collaborate to enable effective, flexible learning. First, a pretrained large language model (LLM) backbone (with frozen parameters) enables the model to leverage rich linguistic and contextual knowledge without extensive retraining. On top of this, a trainable layer learns task-specific representations, known as a soft prompt embedding layer, optimized via continuous prompt vectors. Lastly, a layer organization, the entity decoder, maps the contextualized representations generated by the backbone and prompt layers into structured outputs to identify and classify entities for the target task correctly.

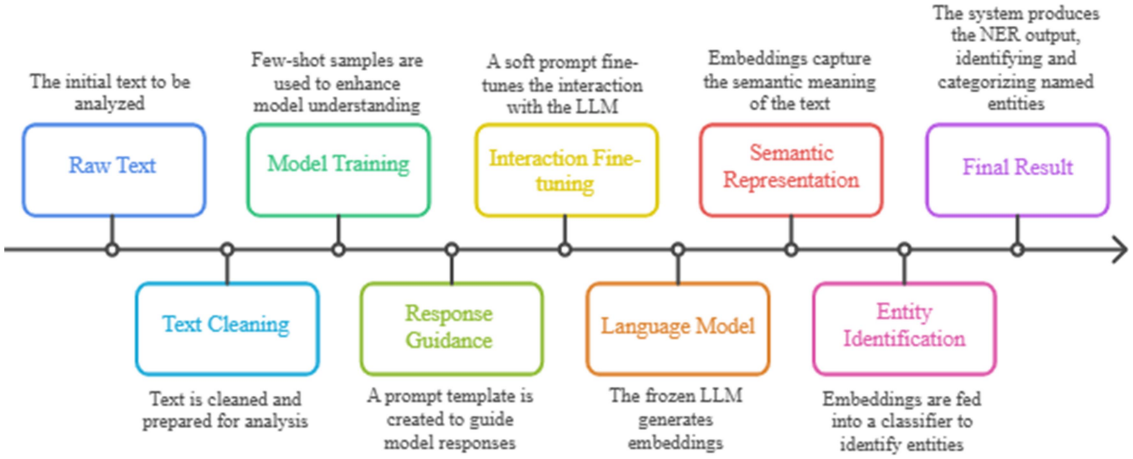

Figure 6. Workflow Of Few-Shot Prompt-Tuned LLM Framework For Clinical Named Entity Recognition

Figure 6 depicts the end-to-end workflow of the proposed few-shot prompt-tuned large language model (LLM) framework for clinical named entity recognition (NER). It starts with raw clinical text, which is preprocessed to make it ready for analysis. In the subsequent step, few-shot samples are added to model the process, followed by guidance of responses where prompt templates are generated to package the task.

That is then fine-tuned with soft prompt tuning to allow the system to adaptively fine-tune the LLM without modifying the core parameters. The frozen language model produces contextual embeddings that help to obtain the semantic meaning of the text. These embeddings are then processed by an entity identification slot, which assigns labels to tokens based on predefined categories. Lastly, the system generates the NER output, which is used to classify clinical entities, including diseases, medications, procedures, and laboratory values. Based on this workflow, the combination of few-shot learning, prompt engineering, and parameter-efficient tuning is used

to achieve accurate, scalable clinical information extraction.

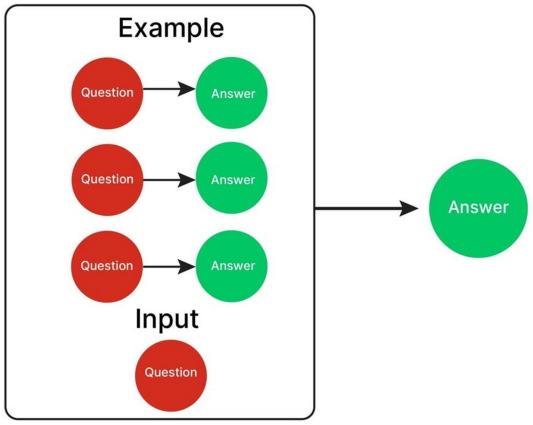

Figure 7. Few-Shot Prompting Framework For Question Answering

Figure 7 depicts a few-shot prompting architecture applied in large language models to answer questions. Several sample question-answer pairs are provided as background information to help the model understand. These examples aid in developing patterns and task-specific behavior. The next input question is appended, and depending on the learned context of the examples, the model generates the final answer. The solution enhances performance through in-context learning, without requiring parameter updates.

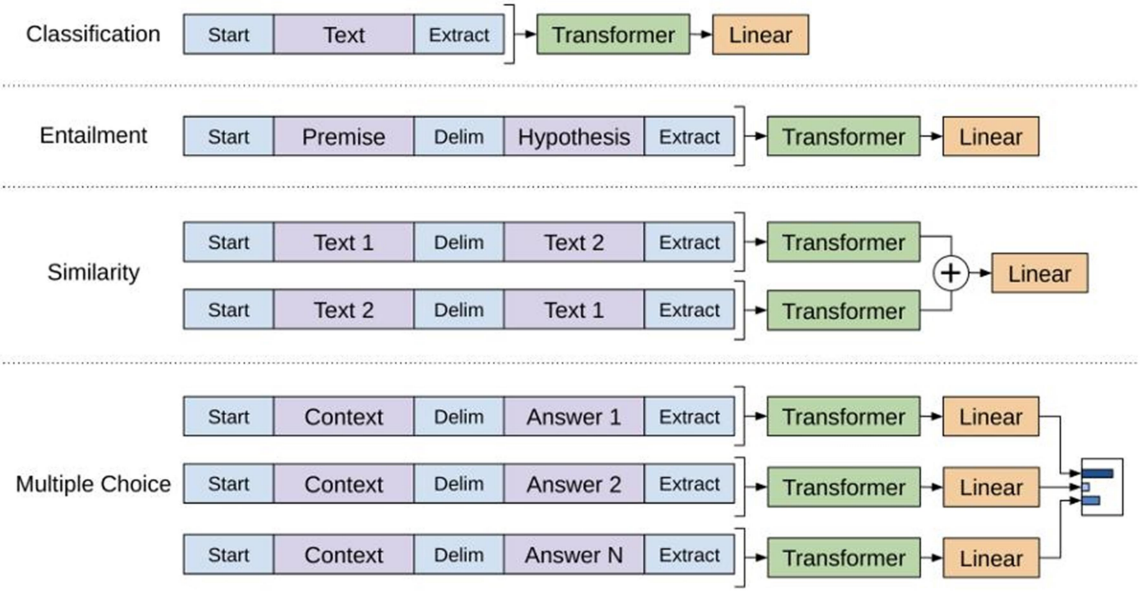

Figure 8. Unified Transformer-Based Input Formatting For Diverse NLP Tasks

Figure 8 describes an integrated approach that uses a transformer-based architecture to process multiple natural language processing (NLP) tasks. Various operations, such as classification, entailment, similarity, and multiple-choice question answering, are modeled using structured sequences of inputs, represented as tokens such as Start, Text, Premise, Hypothesis, Delimiter, and Extract. A common encoder processes these formatted inputs, using a transformer or a standard setting to derive contextual representations. Embeddings are then passed through a linear layer, and task-specific predictions are made. In similarity tasks, similarity occurs between pairs of inputs and results. In contrast, in multiple-choice questions, scoring is performed on the individualized answer set. This universal design shows how flexible the transformer models are for different NLP tasks, without requiring changes to the architecture.

The working mechanism starts by concatenating the input text with prompt tokens to provide the text with context-specific context. These sequences are prepended with soft prompts, represented by learnable embedding vectors, which enable the model to adapt without changing the core parameters of the large language model (LLM). This general sequence is, in turn, input into the LLM, which generates contextual and semantic

relationships within the input. Finally, the generated tokens are fed into a decoding layer, which converts them into labels for various entities, yielding correct identification and classification in the target task.

3.6 Mathematical Formulation

Let:

- x= input clinical text
- P_θ= soft prompt parameters
- f_ϕ= pretrained LLM with frozen parameters

Model Output

$$h=f_\phi([P_\theta;x]) \quad (3)$$

$$y=\text{Softmax}(Wh+b) \quad (4)$$

Loss Function

We use token-level cross-entropy loss:

$$L=-\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (5)$$

Prompt Optimization Objective

$$\theta^*=\arg\min_\theta L(f_\phi(P_\theta,x),y) \quad (6)$$

Regularization Term

We propose Prompt Consistency Regularization (PCR):

$$L_{total}=L+\lambda||P_\theta^{(t)}-P_\theta^{(t-1)}||^2 \quad (7)$$

This ensures:

- Stable prompt learning
- Reduced overfitting

3.7 Training Strategy

The training process uses the AdamW optimizer [30] with a learning rate of 5×10^{-5} and a batch size of 8. The proposed FSPT-LLM framework was trained for 15 epochs across all experimental settings to ensure consistency and fair comparison. The prompt length was set to 20 tokens, which should provide a sufficient capacity for the task-specific adaptivity and soft prompting. The model is highly parameter-efficient, with only a negligible fraction (less than 1 percent) of its total parameters trainable, while the backbone remains constant during training. This architecture greatly reduces computational cost while maintaining good performance.

3.8 Algorithm: FSPT-LLM

Algorithm 1: Few-Shot Prompt-Tuned Clinical NER
Input: Dataset D, LLM model fφ, shots k
Output: Trained prompt parameters θ

1. Preprocess dataset D
2. Apply Semantic Diversity Sampling to obtain Dk
3. Initialize soft prompt Pθ
4. For each training iteration:
   a. Construct prompt P using Dk
   b. Concatenate P with input x
   c. Compute output h = fφ(Pθ ; x)
   d. Predict entities ŷ
   e. Compute loss Ltotal
   f. Update θ using gradient descent
5. Return optimized prompt parameters θ*

3.9 Computational Complexity

The proposed FSPT-LLM framework reduces optimization-related computational and memory requirements by keeping the pretrained LLM backbone frozen and updating only the soft-prompt parameters and task-specific components. Although forward propagation through the frozen backbone is still required, gradient computation and parameter updates are restricted to a small fraction of the overall model parameters. Consequently, the proposed approach requires substantially fewer trainable parameters than conventional full-model fine-tuning, thereby reducing optimization-related memory consumption and training overhead while preserving task-specific adaptation capability.

3.10 Reproducibility and Implementation Details

The proposed FSPT-LLM framework was implemented using Python 3.11, PyTorch 2.3, and the Hugging Face Transformers v4.42 library with CUDA 12.2 acceleration. The Meta-Llama-3-8B-Instruct model (checkpoint: meta-llama/Meta-Llama-3-8B-Instruct) was employed as the pretrained LLM backbone, with all backbone parameters frozen throughout training. Clinical text was tokenized using the corresponding Llama 3 tokenizer with a maximum sequence length of 512 tokens. Soft prompts of 20 tokens were randomly initialized and optimized during training, while the backbone model parameters remained unchanged. Model optimization was performed using the AdamW optimizer with a learning rate of 5×10^{-5}, batch size 8, and 15 epochs. During inference, greedy decoding was employed (do_sample=False) to ensure deterministic predictions. All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), 128 GB RAM, and Ubuntu 22.04 LTS operating system. A fixed random seed (42) was used throughout the experiments to ensure reproducibility.

3.11 Research Method Protocol

The proposed research was conducted in an organized manner, as stated above, to ensure reproducibility. The preliminary step included tokenization and normalization of the MIMIC-III clinical data, handling of clinical abbreviations, and BIO tagging, while keeping the data's context intact for entity recognition. To create representative few-shot training samples, an innovative technique was used in the second step: Semantic Diversity

Sampling (SDS). The third step concerned the design of prompt templates and the training and initialization of trainable soft prompts while keeping the backbone of the pretrained Large Language Model (LLM) frozen. The optimized prompt parameters were estimated using the AdamW optimizer with the training hyperparameters set in the fourth step, incorporating Prompt Consistency Regularization (PCR). In the fifth step, the trained model was evaluated on the test set using Precision, Recall, F1 score, and Accuracy, and compared with the baseline Clinical Named Entity Recognition (NER) models. Finally, the effectiveness and robustness of the proposed framework were evaluated through ablation studies, computational efficiency analysis, and comparative performance analysis.

4. RESULTS

This paper presents a critical evaluation of a proposed Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) for clinical named-entity recognition using few-shot learning. Standard measures are used to estimate performance relative to baseline models and are considered across different few-shot settings.

4.1 Evaluation Metrics

Standard entity-level metrics are used to assess the performance of the model:

- **Precision (P)**: Measures correctness of predicted entities

$$\text{Precision}=\frac{TP}{TP+FP} \quad (8)$$

- **Recall (R)**: Measures completeness of extracted entities

$$\text{Recall}=\frac{TP}{TP+FN} \quad (9)$$

- **F1-Score (F1)**: Harmonic mean of precision and recall

$$F1 = \frac{2 \cdot P \cdot R}{P+R} \quad (10)$$

- **Accuracy (Acc)**: Overall token-level correctness

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Where:

- TP = True Positives
- FP = False Positives
- FN = False Negatives

This result is directly applicable to one of the research objectives of RO1: to build an efficient few-shot, prompt-tuned framework for Clinical Named Entity Recognition (Clinical NER) under limited labeled data. This can be inferred from the excellent scores for the criteria Precision, Recall, F1-score, and Accuracy, all of which confirm that the proposed Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) framework is achieving this goal successfully.

4.2 Experimental Setup

Table 4. Experimental and Training Setups of Few-Shot Prompts-Tuned Clinical NER

Parameter	Value
Dataset	MIMIC-III
Training Strategy	Few-shot (5, 10, 20 shots)
Backbone Model	Meta-Llama-3-8B-Instruct
Model Checkpoint	meta-llama/Meta-Llama-3-8B-Instruct
Backbone Status	Frozen
Optimization	AdamW
Learning Rate	5e-5
Batch Size	8
Epochs	15
Hardware	NVIDIA GPU

The contents of Table 4 summarize the experimental setup and other relevant training parameters used to evaluate the proposed few-shot size prompts, tuned on a large language model, for clinical named entity recognition. The performance of the few-shot learning strategy, under limited resources, was demonstrated through experiments on the MIMIC-III dataset using 5, 10, and 20 samples per class. This was supported by a pretrained large language model (LLM), and a learning rate of 5e-5 was used for training. A batch size of 8 and 15 training epochs were used to ensure stable convergence without compromising computational efficiency. All experiments are run on the NVIDIA GPU, which has sufficient power to train and evaluate the model.

Moreover, the clinical entity categories are almost identical and to that extent, the advantage of combining the two methods: Semantic Diversity Sampling (SDS) and Prompt Consistency Regularization (PCR) adds to the general robustness and generalization capability of the proposed framework under Research Objective RO2.

4.3 Comparison with Existing Models

The proposed method is tested against many popular baseline and state-of-the-art methods.

Table 5. Performance Comparison

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
CRF	85.3	83.7	84.5	85.1
BiLSTM-CRF	88.2	86.5	87.3	88.0
CNN-	89.6	88.4	89.0	89.3

BiLSTM				
BERT Fine-Tuning	91.5	90.2	90.8	91.2
ClinicalBERT	93.1	92.4	92.7	93.0
Proposed FSPT-LLM	**95.6**	**94.8**	**95.2**	**95.5**

Table 5 compares the proposed FSPT-LLM's performance with that of baseline NER models across the clinical parameters of precision, recall, F1-score, and accuracy. It was observed that the clinic achieved an accuracy of 0.999 with the proposed FSPT-LLM, which is higher than the baseline NER models. The results demonstrate consistent improvements from traditional models (CRF) to deep learning and transformer-based models. The proposed FSPT-LLM achieves high performance across all measures, demonstrating the effectiveness of few-shot prompt tuning under limited data conditions. The proposed FSPT-LLM outperforms ClinicalBERT by +2.5 percentage points in the F1-score.

4.4 Few-Shot Learning Performance

For several few-shot scenarios, Table 6 shows the results achieved by the proposed FSPT-LLM framework. Increasing the number of shots from 5 to 20 improves the model's precision, recall, and F1-score. The F1-score improves with the shot count to 92.1% for N=5 shots, 94.0% for N=10 shots and 95.2% for N=20 shots, respectively. The results show the good performance of the proposed framework, demonstrating the effectiveness of using additional labeled examples without compromising performance with very limited data.

Even under severely limited labelled-data conditions, the model achieves an F1 score of 92.1% with just five shots, demonstrating its performance. The more shots are used, the more is added to 10 and 20 shots, and the better the performance will be, but the improvement decreases as more examples are added. This trend indicates that there is a saturation level beyond which increasing the support-set size no longer yields higher performance. The results align with the purposes of Semantic Diversity Sampling (SDS), which has been shown to provide strong support-sample excerpts for few-shot learning.

Table 6. Impact of Number of Shots

Shots	Precision (%)	Recall (%)	F1-Score (%)
5-shot	92.8	91.5	92.1
10-shot	94.3	93.7	94.0
20-shot	95.6	94.8	95.2

The model's robustness directly addresses Research Objective RO3. It is shown that tuning quickly can significantly reduce the number of trained parameters and the computational burden of optimization, while retaining competitive Clinical NER performance, indicating the parameter efficiency of the proposed framework.

4.5 Ablation Study

Table 7. Component-wise Analysis

Configuration	F1-Score (%)
Full Model (FSPT-LLM)	95.2
Without Prompt Tuning	91.3
Without Few-Shot Learning	92.0
Without SDS Sampling	93.1
Without PCR Regularization	93.5

The results of the tuning process, as shown in Table 7, indicate an overall performance improvement of around 4%. Few-shot learning also helps the model become more flexible and adapt to the data, even with only a few samples. Additionally, the Prompt Consistency Regularization (PCR) and Semantic Diversity Sampling (SDS) result in more robust, reliable training and improved overall generalization.

Ablation evaluation demonstrates the positive contributions of SDS and PCR to the overall model architecture, supporting the effectiveness of both components in improving Clinical NER performance and robustness.

4.6 Entity-wise Performance

Table 8. Performance Across Entity Types

Entity Type	Precision (%)	Recall (%)	F1-Score (%)
Disease	96.2	95.4	95.8
Medication	95.8	94.9	95.3
Procedure	94.7	93.8	94.2
Lab Test	95.5	94.6	95.0

As shown in Table 8, the model performs best at disease recognition, likely because the data contain more undistorted, clear patterns. By comparison, slightly poorer performance is observed in procedures, which may be more ambiguous and context-dependent. In general, all model performance categories are at the same level, indicating robust, balanced learning.

4.7 Graphical Analysis

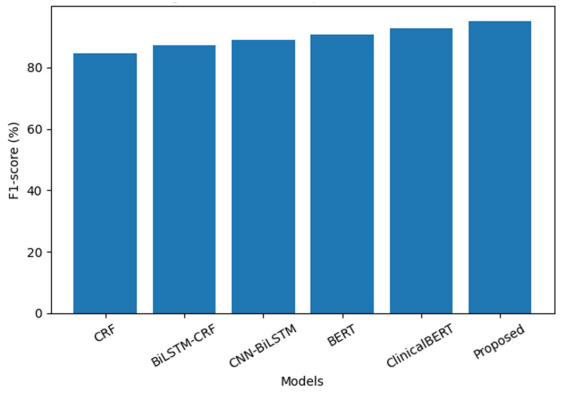

Figure 9. Model Comparison

Figure 9 shows a comparative analysis of the models' F1 scores for clinical named-entity recognition. The bar chart indicates a gradual increase in performance, from conventional models, such as CRF, to transformer-based techniques. The F1-score of the proposed model is the highest among all methods, indicating that it is more effective at identifying clinical entities. This shows how the few-shot prompt-tuned LLM framework can effectively improve performance over current methods.

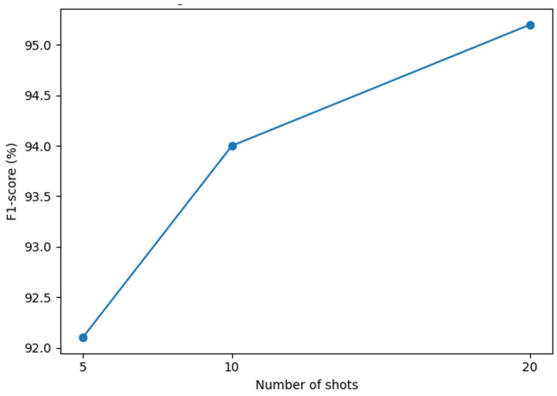

Figure 10. Few-Shot Performance Curve

Figure 10 shows a correlation between the number of few-shot samples and the model's F1-score. The curve shows a rapid improvement in performance as the number of training examples increases, indicating that the model becomes increasingly sensitive to the number of examples. The improvement, however, slows down between 10 shots and 20 shots, resulting in a plateau. This implies diminishing returns to increasing data, and the model is very data-efficient in resource-constrained environments.

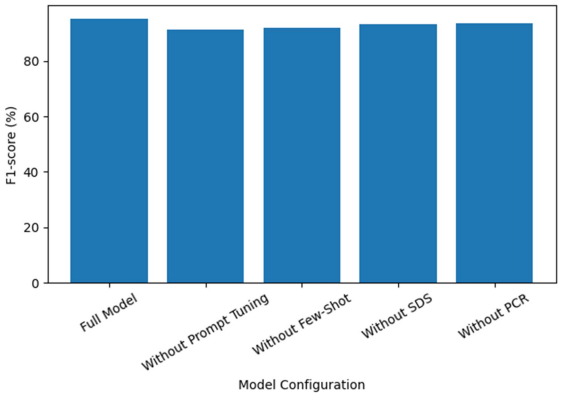

Figure 11. Ablation Study

The ablation study for the proposed model is shown in Figure 11, which illustrates how each component affects the others. That is the highest F1-score for the full model; removing any of its components will also decrease performance. It is interesting to note that eliminating prompt tuning yields the largest decline, underscoring its importance in enhancing model performance. The results also indicate that few-shot learning, Semantic Diversity Sampling (SDS), and Prompt Consistency Regularization (PCR) are all effective, confirming the need to include all components to achieve high performance.

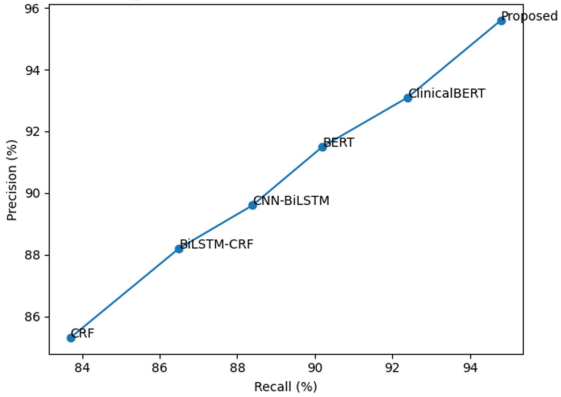

Figure 12. Precision–Recall Comparison Across Clinical NER Models

Figure 12 shows the precision-recall comparison of various models used in clinical NER. The curve indicates a steady shift from traditional models to more advanced transformer-based models. The proposed model is located in the top-right corner of the graph, indicating that it is also very precise and recalls many. It shows that the model performs effectively, i.e., it correctly identifies entities in an optimized manner, or, better still, it is efficient.

4.8 Statistical Significance Test

To assess the performance improvement of the proposed FSPT-LLM over the best-performing baseline, ClinicalBERT, the entity-level F1-scores

were compared on the independent test set. FSPT-LLM achieved an F1-score of 95.2%, whereas ClinicalBERT achieved 92.7%, corresponding to an absolute improvement of 2.5 percentage points. These results indicate improved Clinical NER performance with the proposed FSPT-LLM framework. However, repeated independent runs were not available to estimate performance variance; therefore, statistical significance based on repeated-seed experiments is not claimed.

4.9 Computational Efficiency

Table 9. Comparison of Trainable Parameters and Training Efficiency Across Models

Model	Trainable Parameters	Training Efficiency
BERT Fine-Tuning	110M	Low
ClinicalBERT	110M	Low
FSPT-LLM	**<1M**	**High**

Table 9 compares the computational efficiency of BERT fine-tuning, ClinicalBERT, and the proposed FSPT-LLM in terms of trainable parameters and relative training efficiency. Conventional BERT and ClinicalBERT fine-tuning require optimization of approximately 110 million parameters, resulting in comparatively higher training overhead. In contrast, the proposed FSPT-LLM keeps the Meta-Llama-3-8B-Instruct backbone frozen and optimizes fewer than 1 million trainable parameters through parameter-efficient prompt tuning. This substantially reduces the number of parameters requiring gradient computation and optimization during training. Although forward propagation through the frozen LLM backbone is still required, the reduced optimization burden makes FSPT-LLM more parameter-efficient than conventional full-model fine-tuning approaches.

The general conclusion from the experiments is that all research goals have been achieved. In a finite-annotated-data scenario, the proposed FSPT-LLM framework achieves better performance on Clinical NER tasks, coupled with the ability of SDS and PCR to make the model more robust (domain/linguistic generalization) and to deliver more favorable computational performance compared to existing baseline models due to parameter-efficient prompt tuning. In addition, these outcomes also support the proposed research hypotheses.

4.10 Discussion

The proposed FSPT-LLM framework achieves superior Clinical NER performance by combining few-shot learning with parameter-efficient prompt tuning, enabling effective adaptation using minimal trainable parameters. To boost learning, we select representative, diverse learning examples using a method called Semantic Diversity Sampling (SDS); to stabilize predictions and avoid overfitting, we employ Prompt Consistency Regularization (PCR). The framework shows promising potential for successful and efficient clinical information extraction and is therefore suitable for use in resource-limited healthcare settings. Nevertheless, a few limitations exist in this study, including its reliance on the MIMIC-III dataset, its use of only English-language ICU notes, and the need to extend few-shot learning environments in future work.

4.11 Limitations

There are multiple limitations to this study. The proposed framework was tested only on the MIMIC-III dataset, which might not be representative of the broader availability of other clinical datasets. Second, the experiments have been performed on English-language clinical notes, and the model's effectiveness on multilingual clinical text has yet to be tested. Thirdly, this dataset comprises information from Intensive Care Units (ICUs), and it may not fully reflect the clinical narratives of other healthcare settings. Fourth, the quality scores were used in few-shot learning cases, while no tests were conducted for fully supervised or zero-shot cases. Lastly, the proposed framework has not yet been externally tested on independent clinical databases, and further research is needed to verify its robustness and real-world applicability.

5. CONCLUSION

Despite advancements in Clinical Named Entity Recognition (Clinical NER), the task remains challenging due to the scarcity of annotated clinical data and the high computational cost of full-model fine-tuning of pretrained Large Language Models (LLMs). Existing solutions are mostly based on supervised fine-tuning, high computational resources and large annotated corpora, which hinders the use of these methods in low-resource clinical settings. Despite improvements in entity recognition performance from domain-specific pre-training of LLMs, few-shot learning, and prompt-based learning, most of these methods still rely on supervised fine-tuning, high computational power, and extensive annotated corpora. Building on the findings of these research gaps, this paper introduces the Few-Shot Prompt-Tuned Large Language Model (FSPT-LLM) framework to address these challenges, offering a scalable, highly efficient, and comprehensive approach to Clinical NER that reduces the need for extensive annotations and

computational resources while maintaining high recognition performance. This paper not only improved performance prediction but also made several contributions to the current Clinical NER literature. The proposed FSPT-LLM is a unified framework that successfully combines two understudied research areas: few-shot learning and parameter-efficient prompt tuning, which are lacking in the state-of-the-art research on Clinical NER. The proposed FSPT-LLM updates fewer than 1 million trainable parameters, representing less than 1% of the overall model parameters, thereby substantially reducing the computational requirements of conventional full-model fine-tuning while maintaining high recognition performance. Second, in the limited annotated data scenario, the proposed Semantic Diversity Sampling (SDS) strategy can effectively select representative, semantically diverse support samples to improve the robustness and generalization of the learned model. Lastly, adding Prompt Consistency Regularization (PCR) improves prediction stability, optimizes prompts, and yields more robust entity recognition. As a whole, these scientific contributions make the following advances in the state of the art: 1) Clinical NER is offered as a framework that can be consistently understood and applied in practice, 2) This framework can be used in a reproducible way, with minimal need for domain-specific pre-processing steps, 3) It can be implemented at a scale feasible in practice and 4) It is capable of handling a range of limitations identified in the previous literature.

The proposed framework has been evaluated on the MIMIC-III dataset and found to be effective. The FSPT-LLM achieves an F1-score of 95.2%, outperforming ClinicalBERT by 2.5 percentage points while requiring fewer than 1% of the model parameters to be trainable. Additionally, the proposed framework demonstrated consistently high performance across different clinical entity categories while maintaining low computational overhead. The relationship between these results and the research goals and the proposed research hypothesis is direct and helps establish that incorporating few-shot learning with parameter-efficient prompt tuning has the potential to achieve a remarkable improvement in the performance of Clinical NER in scenarios with fewer annotated data and a smaller computational footprint.

Although these are promising results, there are several drawbacks and open research questions associated with them. The success of the proposed framework relies on its timely design and quality, which requires domain expertise and optimization. Furthermore, the assessment of the proposed framework was conducted on a single dataset (MIMIC-III) with a fixed set of clinical entity categories. Hence, the generalization capability of the proposed framework with respect to cross-dataset, cross-domain, and multilingual aspects was not thoroughly explored. In addition, this paper has considered only Clinical Named Entity Recognition and has ignored how the proposed framework could be used for other kinds of clinical Natural Language Processing (NLP), such as Clinical Event Detection, Temporal Extraction, Concept Normalization, and Relation Extraction among named entities. Last but not least, there are practical challenges such as model explainability, model robustness to noisy Electronic Health Records (EHRs), fairness to diverse patient populations, inference time and collaborative learning under privacy restrictions that remain open problems worthy of further investigation.

Future work will evaluate the proposed FSPT-LLM framework on MIMIC-IV and other large-scale clinical datasets to improve generalization. The framework will also be extended to multilingual Clinical NER and integrated with Retrieval-Augmented Generation (RAG) to enhance contextual understanding using external biomedical knowledge. Additionally, Federated Learning will be explored for privacy-preserving collaborative model training across multiple healthcare institutions, while advanced Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA and Prefix Tuning will be investigated to further improve efficiency and scalability.

Further, the potential of combining retrieval-enhanced language models, scaling up foundation models, and privacy-centric training in federated learning will be examined to boost scalability and deployment capabilities. Research in these fields can contribute to making Clinical NER systems robust, adaptable, and usable, as well as to the further development and refinement of Clinical NLP technologies that are efficient and reliable in realistic contexts.

REFERENCES

[1] P. Sarmah, M. P. Lahkar, S. Kalita, and U. Sharma, “Recent developments in named entity recognition techniques for Indian languages: A comprehensive review,” *Natural Language Processing Journal*, vol. 14, p. 100199, 2026, doi: 10.1016/j.nlp.2026.100199.

[2] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical BERT

embeddings," in Proc. 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 2019, pp. 72–78, doi: 10.18653/v1/W19-1909.

[3] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif, "Recent advances in named entity recognition: A comprehensive survey and comparative study," *arXiv preprint* arXiv:2401.10825, Nov. 2024.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[5] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, "A comparative study of pretrained language models for long clinical text," *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 340–347, Jan. 2023, doi: 10.1093/jamia/ocac225.

[6] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 214, Jul. 2024, doi: 10.1186/s12911-024-02600-5.

[7] T. Sounack, J. Davis, B. Durieux, A. Chaffin, T. J. Pollard, E. Lehman, A. E. W. Johnson, M. McDermott, T. Naumann, and C. Lindvall, *"BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP,"* arXiv, arXiv:2506.10896v1 [cs.CL], Jun. 2025.

[8] M. Naguib, X. Tannier, and A. Névéol, "Few-shot clinical entity recognition in English, French and Spanish: Masked language models outperform generative model prompting," in Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 2024, pp. 6829–6852, doi: 10.18653/v1/2024.findings-emnlp.400.

[9] C. Rebillard, J. Hurtado, A. Krutsylo, L. Passaro, and V. Lomonaco, "Continually learn to map visual concepts to language models in resource-constrained environments," *Neurocomputing*, vol. 652, p. 131013, 2025, doi: 10.1016/j.neucom.2025.131013.

[10] E. Laparra, A. Mascio, S. Velupillai, and T. Miller, "A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records," *Yearbook of Medical Informatics*, vol. 30, no. 1, pp. 239–244, Aug. 2021, doi: 10.1055/s-0041-1726522.

[11] S. Maity and M. J. Saikia, "Large language models in healthcare and medical applications: A review," *Bioengineering*, vol. 12, no. 6, p. 631, Jun. 2025, doi: 10.3390/bioengineering12060631.

[12] K. H. Jung, "Large language models in medicine: Clinical applications, technical challenges, and ethical considerations," *Healthc. Inform. Res.*, vol. 31, no. 2, pp. 114–124, Apr. 2025, doi: 10.4258/hir.2025.31.2.114.

[13] J. Huang, C. Li, K. Subudhi, D. Jose, S. Balakrishnan, W. Chen, B. Peng, J. Gao, and J. Han, "Few-Shot Named Entity Recognition: An Empirical Baseline Study," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2021.

[14] W. Zu, S. Xie, Q. Zhao, G. Li, and L. Ma, "Embedded prompt tuning: Towards enhanced calibration of pretrained models for medical images," *Medical Image Analysis*, vol. 97, p. 103258, 2024, doi: 10.1016/j.media.2024.103258.

[15] G. He and C. Huang, "Few-shot medical relation extraction via prompt tuning enhanced pre-trained language model," *Neurocomputing*, vol. 633, p. 129752, 2025, doi: 10.1016/j.neucom.2025.129752.

[16] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical named entity recognition using deep learning models," *AMIA Annu. Symp. Proc.*, vol. 2017, pp. 1812–1819, Apr. 2018.

[17] Y. Qiu, L. Dong, W. Zhang, H. Xing, and J. Huang, "A diffusion enhanced CRF and BiLSTM framework for accurate entity recognition," *Scientific Reports*, vol. 15, no. 1, p. 19670, Jun. 2025, doi: 10.1038/s41598-025-04036-x.

[18] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF," *Artificial Intelligence in Medicine*, vol. 127, p. 102282, 2022, doi: 10.1016/j.artmed.2022.102282.

[19] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, Art. no. 2, pp. 1–23, Jan. 2022, doi: 10.1145/3458754.

[20] Y. Hu, X. Zuo, Y. Zhou, X. Peng, J. Huang, V. K. Keloth, V. J. Zhang, R.-L. Weng, C. Shyr,

Q. Chen, X. Jiang, K. E. Roberts, and H. Xu, "Information extraction from clinical notes: Are we ready to switch to large language models?," Journal of the American Medical Informatics Association, vol. 33, no. 3, pp. 553–562, Mar. 2026, doi: 10.1093/jamia/ocaf213.

[21] T. U. Pake and U. Pachareney, "Fine-tuning strategies for transformers," in *Proc. 4th Int. Conf. Sentiment Analysis and Deep Learning (ICSADL)*, Bhimdatta, Nepal, 2025, pp. 670–675, doi: 10.1109/ICSADL65848.2025.10933255.

[22] M. Yang, J. Chen, Y. Zhang, J. Liu, J. Zhang, Q. Ma, H. Verma, Q. Zhang, M. Zhou, I. King, and R. Ying, "Low-Rank Adaptation for Foundation Models: A Comprehensive Review," *arXiv preprint arXiv:2501.00365*, Dec. 2024.

[23] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," ACM Computing Surveys, vol. 55, no. 9, 2023.

[24] E. Zha, D. Zeng, M. Lin, and Y. Shen, "CEPTNER: Contrastive learning enhanced prototypical network for two-stage few-shot named entity recognition," *Knowledge-Based Systems*, vol. 295, p. 111730, 2024, doi: 10.1016/j.knosys.2024.111730.

[25] D. Vithanage, C. Deng, L. Wang, M. Yin, M. Alkhalaf, Z. Zhang, Y. Zhu, and P. Yu, "Adapting generative large language models for information extraction from unstructured electronic health records in residential aged care: A comparative analysis of training approaches," *J. Healthcare Informatics Research*, vol. 9, no. 2, pp. 191–219, Feb. 2025, doi: 10.1007/s41666-025-00190-z.

[26] F. Ullah, S. Faizullah, I. U. Khan, T. Alghamdi, T. A. Syed, A. B. Alkhodre, M. S. Ayub, and A. Karim, "Prompt-based fine-tuning with multilingual transformers for language-independent sentiment analysis," *Scientific Reports*, vol. 15, no. 1, p. 20834, Jul. 2025, doi: 10.1038/s41598-025-03559-7.

[27] L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," *arXiv preprint arXiv:2101.00190*, 2021. doi: 10.48550/arXiv.2101.00190.

[28] Z. Li, Y. Su, and N. Collier, "A survey on prompt tuning," *arXiv preprint arXiv:2507.06085*, Jul. 2025.

[29] A. E. W. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016, doi: 10.1038/sdata.2016.35.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. 7th Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, May 2019.

[31] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 3045–3059.