Swansea University Medical School
Ysgol Feddygaeth Prifysgol Abertawe

Swansea University
Prifysgol Abertawe

Assignment for PMIM-702 Dissertation

Module number:	PMIM-702
Module name:	Dissertation
Title of assignment:	*Literature Review*
Student ID number:	2311233
Word count:	5000
Declaration:	I understand the following conditions which apply throughout this course: 1. I confirm that I am the sole author of this work. 2. I understand that proof reading by a third party is discouraged, but if used, records should be available as per guidelines. 3. I understand the need for academic integrity and that all my submitted work will adhere to its principles. 4. I understand that the teaching team will take measures to deter, detect and report any academic misconduct. 5. I agree to my work being submitted to the Turnitin academic database. 6. I understand the importance of assignment deadlines and the need to seek help in good time where personal circumstances interrupt my work.

Please copy and paste this declaration onto the front of the submission.

Literature review: “Assessment of fine-tuned open-source large language models on gold standard synthetic epilepsy clinical notes with focus on extracting Seizure Frequency and Prescriptions”

ABSTRACT

Background

Seizures affects global population that results in epilepsy. There has been extensive study to understand seizure freedom, improve clinician-patient engagement, effectiveness of specific prescription drugs in controlling seizure but with lack of data to actualize it. The clinical notes a documentation managed by the clinician contains the seizure frequency and prescriptions details. Extraction of seizure frequency and prescriptions an application of clinical NER using NLP in the health care domain could be a possible solution. NLP has progressed and reached a new level with advent of Large Language models (LLMs). It needs to be experimented if LLMs could outperform other implementations to extract the desired entities.

Method

A systematic review with thematic analysis with the results aimed to extract themes from the previous similar studies was the methodology in practice. The literature search was conducted on 4 search tools, they are INSPEC, IEEE, PubMed, and Google Scholar. A consult diagram is created to represent the inclusion-exclusion criteria, organization, and formulation of search terms, and to guide and document the count of total number of papers found, selected and chain cited.

Results

From the papers selected in the literature review, themes were extracted to understand the prominent methodologies that were used in NLP to perform clinical NER on the clinical notes. There are 6 unique identified themes they are Knowledge graph, Machine Learning, Rule based approach, Statistics, Deep Learning, and Prompt Tuning in Language models. Further themes could be identified as combination of the above mentioned 6 themes. The identified themes could be segregated to pre-LLM era and LLM era to indicate the exponential progress attained due to the introduction of language models.

Conclusion

Vast studies have discussed and innovated on clinical NER to extract concepts for example diagnosis, prescriptions, patient history. Extracting seizure frequency from epilepsy clinical notes is a niche area. The review of the studies provided a comprehensive idea about the diverse methods used for application of clinical NER. However, extraction of seizure frequency and prescription with large language models is a potential area of research that could pave the way for research and development. Presently, clinical NER with LLMs is at a native stage and is a promising tool to achieve milestones and become state of the art solutions for the coming decades.

INTRODUCTION

Epilepsy

Seizures are known to affect 10% of the global population out of which 1-2% result in epilepsy (Falco-Walter, 2020). Epilepsy was defined in 2005 by the International League Against Epilepsy (ILAE). As per the operational definition, there are three conditions that needs to be met in-order to correctly identify an individual suffering from epilepsy. The first criteria states that an individual who would have experienced at least two unprovoked (or reflex) seizures more than 24 hours apart. The second criteria states that an individual who would have experienced or has one reflex seizure in addition to the probability of seizure recurrence risk greater than 60% after the first condition have experienced. The second condition could occur over the span of 10 years after the first criteria is satisfied. The third criteria states that an individual who satisfies first two criteria and should have a confirmed diagnosis of epilepsy syndrome (Fisher et al., 2014). Epilepsy is a common medical condition that affects people irrespective of age, race, geographical locations, and social class (Asadi-Pooya et al., 2023). Studies suggest that in the United Kingdom, the point prevalence of epilepsy is 9.37 per 1000 persons; overall reported estimated incidence rate is 42.68 per 100, 000 person-years. England reported the lowest estimated incidence rate of 37.41 per 100, 000 person years and prevalence of 8.85 per 1000 persons while Wales reported the highest estimated incidence rate of 54.84 per 100, 000 person years and prevalence of 11.40 per 1000 persons amongst the four countries in the United Kingdom (Wigglesworth et al., 2023). Epilepsy is not a single disease entity. Thus, it needs to be classified into types (Thijs et al., 2019). The ILAE provides three level classification system for classification and specification of epilepsy (Scheffer et al., 2017). During patient visits and care, the information about the condition is recorded by the doctors in the clinical notes. Outpatient clinical notes are tools to document the patient condition and enhance care to the patient (Wood, 2001). Thus, it is important that the quality of the clinical notes is maintained (Hanson et al., 2012; Soto et al., 2002).

Seizure frequency & Prescriptions

Seizure frequency is associated to number of times a patient experienced seizure over a period in the patient history. It is used to identify the severity of epilepsy (Viteva, 2014), adjust the dosage of the medication (Choi et al., 2014) and derive seizure freedom that associate it to health-related quality of life (Baker et al., 1998; Leidy et al., 1999). There are studies that indicate that seizure frequency have correlation with suicidal tendencies and depression (Liu et al., 2020). Due to this importance of seizure frequency, researchers have focused on extracting it from clinical notes to improve the clinical decision-making, and pave the way for large scale retrospective research and improve health-related quality of life for patients with epilepsy. Despite the importance of Seizure frequency, it is rarely audited and poorly documented aspect of the clinical documentation (Decker et al., 2022; Fonferko-Shadrach et al., 2019; Xie et al., 2022, 2023).

Natural Language Processing

Natural Language Processing (NLP) is a field in computer science that is formed by the intersection of artificial intelligence and linguistics domains. NLP typically tries to solve human language related problems like translation, context understanding, entity or label extraction etc that could be used to derive insights or predict information on text-based documents (Nadkarni et al., 2011; Ramsay & Ramsay, 1987; Schank, 1986). Applications of NLP includes interaction between knowledge and user using the implementation such as chatbots (Jain et al., 2024), clinical text mining, an active field of research that uses text documents to develop solutions to enhance patient care (Katyan et al., 2024), aid in clinical decision support, and its interventions to standardize formats, represent clinical knowledge, and leverage clinical narrative to predict estimations. Radiology reports, biomedical reports, outpatient clinical notes are few examples of input data sources that could be used for NLP systems (Demner-Fushman et al., 2009).

Transformers and Large Language models

With the introduction of transformers architecture and the attention mechanism, a revolution came about in the field of machine learning especially in NLP. NLP applications with transformers outperformed all the existing models and implementation (Vaswani et al., 2017). Availability of large data, advent of GPU and compute power and the rise of the transformers have led to Foundation Models (FM). They incorporate self-supervised techniques along with attention mechanism and have demonstrated extra-ordinary proficiency in addressing challenging tasks in NLP (Dean, 2022; Myers et al., 2024).

Clinical NER with LLMs

Information Extraction (IE) is a subfield of NLP where the primary tasks comprise of NER, Entity-Relation (ER) extraction and template completion. The task could be solved using classical pattern matching, heuristics or machine learning based approaches like Naïve Bayes, Support Vector Machines or ensemble machine learning models like Random Forest, Gradient Boosting or deep learning models, transformers and the most recent LLMs (Reichenpfader et al., 2023, Nasar et al., 2022). NER task is performed at word level. A word in NLP could be represented as a token. The NER is to then understand the semantic and syntactic importance of the word with respect to the sentence and the context, in this case a clinical letter. Then the model learns to classify the token as an entity with a probability score.

Medication information comprised of a patient's medication status is one of the most crucial pieces of information in the EHR and is critical for patient's healthcare safety and quality (Wang et al., 2015). In the field of clinical IE, extracting prescriptions from clinical notes is a well-founded and active research area since the introduction of EHR to boost adaptation of automated systems and to improve patient phenotyping performance (Vulpius et al., 2023; Wang et al., 2018; W.-Q. Wei et al., 2016).

Clinical NER is widely considered to be a hard problem to solve. Recent technological advancements in NLP and LLMs have increased the performance and accuracy of the systems to extract information from the clinical notes (Hu et al., 2024). LLMs in biomedicine application have demonstrated 15-20% improvement in NER task (Monajatipoor et al., 2024). Thus, LLMs proves to be a promising tool for NER task on clinical notes.

Few-shot clinical NER with LLMs

Few-shot learning is a data efficient approach where techniques and methods are implemented to compensate the lack of data while providing standard and sometimes state of the art results from the model. (Naguib et al., 2024) discuss the implementation of few-shot learning approach that address the issues. The few-shot learning approaches are applied to two types of LLM models, namely the Masked Language Models (MLM) and the Casual Language Model (CLM). MLM is a type of model that uses one of the robust NER systems, where, the model is pre-trained to predict to select random masked words from a large corpora using dense vector representations (Devlin et al., 2019; Peters et al., 2018) of every token in the text. The training involves linear projection to map vector representations to the corresponding NER labels of the sentence, by fine-tuning the parameters of the model preparing for the downstream NER task. CLM are large models trained on the corpora as generative and autoregressive models. The input to the model is a series of tokens or prompt as input, and the model estimates the most probable following a series of tokens. This type of model relies heavily on the formulation of prompt structure and any change in the phrasing results to unexpected results. The process of model training from the prompts is termed as In Context Learning (ICL) (Brown et al., 2020).

Prompt Tuning

Prompt tuning is one of the two primary strategies to customize a pre-trained LLM on a specific task. Prompt tuning approach trains a model using the concept of ICL by customizing the prompts without changing the internal parameters of the model (Gu et al., 2022; Liu et al., 2022). The prompts and phrases are designed as per the target task and use the existing knowledge encoded in the model to obtain desired responses. In prompt

tuning, the model parameters are not adjusted, relatively small dataset is required, computational cost is low, and the performance could be competitive, especially when compared with few-shot learning. However, the prompt tuned model is not persistent, lacks flexibility and correctness in terms of specialized knowledge as the result is derived from pre-trained knowledge.

Fine Tuning

Fine tuning is one of the two primary strategies to customize a pre-trained LLMs on a specific task. (J. Wei et al., 2022). Fine tuning involves adjusting the model's internal parameters (weights) on a new dataset specific to the target task. The process is like retraining a deep learning model with the task-specific dataset, that allows the model to adjust parameters, learn new patterns and representations. In fine tuning, large and task-specific dataset is required depending on the number of parameters to update, that is the size of the model. Typically, large models require large amount of data to effectively provide best performance. But zero-shot and few-shot techniques have nullified this requirement to some extent. The computational cost for fine tuning is higher. The model is flexible and performs often better than prompt tuning approach.

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is a technique, that combines the strengths of large language models and information retrieval. It is one of the technique employed to control hallucinations (Reddy et al., 2024) in LLMs and provide more customized response using a separate knowledge base as an extension. RAG is a system where when the user provides a query, the relevant information is retrieved from the respective knowledge base. This retrieved information is then combined with the user query to form an augmented prompt, thus the term augmentation. The augmented prompt is then provided as a input to the LLM which generates text with respect to the additional context provided from the retrieved text. (Fan et al., 2024; Gao et al., 2024; Xiong et al., 2024) states that RAG improved accuracy as the LLMs responses is checked against information to reduce hallucinations. It also enables the LLMs to access and incorporate latest information ensuring that the responses are relevant in the context of time. RAG can provide evidences and references for the LLM responses generated text by citing the retrieved sources, thereby increasing the reliability of the model's output.

METHOD

The search strategy for reviewing the literature was carried out using four search tools. The search tools comprised of INSPEC, IEEE, PubMed databases and Google Scholar. The search was organized based on topics and sub-topics of interest. A hierarchical organization consult diagram was designed to represent the search strategy, organize the search as per topics, and document results of the search, that is the total number of results returned, number of papers selected and the number of papers considered for chain citation. The search strategy was carried out in three steps.

First, the search terms were identified. According to the research topic, the root search term was split into two, specifically 'NLP' and 'Epilepsy.' The sub-topics were considered as the children search terms of the root search term. The child search term was identified based on the contextual relevance of the topics and literature thus each of the search term were contextually added in a forward step addition approach during the literature search.

Second, the keywords were identified and compiled to form a query. Each term was conjoined to form a query using boolean conjunctions 'AND' and 'OR.' For example, for the keywords, 'large language models' and 'epilepsy,' the query is formed as "(abstract (large language models) AND abstract (epilepsy))." The query syntax that is used for searching the keywords changed as per the search tool. The search space for the keywords in the query were set to 'abstract' so that the query execution is faster in the respective search tools.

Third, the results obtained from the search keywords where collated from the respective search tools. Only the top 10 results were selected from the results from each search tool. The selected results from each search tool when reviewed were first sorted as per relevance, best match, and date. The relevance was given highest importance and the relevance of each research were understood after reading the abstract. During collation of the results, papers that were common across the search tools were considered as duplicates and removed after including only one unique result. Once the collated list was formed, the list was filtered to relevant list of papers by considering the full text of the research.

The literature was also searched using the synonyms of the keywords in the relevant search queries. For example, the term 'Named Entity Recognition' could also be known and widely accepted as 'Clinical Concept Extraction' or 'Clinical Named Entity Recognition.' This was to ensure robust search strategy is implemented for the literature search process. The literature was searched from the period of 1st June, 2024 to 1st August 2024. Any research paper added after the date is not added in the scope or the narrative of this literature review. There was no date filter applied for selection of the research paper.

Fig 1 represents the consult diagram to explain the analysis, filter strategy, organisation, and hierarchical connection between topic and its sub topics. The atomization of the topic helped in realizing the gap and the overlap between the topics. The sub topics were also considered as a standalone key term and derivations of the parent topic. The derivations were put together in combination to later form into queries. For example, LLMs are further split into the derived keywords as per relevance such as 'Transformers,' 'Foundation models,' 'LLMs in health.' A further split in 'LLMs in health' was made to form 'NER with LLMs.' Now, the search query was constructed considering the parent context for the final node as (abstract(health) AND abstract (large language models)) OR (abstract (named entity recognition) AND abstract (large language models)). While each of the search tools had advanced search option, the search query input method was a little bit different for each search tool.

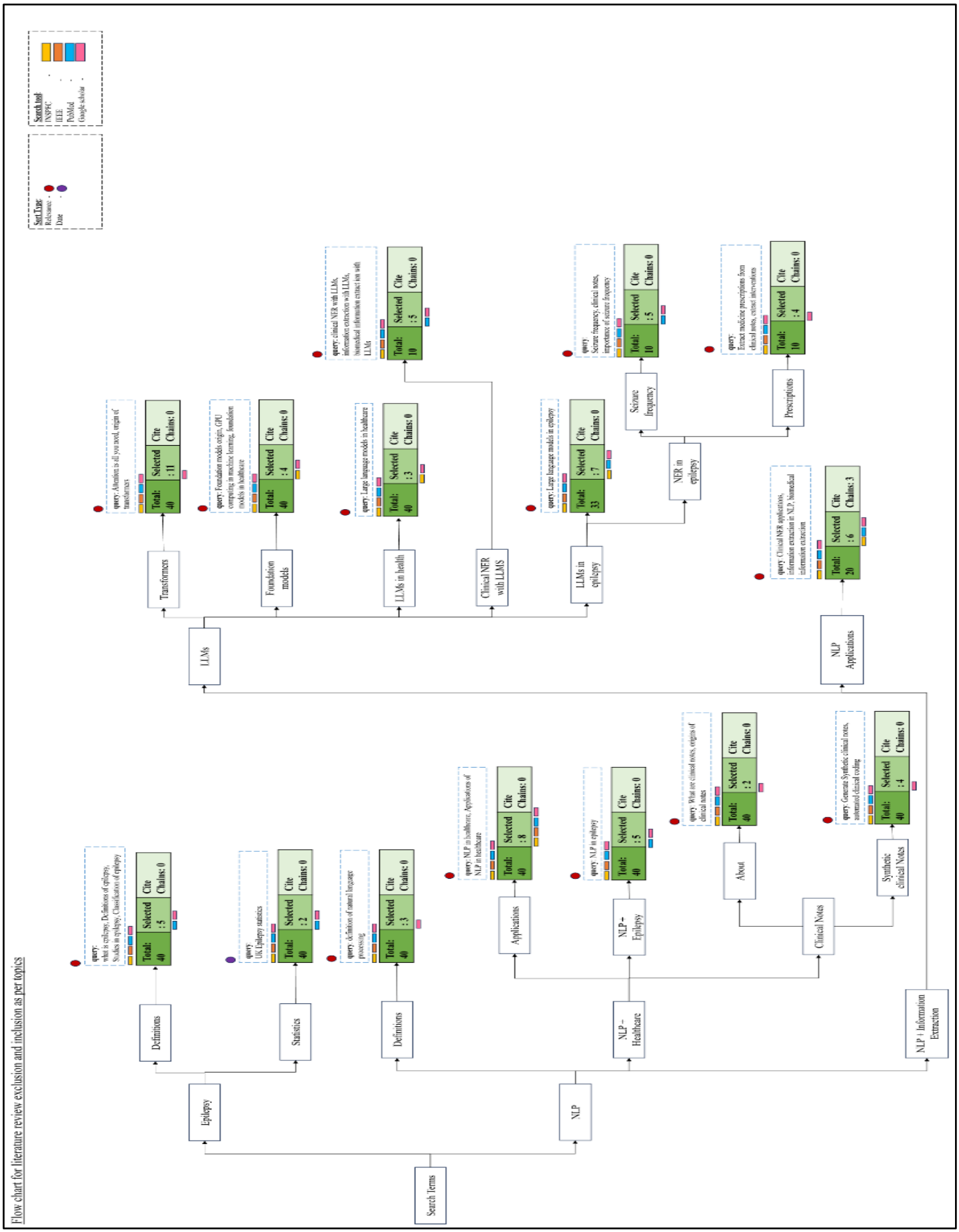

Fig 1: Organization Chart for paper research

The set of keywords used in the literature search are 'ICD clinical coding', 'epilepsy', 'natural language processing', 'large language models', 'named entity recognition', 'information retrieval', 'synthetic clinical notes', 'clinical notes overview', 'named entity recognition overview', 'large language models overview', 'epilepsy seizure frequency extraction', 'epilepsy prescriptions extraction', 'epilpesy interventions', 'prescription extraction' and the combination of the individual terms.

RESULT

From the review of the research, themes were identified. Fig 2 shows the overview of the identified implementation themes as solutions for clinical NER.

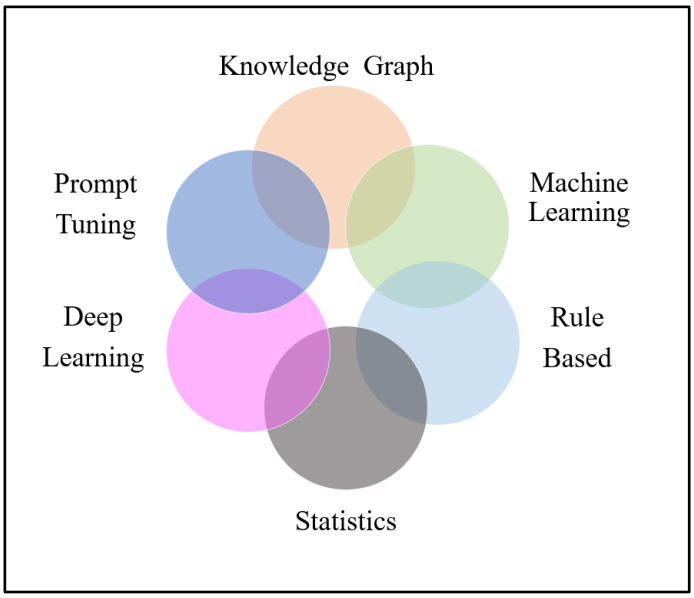

Fig 2: Implementation themes for clinical NER

These themes could be segregated into two different era on a timeline. The first era is the pre-LLMs era and the second is the LLMs era.

Before a deep dive into implementation of themes it is important to understand the nature of the data that are used as input to the systems. An effective approach to represent the clinical notes are important when the notes are the resourced to develop NLP systems. In NLP, the machine learning models could be used to perform prognosis or diagnosis given the history of the patient. The quality of the machine learning model is directly proportional to the quality of the data (Dubois et al., 2018). Thus, high quality clinical notes are required for such projects. As clinical notes contain personal information, the access to it is restricted to limited researchers. For such situation where sensitive data could only be used in siloed research, synthetic clinical notes are preferred alternatives where the clinical notes are generated using advanced generative Artificial Intelligence (AI) algorithms (Melamud & Shivade, 2019; Zhou et al., 2022). The quality of the synthetically generated notes could be improved through post generation screening and annotation process which involves experts from multi-disciplinary domains or leveraging latest generative models for validation purposes (Biswas & Talukdar, 2024). The output of such validated datasets is termed as gold standard data. (Fonferko-Shadrach et al., 2023) gives one such example of gold standard dataset containing 200 synthetic clinical notes. To perform supervised machine learning with NLP, large dataset with relevant labels are required (Schank, 1986). However, manual dataset annotation is identified as one of the bottleneck and the toughest challenge in the machine learning field due to its cost factor (Spasic & Nenadic, 2020). Thus, automated clinical coding using synthetic clinical notes is one such application of NLP where clinical medical term also referred as entities or medical concepts are recognized, tagged, and extracted (Fonferko-Shadrach et al., 2023). The clinical concepts could be associated to Unified Medical Language Systems (UMLS). The purpose of UMLS is to integrate the terminologies into a single paradigm that could be used across different contexts and thereby remove disparity amongst departments in the healthcare industry. UMLS is maintained as per the region with frequent updates (Campbell et al., 1998).

Pre-LLMS Era

In the pre-LLMs era, the clinical NER tasks were solved by creating customized framework using combination of rule-based systems, classical machine learning and deep learning-based techniques and early implementation of transformers (see Table 1.).

Models By	Architecture	Detail	Performance	Scope
(Xie et al., 2022)	Masked Language Modelling (MLM) + BERT	Pretrained Deep Bi-Directional Transformers	Accuracy: 80% + F1 Score: 80% +	Epilepsy
(Xie et al., 2022)	MLM + BioClinicalBert	BERT + pretraining on clinical text	Accuracy: 80% + F1 Score: 80% +	Epilepsy
(Xie et al., 2022)	MLM + RoBERTa	BERT + improved training objectives and hyperparameters	Accuracy: 80% + F1 Score: 80% +	Epilepsy
(Harnoune et al., 2021)	Transformers + CRF	Knowledge Graph + BERT + CRF	Accuracy: 90.7%	MIMIC-III
(Zhang et al., 2020)	Transformers	BERT-XML	Macro AUC: 0.933	ICD -10
(Fonferko-Shadrach et al., 2019)	General Architecture for Text Engineering (GATE) – ExECT	Rule based + Statistical techniques	Precision: 91.4% Recall: 81.4% F1 score: 86.1%	Epilepsy
(Zhu et al., 2018)	Recurrent Neural Network (RNN)	ELMo + Bi-directional LSTM + CRF	Precision: 89.34% Recall: 87.87% F1 Score: 88.60%	2010 i2b2/VA
(Chalapathy et al., 2016)	RNN	GloVe / word2vec + Bi-directional LSTM + CRF	Precision: 84.36% Recall: 83.41% F1 score: 83.88%	2010 i2b2/VA
(Savova et al., 2010)	cTAKES	Rule-based + ML	F1 score exact: 71.5% F1 score overlap: 82.4%	Mayo Clinic EMR

Table 1. Pre-LLMs period applications in clinical NER tasks

Rule-based Mix Approach

The systems are created by integrating two approaches: primarily rule-based, with either statistical techniques or machine learning as secondary methods. (Savova et al., 2010) developed cTAKES, an open-source system for extracting information from clinical free-text in electronic medical records. This system, which combines rules and machine learning, has become the foundation for many information extraction (IE) systems in clinical records. It focuses on five components: sentence boundary detection, tokenization, part-of-speech tagging, named entity recognition (NER), and shallow parsing. The NER component uses a terminology-agnostic lookup algorithm within a noun phrase window, searching concept mappings from UMLS, SNOMED CT, and RxNorm. Similarly, (Fonferko-Shadrach et al., 2019) created an automated IE system for epilepsy clinical text using rule-based and statistical techniques, based on the GATE framework. This system targets nine categories of epilepsy information and clinical dates, with precision and recall calculated for each category.

Deep Learning Mix Approach

The systems primarily use deep learning models. (Chalapathy et al., 2016) combined Bi-directional LSTM with Conditional Random Fields (CRF) and general-purpose word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) for clinical concept extraction. This task involves identifying medical concepts such as diagnoses, tests, interventions, and treatments in clinical records. The study highlights the limitations of rule-based systems, particularly their low recall, and emphasizes the advantages of neural networks for feature engineering, which is typically labour-intensive and requires expert knowledge. However, the model's performance was only evaluated on the i2b2 2010 challenge dataset (Uzuner et al., 2011). (Zhu et al., 2018) used a similar approach but replaced general-purpose embeddings with Embeddings

from Language Models (ELMo), which improved performance due to contextual embeddings from large models trained on medical text. This model's performance was also limited to the i2b2 2010 dataset. (Zhang et al., 2020) developed BERT-XML, a machine learning model using transformers for large-scale ICD-10 classification, trained on EHR notes, and capable of predicting thousands of ICD codes with BERT's multi-label attention mechanism. (Xie et al., 2022) fine-tuned three BERT-based models (BERT, BioClinicalBERT, Roberta) to extract seizure frequency, date of last seizure, and classify patients as seizure-free, using 1,000 annotated clinical notes. These models reportedly outperformed human capabilities in extracting seizure frequency and classifying patients.

Knowledge graph Mix approach

The system is built with a knowledge graph or base to improve the model performance in combination with the other methods. (Harnoune et al., 2021) combined the concept of Knowledge Graph (KG), BERT and CRF to extract knowledge, relations and analyse interactions between biomedical clinical entities to propose a novel end-to-end system for construction of biomedical KG. NLP models for NER and relationship extraction were used to create biomedical KG. The generated biomedical KG was then used in a question-answering system. With the rise of the transformers, question-answering interfaced systems for knowledge discovery and extraction were becoming popular. The accuracy of the system was measured and categorized under NER and Relation Extraction (RE).

LLMs ERA

In the LLMs era, the clinical NER tasks were solved by creating frameworks with combination of LLMs, deep learning, machine learning, statistics and, knowledge base (see Table 2.).

Models By	Architecture	Detail	Performance	Scope
(Monajatipoor et al., 2024)	DiRAG + GPT 3.5 / 4 Turbo	Prompt Tuning with TANL + DICE, DiRAG	M/T analysis: 53.1/62.8; 61.0/66.2; 51.1/55.0	i2b2 / NCBI disease / BC2GM
(Hu et al., 2024)	GPT 3.5 / 4	Prompt Tuning / ICL	F1 Score: 0.794, 0.861 [MTSamples] F1 Score: 0.676, 0.736 [VAERS]	MTSamples, VAERS
(Munnangi et al., 2024)	GPT 3.5 / 4, Claude 2, Llama 2	ICL + Definition Augmentation	-	BigBIO
(Naguib et al., 2024)	**MLM**: mBERT, XLM-R-large, BERT-large, MedBERT, RoBERTa-large, Bio_clinicalBert, Bert variants, **CLM**: Llama 2-70B, Mistral-7B, BLOOM-7B1, Falcon-40B, GPT, OPT, Vicuna, Medalpaca – 7B, Vigogne-13B	Prompt Tuning / ICL on MLM + CLM models	-	WikiNER, CoNLL2003, E3C, n2c2, NCBI

Table 2. LLMs era applications in clinical NER tasks

The foundation large language models (FLLMs) are versatile and applications range from technology, finance, healthcare, and education. FLLMs are known to have boosted the accuracy of NLP applications in the medicine domain (Khan et al., 2024). FLLMs have paved the way for development of Large Language Models (LLMs) such as Llama (Touvron et al., 2023) , GPT (Xu, n.d.) and others. LLMs are basically very-deep deep

learning models with billions of parameters that learn in a self-supervised or unsupervised way (Birhane et al., 2023). LLMs in healthcare are used for CDS, information extraction from medical records like Electronic Health Records (EHR), healthcare question and answer systems, healthcare education, medical research and drug discovery, sentiment analysis, frequency analysis, thematic analysis (A. Latif & J. Kim, 2024; H. Ali et al., 2023; S. Zou & J. He, 2023; Thirunavukarasu et al., 2023; Yew et al., 2023), classification of radiology reports (Bayrak et al., 2022), leverage generative AI for task automation (Landais et al., 2024), diagnose and monitor patient condition (van Diessen et al., 2024), generate data for simulating randomized clinical trial (Goldenholz et al., 2024) and others in epilepsy care.

Prompt Tuning

The applications in clinical NER tasks mainly revolved around large language foundation models (Khan et al., 2024; Myers et al., 2024) as the core of the NER system and extensions like medical retrieval augmented generation and its variations are built for improved accuracy of the models. (Xiong et al., 2024). The extensions are required as a mitigation strategy for the problems like hallucinations (Reddy et al., 2024) in the LLMs. (Naguib et al., 2024) provide a comparative study of two types of LLMs, namely MLM and CLM, for few-shot clinical NER in three languages. They identified, and reported three main forms of prompt engineering. Constrained prompting approach attempts to better formulate the NER task, by constraining the generation to fill in specific and hand-crafted templates usually adapted in MLMs. Listing prompts approach where the prompts are designed to instruct the LLM model to list the respective entities in a list. Tagging prompts where the entities are highlighted by specialized tags to provide emphasis to the LLMs. The model performance is measured in micro-F1 score and the carbon-footprint left behind the model computation.

Prompt Tuning Mix approach

The approach primarily focuses on prompt tuning, enhanced by additional strategies. (Munnangi et al., 2024) advanced this by using sophisticated prompting techniques to boost LLM performance on clinical NER tasks. Their study assessed model performance on the BigBIO dataset for clinical NER using both closed and open-source LLMs, employing advanced prompting strategies. These strategies included single input and iterative prompting, with outputs in a specific JSON format. They tested three setups: zero-shot, few-shot, and fine-tuning. A novel aspect was incorporating clinical NER concept definitions from knowledge bases like UMLS during evaluation, allowing the model to self-correct based on these definitions. An external SciSpacy NER linker connected entities to concept definitions. Models tested included GPT 3.5/4, Claude 2, and Llama 2. The study found that using definition augmentation, LLMs could outperform RAG and REALM approaches in clinical NER. (Hu et al., 2024) created a prompt framework for clinical NER tasks with zero-shot and few-shot methods, divided into four components: baseline prompts, annotation guideline-based prompts, error analysis instructions, and annotated samples for few-shot learning. Despite its innovative design, this framework did not surpass BioClinicalBERT in supervised settings. (Monajatipoor et al., 2024) proposed prompts in few-shot settings and a knowledge base approach inspired by RAG in zero-shot settings, improving performance by about 15% in clinical NER tasks. They used TANL and DICE text-to-text formats for model training and prompt design, with Dictionary infused RAG (DiRAG) enhancing text quality and relevance using UMLS. The study concluded that ICL and contextually significant examples lead to better performance in clinical NER tasks.

DISCUSSION

Before the advent of LLMs, clinical NER tasks were tackled using rule-based, statistical, and machine learning systems, or a combination of these methods (see Table 1.). These systems proved to be the state-of-the-art implementations for its period. The reason it is called as systems because the solution is not designed fully from one domain, but it is designed as combination of concepts from different domains and thus it becomes a

customized solution for the respective dataset. The systems were a success in its implementation and were able to solve at-least 50% of the problem. It paved the way for new ideas in research that could help improve the systems and overcome the limitations.

Despite the considerable accuracy and performance of the implementations, there were many limitations to such systems. The limitations of the systems are due to the in-ability to contextually generalize and produce consistent results. The results reported in the paper are good given the quality and size of the dataset, but the model quality is limited to quality of the dataset. The rule-based system designed on a smaller dataset could underperform when unknown new patterns are encountered. For such situations, the rules need to be manually redesigned and statistical analysis needs to be performed again to understand the nature of the new pattern. For rule-based and classical machine learning setting, feature engineering is often a necessary manual task as and when the dataset is updated. Machine learning systems could face data drift issues where the model refresh is required, that is, updating the model with new data or model retraining. The model adaptability factor remained unresolved and became a necessity to be engineered in the future score at that period. The data has changed over the time. It has evolved in its nature and form. Due better data acquisition and capturing technologies the complexity of the dataset has massively increased.

So, for the thought on how relevant these systems are in the present times, the response could be considered as "somewhat relevant." The systems would perform as good as it was on the data that was used to design the logic. Thus, these systems could be considered as baseline models that could be used as a reference for development. As the data evolves the systems need to be re-designed to model the complexity of the data. Deep learning implementation has an added advantage where the feature engineering is automated and is performed within the deep layers of the neural networks. Only manual activity is to perform data transformation in-order to convert and correctly represent raw data to vector representations. As vectors or arrays are the primary form of data that are compatible with the deep learning models.

The emergence of transformers, large text datasets, and GPUs has revolutionized natural language processing (NLP) applications. Large language models (LLMs) often are trained on large datasets thus the possibility of generalization and less frequent model refresh or update is required. LLMs are basically very deep learning models. The systems designed with LLMs have outperformed all the benchmarks set by the systems of the past. It has become more capable of capturing more diverse dataset, generalization and perform well on the holdout dataset that is out of true distribution. LLMs have outperformed in applications, however, many studies highlight that LLMs often underperform in clinical NER tasks due to insufficient embedded medical expertise and lack of open-source medical data. Medical data usually contains personal protected information (PPI) that is either siloed or protected in vault with controlled access. To address this, researchers have employed advanced prompt tuning techniques, RAG and RAG-inspired methods, which use knowledge bases like UMLS to incorporate medical concepts, data augmentation to enhance prompt structure, and knowledge graphs and external entity linkers to add information on entity relationships and medical contexts.

The direct dependency of the data always is the major dependency. As the dataset grows the complexity grows. Despite the improvement in the systems to perform in zero shot and few shot cases, the model performance is directly associated to quality and size of the dataset. With the privacy policies and governing rules placed in-order to protect the clinical data and the cost of annotation of the data to create gold standard data are the main hurdles to attain a higher quality of data. Even if the gold standard data is available there would be a need for validation of the dataset to examine and reduce errors introduced due to annotator fatigue. This is largely a problem with text dataset in the NLP field as language in general is a complex unstructured data. Customized prompt tuning frameworks focus on in-context learning (ICL) with various settings such as zero-shot, few-shot, N-shot, and random sample training. Some studies also use fine-tuning methods, adjusting model weights based on new patterns in smaller datasets, though this approach has limitations.

Given the exponential growth of technology, science and computation, the unstructured data captured at an individual level would be more precise and denser. This would lead to more customized care as per patient medical history and needs. The solution would be multi-modal and NLP would play a vital role in converting unstructured data into insights and as a foundation to enrich the dataset. NER for extracting prescriptions and

dosages from clinical text is common and has established benchmarks. Extracting seizure frequency from clinical text is rare and crucial for improving documentation, patient care, and research. Few studies focus on extracting both epileptic seizure frequencies and prescriptions using LLMs.

The literature on NER entity extraction for seizure frequency and prescriptions as a generation task is sparse. There are other aspects of the clinical notes like extraction and study of diagnosis targeted by the NLP. The extracted diagnosis details of a patient from a document could be mapped to respective diagnosis code, associated to ICD-10 or other frameworks, and speed up the clinical coding activity. However, this does not undermine the results that have been achieved in the field of NLP. The accuracy of the model is subjective to the scenario. The state-of-the-art models developed in NLP could be used if not as perfect predictors but in an assistance setting. The NLP applications could be used as assistance tools, for example, a tool used for assisting an annotator to speed up the process. This shows that the NLP despite its commendable progress and achievements still need human intervention and supervision post the prediction to govern the results and direct the critical decisions to be made. With the current momentum of progress due to LLMs and its applications in the NLP, it is expected that new milestones will be achieved in the medical domain irrespective of the dataset size due to accumulation of knowledge learned by the models and the ability to perform transfer learning.

REFERENCES:

A. Latif & J. Kim. (2024). Evaluation and Analysis of Large Language Models for Clinical Text Augmentation and Generation. *IEEE Access*, *12*, 48987–48996. https://doi.org/10.1109/ACCESS.2024.3384496

Asadi-Pooya, A. A., Brigo, F., Lattanzi, S., & Blumcke, I. (2023). Adult epilepsy. *The Lancet*, *402*(10399), 412–424. https://doi.org/10.1016/S0140-6736(23)01048-6

Baker, G. A., Gagnon, D., & McNulty, P. (1998). The relationship between seizure frequency, seizure type and quality of life: Findings from three European countries. *Epilepsy Research*, *30*(3), 231–240. https://doi.org/10.1016/S0920-1211(98)00010-2

Bayrak, S., Yucel, E., & Takci, H. (2022). Epilepsy Radiology Reports Classification Using Deep Learning Networks. *Computers, Materials & Continua*, *70*(2), 3589–3607. Inspec®. https://doi.org/10.32604/cmc.2022.018742

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Biswas, A., & Talukdar, W. (2024). Enhancing Clinical Documentation with Synthetic Data: Leveraging Generative Models for Improved Accuracy. *International Journal of Innovative Science and Research Technology (IJISRT)*, 1553–1566. https://doi.org/10.38124/ijisrt/IJISRT24MAY2085

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Campbell, K. E., Oliver, D. E., & Shortliffe, E. H. (1998). The Unified Medical Language System: Toward a Collaborative Approach for Solving Terminologic Problems. *Journal of the American Medical Informatics Association*, *5*(1), 12–16. https://doi.org/10.1136/jamia.1998.0050012

Chalapathy, R., Borzeshi, E. Z., & Piccardi, M. (2016). *Bidirectional LSTM-CRF for Clinical Concept Extraction* (arXiv:1611.08373). arXiv. http://arxiv.org/abs/1611.08373

Choi, H., Hamberger, M. J., Munger Clary, H., Loeb, R., Onchiri, F. M., Baker, G., Hauser, W. A., & Wong, J. B. (2014). Seizure frequency and patient-centered outcome assessment in epilepsy. *Epilepsia*, *55*(8), 1205–1212. https://doi.org/10.1111/epi.12672

Dean, J. (2022). A Golden Decade of Deep Learning: Computing Systems & Applications. *Daedalus*, *151*(2), 58–74. https://doi.org/10.1162/daed_a_01900

Decker, B. M., Turco, A., Xu, J., Terman, S. W., Kosaraju, N., Jamil, A., Davis, K. A., Litt, B., Ellis, C. A., Khankhanian, P., & Hill, C. E. (2022). Development of a natural language processing algorithm to extract seizure types and frequencies from the electronic health record. *Seizure: European Journal of Epilepsy*, *101*, 48–51. https://doi.org/10.1016/j.seizure.2022.07.010

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, *42*(5), 760–772. https://doi.org/10.1016/j.jbi.2009.08.007

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dubois, S., Romano, N., Kale, D. C., Shah, N., & Jung, K. (2018). *Effective Representations of Clinical Notes* (arXiv:1705.07025). arXiv. http://arxiv.org/abs/1705.07025

Falco-Walter, J. (2020). Epilepsy-Definition, Classification, Pathophysiology, and Epidemiology. *Seminars in Neurology*, *40*(6), 617–623. https://doi.org/10.1055/s-0040-1718719

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models* (arXiv:2405.06211). arXiv. https://doi.org/10.48550/arXiv.2405.06211

Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., Engel Jr, J., Forsgren, L., French, J. A., Glynn, M., Hesdorffer, D. C., Lee, B. i., Mathern, G. W., Moshé, S. L., Perucca, E., Scheffer, I. E., Tomson, T., Watanabe, M., & Wiebe, S. (2014). ILAE Official Report: A practical clinical definition of epilepsy. *Epilepsia*, *55*(4), 475–482. https://doi.org/10.1111/epi.12550

Fonferko-Shadrach, B., Lacey, A. S., Roberts, A., Akbari, A., Thompson, S., Ford, D. V., Lyons, R. A., Rees, M. I., & Pickrell, W. O. (2019). Using natural language processing to extract structured epilepsy data from unstructured clinic letters: Development and validation of the ExECT (extraction of epilepsy clinical text) system. *BMJ Open*, *9*(4), e023232. https://doi.org/10.1136/bmjopen-2018-023232

Fonferko-Shadrach, B., Strafford, H., Jones, C., Khan, R., Brown, S., Edwards, J., Hawken, J., Shrimpton, L., White, C. P., Powell, R., Sawhney, I., Pickrell, W. O., & Lacey, A. (2023). *Gold standard annotation of epilepsy clinic letters for the development of information extraction tools* (Version version 1) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8381080

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Goldenholz, D. M., Goldenholz, S. R., Habib, S., & Westover, M. B. (2024). Inductive reasoning with large language models: A simulated randomized controlled trial for epilepsy. *medRxiv: The Preprint Server for Health Sciences*, 2024.03.18.24304493. https://doi.org/10.1101/2024.03.18.24304493

Gu, Y., Han, X., Liu, Z., & Huang, M. (2022). *PPT: Pre-trained Prompt Tuning for Few-shot Learning* (arXiv:2109.04332). arXiv. https://doi.org/10.48550/arXiv.2109.04332

H. Ali, J. Qadir, T. Alam, M. Househ, & Z. Shah. (2023). ChatGPT and Large Language Models in Healthcare: Opportunities and Risks. *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, 1–4. https://doi.org/10.1109/AIBThings58340.2023.10291020

Hanson, J. L., Stephens, M. B., Pangaro, L. N., & Gimbel, R. W. (2012). Quality of outpatient clinical notes: A stakeholder definition derived through qualitative research. *BMC Health Services Research*, *12*(1), 407. https://doi.org/10.1186/1472-6963-12-407

Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., & El Asri, B. (2021). BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, *1*, 100042. https://doi.org/10.1016/j.cmpbup.2021.100042

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., & Xu, H. (2024). Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, ocad259. https://doi.org/10.1093/jamia/ocad259

Jain, M., Nathe, P., Rathod, K., Tiwari, N. K., Dedgaonkar, S., & Shewale, C. (2024). AI HealthCare Chatbot. *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*, 1–6. https://doi.org/10.1109/MITADTSoCiCon60330.2024.10575622

Katyan, D., Gulati, G., & Upreti, G. (2024). Utilising NLP for Enhanced Clinical Text Mining. *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 883–889. https://doi.org/10.1109/ICAAIC60222.2024.10575794

Khan, W., Leem, S., See, K. B., Wong, J. K., Zhang, S., & Fang, R. (2024). A Comprehensive Survey of Foundation Models in Medicine. *arXiv*. Inspec®. https://www.proquest.com/undefined/comprehensive-survey-foundation-models-medicine/docview/3072178954/se-2?accountid=14680

Landais, R., Sultan, M., & Thomas, R. H. (2024). The promise of AI Large Language Models for Epilepsy care. *Epilepsy & Behavior: E&B*, *154*, 109747. https://doi.org/10.1016/j.yebeh.2024.109747

Leidy, N. K., Elixhauser, A., Vickrey, B., Means, E., & Willian, M. K. (1999). Seizure frequency and the health-related quality of life of adults with epilepsy. *Neurology*, *53*(1), 162–162. https://doi.org/10.1212/WNL.53.1.162

Liu, X., Chen, H., & Zheng, X. (2020). Effects of seizure frequency, depression and generalized anxiety on suicidal tendency in people with epilepsy. *Epilepsy Research*, *160*, 106265. https://doi.org/10.1016/j.eplepsyres.2020.106265

Liu, X., Ji, K., Fu, Y., Tam, W. L., Du, Z., Yang, Z., & Tang, J. (2022). *P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks* (arXiv:2110.07602). arXiv. https://doi.org/10.48550/arXiv.2110.07602

Melamud, O., & Shivade, C. (2019). *Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models* (arXiv:1905.07002). arXiv. http://arxiv.org/abs/1905.07002

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality* (arXiv:1310.4546). arXiv. http://arxiv.org/abs/1310.4546

Monajatipoor, M., Yang, J., Stremmel, J., Emami, M., Mohaghegh, F., Rouhsedaghat, M., & Chang, K.-W. (2024). *LLMs in Biomedicine: A study on clinical Named Entity Recognition* (arXiv:2404.07376). arXiv. http://arxiv.org/abs/2404.07376

Munnangi, M., Feldman, S., Wallace, B. C., Amir, S., Hope, T., & Naik, A. (2024). *On-the-fly Definition Augmentation of LLMs for Biomedical NER* (arXiv:2404.00152). arXiv. https://doi.org/10.48550/arXiv.2404.00152

Myers, D., Mohawesh, R., Chellaboina, V. I., Sathvik, A. L., Venkatesh, P., Ho, Y.-H., Henshaw, H., Alhawawreh, M., Berdik, D., & Jararweh, Y. (2024). Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts. *Cluster Computing*, *27*(1), 1–26. https://doi.org/10.1007/s10586-023-04203-7

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Naguib, M., Tannier, X., & Névéol, A. (2024). *Few shot clinical entity recognition in three languages: Masked language models outperform LLM prompting* (arXiv:2402.12801). arXiv. https://doi.org/10.48550/arXiv.2402.12801

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, February 15). *Deep contextualized word representations*. arXiv.Org. https://arxiv.org/abs/1802.05365v2

Ramsay, A. ., & Ramsay, A. (1987). What we say and what we mean. *Artificial Intelligence Review.*, *1*(3).

Reddy, G. P., Pavan Kumar, Y. V., & Prakash, K. P. (2024). Hallucinations in Large Language Models (LLMs). *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1–6. https://doi.org/10.1109/eStream61684.2024.10542617

Reichenpfader, D., Müller, H., & Denecke, K. (2023). Large language model-based information extraction from free-text radiology reports: A scoping review protocol. *BMJ Open*, *13*(12), e076865. https://doi.org/10.1136/bmjopen-2023-076865

S. Zou & J. He. (2023). Large Language Models in Healthcare: A Review. *2023 7th International Symposium on Computer Science and Intelligent Control (ISCSIC)*, 141–145. https://doi.org/10.1109/ISCSIC60498.2023.00038

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 507–513. https://doi.org/10.1136/jamia.2009.001560

Schank, A. ., R. ;Kass. (1986, January 1). Natural language processing: What's really involved? *TINLAP-3. Theoretical Issues in Natural Language Processing-3. Position Papers*.

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., Hirsch, E., Jain, S., Mathern, G. W., Moshé, S. L., Nordli, D. R., Perucca, E., Tomson, T., Wiebe, S., Zhang, Y.-H., & Zuberi, S. M. (2017). ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 512–521. https://doi.org/10.1111/epi.13709

Soto, C. M., Kleinman, K. P., & Simon, S. R. (2002). Quality and correlates of medical record documentation in the ambulatory care setting. *BMC Health Services Research*, *2*(1), 22. https://doi.org/10.1186/1472-6963-2-22

Spasic, I., & Nenadic, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, *8*(3), e17984. https://doi.org/10.2196/17984

Thijs, R. D., Surges, R., O'Brien, T. J., & Sander, J. W. (2019). Epilepsy in adults. *Lancet (London, England)*, *393*(10172), 689–701. https://doi.org/10.1016/S0140-6736(18)32596-0

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 552–556. https://doi.org/10.1136/amiajnl-2011-000203

van Diessen, E., van Amerongen, R. A., Zijlmans, M., & Otte, W. M. (2024). Potential merits and flaws of large language models in epilepsy care: A critical review. *Epilepsia*, *65*(4), 873–886. https://doi.org/10.1111/epi.17907

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Viteva, E. I. (2014). Seizure frequency and severity: How really important are they for the quality of life of patients with refractory epilepsy. *Annals of Indian Academy of Neurology*, *17*(1), 35. https://doi.org/10.4103/0972-2327.128544

Vulpius, S. A., Werge, S., Jørgensen, I. F., Siggaard, T., Hernansanz Biel, J., Knudsen, G. M., Brunak, S., & Pinborg, L. H. (2023). Text mining of electronic health records can validate a register-based diagnosis of epilepsy and subgroup into focal and generalized epilepsy. *Epilepsia*, *64*(10), 2750–2760. https://doi.org/10.1111/epi.17734

Wang, Y., Steinhubl, S. R., Defilippi, C., Ng, K., Ebadollahi, S., Stewart, W. F., & Byrd, R. J. (2015). Prescription Extraction from Clinical Notes: Towards Automating EMR Medication Reconciliation. *AMIA Summits on Translational Science Proceedings*, *2015*, 188–193.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, *77*, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). *Finetuned Language Models Are Zero-Shot Learners* (arXiv:2109.01652). arXiv. https://doi.org/10.48550/arXiv.2109.01652

Wei, W.-Q., Teixeira, P. L., Mo, H., Cronin, R. M., Warner, J. L., & Denny, J. C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, *23*(e1), e20–e27. https://doi.org/10.1093/jamia/ocv130

Wigglesworth, S., Neligan, A., Dickson, J., Pullen, A., Yelland, E., Anjuman, T., & Reuber, M. (2023). The incidence and prevalence of epilepsy in the United Kingdom 2013–2018: A retrospective cohort study of UK primary care data. *Seizure: European Journal of Epilepsy*, *105*, 37–42. https://doi.org/10.1016/j.seizure.2023.01.003

Wood, D. L. (2001). Documentation guidelines: Evolution, future direction, and compliance. *The American Journal of Medicine*, *110*(4), 332–334. https://doi.org/10.1016/s0002-9343(00)00748-8

Xie, K., Gallagher, R. S., Conrad, E. C., Garrick, C. O., Baldassano, S. N., Bernabei, J. M., Galer, P. D., Ghosn, N. J., Greenblatt, A. S., Jennings, T., Kornspun, A., Kulick-Soper, C. V., Panchal, J. M., Pattnaik, A. R., Scheid, B. H., Wei, D., Weitzman, M., Muthukrishnan, R., Kim, J., … Roth, D. (2022). Extracting seizure frequency from epilepsy clinic notes: A machine reading approach to natural language processing. *Journal of the American Medical Informatics Association*, *29*(5), 873–881. https://doi.org/10.1093/jamia/ocac018

Xie, K., Gallagher, R. S., Shinohara, R. T., Xie, S. X., Hill, C. E., Conrad, E. C., Davis, K. A., Roth, D., Litt, B., & Ellis, C. A. (2023). Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. *Epilepsia*, *64*(7), 1900–1909. https://doi.org/10.1111/epi.17633

Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). *Benchmarking Retrieval-Augmented Generation for Medicine* (arXiv:2402.13178). arXiv. http://arxiv.org/abs/2402.13178

Xu, Z. (n.d.). *The Mysteries of Large Language Models: Tracing the Evolution of Transparency for OpenAI's GPT Models*.

Yew, A. N. J., Schraagen, M., Otte, W. M., & van Diessen, E. (2023). Transforming epilepsy research: A systematic review on natural language processing applications. *Epilepsia*, *64*(2), 292–305. https://doi.org/10.1111/epi.17474

Zhang, Z., Liu, J., & Razavian, N. (2020). *BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining* (arXiv:2006.03685). arXiv. http://arxiv.org/abs/2006.03685

Zhou, N., Wu, Q., Wu, Z., Marino, S., & Dinov, I. D. (2022). DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes. *Journal of Medical Systems*, *46*(12), 96. https://doi.org/10.1007/s10916-022-01880-6

Zhu, H., Paschalidis, I. C., & Tahmasebi, A. (2018). *Clinical Concept Extraction with Contextual Word Embedding* (arXiv:1810.10566). arXiv. http://arxiv.org/abs/1810.10566